

# SCALAR IMPLICATURES

EDITED BY: Penka Stateva and Anne Reboul

PUBLISHED IN: Frontiers in Psychology and Frontiers in Communication





# frontiers

## Frontiers Copyright Statement

© Copyright 2007-2019 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714

ISBN 978-2-88963-134-6

DOI 10.3389/978-2-88963-134-6

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: [researchtopics@frontiersin.org](mailto:researchtopics@frontiersin.org)

# SCALAR IMPLICATURES

Topic Editors:

**Penka Stateva**, University of Nova Gorica, Slovenia

**Anne Reboul**, CNRS UMR 5304, France

Scalar implicatures have enjoyed the status of one of the most researched topics in both theoretical and experimental pragmatics in recent years. This Research Topic presents new developments in studying the comprehension, as well as the production of scalar inferences, suggests new testing paradigms that trigger important discussions about the methodology of experimental investigation, explores the effect of prosody and context on inference rates. To a great extent the articles reflect the state of the art in the domain and outline promising paths for future research.

**Citation:** Stateva, P., Reboul, A., eds. (2019). *Scalar Implicatures*. Lausanne: Frontiers Media. doi: 10.3389/978-2-88963-134-6

# Table of Contents

## **05 Editorial: Scalar Implicatures**

Anne Colette Reboul and Penka Stateva

## **SECTION 1**

### **THE ROLE OF PROSODY AND CONTEXT**

#### **08 Determining the Types of Contrasts: The Influences of Prosody on Pragmatic Inferences**

I-Hsuan Chen, Chu-Ren Huang and Stephen Politzer-Ahles

#### **17 Processing Presuppositions and Implicatures: Similarities and Differences**

Cory Bill, Jacopo Romoli and Florian Schwarz

#### **39 Context-Sensitivity and Individual Differences in the Derivation of Scalar Implicature**

Xiao Yang, Utako Minai and Robert Fiorentino

## **SECTION 2**

### **LINKING RESPONSE OPTIONS TO PRAGMATIC INFERENCES**

#### **53 Linking Hypothesis and Number of Response Options Modulate Inferred Scalar Implicature Rate**

Masoud Jasbi, Brandon Waldon and Judith Degen

#### **67 Social Context Modulates Tolerance for Pragmatic Violations in Binary but not Graded Judgments**

Les Sikos, Minjae Kim and Daniel J. Grodner

#### **78 Believing What You're Told: Politeness and Scalar Inferences**

Diana Mazzarella, Emmanuel Trouche, Hugo Mercier and Ira Noveck

## **SECTION 3**

### **SCALE DIVERSITY**

#### **90 Scalar Diversity, Negative Strengthening, and Adjectival Semantics**

Nicole Gotzner, Stephanie Solt and Anton Benz

#### **103 A Link Between Local Enrichment and Scalar Diversity**

Chao Sun, Ye Tian and Richard Breheny

#### **115 Development of Quantitative and Temporal Scalar Implicatures in a Felicity Judgment Task**

Walter Schaeken, Bojoura Schouten and Kristien Dieussaert

#### **129 Cross-Linguistic Variation in the Meaning of Quantifiers: Implications for Pragmatic Enrichment**

Penka Stateva, Arthur Stepanov, Viviane Déprez, Ludivine Emma Dupuy and Anne Colette Reboul



## SECTION 4

### NEW DIRECTIONS FOR EXPLORATION

**147** *Some Pieces are Missing: Implicature Production in Children*

Sarah F. V. Eiteljoerge, Nausicaa Pouscoulous and Elena V. M. Lieven

**163** *Competition and Symmetry in an Artificial Word Learning Task*

Brian Buccola, Isabelle Dautriche and Emmanuel Chemla



# Editorial: Scalar Implicatures

Anne Colette Reboul<sup>1</sup> and Penka Stateva<sup>2\*</sup>

<sup>1</sup> UMR5304, Institut des Sciences Cognitives Marc Jeannerod, Bron, France, <sup>2</sup> Center for Cognitive Science of Language, University of Nova Gorica, Nova Gorica, Slovenia

**Keywords:** scalar implicatures, variability, neo-Griceans, post-Griceans, grammatical approach

## Editorial on the Research Topic

### Scalar Implicatures

In 1975, Grice introduced the notion of implicature, arguing that it was more appropriate to account for a class of apparent lexical ambiguities through pragmatic processes than by multiplying lexical meanings (Modified Ockham's razor: *Do not multiply meanings beyond necessity*; Grice, 1975). His aim was to defend the idea that logical terms (*and*, *or*, *if... then*, quantifiers, etc.) do not have a meaning specific to their use in natural language. Rather, or so he argued, logical terms in natural language mean exactly what they mean in logic and their lexical meaning can be read off their logical truth tables. What gives the illusion that they acquire a different meaning in natural language is that their use in conversation frequently gives rise to implicatures. The following theoretical debate centered on how the pragmatic inferences necessary to access these implicatures were produced: neo-Griceans insisted on the specificity of scalar implicatures and on the importance of lexical scales (Horn, 1984; Levinson, 2000); post-Griceans rejected the idea that there was anything specific about scalar implicatures and emphasized the role of pragmatic processes (Sperber and Wilson, 1995; Noveck and Sperber, 2007).

For the past 20 years, experimental approaches have superseded purely theoretical ones, with mixed results. Paradigms using verification tasks on infelicitous sentences, with rate of pragmatic answers and reaction time as measures, have generally concluded in favor of the post-Gricean views (Bott and Noveck, 2004; Noveck and Reboul, 2008). However, some recent studies discuss additional factors affecting implicature processing and have introduced new paradigms which suggest a different conclusion (Katsos and Bishop, 2011; Breheny et al., 2013; Degen and Tanenhaus, 2015; Foppolo and Marelli, 2017; Bill et al.; Jasbi et al.; Sikos et al.). In addition, current research has shown that lexical scales may play a role in the process in keeping with neo-Gricean views (Doran et al., 2009; van Tiel et al., 2016; Gotzner et al.; Sun et al.). Furthermore, scales may vary in their potential to trigger pragmatic interpretations cross-linguistically. One possible explanation is that part of the variation may be due to the employment of different processes of pragmatic strengthening in different languages (Stateva et al.). Consequently, one might expect some more cases of cross-linguistic variation, notably among logical words (*or*, *if... then*, quantifiers, etc.).

This Frontiers topic is a collection of 12 contributions in experimental pragmatics focusing on different aspects of child and adult processing of implicatures, factors affecting their rate, relevance of testing paradigms, scale diversity, cross-linguistic differences, and variation in triggers.

A substantial part of the reported research examined various factors affecting the rates of pragmatic inferences, as well as their content. The role of prosody on restricting the relevant set of alternatives was given central attention in Chen et al. The study also investigated how context interacts with prosody. How prosodic stress on the scalar trigger influences pragmatic rates was also evaluated in one of the experiments reported in Bill et al. Two more studies investigate the effect of context on rates of pragmatic inferences. Yang et al.'s article argues for a relation between

## OPEN ACCESS

### Edited and reviewed by:

Manuel Carreiras,  
Basque Center on Cognition, Brain  
and Language, Spain

### \*Correspondence:

Penka Stateva  
penka.stateva@ung.si

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 10 July 2019

**Accepted:** 15 July 2019

**Published:** 31 July 2019

### Citation:

Reboul AC and Stateva P (2019)  
Editorial: Scalar Implicatures.  
Front. Psychol. 10:1767.  
doi: 10.3389/fpsyg.2019.01767

individual cognitive resources, personality-based pragmatic abilities and language abilities, on the one hand, and sensitivity to context, on the other, which in turn, affects positively pragmatic rates. In their study, Sikos et al. manipulate social contexts to conclude that speaker's tolerance to pragmatic violations in the sense of Katsos and Bishop (2011) is affected in binary judgment task but not in graded judgments tasks. That study reveals another factor affecting rates of inferences: the number of response options in implicature comprehension studies. Whether the number of possible answers affects pragmatic rates is the main research question also in Jasbi et al. In its turn, this question raises important methodological considerations related to experimental designs in pragmatic studies and consequently the validity of the result interpretations. In line with Katsos and Bishop's (2011) evidence that a binary option task can mask children's ability to compute scalar implicatures, Jasbi et al. argue that a graded judgment design is more informative in evaluating rates of pragmatic inferences also in studies with adult speakers. However, designs involving a multiplicity of options necessitates careful effort in formulating the hypothesis that links the pragmatic inferences with the choice of provided answers. In addition to Jasbi et al.'s discussion, this volume includes an article on the role of politeness in the comprehension of scalar implicatures which bears on the "linking hypothesis." Mazzarella et al. distinguish between "comprehension" and "epistemic assessment" of communicated information. Their study reveals that it is possible to observe a discrepancy between rates of pragmatic answers and actually drawn inferences if the participants' evaluation of the truth of the potential inference is taken into consideration.

Scale diversity, as a major factor affecting pragmatic rates, and the source of the different potential of scalar triggers to incur inferences is discussed in Gotzner et al.; Sun et al.; Schaeken et al.; Stateva et al. and Gotzner et al. argue that scale structure related to a scalar item affects that item's potential to trigger scalar implicatures. In other words, properties (like gradability) of scale structures are a prerequisite for pragmatic strengthening not only by scalar implicatures but also by other kinds of inferences which can obscure each other's availability. Stateva et al. extend the

topic of scale diversity and interaction of pragmatic enrichment processes to give it a cross-linguistic dimension. Schaeken et al. discuss scale diversity from the point of language acquisition. The study reveals different patterns of pragmatic rates in inferences related to quantitative vs. temporal scales. Sun et al. also explore potential factors responsible for the different implicature rates of scalar triggers and relate them to the susceptibility of different lexical items to local enrichment. This opens the door for an enlightening comparison between the grammatical theory of pragmatic enrichment and dual route theories. Evaluating the descriptive adequacy of different theories is also a topic of major interest in Bill et al. The article explores parallels and differences between scalar implicatures and presuppositions in patterns of processing. The results pave the way for further discussion in view of current proposals to subsume presuppositions under the umbrella of scalar inferences.

Buccola et al. offer an artificial word learning paradigm to examine competition which is at the core of pragmatic processes like computing scalar implicatures. The study demonstrates that symmetry among alternatives is another factor affecting the rate of inferences.

The corpus study reported Eiteljoerge et al. is one of the few available production studies of scalar implicatures. Its major contribution that children as young as 3 years of age can produce scalar inferences at rates comparable to their adult caregiver poses a curious puzzle in view of the acquisition delay observed in implicature comprehension studies (Noveck, 2001).

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

## FUNDING

PS acknowledges the financial support from the Slovenian Research Agency (research core funding No. P6-0382).

## REFERENCES

- Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Memory Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006
- Breheny, R., Ferguson, H. J., and Katsos, N. (2013). Taking the epistemic step: towards a model of on-line access to conversational implicatures. *Cognition* 126, 423–440. doi: 10.1016/j.cognition.2012.11.012
- Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171
- Doran, R., Baker, R. E., McNabb, Y., Larson, M., and Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *Int. Rev. Pragmat.* 1, 1–38. doi: 10.1163/187730909X12538045489854
- Foppolo, F., and Marelli, M. (2017). No delay for some inferences. *J. Semant.* 34, 659–681. doi: 10.1093/jos/ffx013
- Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics* vol. 3: *Speech Acts*, eds P. Cole and J. L. Morgan (New York, NY: Academic Press), 41–58.
- Horn, L. (1984). "Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature," in *Meaning, Form, and Use in Context: Linguistic Applications*, ed D. Schiffrin (Washington, DC: Georgetown University Press), 11–89.
- Katsos, N., and Bishop, D. V. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and inference. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015
- Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.
- Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicatures. *Cognition* 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1

- Noveck, I. A., and Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends Cogn. Sci.* 12, 425–431. doi: 10.1016/j.tics.2008.07.009
- Noveck, I. A., and Sperber, D. (2007). “The why and how of experimental pragmatics: the case of scalar inference,” in *Advances in Pragmatics*, ed N. Burton-Roberts (Basingstoke: Palgrave), 184–212.
- Sperber, D., and Wilson, D. (1995). *Relevance: Communication Cognition*. Oxford: Blackwell.
- van Tiel, B., van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *J. Semant.* 33, 107–135. doi: 10.1093/jos/ffu017

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Reboul and Stateva. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Determining the Types of Contrasts: The Influences of Prosody on Pragmatic Inferences

I-Hsuan Chen\*, Chu-Ren Huang and Stephen Politzer-Ahles

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong

## OPEN ACCESS

### Edited by:

Anne Colette Reboul,  
Claude Bernard University Lyon 1,  
France

### Reviewed by:

Christelle Declercq,  
Université de Reims Champagne  
Ardenne, France  
John E. Drury,  
Stony Brook University, United States

### \*Correspondence:

I-Hsuan Chen  
ihuan.chen@polyu.edu.hk;  
ihcucb@gmail.com

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 31 March 2018

**Accepted:** 12 October 2018

**Published:** 08 November 2018

### Citation:

Chen I-H, Huang C-R and  
Politzer-Ahles S (2018) Determining  
the Types of Contrasts: The Influences  
of Prosody on Pragmatic Inferences.  
*Front. Psychol.* 9:2110.  
doi: 10.3389/fpsyg.2018.02110

This study explores the issues involving pragmatic inferences with prosodic cues. Although there is a well-established literature from multiple languages demonstrating how different pragmatic inferences can be applied to the same syntactic structure, few studies discuss whether prosody can determine types of alternative sets based on the same syntactic structure. In Mandarin Chinese, the same sentence containing a numeral-classifier phrase as a negative polarity item can be employed for two types of scalar inferences based on either the numeral or the noun. The sentence *wo yi zhi mayi dou mei kan dao* ("I didn't even see one ant") can induce two different scalar inferences: Quantity-contrast ('I did not see one ant, much less two ants, three ants, and so on' by drawing a contrast against the minimal quantity of one), and Type-contrast ('I did not see an ant, much less a dog, a cat, a human being, and so on' by drawing a contrast against the minimally surprising type, that of ants). Taking advantage of similar sentences with the syntactic structure and lexical items, our study examines whether prosodic conditions can guide people to choose pragmatic inferences from a set of options based on the same syntactic structure. The experiments of this study are designed to answer whether prosody interacts with contextual information in this grammatical structure. The results suggest that Mandarin speakers can use sentence prosody to determine which inference is intended, at least in experimental contexts that directly probe explicit awareness of prosody. Prosody does play a role in inducing scalar inferences, but contextual information can override the effects of prosody. Each prosodic pattern can evoke a specific set of scalar inferences, but quantity-contrast inferences are favored over type-contrast inferences. Our experiments show that prosodic prominence can serve as a linguistic cue to pragmatic inferences.

**Keywords:** prosody, scalar inferences, numeral-classifier phrases, negative polarity items, intonation

## INTRODUCTION

Pragmatics is the study of how signs are used and interpreted in context by language users and their interlocutors (Morris, 1938). The studies of pragmatics focus on the context-dependent meanings which are systematically abstracted from the logical form or the content of a construction concerned in syntax and semantics (Grice, 1989; Horn and Ward, 2005). In order to interpret information from a speaker, the hearer has to take the interaction of grammatical structure and context into consideration. The scalar inferences discussed in this study are cases showing that the hearer evokes a mental scalar model from a grammatical construction and context

(Fillmore et al., 1988). The inferences from a scalar model compare the possibility of all alternatives on a defined scale.

This study investigates whether prosody influences pragmatic inferences by examining the types of inferences inferred from negative polarity items (NPIs) in Chinese. As a tonal language, Chinese has both syllable-level lexical tones and sentence-level intonation. The syllable-level lexical tones have been described as “small ripples riding on large waves of intonation” (Chao, 1933). Intonation interacts with syllabic tones without canceling their acoustic effects. The prominence of intonation is regarded as expanding pitch range. For example, prominent words have larger pitch range, longer duration, and higher intensity in prosody (Shih, 1988). Particularly, contextually focused words in a sentence are prominent in pitch height and intensity (Yuan, 2004). The present study examines how sentence-level intonation, particularly focus, influences the interpretation of NPIs.

Negative polarity items are expressions that are only grammatical under certain semantic contexts, such as negation and other forms of downward entailing contexts (Giannakidou, 2011; Israel, 2011). For example, in English, *I haven't ever been to France* is grammatical but *\*I have ever been to France* is not; *ever* is an NPI which is only grammatical in NPI-licensing contexts. NPIs have been observed across many languages (Haspelmath, 1997). They are often words referring to very small amounts, e.g., *I didn't sleep a wink*, *He won't spend a red cent*, *They don't give a rat's ass about this topic*. In such cases, the negation of such a small amount allows the hearer to infer that larger amounts are also not true: e.g., if somebody did not sleep “a wink” then they surely did not sleep for a long time either. These types of small-quantity expressions which occur in environments related to negation are called *minimizers*, and are a type of NPI. Across languages, minimizers are widely employed for pragmatic emphasis, due to their robust scalar inferences (Giannakidou, 2011; Israel, 2011). Minimizers induce scalar reasoning because they evoke a mental scalar model with all the alternatives ranked for contrasting (Israel, 2011). Since minimizers refer to an endpoint of a scale, they can contrast with all the other alternatives along the scale for emphasis (Fauconnier, 1975; Horn, 1989). That is to say, if the smallest or weakest item on the scale (e.g., sleeping a wink) is not true, then all larger or stronger items (sleeping a minute, sleeping an hour, etc.) must also not be true.

The paper reports three experiments regarding scalar implicatures and prosody. In each experiment in this study, all the participants provided their informed consent before they began the survey. Each experiment had both a traditional character version and a simplified character version. The traditional character version was distributed in Taiwan and Hong Kong, while the simplified character version was distributed in Mainland China. When the survey was advertised through the platforms of social media, both links were provided and volunteers could choose based on their preference.

In numeral-classifier languages such as Mandarin Chinese, ‘one’-phrases, which are composed of the numeral ‘one,’ a classifier or a measure word, and a noun, are pervasively used as minimizers, as in example (1) below. Specifically, sortal classifiers

are employed for categorizing a semantically salient perceptual property of a noun which can be individuated (Ahrens and Huang, 2016).

- (1) wo yi zhi cangying dou mei kandao  
I one CLF fly FOC NEG see  
‘I did not see even one fly.’

Just as in the examples above, sentences with numeral-classifier phrases like (1) also elicit inferences about what the phrase is being contrasted with. (In this and other examples, CLF stands for classifiers, FOC for focus markers, and NEG for negation.) Specifically, for a sentence like (1), two types of inferences are possible. The sentence can infer that the speaker saw ‘not even one fly, much less two’ if the minimizer is interpreted as invoking a quantity-based contrast, while it can instead imply that the speaker saw ‘not even one fly, much less one human being’ if the minimizer is interpreted as invoking a type-based contrast. In the quantity-contrast interpretation, the minimal amount that is being invoked is “one,” and this is raised in contrast with greater amounts (“two flies,” and “three flies,” etc.); in the type-contrast interpretation, the minimal amount is some type of noun that has a high probability of occurring in this context. For example, this sentence is uttered in a context where there are likely to be flies, and this is raised in contrast with nouns that are even less likely or prototypical in this context. The quantity-contrast interpretation is straightforward due to the involvement of a numeral phrase, while the type-contrast interpretation is relatively less straightforward since it is relevant to the shared knowledge of the contexts. However, it is clear that the noun chosen for contrasting is the proposition which is assumed to be the most likely one.

In other numeral classifier languages such as Japanese and Korean, the distinction of the two types of inferences is reflected in morphology and word order (Lee, 2006; Nakanishi, 2006). However, in Mandarin, the two sets of inferences occur in the same word order, syntax, and semantics. Native Mandarin speakers thus require other cues to discern the pragmatic differences. It has been noted in studies of NPIs that minimizers are claimed to tend to occur in constructions that can attract people's focus (Israel, 2011). For instance, an expression interpreted as a minimizer carries an emphasized intonation which is different from its other uses. In line with this observation, Mandarin minimizers tend to occur in the preverbal construction as in (1): this sentence has a Subject-Object-Verb word order, which differs from the Subject-Verb-Object word order that is canonical and unmarked in Mandarin. This preverbal position, where “one fly” occurs in sentence (1), has received substantial attention in the literature and has been regarded to carry focus (Zhang, 2000; Tsai, 2004; Huang et al., 2009). It is also noted that ‘one’-phrases may bear a different prosodic stress when they are used as minimizers as opposed to when they are used normally (Chao, 1968). According to these studies, a connection between prosodic stress, focus of attention, and pragmatics can be inferred. However, the issues of how focus is perceived by native speakers and of whether prosodic stress modulates the inferences drawn by speakers in this type of sentence have been barely touched upon.



On the other hand, scalar inferences have been shown to be associated with grammatical structures. For example, Chierchia (2004) and Chierchia et al. (2012) argued that a grammatical well-formedness condition based on pragmatics must be checked during the morphosemantic processing of NPIs and scalar implicatures. Other accounts differ on how or when grammatical information is integrated to process scalar references. However, to the best of our knowledge, the question of whether prosody would also be checked has not been answered. For instance, it is already known that prosody has an immediate impact on the incremental interpretation of an utterance that is unfolding: for example, prosodic focus influences how likely listeners are to commit to interpreting *some* as *not all* (Degen and Tanenhaus, 2015) and *or* as exclusive *or* (Chevallier et al., 2010), and to disambiguate the meaning of sentences with attachment ambiguities like *Tap the frog with the flower* (Snedeker and Trueswell, 2003). In the study of *some* as *not all*, the impact of prosody is whether to apply the inferences, while in the case of the attachment ambiguity the question is whether prosody can help to differentiate the actual differences in syntactic structure. However, it has not yet been empirically demonstrated that prosody has an impact on the inferences elicited by minimizers like those described above. The abovementioned examples are cases where ambiguity derives from the choice whether or not to realize an implicature at all, or the choice between different syntactic structures to build; on the other hand, the interpretational ambiguity in Mandarin minimizers comes from two types of alternative sets and not from syntactic differences or from the presence or absence of an implicature (as the same implicature is made under both interpretations, the implicature is simply applied over different alternative sets).

The experiments of this study are designed to answer this question. The experiments force participants to consider prosodic conditions by using identical, well-formed morphosyntactic structures. In particular, Chinese provides an interesting and challenging environment for testing the role of prosody in scalar inferences. The prosody of Chinese, a tonal language, is an overlaying pattern which modifies pitch ranges and intensities, instead of lexicalizing pitch patterns, as discussed above. Since Chinese prosody does not depend on change of pitch value *per se*, our experiment has the added value of being able to show that it is the linguistic concept of prosody that plays the central role in processing scalar inferences. In particular, the three experiments of the study attempt to show whether prosody interacts with contextual information in the processing of scalar inferences.

The critical stimuli of the three experiments in this study are sentences with the structure exemplified in (2). A prosodic stress is superimposed either on the numeral-classifier constituent or on the noun of the numeral phrase, as shown in the bolded sections. The stimuli were produced by a female native Mandarin speaker, who speaks only Beijing Mandarin without other dialects.

Although other numeral-classifier languages such as Japanese can rely on morphology to distinguish the two types of scalar inferences, it has been noted that the elements attached by a scalar particle, such as the noun or the numeral-classifier unit of a numeral phrase, carry an emphatic prosody (Nakanishi, 2006). In the setting of a quantity contrast, the numeral-classifier unit is stressed; in the setting of a type contrast, the noun is stressed. Therefore, our intuition suggests that the prosody in (2)a should be more likely to evoke a quantity contrast (i.e., an interpretation like “I didn’t even see one cat, let alone two cats, three cats, etc.”), whereas the prosody in (2)b should be more likely to evoke a type contrast (i.e., an interpretation like “I didn’t even see one cat, let alone one person, one bird, etc.”). The purpose of the present study was to see whether this intuition is supported by empirical data from naïve listeners.

Each experiment has a different task for the participants to respond to the stimuli. In Experiment 1, the participants were asked to judge whether the sentence which they heard from an audio clip was consistent with a paragraph they read previously, which set up a context consistent with either a quantity contrast or a type contrast. The design of Experiment 2 is the same as that of Experiment 1, but the participants were asked to give consistency ratings on a Likert scale rather than binary judgments of consistency. In Experiment 3, the participants read a context and then heard two auditory versions of the sentence with different prosody, and were instructed to select the version that better fit the context. The three tasks were made to test whether prosody is a determinant of the types of scalar reasoning and how much Mandarin speakers are aware of prosody. The results can help to validate the associations of the unconnected pieces in the literature of focus, prosody, and pragmatic inferences.

## EXPERIMENTS

We performed three experiments involving reading and listening to texts, with slightly different procedures, to test how participants evoke scalar implicatures based on the available information.

### Experiment 1: Matching Scalar Inferences

The first experiment is designed to test whether prosody would help Mandarin native speakers to determine types of scalar inferences when there is no distinction in the grammatical form. In the experiment, the participants had to read a short paragraph and to listen to a sentence. They had to judge whether the sentence they heard matched the provided context.

#### Participants

Sixty-nine native speakers of Mandarin (60 users of traditional Chinese characters and 9 users of simplified Chinese characters)

(2)	(a)	jintian	maomi	kafeiguan	mei	kai,	<b>yi</b>	<b>zhi</b>	maomi	dou	mei	you
	(b)	jintian	maomi	kafeiguan	mei	kai,	yi	zhi	<b>maomi</b>	dou	mei	you
		today	cat	café	NEG	open	one	CLF	cat	FOC	NEG	exist

‘The cat café is closed today. There isn’t even one cat.’

were included in the first experiment. Two were removed from analysis because they did not correctly respond to baseline questions (see section “Procedure”), leaving 67 participants (aged 20–60, mean 31) in the final analysis.

## Materials

The experimental stimuli comprised 12 sentences along with 16 fillers. A short paragraph was provided to set up the relevant context for each stimuli sentence, as shown in (3). The sentence always referred to some set that did not have some property. In (3), for example, the dog park does not have dogs, which should be expected to be most likely encountered in the defined setting. Another type of noun, human being, is involved as an alternative to be contrasted with dogs in this setting. Each context paragraph either indicated that the most likely property is not present but the other one is (e.g., the park did not have dogs but did have people, as in 3a), or that both properties are not present (e.g., the park had neither dogs nor people there). Finally, it introduced a speaker about to say the critical sentence.

- (3) *Zhangwei dao le youmingde liugou gongyuan, pingchang zheli henduo gou.* ‘Zhangwei went to a famous dog park. Usually there were a lot of dogs.’

- (a) *Jintian meiyou gou que you ren zai gongyuan li sanbu*  
‘Today there were no dogs, but there were people walking in the park.’ [Yes-context]  
(b) *Jintian gongyuan li meiyou gou ye meiyou ren*  
‘Today there were neither dogs nor people in the park’ [No-context]

*Zhangwei huilei hou gen ni shou:* ‘Zhangwei came back and told you:’

The experimental stimuli appeared in the syntactic format of (1). The context paragraph (3) was presented in Chinese characters, and the critical sentence (4) presented auditorily afterward:

- (4) *Wo jiantian qu le liugou gongyuan,* ‘I went to the dog park today.’

- (a) **yi zhi** gou dou mei kandao  
one CLF dog FOC **NEG** see  
(b) yi zhi **gou** dou mei kandao  
one CLF dog FOC **NEG** see  
‘I did not see even one dog.’

The two versions of the audio files both express the lack of a specific property which is the most expected in the defined set, e.g., the speaker did not see even one dog at the dog park. The only difference is that one has the prosodic stress on the numeral-classifier combination (4a), while the other stresses the noun (4b). For ease of reference, we refer to the former as quantifier stress, and the latter as noun stress. Based on the four conditions, the experiment followed a  $2 \times 2$  design: PROSODIC STRESS (*noun stress* vs. *quantifier stress*)  $\times$  CONTEXT (*type alternative present* vs. *type alternative absent*). The items were organized into four lists in a Latin square design.

The fillers can be divided into two groups. The first group contains six sentences which mismatch the content from the audio files. There are three types of mismatches including number, quantity, and location. One example of number mismatch is provided in (5), where the context and the critical sentence are unambiguously semantically inconsistent. The fillers both serve as a check that the received data are valid, and to distract participants from the experimental manipulation.

- (5) Reading context: *Mama qie le san ge pingguo, danshi meiyou chi. Baba gen ni shuo:* ‘Mom cut three apples, but didn’t eat them. Dad told you.’  
Audio context: *Mama yi ge pingguo dou mei qie.* ‘Mom did not cut even one apple.’

The other group of fillers consists of 10 sentences from another experiment for investigating the scalar implicatures from Mandarin *youxie* ‘some.’ The full list of stimuli is available at <https://osf.io/nsqfv/>.

## Prediction

In the context where the type alternative is present (3)a, we expected that the critical sentence with prosodic focus on the noun, compared to the critical sentence with prosodic focus on the numeral and classifier, would be less consistent with the context. This is because prosodic focus on the noun (i.e., “I didn’t even see one *dog*”) should license the inference that the speaker didn’t see anything else either, including the type alternative (i.e., “I didn’t even see one *dog*, let alone one person”). Thus, we expected a difference in consistency ratings between the two prosodic conditions in this context. On the other hand, in the context where the type alternative is absent (3)b, we expected no difference in consistency ratings between the two prosodic conditions, since both inferences (i.e., “I didn’t even see one dog, let alone one person” and “I didn’t even see one dog, let alone two”) are consistent with the context in which there are neither dogs nor people in the park.

## Procedure

This experiment was administered online via Ibex Farm (<http://spellout.net/ibexfarm/>). At the beginning of the experiment, participants indicated their consent to participate, provided demographic information (age, sex, and native language), and answered two questions about the experiment meant to probe whether they had read the instructions. One question was which university the experiment was being run by, and the other question was how many trials there would be in the experiments. The 12 items along with 16 fillers were then randomly presented in a Latin square design after three practice trials. For each trial, participants read a short Mandarin paragraph which either established a context where a contrasting type is present (e.g., (3)a, in which there are no dogs but there are people), or a context where the contrasting type is absent (e.g., (3)b, where there are neither people nor dogs in the park). When they finished reading at their own pace, they then clicked a button to listen to a sentence relevant to this setting with either a stressed noun or a stressed numeral-classifier combination. The task for the participants was to judge whether what they heard



could fit the context that they read. They were asked to click either *consistent* or *inconsistent* based on their own judgments. After submitting the answer, a participant could move onto the next question. The whole survey was self-paced. It took less than 30 min for the participants to finish the survey.

## Results

The full dataset and analysis code (for the R statistical programming environment) are available at <https://osf.io/nsgfv/>. Overall, in contexts where the type alternative referent was present, participants accepted 85.6% of sentences with quantifier stress and 82.1% of items with noun stress, a difference in the expected direction; also consistent with the predictions, they showed less difference in acceptance of different prosody in the context where the alternative referent is also not available, accepting 89.6% of sentences with quantifier stress and 90.0% of sentences with noun stress. **Figure 1** shows the variability of the effect across items (by-subject aggregates are not plotted; since each participant only saw a small number of items and thus only had a small number of possible outcomes per condition [0, 33, 66, or 100%], there is little subject-wise variability to be seen). In **Figure 1**, because the prediction was that there would be a larger prosody effect (in the negative direction) in these context than in contexts where the alternative is available, this means that points below the diagonal represent items showing effects

consistent with the prediction, and points above the diagonal are inconsistent with the prediction.

The results were statistically analyzed using generalized (binomial) mixed-effects models with crossed random effects for subjects and items (Baayen et al., 2008). The predictors PROSODIC STRESS (*noun stress* vs. *quantifier stress*) and CONTEXT (alternative type present vs. alternative type absent) were sum-coded (as 0.5 and -0.5) and used as fixed predictors, along with their interaction; random effects of these three parameters were also fit for items (Barr et al., 2013), but not for subjects, since each subject had too few trials to fit this complex structure well. The significance of the crucial PROSODIC STRESS \* CONTEXT interaction was assessed with a log-likelihood test comparing this model to a maximally similar model without the fixed interaction effect. The interaction did not reach significance in this comparison [ $\chi^2(1) = 0.14, p = 0.707$ ].

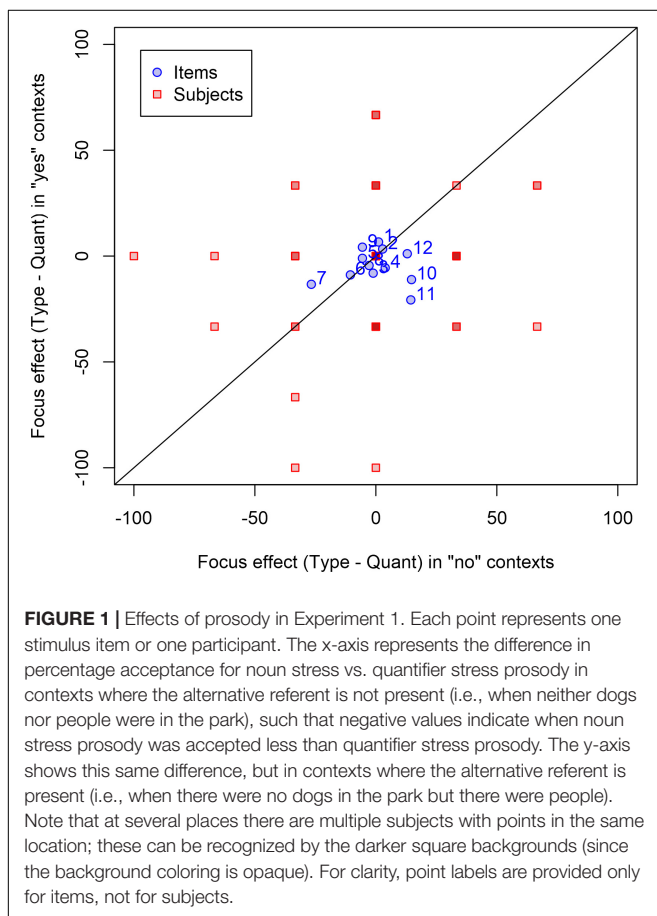
The results of the experiment showed a numerical trend in the predicted direction, such that prosody influenced sentence acceptability in contexts where the alternative type was present and less so in contexts where the alternative type was absent. However, this trend was not statistically significant. Furthermore, even in contexts where prosody should have elicited an inference that does not fit the context (i.e., “I didn’t even see one *dog* [let alone one person],” in a context where there were no people in the park), sentence acceptance was still quite high, over 80%; this suggests that participants were not influenced very much by prosody, as long as the lexico-semantic content of the sentence fit the context. For this reason, we attempted to conceptually replicate the experiment, while making changes to potentially increase the size of the effect. We suspected that the binary nature of the acceptability judgment may have forced participants to ‘accept’ sentences even when they were aware of slight inconsistencies; thus, in this experiment we instead had participants rate sentences on a six-point Likert scale, which we predicted might allow them to register their awareness of the prosodic mismatch and thus might increase the chances of observing a prosodic effect. Otherwise, the predictions for Experiment 2 are the same as for Experiment 1: we expect worse ratings for noun-stress prosody than for quantifier-stress prosody in contexts where the alternative type is present, but not in contexts where the alternative type is absent.

## Experiment 2: Rating Scalar Inferences

The procedure of the second experiment is the same as that of the first experiment, except that in this experiment participants had to rate to what extent the inferences from the audio contents were consistent with the provided contexts, rather than making a binary judgment. The predictions are the same as in Experiment 1.

## Participants, Materials, and Procedure

Seventy-eight native speakers of Mandarin (63 users of traditional Chinese characters, 15 users of simplified Chinese characters) took part in this experiment. Ten were excluded for answering baseline questions incorrectly; the exclusion criteria and data collection stopping rule were pre-registered at <https://osf.io/bz6c2/register/5771ca429ad5a1020de2872e>. This



left 68 participants (aged 18–70, mean 42) in the final analysis. The materials are the same as those from Experiment 1.

The procedure was the same as in Experiment 1, except that the task for the participants in Experiment 2 was to rate the consistency between what they heard and what read based on a 1–6 scale. 1 stood for *completely inconsistent*, while 6 stood for *completely consistent*. The participants were guided to go through two practice trials: one practice is an example of *completely inconsistent*, while the other is a practice of *completely consistent*, before starting the experiment. The example of *completely consistent* is provided in (6), where the audio content emphasized the quantity of water and has no conflicts with the written content.

- (6) Written content: *Huang laoshi tongchang he henduo sui. Ta jintian hen mang. Ta mei he sui ye mei he kele. Ta de xuesheng gen ni shuo: 'Mr. Huang usually drank a lot of water. He was very busy today. He drank neither water nor coke. His student said:'*

Audio content: *Ta yi di sui dou mei he. 'He didn't eat even one drop of water.'*

## Results and Discussion

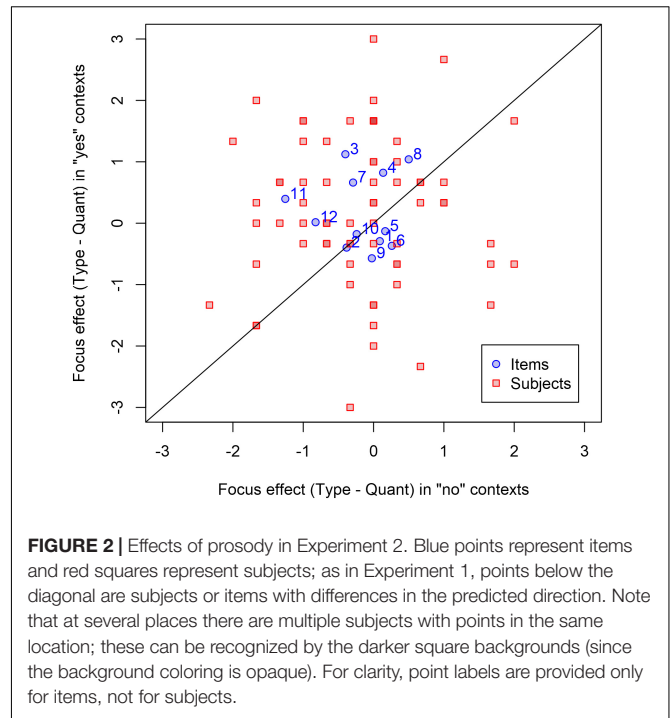
The full dataset and analysis code are available at <https://osf.io/nsgfv/>.

In contexts where the alternative type is present, the mean consistency rating was 4.9 for sentences with noun-stress prosody and 4.7 for sentences with quantity-stress prosody; this difference is opposite the predicted direction. In contexts where the alternative type is absent, consistency ratings were 5.2 for noun-stress prosody and 5.4 for quantity-stress prosody. In both contexts, the mean consistency rating was fairly high. The distribution of differences by subjects and items is shown in **Figure 2**. Since the effects were opposite the predicted direction, inferential statistics were not conducted.

Experiment 2 failed to replicate the trend observed in Experiment 1. We suspected that the effects of prosody may have been weakened or obscured in these experiments by two factors. First, the experimental contexts did not particularly draw participants' attention to prosody, and in fact may have drawn their attention more to lexico-semantic factors. Since the experiment included fillers in which the target sentence clearly mismatched the context based on basic semantics as in (7), many participants' attention may have been focused more on these issues. This type of filler may have become the standard for *completely inconsistent* for participants. Therefore, participants may have considered sentences with inconsistent prosody but consistent semantics to be fairly acceptable by comparison.

- (7) Reading content: *Wangfang shi chuan le liang jian yifu, zuihou meiyou mai. Dianyuan gen ni shuo: 'Fang Wang tried on two pieces of clothes. She didn't buy any. The shop assistant told you:'*

Audio content: *Ta yi jian yifu dou mei shi chuan. 'She didn't try on even one piece of clothes.'*



**FIGURE 2 |** Effects of prosody in Experiment 2. Blue points represent items and red squares represent subjects; as in Experiment 1, points below the diagonal are subjects or items with differences in the predicted direction. Note that at several places there are multiple subjects with points in the same location; these can be recognized by the darker square backgrounds (since the background coloring is opaque). For clarity, point labels are provided only for items, not for subjects.

Secondly, Experiments 1 and 2 tested the effects of prosody on inferences indirectly, by testing whether prosody engenders an inference which mismatches a context (rather than by directly testing whether prosody engenders a given inference at all). In these two experiments, both prosody and lexico-semantic contents might influence the participants' judgments. Thus, the results are not merely reflective of prosody. In Experiment 3 we attempted to address these issues by using a more direct approach, and by using a design meant to explicitly draw participants' attention to prosody.

## Experiment 3: Comparing Types of Intonation

The experiment is designed to force participants to focus on prosody by providing different prosodic patterns and minimizing contextual information. In this experiment, participants had to listen to two sentences which differ in intonation. Afterward, they had to choose which sentence matched the provided context.

### Participants, Materials, and Procedure

Sixty-four native speakers of Mandarin (63 users of traditional Chinese characters, 1 user of simplified Chinese characters) attended this experiment. Eleven were excluded for answering baseline questions incorrectly, and four for having low accuracy in the unambiguous filler trials. This left 49 participants (aged 18–60, mean 26) in the final analysis.

The experiment consists of 12 critical sentences along with 6 fillers. The critical sentences are in the format of (2). The participants were asked to listen to the same sentence in two kinds of prosodic patterns: one with stress on the noun (e.g., *In the cat café I didn't see even one cat*), one with stress on the

numeral-classifier combination (e.g., *In the cat café I didn't see even **one** cat*). Afterward, they were asked to choose the most appropriate answer to be the first clause of a two-clause sentence. The question appears in the format as (8)a or (8)b. (8)a provides an alternative in the category of types, whereas (8)b offers an alternative in the domain of quantity.

- (8) (a) \_\_\_\_\_, gengbieshuo you guke le  
much less there be customer ASP  
'\_\_\_\_\_, much less customers.' [an alternative in type]  
(b) \_\_\_\_\_, gengbieshuo you yi qun maomile  
much less there be one CLF cat ASP  
'\_\_\_\_\_, much less a group of cats.' [an alternative in quantity]

This version of the experiment only had two conditions: follow-up contexts which stress the type alternative, and follow-up contexts which stress the quantity alternative. We predicted that sentences with stress on the noun (consistent with type focus) would be selected more often when the follow-up sentence stresses the type alternative (8)a than when it stresses the quantity alternative (8)b. The items were organized into two lists in a Latin square design. There were six fillers, which also appear in the same format.

Among the fillers, three of them were in positive polarity environments, and three of them are in negative polarity environments. For each trial, two audio files were provided: one option matches the follow-up context (9)a, while the other mismatches the follow-up context (9)b.

- (9) \_\_\_\_\_, genbieshuo xiao gongyu le.  
'\_\_\_\_\_, much less a small apartment.'  
Audio files:  
(a) Match: *Ta mai de qi chengshi li de da haozhai...* 'He can afford a mansion in the city...'  
(b) Mismatch: *Ta chi de qi niupai...* 'He can afford steaks...'

The fillers, which were unambiguous, also served to check the validity of the responses. The full list of stimuli is available at <https://osf.io/nsgfv/>.

This experiment was administered via Ibex Farm. The 12 critical items along with 6 fillers were presented in a fully random order after two practice trials. The practices were designed to direct participants' attention to prosodic differences. As in (10), the two audio files have the same format, but the placement of a contrastive stress determined the item to be contrasted. According to the written context provided in (10), only the prosody of (10)a can match the follow-up sentence.

- (10) \_\_\_\_\_, bu shi Xiaohan hui.  
'\_\_\_\_\_, not Xiaohan who is able to.'  
Audio files:  
(a) *wo zhidao **ta** hui tiaowu* 'I know it is she who is able to dance.' [contrastive stress on *ta* 'she']  
(b) *wo zhidao ta **hui** tiaowu* 'I know it is cooking that she is able to do.' [contrastive stress on *hui* 'be able to dance']

For each trial, with a written context sentence and two audio clips occurred on the screen at the same time. The task for the

participants is to choose one of two audio clips to complete the sentence shown on the screen, which only the second clause of a two-clause sentence is provided. Participants could play the audio clips more than one time. The self-paced survey took less than 30 min to finish.

## Results

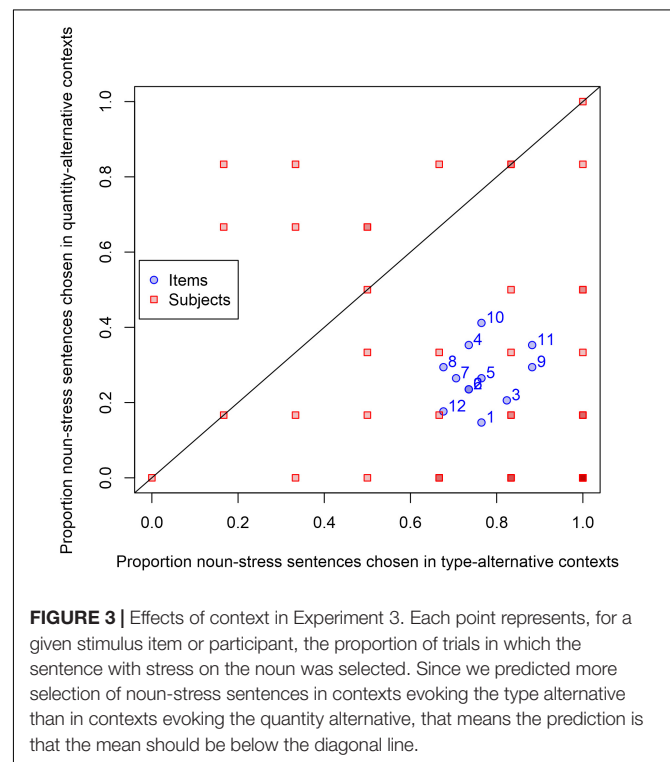
The data and analysis code are available at <https://osf.io/nsgfv/>.

As shown in **Figure 3**, sentences with noun stress were chosen more often in contexts that evoked the type alternative than in contexts that evoked the quantity alternative; conversely, sentences with quantifier stress were chosen more often in contexts that evoked the quantity alternative than contexts that evoked the type alternative.

The context effect was analyzed with a generalized (binomial) mixed-effects model regressing the binary response (coded with quantifier stress as the baseline level) on the fixed effect of context (dummy-coded with quantity-alternative contexts as the baseline level) and maximal random effects for subjects and items. This model revealed a significant effect of context ( $b = 358$ ,  $z = 5.83$ ,  $p < 0.001$ ), indicating that the likelihood of selecting a sentence with stress on the noun was significantly higher in type-alternative contexts than in quantity-alternative contexts.

## GENERAL DISCUSSION

This study tested whether prosody plays a role in pragmatic judgments, especially in terms of differentiating interpretational ambiguity of scalar implicatures. The connections among focus,



prosody, and pragmatic inferences have been hinted in different literature (Chao, 1968; Haspelmath, 1997; Zhang, 2000; Tsai, 2004; Lee, 2006; Nakanishi, 2006; Israel, 2011), and prosody is known to influence utterance interpretation in other kinds of structures (e.g., Snedeker and Trueswell, 2003; Chevallier et al., 2010; Degen and Tanenhaus, 2015; among others), but the relations have not yet been specified for ambiguous alternative sets invoked by minimizers. In order to find empirical evidence for such an influence, we tested whether the type-contrast and quantity-contrast prosodic patterns can guide Mandarin native speakers to the correspondent scalar inferences. The results of the experiments suggest that Mandarin native speakers may use prosody to inform their interpretations of minimizers, but not necessarily in all contexts. The stimuli appear in the same syntactic structure, which has the numeral 'one' and the classifier specified. This syntactic pattern inherently entails the semantics of quantity. According to the participants' responses, they tend to use the quantity contrast for this syntactic structure if the experiment design does not strongly draw their awareness to the prosodic changes. The expected effects of prosody were not strong in Experiment 1, and not present at all in Experiment 2; this may have been because in this setting the participants' attention was drawn to syntax and semantics than to prosody. In this case, the participants judged the consistency between what they heard and what they read based on the quantity-contrast inferences. The pattern of results in the first two experiments also suggests that this minimizer structure, with the quantity specified, strongly prefer a quantity-contrast interpretation. However, when the role of prosody was tested in a design that more directly addressed alternative interpretations of the minimizer and that draw participants' attention more explicitly to prosody as in Experiment 3, then participants' judgments of scalar inferences were heavily influenced by the patterns of prosody, in the direction we had predicted. This suggests that prosody is a factor which Mandarin speakers use to identify alternative sets when interpreting minimizers. These results suggest that prosody has an influence not only on structural disambiguation (in cases where utterances may be parsed into multiple syntactic or semantic structures) and the choice of whether to apply an implicature at all (in cases where utterances may be interpreted with or without a conversational implicature), but also on what alternatives the same implicature operates over.

The experiments also provide the evidence for the observed connection between prosodic stress and minimizers in the literature. The occurrence of a prosodic stress contributes to inducing a set of pragmatic inferences coherent with contexts. The placement of a prosodic stress is an indicator of where the attention of the native Mandarin speakers would be. This relation between prosodic prominence and loci of attention helps to account for the concept of focus in the syntax in Mandarin Chinese.

## CONCLUSION

The present study provided evidence that different prosodic patterns can guide hearers to induce different scalar reasoning.

It has been observed that 'one'-phrases minimizers in numeral-classifier languages have two types of scalar inferences due to the structure as a numeral phrase (Lee, 2006; Nakanishi, 2006). The two types of inferences, quantity-contrast and type-contrast, are reflected in morphology in other numeral-classifier languages, but not in Mandarin.

It has been noted that the numeral 'one' in Mandarin minimizers may bear a stress, but the actual loci of the stress and the purpose the stress were not specified. The experiments provide evidence that the locus of a prosodic stress can carry pragmatic information play a role in evoking alternatives during sentence comprehensions. In conditions where syntax, semantics, and morphology do not differentiate types of scalar inferences, prosody can help native Mandarin speakers to determine the entailed conceptual scale. However, according to the results of Experiments 1 and 2, the role of contextual information may sometimes override that of prosody in determining scalar inferences.

Although the placement of a prosodic stress specifies the types of inferences, the induced scalar inferences are asymmetrical as shown in the results of experiments. The quantity-contrast inferences involve choosing from a set of alternatives that are already lexically entailed by the minimizer. On the contrary, type-contrast inferences require choosing from an open set of nouns and an open set of conceptual scales, which is highly dependent on the context: i.e., there is no natural ordered ranking or entailment relationship between cats and people; in some contexts cats may be less likely than people (and the presence of cats may entail the presence of people), in other contexts the opposite may be true, and in still other contexts they may have no such relationships at all. Thus, the role of prosody in the contexts of type-contrasts is more difficult to test in a controlled fashion because it is difficult to predict which specific alternatives will be ruled out by this interpretation across different interlocutors and contexts.

In terms of our experimental results, what is the precise role of prosody in the processing of scalar implicatures in Mandarin Chinese? Note that Experiments 1 and 2 suggest that scalar implicatures are strongly defaulted to quantity type inferences regardless of the potential ambiguity. Hence Experiment 3 is the critical one that shows the effect of prosody on interpretation and the effect is the over-riding of the default. As there is no reason to believe that the prosodic stress directly encodes either interpretation, a likely explanation is that stress brings attention to a typically less likely interpretation, such as flagging or underlining parts of a text. It is possible that type-contrast is more cognitively costly to realize, as it requires generating a context-dependent set of alternatives, as opposed to the lexically-encoded set of alternatives (i.e., "not one" entails "not two," "not three," "not four," etc.) used for quantity-contrast. If that is the case, participants may avoid realizing a type contrast unless either the contrast is made less cognitively costly [e.g., if specific alternatives are made salient in the preceding contrast; relatedly, experiments have suggested that *ad hoc* scalar implicatures can be realized with little processing cost if the *ad hoc* scale is already contextually salient (Breheny et al., 2013; Politzer-Ahles and Fiorentino, 2013)] or if additional cues give



them evidence that this contrast is particularly relevant and thus worth the effort. If this is the case, the prosodic cues may act to trigger the additional processing of potential type inferences.

The experiments of this study show that prosody can play a role in influencing the kind of scalar inferences that are induced by a minimizer in Mandarin. The prosodic conditions are considered when the non-default type inference needs to be processed. Hence, the effects of prosody on determining types of scalar inferences can be diminished by contextual information. The types of scalar inferences in Mandarin are determined by how prosody and contextual information interact.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Hong Kong Polytechnic University, Human Subjects Ethics Sub-committee. The protocol was approved by the Hong Kong Polytechnic University, Human Subjects Ethics

Sub-committee. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

## AUTHOR CONTRIBUTIONS

I-HC and SP-A contributed to conception and design of this study. I-HC executed the experiments and analyzed the results. SP-A advised the design of experiments, supervised the procedure of experiments, and performed the statistical analysis. C-RH contributed to the linguistic theory and interpretation of the results of this study. I-HC wrote the first draft of the manuscript. All authors contributed to manuscript revision and approved the submitted version.

## FUNDING

This work was partially supported by the following grants from the Hong Kong Polytechnic University: 1-YW1V and G-UAET.

## REFERENCES

- Ahrens, K., and Huang, C. R. (2016). "Classifiers," in *A Reference Grammar of Chinese*, eds C. R. Huang and D. X. Shi (Cambridge, MA: Cambridge University Press), 169–198. doi: 10.1017/CBO9781139028462.008
- Baayen, H., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005
- Barr, D., Levy, R., Scheepers, C., and Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001
- Breheny, R., Ferguson, H., and Katsos, N. (2013). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition* 126, 423–440. doi: 10.1016/j.cognition.2012.11.012
- Chao, Y. R. (1933). Tone and intonation in Chinese. *Bull. Instit. History Philol. Acad. Sin.* 4, 121–134.
- Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley: Press.
- Chevallier, C., Bonnefond, M., Van der Henst, J.-B., and Noveck, I. (2010). Using ERPs to capture prosodic stress and inference making. *Ital. J. Linguist.* 22, 125–152.
- Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Struct. Beyond* 3, 39–103.
- Chierchia, G., Fox, D., and Spector, B. (2012). "Scalar implicature as a grammatical phenomenon," in *Semantics: An International Handbook of Natural Language Meaning*, Vol. 3, eds C. Maienborn, K. von Stechow, and P. Portner (Berlin: Mouton de Gruyter), 2297–2331.
- Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171
- Fauconnier, G. (1975). Pragmatic scales and logical structure. *Linguist. Inq.* 6, 353–375.
- Fillmore, C. J., Kay, P., and O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language* 64, 501–538. doi: 10.2307/414531
- Giannakidou, A. (2011). "Positive polarity items and negative polarity items: variation, licensing, and compositionality," in *Semantics: An International Handbook of Natural Language Meaning*, eds C. Maienborn, K. von Stechow, and P. Portner (Berlin: Mouton de Gruyter), 1660–1712.
- Grice, P. (1989). *Studies in the Way of Words*. Cambridge: Harvard University Press.
- Haspelmath, M. (1997). *Indefinite Pronouns*. New York, NY: Oxford University Press.
- Horn, L. R. (1989). *A Natural History of Negation*. Chicago: The University of Chicago Press.
- Horn, L. R., and Ward, G. (2005). *The Handbook of Pragmatics*. Hoboken, NJ: Wiley-Blackwell. doi: 10.1111/b.9780631225485.2005.x
- Huang, C. T., Li, Y. H., and Li, Y. (2009). *Syntax of Chinese*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139166935
- Israel, M. (2011). *The Grammar of Polarity: Pragmatics, Sensitivity and the Logic of Scales*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511975288
- Lee, C. (2006). "Contrastive topic/focus and polarity in discourse," in *Where Semantics Meets Pragmatics*, eds K. von Stechow and K. Turner (Oxford: Elsevier Science), 381–420.
- Morris, C. (1938). *Foundations of the Theory of Signs*. Vol. 1, Chicago: University of Chicago Press.
- Nakanishi, K. (2006). "The semantics of even and negative polarity items in Japanese," in *Proceedings of the 25th West Coast Conference in Formal Linguistics (WCCFL 25)*, eds D. Baumer, D. Montero, and M. Scanlon (Somerville, MA: Cascadia Press), 288–296.
- Politzer-Ahles, S., and Fiorentino, R. (2013). The realization of scalar inferences: context sensitivity without processing cost. *PLoS One* 8:e63943. doi: 10.1371/journal.pone.0063943
- Shih, C. (1988). Tone and intonation in Mandarin. *Work. Pap. Cornell Phon. Lab.* 3, 83–109.
- Snedeker, J., and Trueswell, J. (2003). Using prosody to avoid ambiguity: effects of speaker awareness and referential context. *J. Mem. Lang.* 48, 103–130. doi: 10.1016/S0749-596X(02)00519-3
- Tsai, W. T. D. (2004). On formal semantics of zhi and lian in Chinese. *Zhouguo Yuwen* 2, 99–111.
- Yuan, J. (2004). *Intonation in Mandarin Chinese: Acoustics, Perception, and Computational Modeling*. Ph.D. dissertation: Cornell University: Ithaca, NY.
- Zhang, N. (2000). Object shift in Mandarin Chinese. *J. Chin. Linguist.* 28, 201–246. doi: 10.1007/s10936-015-9394-y

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Chen, Huang and Politzer-Ahles. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Processing Presuppositions and Implicatures: Similarities and Differences

Cory Bill<sup>1,2</sup>, Jacopo Romoli<sup>3\*</sup> and Florian Schwarz<sup>4</sup>

<sup>1</sup> Department of Cognitive Science, ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Sydney, NSW, Australia, <sup>2</sup> Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), Berlin, Germany, <sup>3</sup> School of Communication and Media, Ulster University, Belfast, United Kingdom, <sup>4</sup> Linguistics Department, University of Pennsylvania, Philadelphia, PA, United States

## OPEN ACCESS

### Edited by:

Penka Stateva,  
University of Nova Gorica, Slovenia

### Reviewed by:

Chris Cummins,  
University of Edinburgh,  
United Kingdom  
Emmanuel Chemla,  
UMR8554 Laboratoire de Sciences  
Cognitives et Psycholinguistique  
(LSCP), France  
Raj Singh,  
Carleton University, Canada

### \*Correspondence:

Jacopo Romoli  
j.romoli@ulster.ac.uk

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Communication

**Received:** 01 May 2018

**Accepted:** 19 September 2018

**Published:** 11 October 2018

### Citation:

Bill C, Romoli J and Schwarz F (2018)  
Processing Presuppositions and  
Implicatures: Similarities and  
Differences. *Front. Commun.* 3:44.  
doi: 10.3389/fcomm.2018.00044

Presuppositions and scalar implicatures are traditionally considered to be distinct phenomena, but recent accounts analyze (at least some of) the former as the latter. All else being equal, this “scalar implicature approach to presuppositions” predicts uniform behavior for the two types of inferences. Initial experimental studies comparing them yielded conflicting results. While some found a difference in the Response Time (RT) patterns of scalar implicatures and presuppositions, others found them to be uniform. We argue that the difference in outcomes is attributable to a difference in the type of response being measured: RTs associated with acceptance and rejection responses seem to pattern in opposite ways. Next, we report on a series of experiments to support this, and to compare the behavior of the two inferences more comprehensively. Experiments Ia and Ib look at both acceptance and rejection responses for both inference types, and find uniform patterns once the acceptance vs. rejection variable is factored in. Experiment II adds a new dimension by testing for the influence of prosody on the two inference types, and in this regard a clear difference between them emerges, posing a first substantive challenge to the scalar implicature approach to presuppositions. A third set of experiments investigates yet another prediction of this approach, according to which the presuppositional inference is introduced as a simple entailment in affirmative contexts. This predicts that these presuppositional inferences behave parallel to other entailments. Experiment IIIa compares rejections of affirmative sentences based on either their presuppositional inference or their entailed content and finds that they differ, with greater RTs for the former. As an additional control, Experiments IIIb and IIIc test for parallel differences between two entailments associated with *a/ways*, which yield uniform results. In sum, while Experiments Ia and Ib are in line with previous findings that presuppositions and scalar implicatures under negation show uniform response time patterns, the differences found in Experiments II and IIIa-c pose a substantial challenge to approaches assimilating the two phenomena, while being entirely in line with the traditional perspective of seeing them as distinct.

**Keywords:** scalar implicature, presupposition, inference, processing, semantics, pragmatics

# 1. INTRODUCTION

This paper experimentally compares two central linguistic inference types, namely Presuppositions (Ps) and Scalar Implicatures (SIs). Traditional approaches treat these as entirely distinct categories (Heim, 1982; van der Sandt, 1992; Beaver, 2001, among many others). But recent approaches, building on a line of work going back to Gazdar (1979) and Wilson (1975) (among others), analyze at least certain presuppositions as scalar implicatures, largely motivated by the need to account for varying behavior of different presupposition triggers (Simons, 2001; Abusch, 2002, 2010; Chemla, 2009, 2010; Abrusán, 2011; Romoli, 2012, 2015)<sup>1</sup>. We begin by sketching one form of this overall approach, directly assimilating scalar implicatures and presuppositions, which we refer to as the “SI approach to Ps,” and whose two core properties are schematized in (1-a) and (1-b)<sup>2</sup>.

## (1) Properties:

- a. In affirmative contexts, Ps are simply entailments<sup>3</sup>.
- b. In all other contexts (e.g., under negation), Ps are derived as SIs.

To illustrate (1-a), the presuppositional inference in (2-b) arising from (2-a), is a simple entailment according to this approach, just as (3-b) is an entailment of (3-a)<sup>4</sup>.

- (2) a. John stopped going to the movies.  
b.  $\leadsto$  *John used to go to the movies*
- (3) a. John always went to the movies.  
b.  $\leadsto$  *John sometimes went to the movies*

Turning to the property in (1-b), the inference in (4-b), arising from the sentence in (4-a), is derived as an SI in contexts like negation, parallel to the derivation of (5-b) from (5-a).

- (4) a. John didn’t stop going to the movies.  
b.  $\leadsto$  *John used to go to the movies*
- (5) a. John didn’t always go to the movies.  
b.  $\leadsto$  *John sometimes went to the movies*

Two predictions that follow from the properties above are (6-a) and (6-b):

- (6) **Predictions:** All else being equal,  
a. in affirmative contexts, Ps and entailments should display uniform behavior.  
b. in all other contexts, Ps and SIs should display uniform behavior.

We tested these predictions by comparing Ps to simple entailments, on the one hand, and to SIs, on the other. Specifically, we focus on the predictions in (6), in order to answer the question in (7). A positive answer to this question would be challenging for a unified approach to SIs and Ps, at least in its strongest version<sup>5</sup>.

- (7) **Main question:** Do behavior patterns yield evidence for a distinction between Ps and entailments in affirmative contexts and between Ps and SIs in other contexts?

Previous studies in the literature have focused on the prediction in (6-b), comparing SIs and Ps directly, and have produced results that run against this prediction, based on delays in Response Times (RTs) found for SIs (Bott and Noveck, 2004 and much subsequent work) on the one hand, and recent reports of the

<sup>1</sup>Note that such approaches commonly differentiate between different types of presupposition triggers, and only propose to treat the inferences of a sub-class of traditional presupposition triggers as implicatures. Given our focus on triggers in the relevant sub-class, we simply refer to them as Ps here.

<sup>2</sup>Many of the proposals in the literature mentioned above depart from this strong version of the approach to some extent, by re-introducing some elements of difference between implicatures and presuppositions (for instance, Chemla, 2010 assumes that they differ in the alternatives they involve and their discourse properties, while Romoli, 2015 argues that there is a difference between the two in terms of obligatoriness of the inference). These elements might affect the predictions in relation to the properties in (1-a) and (1-b) in different ways. We think that it is nonetheless useful to test experimentally the prediction of the strongest and most ambitious version of the approach and then take the results of that as a quantitative base to evaluate if and where a departure is needed from simply assimilating scalar implicatures and presuppositions. Recent pragmatic accounts to presuppositions like that in Schlenker (2008) also derive them in terms of conversational reasoning, though not equating them with scalar implicatures. This type of account makes non-trivial predictions in relation to the processing of presuppositions. Despite this distinction, we group it with the “traditional approach” here and leave explorations of these predictions for further research.

<sup>3</sup>Traditional accounts are compatible with the assumption that presuppositional inferences in affirmative contexts are entailments, in addition to being presupposed, though this isn’t necessarily extended to all presupposition triggers (see Sudo, 2012 for discussion).

<sup>4</sup>The entailment from (3-a) to (3-b) actually involves some complications: in order for it to go through one has to assume that the restrictor of the universal quantifier *always* is non-empty. We leave this aside here, as it is orthogonal to our purposes; for discussion see Heim and Kratzer (1998, chapter 6).

<sup>5</sup>Let us emphasize here the “all else being equal” element of these predictions. That is, these predictions are only claimed to apply in situations where the properties of the relevant meanings are as close to each other as possible. This is important as it increases the likelihood that any difference in the behavior patterns of the inferences is genuinely a result of the inferences being of different types. In line with this, we compared triggers that are as similar to each other as possible. Moreover, we would note that in our experiments the nature of the *uniformity* predicted in (6-a) and (6-b) varies somewhat depending on how close the situation is to the ideal of *all else being equal*. For example, in Experiment Ia and Ib we compare the processing profiles of three inferences that, according to the SI approach to Ps, are all derived as SIs. Despite this common derivational mechanism, there are other dimensions on which the relevant triggers vary (e.g., presence of negation), as a result, we take the “uniformity” predicted by this approach to hold at a fairly general level. Specifically, for these experiments we test the prediction that, for each trigger, there will be uniformity in the general processing pattern produced when comparing responses motivated by an inference-based interpretation to responses based on a literal interpretation. At the beginning of each experiment we identify and justify the degree of behavioral uniformity predicted by the SI approach to Ps for the situation under investigation. Finally, in connection to the qualifications above, we also should make note of work on “scalar diversity” in the implicature literature, which has found differences across different scalar terms (Van Tiel et al., 2016, among others). The differences that have been found so far have chiefly been in the realm of inference derivation rates, but it is in principle possible for there to be within-inference variation in regards to other aspects of behavior as well. Nonetheless, when considering the strong version of the SI approach to Ps, outlined above, the differences we do find between SIs and Ps are not readily explained by scalar diversity. We will return to this later when discussing one such result, which is generated by Experiment II.

opposite pattern for Ps (Chemla and Bott, 2013). We begin our discussion below with a review of these findings and contrast them with some other recent results reported by Romoli and Schwarz (2015), which found uniform RT patterns for Ps and SIs. We then argue, following a similar point made by Cremers and Chemla (2014), that the source of the difference in the results on Ps could well be due to a confound, namely a difference in terms of the types of responses—acceptances vs. rejections—being measured.

This motivates the first series of experiments reported here, which further extend the comparison between SIs and Ps. The results from Experiments Ia and Ib reconcile the conflicts between previous findings and show that once we look systematically at both acceptance and rejection responses, the evidence for a difference between Ps and SIs in RTs disappears. Thus, comparisons of RT patterns of the sort first employed in the study of SIs, testing the prediction in (6-b), do not challenge the SI approach to Ps. However, Experiment II clearly differentiates the two inference types by looking at the impact of prosodic stress on the inference-triggering expressions, which yields opposite effects for SIs and Ps. This poses a first challenge to the SI approach to Ps. An additional finding from our RT studies is that we do not replicate the previously reported general delays associated with SIs (e.g., Bott and Noveck, 2004).

We then shift our attention to the prediction in (6-a) and report a third series of experiments that follow an approach presented in Kim (2007) and Schwarz (2016b). That is, these experiments look at rejections of sentences based on either their presuppositional inferences or their entailments. We find longer RTs for the former, which runs against the prediction in (6-a) and poses a second challenge to the SI approach to Ps.

In sum, the results of Experiment II and those of Experiment IIIa-c challenge the SI approach to Ps by revealing differences between them where this approach predicts uniform behavior. This is further corroborated by differences between SIs and Ps found in previous work on language acquisition and language disorders (Kennedy et al., 2014; Bill et al., 2016). The overall evidence, then, is not in line with the predictions of the SI approach to Ps, as outlined in (6-a) and (6-b).

The paper is organized as follows. In section 2, we present the theoretical background on SIs, Ps, and the SI approach to Ps. In section 3, we discuss previous work on the processing of SIs and Ps and in particular those results taken as evidence for a difference between Ps and SIs. In section 4, we report our new series of experiments and in section 5 we discuss their implications for our main question and the processing of SIs and Ps. Section 6 closes the paper with some general conclusions.

## 2. BACKGROUND

### 2.1. The Phenomena

Ps and SIs are inferences associated with certain expressions that go beyond the core lexically encoded, truth-conditional meaning. (8) and (9), repeated from above, illustrate inferences that are traditionally analyzed as Ps and SIs, respectively.

- (8) a. John didn't stop going to the movies.

- b.  $\leadsto$  John used to go to the movies

- (9) a. John didn't always go to the movies.

- b.  $\leadsto$  John sometimes went to the movies

We focus on cases like (8) and (9) in particular, as they are maximally parallel, at least on the surface, in involving negation. But we also consider more standard cases of SIs in affirmative sentences such as (10). Sometimes the SIs in (9) and that in (10) are distinguished terminologically as “indirect” and “direct” ones (Chierchia, 2004), and we will adopt this terminology<sup>6</sup>.

- (10) a. John sometimes went to the movies.

- b.  $\leadsto$  John didn't always go to the movies

One shared property of all these inferences is that they are not obligatorily present. In other words, in addition to “inference readings” illustrated above, all these sentences can have a “no-inference” reading as well, where the inference is absent. Consider (11) as compared to (8): the felicity of the continuation illustrates that the inference that John used to go to the movies is not necessarily present. The same goes for (12) and (13) and their inferences that John sometimes went to the movies and that he didn't always go, respectively.

- (11) John didn't stop going to the movies ... he never went!

- (12) John didn't always go to the movies ... (in fact) he never went!

- (13) John sometimes went to the movies ... (in fact) he always went!

This property, of course, is not shared by all inferences: in the case of a regular entailment like (14-b) of the sentence in (14-a), any attempt to suspend the inference, as in (15), results in infelicity, and the sentence sounds contradictory.

- (14) a. John and Mary went to the movies.

- b.  $\leadsto$  John went to the movies

- (15) John and Mary went to the movies ... # (in fact) John didn't go!

In light of this property any theory of SIs and Ps, unified or not, requires an account of (i) how these inferences arise to account for the inference readings, while (ii) also allowing for no-inference readings. In the next section, we briefly sketch how traditional approaches handle this challenge for SIs and Ps.

### 2.2. The Traditional Approach

In sketching standard analyses of Ps and SIs, we focus on the traditional approach, but for present purposes any account, old or new, which treats presuppositions and scalar implicatures as different falls in same class as the traditional perspective.

<sup>6</sup>Roughly, the distinction is as follows: a direct SI is an SI arising from a weak scalar term in an upward entailing context and an indirect SI is one arising from a strong scalar term in a downward entailing context, such as the scope of negation. As we will see below, this distinction is purely terminological, as all theories of SIs that we know of treat direct and indirect SIs in the same way.



### 2.2.1. Presuppositions

Considering Ps first: the traditional approach is to analyse them as definedness conditions on admissible conversational contexts for the sentence carrying the presupposition. The gist of the idea is that a sentence like (16-a) is only felicitous in a context in which the presupposition in (16-b) is already assumed to be mutually accepted by the discourse participants (Karttunen, 1974; Stalnaker, 1974; Heim, 1982, 1983; see also Beaver and Geurts, 2012; Schwarz, 2015; Romoli and Sauerland, 2017 for an introduction to presuppositions).

- (16) a. John stopped going to the movies.  
b.  $\leadsto$  *John used to go to the movies*

In addition, an account of the so called “projection” behavior of presuppositions is needed to explain how the presupposition of a sentence like (16-a) appears to be “inherited” by more complex sentences containing (16-a) such as (17), repeated from above.

- (17) John didn’t stop going to the movies.

Note that (16-a) and its negation in (17) both have the same presupposition that John used to go to the movies; in the traditional terminology, the presupposition of (16-a) in (16-b) “projects” from the scope of negation in (17). Projection is not limited to negation, but is a general pattern involving all sorts of complex embeddings. For instance, the presupposition of (16-a) is also inherited by conditional sentences containing (16-a) in their antecedent, as well as questions or modal embedding (16-a): all of (18)–(20) standardly give rise to the inference that John used to go to the movies. In contrast, none of them convey that John is not going to the movies now, as entailments are interpreted relative to the embedding operators.

- (18) If John stopped going to the movies, he must have gone to the gym more regularly.  
(19) Did John stop going to the movies?  
(20) John might have stopped going to the movies.

There are various well-developed proposals for accounting for presupposition projection in traditional terms, but we will not review these here in any detail for reasons of space. What is crucial for us, as before, is that all of these accounts treat presuppositions in a way that is very different from their treatment of SIs.

Finally, notice that traditional approaches quite generally assume presuppositions to be conventionally encoded in the lexical entries of the relevant expressions. This means that sentences containing a presupposition trigger necessarily introduce the corresponding presupposition. In order to reconcile this with cases of apparent suspension of presuppositions, as in (21), a further mechanism is assumed, e.g., one that “accommodates” the presupposition locally, which results in the absence of any contextual constraints at the sentence level (Heim, 1983; see also Von Stechow, 2008). This gives rise to the meaning paraphrased in (22), which is compatible with the continuation of (21), asserting that John never went to the movies.

- (21) John didn’t stop going to the movies ... he never went!  
(22) It’s not true that (John used to go to the movies and stopped)  
( $\approx$  Either John didn’t use to go to the movies or he didn’t stop).

### 2.2.2. Scalar Implicatures

The traditional approach to SIs, which sees them as distinct from Ps, goes back to Grice (1975) and Horn (1972). On this approach, SIs can be understood as arising from the hearer reasoning about the speaker’s communicative intentions. Take the inference in (23-b) based on (23-a).

- (23) a. John sometimes went to the movies.  
b.  $\leadsto$  *John didn’t always go to the movies*

In brief, the idea is that the hearer reasons that the speaker said (23-a), rather than something else, and in particular the more informative sentence in (24). Assuming that (24) is relevant to the purposes of the conversation, and that speakers are assumed to be committed to conveying the most informative relevant information at their disposal, the hearer will infer that the speaker’s reason for not saying (24) is that the speaker believes (24) to be false. Therefore, the hearer derives the inference (23-b)<sup>7</sup>.

- (24) John always went to the movies.

A parallel line of reasoning, can be used to derive the indirect SI in (25-b) from (25-a). The hearer reasons that the speaker said (25-a), rather than the relevant and more informative (26). Therefore, the hearer infers that (26) is false, i.e., (25-b).

- (25) a. John didn’t always go to the movies.  
b.  $\leadsto$  *John sometimes went to the movies*  
(26) John didn’t sometimes go to the movies ( $\approx$  John never went to the movies)

This brief review of the traditional perspective on Ps and SIs, while glossing over many intricacies, will suffice for our purposes. We primarily wish to provide a sense of how Ps and SIs are traditionally analyzed in clearly distinct ways. We now turn to more recent accounts of these inferences, in particular the SI approach to Ps.

## 2.3. The Scalar Implicature Approach to Presuppositions

The scalar implicature approach to presuppositions generally attempts to assimilate (certain) presuppositions to implicatures. In particular, some of the accounts within this general approach treat the presupposition associated with verbs like “stop” as scalar implicatures of a sort (Simons, 2001; Abusch, 2002, 2010; Chemla, 2010; Romoli, 2012, 2015). In this section, we briefly sketch the strongest version of this approach focusing on sentences like (27-a) and its associated inference in (27-b):

<sup>7</sup>We are skipping over a variety of details and assumptions here. See Gamut (1991) for a precise discussion of all the assumptions needed here to derive this inference.

- (27) a. John didn't stop going to the movies.  
b.  $\leadsto$  *John used to go to the movies*

Recall that one of the main phenomena to be accounted for is how the presuppositional inference of “stop” arises from both affirmative and negated sentences. As mentioned, the traditional explanation is that (28), by virtue of the lexical entry of “stop,” is associated with the presupposition in (27-b), which then projects from the scope of negation in (27-a).

- (28) John stopped going to the movies.

The SI approach to Ps offers a rather different explanation. First, (27-b) is simply (and only) an entailment of (28) on this account. This is in line with the observation that (27-b) is a non-cancelable ingredient of the overall meaning of (28), as asserting (28) and negating (27-b) sounds contradictory.

- (29) #John stopped going to the movies but in fact he never went.

Assuming that (27-b) is an entailment of (28) is neither novel nor surprising; many accounts of Ps in the traditional approach share the view that the presuppositional inference is entailed in affirmative contexts. What is novel in the SI approach to Ps is to argue that (27-b) is *only* an entailment of (28). The inference in (27-b), standardly associated with negated sentences like (27-a), is derived by this approach as an SI. Therefore, the hearer infers that the speaker believes the latter to be false, which is equivalent to (27-b).

- (30) John didn't use to go to the movies.

If this approach is correct, then the inferences associated with soft triggers such as *stop* are simply entailments when occurring in affirmative contexts, but (indirect) SIs when occurring under negation, leading to the two key predictions in (6-a) and (6-b) above. On this view, verbs like *stop* are completely parallel to strong scalar items like *always*, which give rise to parallel inferences in positive contexts and in the scope of negation.

### 3. THE PROCESSING OF SCALAR IMPLICATURES AND PRESUPPOSITIONS

In this section, we briefly review previous work on the processing of SIs and Ps, focusing in particular on RT experiments<sup>8</sup>.

#### 3.1. The Processing of SIs

In recent years, research on scalar implicatures has undergone what Chemla and Singh (2014) call an “experimental turn.” In particular, investigations of their processing properties have played a central role in the overall theoretical discussion. Most studies have focused direct SIs (DSIs) but some recent studies have started looking at indirect ones, too. In a seminal paper, Bott and Noveck (2004) argue that SIs are associated with a delay in RTs. They investigated sentences

like (31-a) and their direct SI in (31-b), which directly conflicts with common knowledge (as in fact all elephants are mammals). Based on the inference reading of the sentence, (31-a) should thus be judged “false.”<sup>9</sup> As discussed above, however, the sentence also has a no-inference (or “literal”) “some and possibly all” reading, which is compatible with common knowledge, and thus should lead to a “true” judgment.

- (31) a. Some elephants are mammals.  
b.  $\leadsto$  *Not all elephants are mammals*

The logic of the design in Bott and Noveck (2004) then is as follows: since “false” responses are indicative of inference interpretations and “true” responses of no-inference interpretations, measuring RTs for both types of responses should shed light on the time course of the availability of the two interpretations<sup>10</sup>. Their main finding, schematically represented in (32) (with > indicating greater RTs) is that false responses were slower than true responses. They interpret this delay as showing that the computation of scalar implicatures involves additional processing efforts that go beyond those involved in the computation of literal meaning.

- (32) **Bott & Noveck on DSIs**  
inference readings > no-inference readings

One particularly relevant version of their general approach trains participants prior to the main task to respond according to one or the other possible interpretations of the sentence in question. They find that participants that were trained to respond based on the no-inference interpretation were generally faster than those trained on the inference interpretation. Parallel results have been obtained in various similar studies since (Bott et al., 2012, among others), and also for implicatures associated with disjunction (Chevallier et al., 2008). Other methodologies, such as reading times (Breheny et al., 2006) and visual world eye tracking (Huang and Snedeker, 2009 and following work) have yielded comparable results as well<sup>11</sup>.

Cremers and Chemla (2014) extend Bott and Noveck's approach to indirect scalar implicatures (ISIs) by looking at sentences like (33-a), with the inference in (33-b), which is again incompatible with common knowledge.

- (33) a. Not all elephants are reptiles  
b.  $\leadsto$  *Some elephants are reptiles*

<sup>9</sup>Notice that the sentence in (31-a) is generally found to be somewhat odd, as is generally the case when scalar implicatures conflict with common knowledge (Magri, 2010). This feature of the design is however shown not to be important in work replicating the main result of Bott and Noveck (2004), like that of Bott et al. (2012).

<sup>10</sup>There is an obvious potential concern about general difference between the time course of true and false responses, which Bott & Noveck try to address through different variants of their basic design. We will return to this issue when introducing our own study below.

<sup>11</sup>Although other researchers have found different results using visual world eye tracking, which suggest implicatures are immediately available (e.g., Grodner et al., 2010; Breheny et al., 2013; Foppolo and Marelli, 2017).

<sup>8</sup>This section is adapted from Schwarz et al. (2015).

Overall, they argue their findings to be parallel to Bott and Noveck's results, in that training participants to respond based on an inference interpretation vs. a no-inference interpretation gives yields slower responses for responses based on inference-readings than those based on no-inference readings:

- (34) **Cremers and Chemla on ISIs**  
inference > no-inference.

Note, however, that Cremers and Chemla (2014) report two experiments, with *prima facie* conflicting results. In the first one, without training, they actually found opposite results for DSIs and ISIs, as participants' "false" responses were faster than "true" responses for ISIs. However, they argue that this outcome is the result of confounds in the materials. First, subjects may have calculated implicatures for controls as well, due to the specifics of the overall stimuli in the experiment. Secondly, DSIs and ISIs differ in whether they contain "matching" or "mismatching" animal names and categories (e.g., *elephant* paired with *mammals* and *reptiles* respectively). Their second experiment avoided these confounds and statistically controlled for effects of polarity and truth value, and yielded results in line with those for DSIs, leading to the interpretation of their overall results outlined above. We will return to some related issues when discussing the investigation of Ps by Chemla and Bott (2013) below.

In sum, Bott and Noveck found that "false" responses based on inference readings for direct SIs were slower than "true" responses based on no-inference interpretations. Similarly, Cremers and Chemla found that "false" responses based on inference readings for indirect SI were slower in comparison to "true" responses based on no-inference readings. These results are in line with the general uniformity for direct and indirect SIs assumed in the literature, and with the initial interpretation by Bott and Noveck that scalar implicatures are associated with a delay.

### 3.2. The Processing of Ps

The processing of Ps has been studied less than that of SIs. However, a number of recent studies have begun to fill this gap, using various processing measures to investigate Ps (see Schwarz, 2015, 2016a). In this section, we review two recent RT studies on Ps that are directly relevant for our purposes. The first, by Chemla and Bott (2013), uses the paradigm of Bott and Noveck (2004) to look at Ps under negation, and yields results that appear to be very different from those for SIs. The second, by Romoli and Schwarz (2015), compares Ps (under negation) and (indirect) SIs directly and finds uniform RT patterns. These two results appear to be in direct conflict with one another and thus suggest opposite answers to our main question about the relationship between Ps and SIs. We discuss a possible source of the difference in outcomes, which motivates the first set of experiments reported below.

#### 3.2.1. Chemla and Bott, 2013

Chemla and Bott (2013) adapts the paradigm from Bott and Noveck (2004) to investigate Ps. The logic is entirely parallel: subjects judge sentences like (35-a) with the factive verb "realise" (or, in their first experiment, "know"), which gives rise to the

presupposition in (35-b). This presupposition conflicts with common knowledge, and therefore, the sentence in (35-a) is only true on a no-inference reading.

- (35) a. Zoologists did not realize that Elephants are reptiles.  
b.  $\sim \rightarrow$  *Elephants are reptiles*

Comparing the RTs of True vs. False responses provides a measure of comparison between the inference readings and the no-inference readings. *Prima facie*, their results suggest the opposite pattern of that found for SIs by Bott and Noveck (2004): True responses were slower than false responses, i.e., inference readings were faster than no-inference readings:

- (36) **Bott and Chemla on Ps**  
inference readings < no-inference readings

The interpretation proposed by Chemla and Bott (2013) is that the computation of P-inferences, unlike that of SI-inferences, does not incur a delay, suggesting that the inferences involved are different, at least in the way they are processed. This poses a challenge for the SI approach to Ps. Note however, that the confound from the first experiment by Cremers and Chemla (2014) arising for indirect SIs is relevant for the present results for Ps as well: recall that the indirect SI materials involved a mismatch with respect to the relationship between the name of the animals mentioned (e.g., *elephants* paired with *reptiles*), which the authors argue might have hindered acceptance of sentences like (33-a). Recall also, that for direct SIs, the relevant targets instead involve a match between name and category, so conversely this might have facilitated the acceptance of sentences like (37).

- (37) Some elephants are mammals.

Turning back to the experiment in Chemla and Bott (2013), it is entirely parallel with the situation in Cremers and Chemla (2014). That is, unlike in Bott and Noveck, the target sentences in Chemla and Bott (2013), such as (35-a), involve a mismatch between the name and the category. As suggested by Cremers and Chemla (2014) for their own results, this factor could have influenced the results of Chemla and Bott (2013). That is, the increased RTs associated with no-inference readings could have been caused by this mismatch, rather than different derivational mechanisms. The existence of this potential confound means that the results in Chemla and Bott (2013) have to be interpreted with caution, and without implementing the same kinds of control techniques as Cremers and Chemla (2014) use in experiment 2, they do not conclusively establish any difference between SIs and Ps.

#### 3.2.2. Romoli and Schwarz, 2015

Recently, in a study by Romoli and Schwarz (2015) RTs for Ps and SIs under negation were directly compared to one another. In this study, instead of a direct truth-value judgment task, a version of a sentence picture matching task was used (Huang et al., 2013). This paradigm records both response choices and response times as dependent variables. A sentence was presented to participants and they were directed to pick a picture, from a set of three,

that best matched the sentence. Each of the pictures depicted an individual and a 5-day calendar strip, with each day being filled with an iconic representation of an activity that the individual had engaged in on that day (see **Figures 1, 2**). In addition to these two “visible pictures” there was a “Covered picture.” Participants were told that one of the three pictures was a match for the presented sentences like (38). One of the visible pictures was a “Target picture,” which was either consistent or inconsistent with the inference (“+LIT/+INF” vs. “+LIT/-INF” condition)<sup>12</sup>. The second visible picture was a distractor and so was incompatible with both possible interpretations. Participants were told that if neither of the visible pictures were a good match, then they should select the Covered picture.

(38) John didn’t always go to the movies last week.

The +LIT/+INF Target picture depicts the character going to the movies on several days, making it consistent with the “sometimes” implicature of “not always.” In contrast, the +LIT/-INF Target picture depicts the character never going to the movies, making it inconsistent with this implicature. By comparing the RTs associated with Target choices in these two conditions Romoli and Schwarz (2015) were able to compare the processing of different interpretations based on the same type of response<sup>13</sup>.

Similarly, for the *stop* condition, participants would evaluate sentences like (39) against one of the two overt pictures in **Figure 2**, a distractor picture and a Covered picture. Again the +LIT/+INF Target picture was compatible with the inference interpretation of the sentence, while the +LIT/-INF Target picture was only compatible with the no-inference interpretation.

(39) John didn’t stop going to the movies on Wednesday.

Unsurprisingly, the Target picture in the +LIT/+INF condition was chosen at ceiling level, while the +LIT/-INF condition yielded more mixed results. But most importantly, the RT results for Target choices were uniform for Ps and SIs, as schematized in (40), in that RTs in the +LIT/+INF conditions were significantly faster than in the +LIT/-INF conditions, in contrast with the findings discussed above. (Note that while the +LIT/+INF picture could be accepted on either a no-inference or an inference interpretation, the difference in RTs suggests that at least a sizable portion of Target choices was based on the latter; this assumption justifies the use of “inference” and “no-inference” in the schematic illustration below, and will also be utilized in the data analysis of the experiments in the next section.)

<sup>12</sup>Romoli and Schwarz (2015) label the conditions INFERENCE-TRUE and INFERENCE-FALSE respectively; we choose the more transparent labels here to clearly signal that the images shown in the former can in principle be accepted on either a literal or an inference interpretation.

<sup>13</sup>Note that, in principle, selection of the +LIT/+INF Target picture could also be motivated by a no-inference/literal interpretation. However, if all these selections were based on such an interpretation, then we would expect participants’ behavior in these two conditions to be equivalent. Therefore, the fact that Romoli and Schwarz (2015) found substantial variance in the RT results, suggests that, at least a sizable portion of Target picture selections in the relevant condition are motivated by inference interpretations.

- (40) a. **Romoli and Schwarz 2015 on indirect SIs**  
inference < no-inference.  
b. **Romoli and Schwarz 2015 on Ps**  
inference < no-inference.

Note that the results for Ps here seem to be in-line with those in Chemla and Bott (2013), in that inference readings were faster than no-inference readings. The result for indirect SIs, however, is puzzling in that it appears to be exactly the opposite of what Cremers and Chemla (2014) find in their experiment 2. Moreover, with regards to our main question in (7), these results suggest that Ps and SIs (at least indirect ones) do not differ in their RT patterns after all, which would be consistent with a uniform account of SIs and Ps. This raises the question of what is behind these seemingly conflicting findings. One possibility relates to differences in the types of responses that were compared between these studies. As mentioned, previous response time studies generally explored the relevant inferences by comparing “true” responses to “false” responses. And, while Cremers and Chemla (2014) attempted to control for any effect of response-type, the more reliable way of controlling for such an effect is to compare the same kind of responses, which the setup of Romoli and Schwarz (2015) made possible. To put it another way, Romoli and Schwarz (2015) raise the possibility that the method employed by previous studies may have been undermined by a confound. Specifically that, rather than only being influenced by the interpretations of interest, participants’ responses may have also been influenced by the nature of the response provided (i.e., sentence acceptance vs. rejection). The experiments reported in the next sections were designed to investigate this issue by further exploring the relationship between Ps and SIs.

## 4. THE EXPERIMENTS

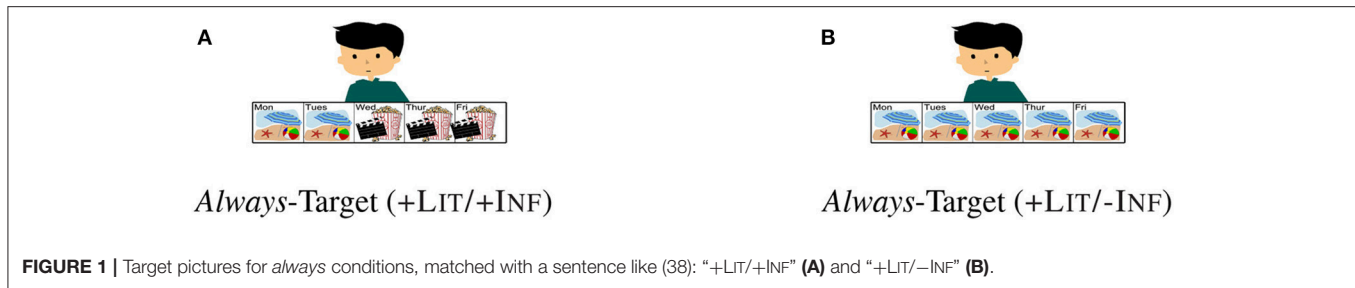
In this section, we report on three series of experiments testing the two predictions of the SI approach to Ps outlined in (6-a) and (6-b).

### 4.1. Experiment Ia

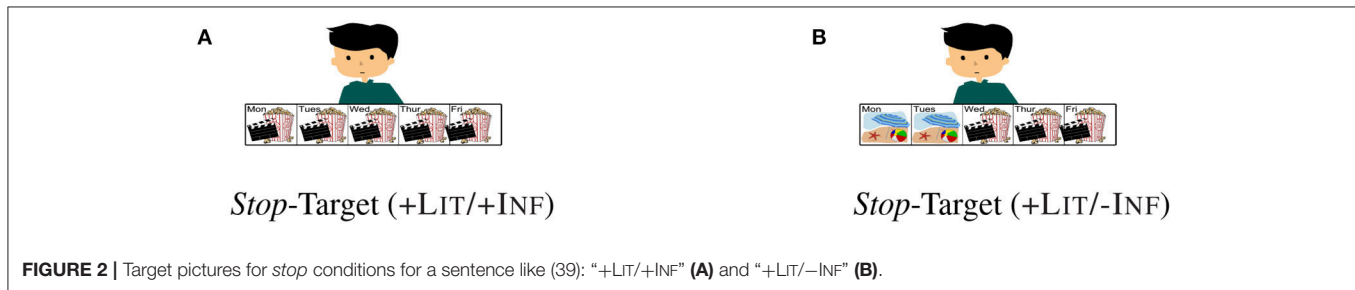
The first experiment adopted the approach taken in Romoli and Schwarz (2015) and applied it to investigating whether there are processing pattern differences between direct and indirect implicatures when we compare alike responses<sup>14</sup>. This allows for a more comprehensive comparison to the results from Bott and Noveck (2004) and Cremers and Chemla (2014) on the one hand, and Romoli and Schwarz (2015) on the other. It also offers a more comprehensive perspective on the role of response type in RT patterns. Note that, for this experiment (and Experiment Ib), the relevant uniformity prediction is that the relative processing patterns of each trigger will be similar. That is, the prediction is not that the RTs will be exactly the same as the relevant triggers differ substantially in other ways; namely, the presence of negation in one and not the other. Instead, the prediction

<sup>14</sup>This experiment was first reported in Schwarz et al. (2015), from which this subsection is adapted.





**FIGURE 1** | Target pictures for *always* conditions, matched with a sentence like (38): “+LIT/+INF” (A) and “+LIT/-INF” (B).



**FIGURE 2** | Target pictures for *stop* conditions for a sentence like (39): “+LIT/+INF” (A) and “+LIT/-INF” (B).

is that the overall RT pattern, created by comparing inference and no-inference interpretations, will be similar. To gain a full comparison, we looked at both target choices (acceptance judgments) based on inference and no-inference interpretations, and Covered picture choices (rejection judgments) based on both types of interpretation.

#### 4.1.1. Methods

##### 4.1.1.1. Materials and Design

Following Romoli and Schwarz (2015), we used the Covered picture paradigm (Huang et al., 2013), with both response choices and RTs as dependent variables. Participants were presented with two pictures, one of which was simply black and was introduced as covering a hidden picture<sup>15</sup>. The instructions provided a detective scenario, where information about a suspect was presented as having been extracted from intercepted communication, and the participant’s task was to decide which of two potential culprits fit the provided description. It was explicitly stated that only one of the two pictures would match the description, so that the Covered picture should only be chosen in situations where the overt picture did not match the sentence. We believed this setup would increase the chance of participants basing their responses on no-inference interpretations for the following reasons: First, the described source of the information remained opaque due to its nature of stemming from intercepted communication, which makes it uncertain whether the speaker of that sentence was fully informed. Secondly, the emphasis that only one picture would match the description provided by the sentence should increase target choices for +LIT/-INF pictures, on the assumption that no-inference interpretations are in principle

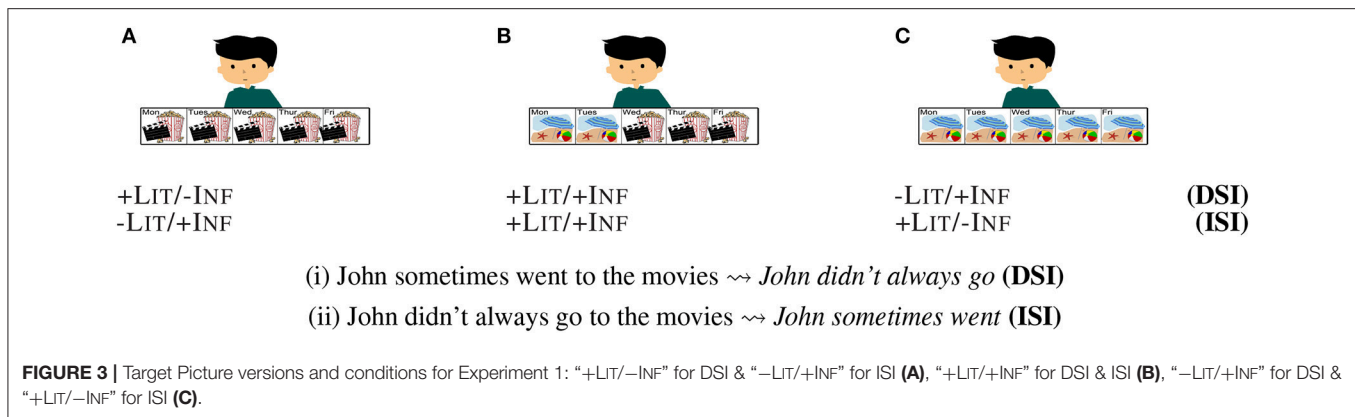
available but generally somewhat dispreferred<sup>16</sup>. That is, as the Covered picture could be completely “False,” if there is a possible reading that makes the Target picture “True” the participant has a good reason to go with that reading, even if it is a dispreferred reading. At the same time, as noted above, having the Covered Picture as a response option ensures that subjects need not feel forced to give a response that they may feel uncomfortable about.

The basic logic of the design was parallel to that of Romoli and Schwarz (2015), in that the overt Target picture either was consistent with a given interpretation or not. More concretely, sentences (i) and (ii) in **Figure 3** were displayed with one of the pictures in **Figure 3** and a Covered picture<sup>17</sup>. For the DSI condition with *sometimes*, the picture in **Figure 3A** is only compatible with a no-inference interpretation, as the depicted person always went to the movies. Target choices in this case must therefore be based on the no-inference interpretation. Covered picture choices for this picture in turn are indicative of inference interpretations. The picture in **Figure 3B** is consistent with an inference interpretation (as well as a no-inference interpretation, since it is entailed by the inference interpretation), so target choices are generally expected here. Finally, the picture in **Figure 3C** is inconsistent with both interpretations, as the

<sup>15</sup>Note that, unlike Romoli and Schwarz (2015), we didn’t include a “distractor picture.” This change was done merely to simplify the material and was not expected to have any substantive effect on the results.

<sup>16</sup>While work such as Van Tiel et al. (2016) has shown considerable variability in this preference between SIs, this work and others (e.g., Noveck, 2001; Papafragou and Musolino, 2003; Foppolo and Marelli, 2017) seems to suggest that, for the SI associated with the “some/all” scale, it is indeed the case that the no-inference interpretation tends to be dispreferred.

<sup>17</sup>Note that the condition labels presented in **Figure 3** relate to the truth-value of the two critical elements of the sentence; namely, the literal content and the inferential content. For example, in the case of the condition “+Lit/-Inf” for the DSI sentence, the picture is consistent with the literal content that *John went to the movies at least once*, but is inconsistent with the inference that *John didn’t always go to the movies*. Moreover, in the case of the “-Lit/+Inf” conditions, the target picture should not be able to be selected, due to it not satisfying the literal content of the relevant sentence, despite the fact that it is consistent with the inference (corresponding to the literal meaning of the paraphrase).



depicted individual never went to the movies, so Covered picture choices are expected here. For purposes of analysis, this design allowed us to compare Target and Covered picture responses to the picture in **Figure 3A** to Target and Covered picture responses in the control conditions in **Figures 3B,C**, respectively. Thus, this set up provides a comparison between inference-based rejections (Covered picture choices for **Figure 3A**) and literal meaning based rejections (Covered picture choices for **Figure 3C**), as well as between no-inference acceptances (target choices for **Figure 3A**) and inference acceptances (target choices for **Figure 3B**, assuming as above that at least a sizable portion of responses here is based on an inference interpretation).

The same general logic applies to the ISI sentences (ii), though with different mappings onto the pictures. The picture in **Figure 3C** serves as a test for no-inference interpretations, as target choices are incompatible with the inference that John sometimes went to the movies. Covered picture choices for this picture in turn must be based on inference interpretations. The picture in **Figure 3B** is consistent with the inference interpretation (as well as a no-inference interpretation, as for DSIs), and the picture in **Figure 3A** is inconsistent with either interpretation. So in the case of ISIs, **Figure 3C** is expected to yield a mix of target and Covered picture choices, depending on the interpretation participants base their judgments on in a given trial, which can be compared to the Covered picture and target choices in the respective control conditions.

Let us expand here on our assumption about the correspondence between responses and the interpretation that they are based on. As pointed out already, in certain conditions, it is not clear whether certain picture selection choices are motivated by an inference or a no-inference interpretation. Specifically, target choices for **Figure 3B** and Covered picture choices for **Figure 3C** could be based on either inference or no-inference interpretations. This is because both interpretations are consistent with **Figure 3B** and inconsistent with **Figure 3C**. However, if we assume consistency in participant's interpretations between conditions, then we can discern whether any of these responses are based on inference interpretations by comparing responses to **Figures 3B,C** to a condition without this ambiguity. For example, in the case of the DSIs condition, **Figure 3A** is only consistent with a no-inference interpretation. Therefore, if the participant group selects more covered pictures when presented with Figures like

**Figure 3A** than with Figures like **Figure 3C**, then it is likely that at least some of the latter Covered picture selections were motivated by inference interpretations. Similarly, Target picture selections of **Figure 3B** can be compared with Target picture selections of **Figure 3A** to determine if any of the former were motivated by no-inference interpretations. A similar comparison between conditions can be done in the ISI condition. (In addition to response patterns, differences in RTs also support this assumption, as noted already for Romoli and Schwarz (2015) above.)

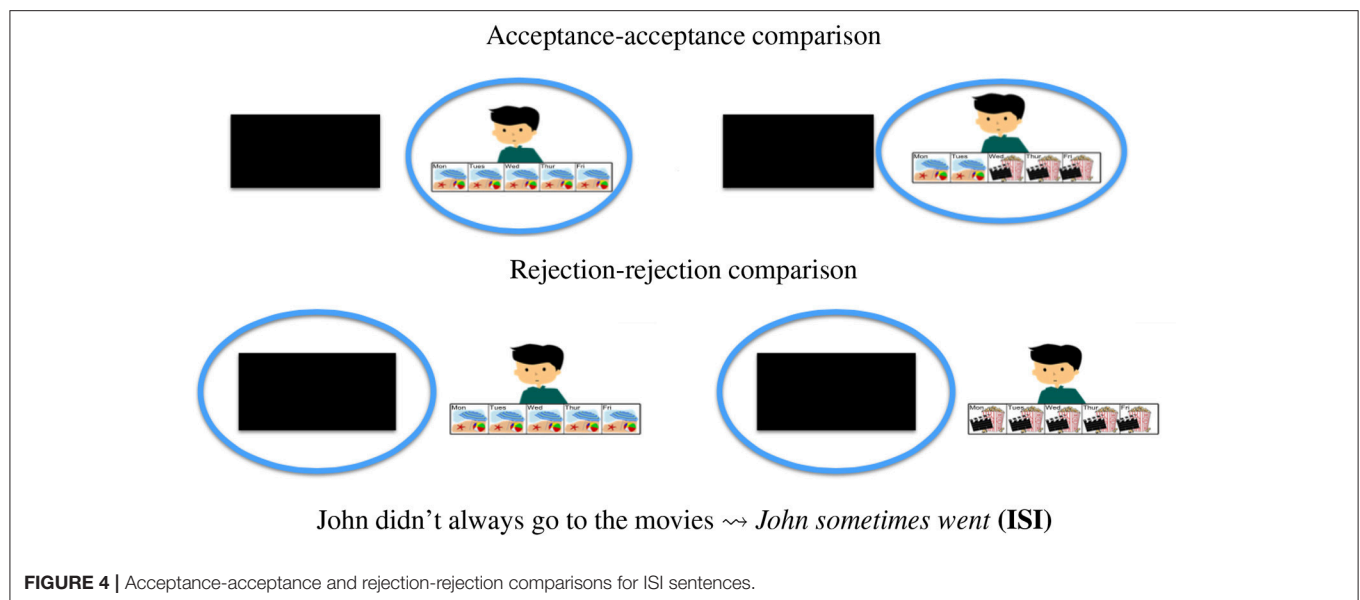
**Figure 4** summarizes the two critical comparisons in the ISI conditions in the display format used in the experiment: no-inference acceptance vs. inference acceptance (“acceptance-acceptance” comparison) and inference-rejection vs. no-inference rejection (“rejection-rejection” comparison).

#### 4.1.1.2. Participants and Procedure

35 undergraduate students from Macquarie University participated in the study. They saw 36 sentence picture pairs of the sort described above, with 6 items for each pairing, counterbalanced across participant groups. In addition, there were a total of 36 filler items; 18 were variants of the experimental items containing *always* without negation, paired with all three picture types to ensure that pictures such as those in **Figures 3A,C** were viable target choices throughout the experiment sufficiently often. There also were 6 items containing plain negation (e.g., *John didn't go to the movies last week.*), again paired with the various picture types to even out choices of types of pictures. Finally, 12 items were from another sub-experiment containing negation and *again*. At the beginning of the experiment, participants were presented with instructions laying out the detective scenario described above. They then were shown some example sentences and pictures, and completed a total of 4 practice trials (none of them resembling the crucial experimental conditions) to ensure they understood the Covered picture setup. Throughout this initial phase, they were free to ask any clarification questions. After this, presentation of the experimental trials began.

#### 4.1.2. Results and Discussion

For purposes of statistical analysis, responses were coded according to whether they were based on their relation to



an inference reading. Target selection of the pictures in **Figure 3A** (DSI) and **Figure 3C** (ISI) clearly indicates a no-inference reading, whereas Covered Picture selection for these pictures unambiguously reflected an inference reading. Accurate responses in the other conditions were compatible with both inference and no-inference readings, but were coded in terms of the strongest reading on which they could be based. For example, acceptance of the Target picture in 3b was coded as an inference response, though of course a positive instantiation of an inference reading entails truth of a no-inference reading as well. The negative response toward the Target picture for the versions in **Figure 3C** (DSI) and **Figure 3A** (ISI), as reflected in selection of the Covered Picture, was coded as a no-inference response, though again, a negative relation of a no-inference reading toward a picture entails a negative relation for the inference reading as well. This coding decision is not crucial for the overall interpretation of the data, but we think it reflects the difference across conditions in terms of whether the two readings are in conflict or not reasonably well. Target choice proportions as well as RTs (measured from the display of the sentence, which was added to the screen 800ms after the picture was first shown) were analyzed.

#### 4.1.2.1. Response rates

Mean target selection rates are provided in **Table 1**. Accuracy in the conditions where both literal and inference interpretations led to the selection of the same image (**Figures 3B,C** for DSIs, **Figures 3A,B** for ISIs) were at ceiling, as expected. Both inference and no-inference (i.e., literal) interpretations occurred in the DSI and ISI +Lit/-Inf conditions, but inference interpretations occurred more often with DSIs than with ISIs, as there were fewer Target picture choices for DSIs. A planned comparison between these two conditions using a logistic regression mixed-effect model revealed this difference in implicature-response rates to be significant ( $\beta = 4.01$ ,  $SE = 0.98$ ,  $z = 4.07$ ,  $p < 0.001$ ).

**TABLE 1 |** Target choice rates in % by condition.

Inference type	+Lit/-Inf (Figure 3A/Figure 3C)	-Lit/-Inf (Figure 3C/Figure 3A)	+Lit/+Inf (Figure 3B)
DSI	22.9	0.005	97.1
ISI	50.9	0.005	95.7

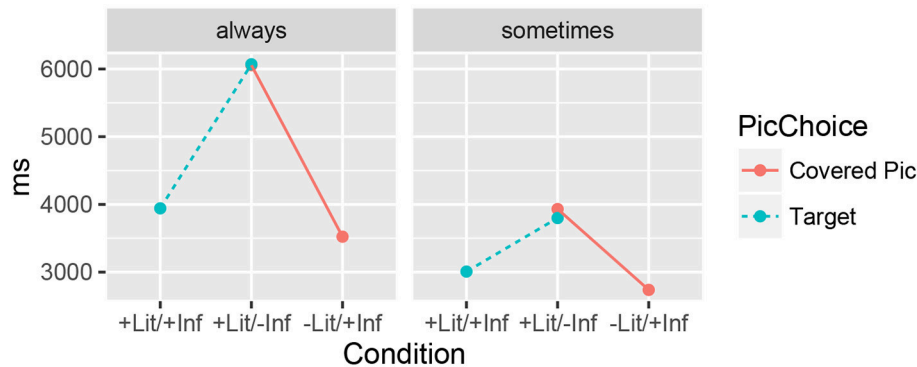
Note also that the difference between the +LIT/+INF and +LIT/-INF responses suggests that at least some of the Target picture selections in the former condition were a result of participants accessing an inference interpretation. That is, if participants were only accessing literal interpretations for our test sentences, you would expect the response rates in these two conditions to be the same<sup>18</sup>.

#### 4.1.2.2. Response Times

The mean RTs for all conditions are illustrated in **Figure 5**. Note that seeing this from the perspective of inference vs. no-inference interpretations as laid out above, yields a cross-over interaction pattern, showing that the relation between RTs for inference and no-inference interpretations depends crucially on whether we look at acceptances in the form of target choices or rejections in the form of Covered picture choices. In the former case, inference interpretations are faster than no-inference ones, while the reverse holds in the latter.

To investigate this result statistically, we analyzed both the DSI and ISI subsets of data as a  $2 \times 2$  interaction design with response (Target vs. Covered picture) and interpretation (inference vs. no-inference) as factors, using mixed-effect models

<sup>18</sup>Similarly, the Covered picture selections between the -LIT/+INF and +LIT/-INF conditions suggests that some of these selections in the former condition were a result of accessing an inference interpretation.



**FIGURE 5** | RTs for responses by picture choice and condition. +Lit/+Inf target choices and +Lit/-Inf Covered picture choices are taken to reflect inference interpretations, and +Lit/-Inf target choices and -Lit/+Inf Covered picture choices no-inference interpretations.

with subjects and items as random effects, as implemented in the *lmer* function of the *lme4* package in *R* (Bates, 2005). Following Barr et al. (2013), we used the maximal random effect structure that would converge, with random effect slopes for each factor, as well as the interaction, if possible. To assess whether inclusion of a given factor significantly improved the fit of the overall model, likelihood-ratio tests were performed that compared two minimally different models, one with the fixed effects factor in question and one without, while keeping the random effects structure identical (Barr et al., 2013). We report estimates, standard errors, and *t*-values for all models, as well as the  $\chi^2$  and *p*-value from the likelihood-ratio test for individual factors. The statistical details are summarized in **Table 2**. The  $2 \times 2$  interactions were highly significant for both ISIs and DSIs, as were the relevant simple effects comparing inference vs. no-inference responses by response type. Schematically, the results can be summarized as follows:

- (41) RT patterns for Scalar Implicatures (for both DSIs and ISIs):
- rejection response**  
inference > no-inference
  - acceptance response**  
inference < no-inference

The results for acceptances (Target-choices), where implicature-based responses were faster than those only compatible with the literal meaning, are entirely in line with the findings by Romoli and Schwarz (2015) for ISIs, but constitute a novel finding for DSIs. The finding that inference-based rejections (Covered Picture-choices) were slower for both types of implicatures *prima facie* seems to be in line with previous findings for DSIs from Bott and Noveck (2004) on, and with the findings by Cremers and Chemla (2014) for ISIs. However, note that the comparison we make is one between a condition where a Covered Picture choice can be unambiguously attributed to an inference interpretation (the equivalent of saying “false” to *Some elephants are mammals.*), and a condition where the literal meaning suffices to lead to a Covered Picture choice, but an inference interpretation would have led to the same result (the equivalent of

saying “false” to *Some elephants are insects.* - B&N’s control T3). Similarly, our acceptance comparison is between acceptances that are unambiguously based on a no-inference reading and ones where inference and no-inference readings yield the same result (parallel to B&N’s T2 control: *Some mammals are elephants.*). The comparison within our data that is truly on par with the crucial comparison of Bott and Noveck (2004) (as well as Cremers and Chemla, 2014) is the one between Covered Picture choices based on an inference interpretation and Target choices based on a no-inference interpretation. But here, we find no significant difference at all.

Now, let us consider these results in light of the SI approach to Ps’ prediction of uniform processing patterns between DSIs, ISIs, and Ps, (i.e., (6-b)). Once we considered the acceptance vs. rejection factor, DSIs and ISI exhibited uniform RT patterns, contrary to initial appearances from Romoli and Schwarz (2015). Next, we turn to Ps considered from the same, more comprehensive perspective, to see whether this uniformity might extend in the manner proposed by the SI approach to Ps.

## 4.2. Experiment Ib: Stop in Negated Sentences

In Experiment Ib, we used the same methods as in Experiment Ia to extend the investigation above to Ps, and in so doing, address the main question of this paper regarding the relationship between Ps and SIs. That is, to test the SI approach to Ps’ prediction that the processing patterns of SIs and the relevant Ps should be uniform. Note that, as in Experiment Ia, the uniformity prediction that we are testing is the expectation that the relative processing patterns of Ps will be the same as SIs, not that the RTs will be exactly the same across these inferences.

### 4.2.1. Methods

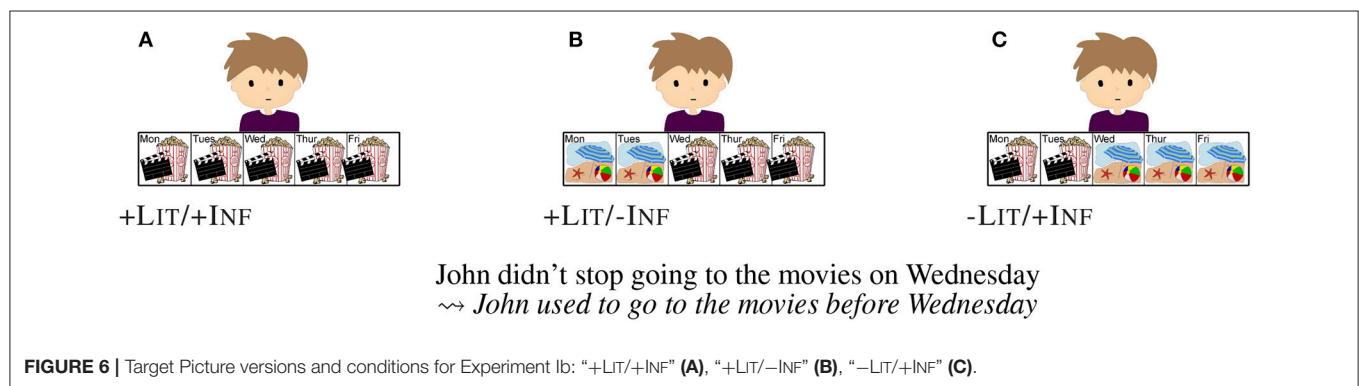
#### 4.2.1.1. Materials and Design

We used the same Covered picture paradigm as in Experiment Ia, with two pictures and both response choices and RTs as dependent variables. The basic logic of the design was also



**TABLE 2 |** Summary of response time analyses: Interaction between Picture Choice and inference status and simple effects for relevant paired factor levels.

DSI's	$\beta$	SE	t	$\chi^2$	p
<b>Interaction</b>	2119.1	563.4	3.76	9.67	<0.01
<b>Simple Effects</b>					
Covered Picture Choices: inference > no-inference	-1418.6	534.8	-2.65	6.38	<0.05
Target Choices: inference < no-inference	666.1	276.5	2.41	5.42	<0.05
ISI's					
<b>Interaction</b>	5902.7	1793.5	3.29	9.67	<0.01
<b>Simple Effects</b>					
Covered Picture Choices: inference > no-inference	-3302.2	881.6	-3.75	7.80	<0.01
Target Choices: inference < no-inference	2197.9	580.2	3.788	11.734	<0.001



identical to that of Experiment 1a, but this time we were looking at presuppositional sentences. The stimuli included both sentences with and without negation. However, as laid out in the introduction, only the case of soft triggers under negation lends itself to a direct comparison with SIs (and specifically ISIs). We therefore focus the discussion in the present section on that case. The case of "stop" in affirmative sentences will be discussed separately in section 4.4. An illustration of the negative conditions is provided in **Figure 6**. The sentence in **Figure 6** was displayed with one of the pictures in **Figure 6** and a Covered picture.

The picture in **Figure 6A**, paired with the negative "stop" sentence, constitutes the Target-selection control, as both the putative presupposition (that John went to the movies before Wednesday) and the asserted part (that he went to the movies from Wednesday on) are true. The picture in **Figure 6C** provides the Covered Picture-selection control, as the asserted part is false (since he did stop going to the movies), although the presupposition is true. **Figure 6B** constitutes the critical case, as the putative presupposition is false, while the assertion is true. If a participant accesses an inference interpretation, the Covered Picture should be chosen. If a participant accesses a no-inference interpretation the Target picture should be selected. As in Experiment 1a, responses to **Figure 6B** were coded as inference and no-inference responses respectively, based on whether the Covered picture or the Target picture was selected. **Figures 6A,C** were taken to provide

controls with the same response for the respective critical trials.

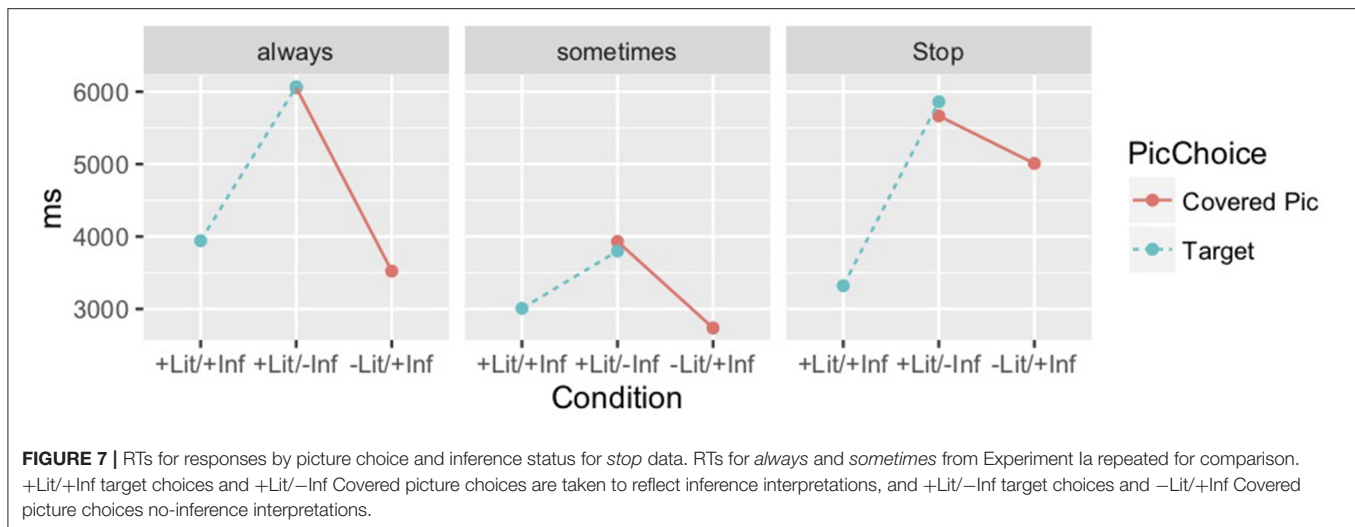
#### 4.2.1.2. Participants and Procedure

34 undergraduate students from the University of Pennsylvania participated in this study for course credit. Each saw 6 sentences in the +LIT/-INF and 6 in the -LIT/+INF conditions, and these were drawn from a total of 24 sentences. The other 12 were shown in the affirmative condition (discussed below), and the condition in which a given item was shown was counterbalanced across four groups of subjects. Another 12 items were presented in the +LIT/+INF condition, again drawn from a total of 24, with counter-balancing between it and an affirmative variant. In addition, there were 21 fillers from another sub-experiment. Instructions and practice trials were as described for Experiment 1a.

### 4.2.2. Results and Discussion

#### 4.2.2.1. Response rates

Unsurprisingly, the Target-selection rates for the control conditions were at ceiling and floor for the respective control conditions. In the critical condition, the Target was selected 62% of the time, which was significantly higher than in the -LIT/+INF control ( $\beta = -4.63$ ,  $SE = 0.82$ ,  $z = -5.63$ ,  $p < 0.001$ ), but also significantly lower than in the +LIT/+INF control ( $\beta = 3.11$ ,  $SE = 0.71$ ,  $z = 4.38$ ,  $p < 0.001$ ).



#### 4.2.2.2. Response times

The RT results are summarized in **Figure 7**. We find a pattern that is generally parallel to that for implicatures, and which corresponds to a cross-over interaction between type of reading (inference vs. no-inference) and type of response (acceptance vs. rejection) when coded as corresponding to inference and no-inference interpretations as described: Target choices compatible with the inference were faster than those only compatible with a no-inference reading, and Covered Picture choices based on the falsity of the inference were slower than Covered Picture choices (which could be) based on the falsity of literal meaning alone. To investigate this result statistically, we analyzed the data as a  $2 \times 2$  interaction design, using the same statistical analyses as detailed for Experiment Ia. The detailed results are summarized in **Table 3**. The  $2 \times 2$  interaction was highly significant, as was the relevant simple effect comparing inference vs. no-inference responses for Target choices. For Covered Picture choices, there was a numerical effect in the same direction as for SIs (Inf > NoInf), but this did not reach significance.

The first finding extends the findings in Romoli and Schwarz (2015) and our Experiment Ia to the domain of presuppositions, as inference interpretations seem to be faster than no-inference ones when looking at acceptance judgments. The direction of the RT effect for Covered Picture responses seems parallel to the SI-results in Bott and Noveck (2004) and Cremers and Chemla (2014), again extended to presuppositional inferences. However, as in the case with SIs, it's worth noting that the more direct comparison with these previous studies would be between Target choices based on a no-inference interpretation and Covered Picture choices based on an inference interpretation, and we find no difference here, parallel to the case of SIs. Thus, our result here differs from both the previous findings for SIs as well as those for Ps by Chemla and Bott (2013), but the results are parallel to our findings for SIs in Experiment Ia. In sum, based on the results from Experiments Ia and Ib, we find no difference in the processing patterns (measured through RTs) of Ps, DSIs or ISIs. This is consistent with the SI approach to Ps' prediction

of uniformity between SIs and Ps (i.e., (6-b)). Next we turn to investigating the effect of one more variable, that of prosody, on these inferences, as a further test of their uniformity.

### 4.3. Experiment II: The Effect of Prosody on Inference Interpretations

It has been observed in the literature that prosodic focus interacts with both SIs and Ps. In particular, in the case of ISIs, stress on the scalar terms trigger has been argued to be necessary for the felicity of a reading without the inference (i.e., also described as “cancellation” of the implicature; see Horn, 1989; Fox and Spector, 2018 and references therein).

(42) John didn't ALWAYS go to the movies.

As for presuppositions, it has also been observed that stress on the trigger changes the availability of the inference reading (see Abusch, 2002; Beaver, 2010; Romoli, 2012; Abrusán, 2016; Simons et al., 2017; Esipova, 2018). In cases of negation like (43), stress on the trigger has also been associated with less inference interpretations.

(43) John didn't STOP going to the movies.

There are ongoing debates about the precise role of prosody in cases (42) and (43) and how it interacts with the mechanisms for deriving implicatures and presuppositions. All that matters for current purposes is that according to the SI approach to Ps, we expect stress to play a parallel role for SIs and (the relevant type of) Ps. That is, on this approach the derivation of (indirect) implicatures and (“projecting”) presuppositions under negation proceeds in entirely parallel ways, and thus should be modulated in the same way by variations of the prosody. A traditional approach, on the other hand, can more easily accommodate a difference in the effect of prosody on the two inferences.

In order to assess this prediction, we conducted an experiment comparing written stimuli to auditory ones, which either had neutral intonation or prosodic stress placed on the expression

**TABLE 3 |** Summary of response time analyses for Experiment 1b: Interaction between Picture Choice and inference status and simple effects for relevant paired factor levels.

P's	$\beta$	SE	t	$\chi^2$	p
<b>Interaction</b>	3088.2	592.1	5.22	19.66	<0.001
<b>Simple Effects</b>					
Covered Picture Choices: inference > no-inference	-772.9	515.5	-1.50	2.16	= 0.14
Target Choices: inference < no-inference	-2340.0	431.7	-5.42	21.55	<0.001

giving rise to the implicature or presupposition. The setup is overall parallel to that above, with a sentence-picture matching task that included a Covered Picture<sup>19</sup>.

### 4.3.1. Methods

#### 4.3.1.1. Materials and Design

The sentences were slight variations of those above, with a more uniform wording for the *always* and *stop*-versions:

- (44) a. John didn't stop going to the movies this week.  
b. John didn't always go to the movies this week.

These were presented along with one of the picture variations in **Figure 8** and a Covered Picture as the alternative choice. As before, the +LIT-INF pictures can only be accepted if the judgment is based on a reading that lacks the respective inferences. In the WRITTEN condition, the sentences in (44) were presented as text on the screen. For the auditory conditions, we used audio recordings of the sentences in (44). In the NO-STRESS condition, a neutral prosody, as would be appropriate in an all-nore context, was used. In the STRESS condition, *always* and *stop* bore the main pitch accent of the sentence.

In addition to 24 critical items, there were 48 fillers, 9 using *stop* with negation and Covered Picture-choices, 15 with affirmative *stop* (8 Target and 7 Covered Picture Choices), as well as 24 items replicating that pattern for *always*.

#### 4.3.1.2. Participants and Procedure

The design was between-groups, so each participant was only exposed to one mode of presentation (WRITTEN, NO-STRESS, STRESS). The NO-STRESS data was collected as part of an eye-tracking experiment, but we only focus on the response patterns here<sup>20</sup>. A total of 97 undergraduate students from the University of Pennsylvania participated in the experiments for course credit (23 in WRITTEN, 27 in STRESS, and 47 in NOSTRESS). Instructions and practice trials were parallel to those for the

previous experiments. Participants saw a total of 72 trials, and the 4 conditions of the 24 critical items were counter-balanced across groups of participants.

### 4.3.2. Results and Discussion

The dependent variable of main interest for this study was response rates, as we were interested in assessing the impact of prosody on the prevalence of inference interpretations. The overall response patterns across conditions are illustrated in **Figure 9**. The key observation is that we find variation in the frequency of target choices in the +LIT-INF condition across different stimulus presentation types. In the NOSTRESS condition with auditory stimuli using neutral prosody, target acceptances seem to be lower than in the WRITTEN condition, indicating a greater prevalence of inference interpretations, for both *always* and *stop*. However, in the STRESS condition, we find the opposite effect for *stop*, as the marked prosody increased the availability of no-inference interpretations.

To assess the main contrasts of theoretical interest statistically, we conducted 2×3 mixed-effect model logistic regression analyses using treatment coding on the data for the +LIT-INF conditions, with varying baselines to assess different simple effects. Comparing the WRITTEN version to the NOSTRESS version confirmed a significant decrease in Target-acceptances for both *stop* ( $\beta = -4.85$ ,  $SE = 1.23$ ,  $z = -3.96$ ,  $p < 0.001$ ) and *always* ( $\beta = -3.98$ ,  $SE = 1.18$ ,  $z = -3.36$ ,  $p < 0.001$ ). The interaction term for this comparison did not reach significance ( $p = 0.12$ ), but there is a significant simple effect with fewer Target acceptances for *stop* than for *always* in the NOSTRESS condition ( $\beta = 1.42$ ,  $SE = 0.40$ ,  $z = 3.53$ ,  $p < 0.001$ ). Turning to a comparison of the WRITTEN condition and the STRESS condition, there was a significant increase in Target acceptances for *stop* ( $\beta = 2.49$ ,  $SE = 1.23$ ,  $z = -2.03$ ,  $p < 0.05$ ), and a marginally significant decrease for *always* ( $\beta = -2.39$ ,  $SE = 1.25$ ,  $z = -1.91$ ,  $p < 0.1$ ). In addition, there was a significant interaction ( $\beta = -4.89$ ,  $SE = 0.69$ ,  $z = -7.07$ ,  $p < 0.001$ ). Comparing the STRESS and NOSTRESS conditions directly revealed more Target acceptances for *stop* sentences in the STRESS condition ( $\beta = 7.35$ ,  $SE = 1.21$ ,  $z = 6.07$ ,  $p < 0.001$ ), while there was no difference between these condition for *always* sentences. Finally, the interaction term for this comparison was also significant ( $\beta = 5.76$ ,  $SE = 0.70$ ,  $z = 8.21$ ,  $p < 0.001$ ).

The outcome pattern for the prosodic manipulations is striking, and entirely unexpected from the perspective of the SI approach to Ps, at least in the strong version we are focusing on here. If presuppositions and implicatures are derived in

<sup>19</sup>Note that this experiment is different from the previous two in that we are no longer looking for uniformity in processing patterns. Instead we are investigating whether there is uniformity in the response of these inferences to prosodic stress, measured through rates of derivation. While the measure is different, the SI approach to Ps' prediction is similar to that made for Experiments 1a and b; namely, that there will be uniform effects of prosodic stress on the pattern of derivation rates. That is, we do not take this approach to be requiring that the effect needs to be to the same extent for both these inferences, just that it needs to be in the same direction.

<sup>20</sup>As will be detailed below, there were very few Target choices in the +LIT-INF condition for *stop* here, which prevented any meaningful eye tracking data analysis for the trials of interest.

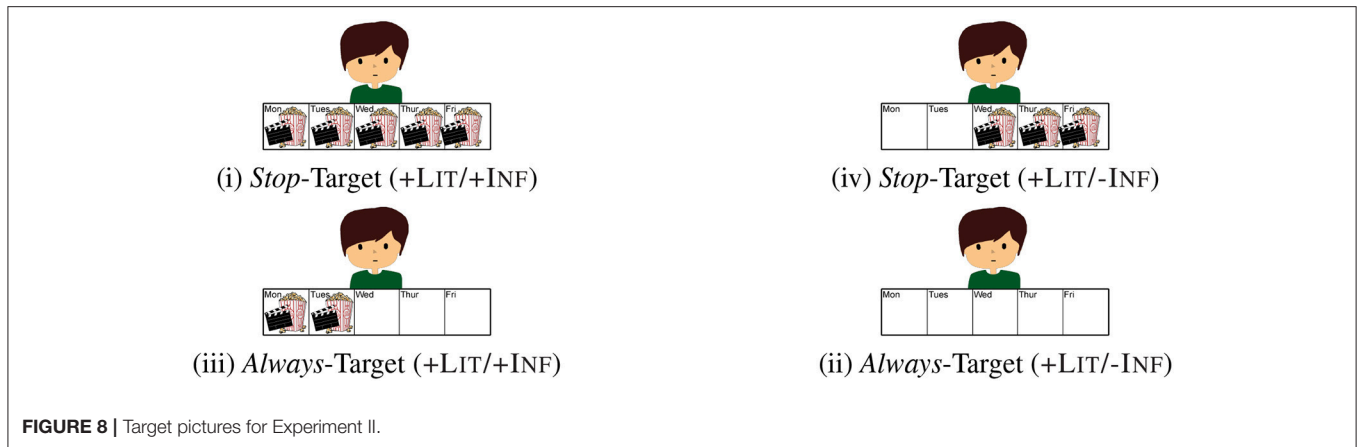


FIGURE 8 | Target pictures for Experiment II.

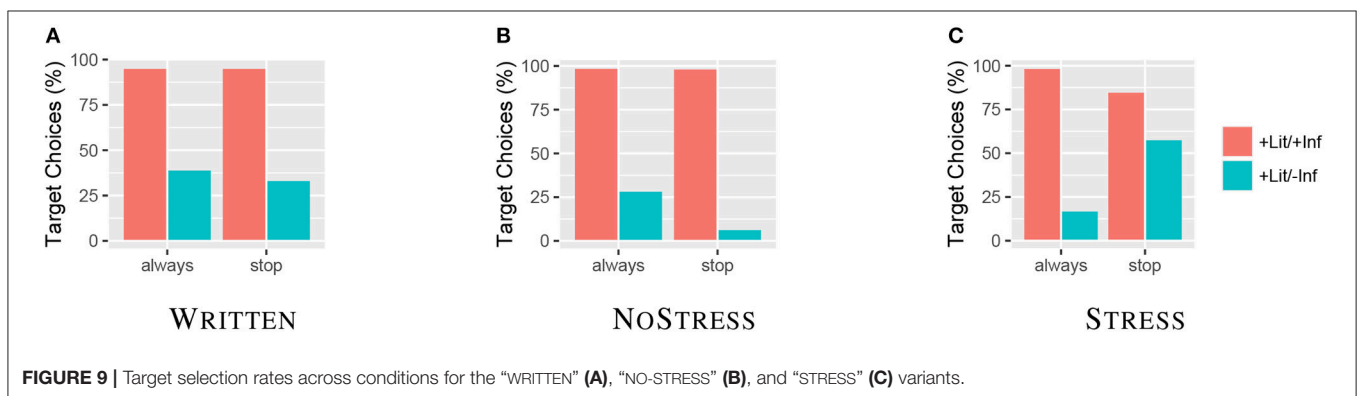


FIGURE 9 | Target selection rates across conditions for the “WRITTEN” (A), “NO-STRESS” (B), and “STRESS” (C) variants.

parallel ways based on reasoning over alternatives, then prosodic stress on the inference-triggering expression should have parallel effects. However, for *always*, we find that auditory stimuli in general increase the availability of inference interpretations. And at least numerically, in our results stress increases the likelihood of inference interpretations for implicature-triggers rather than decreasing it (although this effect did not come out as significant in our analyses)<sup>21</sup>. The effects for *stop*, on the other hand, go in opposite directions based on whether it is stressed or unstressed in the auditory versions. The latter leads to an increase in inference interpretations, whereas the former leads to a decrease. This last result is in line with the observations in the literature mentioned above, about stress on presuppositional trigger leading to an increase in no-inference interpretations. Most important for our purposes is the different effect of prosody on SIs and Ps, which is unexpected by the SI approach to Ps.

This difference in the effect of prosody on SIs and Ps provides a first clear argument against a unified analysis of the derivation of these inferences. In contrast, these results are perfectly compatible with a more traditional view that sees them as theoretically very different cases. The next section presents

further evidence along the same lines, produced as a result of evaluating the other identified prediction made by the SI approach to Ps. Namely, that in affirmative contexts, Ps and entailments should behave uniformly (i.e., (6-a)).

Before that, however, let us mention briefly how these results relate with the work on “scalar diversity” done by Van Tiel et al. (2016) (among others). This work has shown substantial variation in the derivation rates of different scalar implicatures. One might wonder whether the difference we have found between SIs and Ps might “just” be a sign of this scalar diversity, rather than evidence of different derivational mechanisms. However, the fact that the prosodic stress appears to have, not just *different*, but *opposite* effects on the derivation rates of these inferences is more in-line with a qualitative distinction between them (à la different derivational mechanisms), than a quantitative difference (à la scalar diversity).

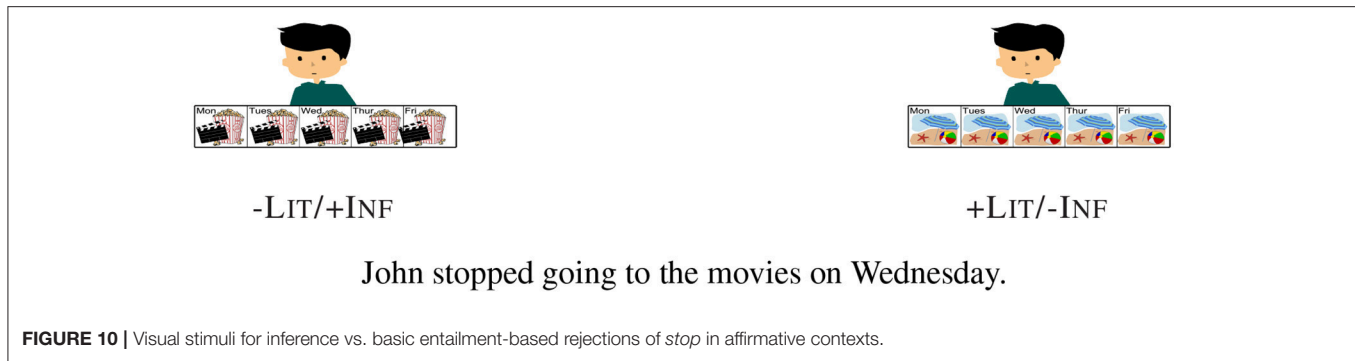
## 4.4. Experiment IIIa: Stop in Affirmative Sentences

### 4.4.1. Motivations

We set out to test the predictions of the SI approach to Ps, as presented in (6-a) and (6-b). Turning to the former, the approach sees Ps as simple entailments. This feature of SI approach to Ps predicts that—everything else being equal—the inference traditionally considered to be a P should be entirely on par with other entailed content. That is, they predict uniformity between

<sup>21</sup>Note however that this result is still compatible with the claim in the literature that stress on the trigger is a necessary but not sufficient condition for the no-inference interpretation to become available.





Ps and simple entailments in affirmative contexts. For example, according to the SI approach to Ps, *stop* in the following sentence is assumed to entail (and only to entail) both of the following:

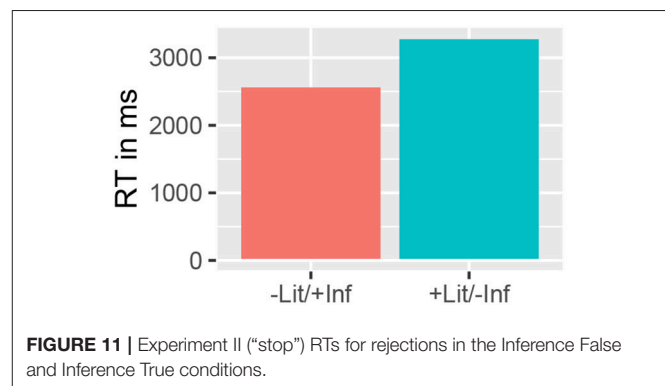
- (45) John stopped going to the movies on Wednesday.
- a. John did not go to the movies from Wednesday on.
  - b. John did go to the movies before.

Both these inferences are derived from the same sentence and, according to the SI approach to Ps, they are equivalent in status (i.e., they are both simply entailed). As a result, we take it that the SI approach to Ps would predict a greater degree of uniformity in the behavior of these inferences, compared to others we have investigated thus far. In particular, we take it that the SI approach to Ps predicts that rejecting a picture based on one of these should be just as fast as for the other. In contrast, traditional accounts posit that while both (45-a) and (45-b) are entailed by (45), (45-b) is also presupposed by (45) and thus differs in status from the first. More precisely, the fact that (45-b) is both entailed and presupposed might lead to different patterns in behavioral data than (45-a), which is simply entailed (see Kim, 2007; Schwarz, 2016b for previous instances of this approach to *only* and *definites*, respectively). We investigated the relationship between rejections based on either one of these two inferences in affirmative sentences.

#### 4.4.2. Methods

##### 4.4.2.1. Materials and Design

The materials of this experiment were part of the same overall experiment reported as Experiment Ib on *stop* in negative sentences above. Affirmative sentences with a presupposition trigger such as *stop* differ from those with DSIs in that they cannot be judged true in a context where the inference of interest (that the relevant activity had been going on before) is false. This renders such sentences unsuitable for a direct comparison with affirmative SI sentences (i.e., DSIs), but they provide a possible angle for assessing the status of the inference. Note first that rejection responses in such contexts are captured on both traditional accounts and the SI approach to Ps, though in different ways: the former sees it as a case of presupposition failure, whereas the latter sees it as a simple rejection based on unmet entailments. The contexts we used are depicted in **Figure 10**. In the -LIT/+INF condition, the overt picture does not



match the sentence based on its simply entailed content, since the movie-going continued past Wednesday, but the inference that John was going to the movies before Wednesday is met. In contrast, in the +LIT/-INF condition, the inference—be it both a presupposition and an entailment, or merely an entailment—is not met, while the simply entailed content, that there was no “movie-going” after Wednesday, does hold.

##### 4.4.2.2. Participants and Procedure

The data stem from the same 34 participants as in Experiment Ib, and the sentence-picture combinations that they saw were variants of the negative versions reported there. In particular, subjects saw 6 sentences in the -LIT/+INF condition and 6 in the +LIT/-INF condition, drawn from a total of 24 sentences, counterbalanced across groups as described above. The Instructions and procedure were as laid out for Experiment Ib, (see section 4.2.1).

##### 4.4.3. Results and Discussion

Unsurprisingly, Covered Picture selections were at ceiling level (over 97% for both conditions). RTs are illustrated in **Figure 11**. Covered Picture choices were slower in the +LIT/-INF condition (3,296 ms) than in the -LIT/+INF condition (2,583 ms). This difference was statistically significant, as confirmed by a mixed-effect regression analysis with random effects for subjects and items, including intercepts and slopes ( $\beta = -689.6$ ,  $SE = 203.1$ ,  $t = -3.40$ ,  $\chi^2 = 9.48$ ,  $p < 0.01$ ).

The observed difference in RTs points to a difference between the two ingredients of meaning at play. This pattern is not predicted by the SI approach to Ps, which would expect uniformity between these conditions, (6-a). On the other hand, it fits quite naturally with a traditional account, where one is presupposed and entailed, while, the other is simply entailed. Previous findings by Kim (2007) and Schwarz (2016b) have shown that rejection of sentences based on presupposed material is slower than rejection based on entailed content, and the present results fits into that picture straightforwardly on the traditional view. The SI approach to Ps does not offer an obvious explanation for this difference, as it sees both aspects of the meaning of (45) as simple entailments. However, one way of potentially saving the SI approach to Ps would be to challenge the assumption implicit in this interpretation of the data, namely that entailments of a sentence (that are generally comparable, specifically with regards to the task at hand), are on par with one another, specifically with respect to behavioral patterns such as those in RT results. An obvious approach to test this in light of our previous comparisons between *always* and *stop* is to look at different falsifying scenarios for the former. If we also find a difference between corresponding entailments associated with sentences containing *always*, then our current result for sentences containing *stop* would be less problematic for the SI approach to Ps.

#### 4.5. Experiment IIIb and c: Rejections of *Always* Based on Different Entailments

When we compared sentences with *always* to ones with *stop* under negation, there were two ingredients of the overall conveyed meaning, which differed in status when occurring under negation:

- (46) John didn't always go to the movies.
- There were times when John did not go to the movies.
  - John sometimes went to the movies.

The inferences in (46-a) and (46-b) are traditionally analyzed as an entailment and an SI, respectively. However, in the case of an affirmative *always* sentence like (47) both (46-b) and the negation of (46-a) (i.e., (47-a)) are entailed. This makes affirmative sentences like (47) a good test for the assumption that different aspects of the entailments of a sentence yield equivalent RT results when providing the grounds for rejection of the sentence.

- (47) John always went to the movies.
- It's not the case that there are times when John did not go to the movies.

Two follow-up experiments looked at rejections of positive *always*-sentences based on pictures corresponding to the two entailments in question. The design is illustrated in Figure 12.

The crucial manipulation was whether the *always* sentence was falsified by an overt picture where the depicted individual sometimes went to the movies or whether they never went to the movies. If the two different aspects of the overall entailments of

the sentences involved an asymmetry parallel to that found for the two ingredients of *stop*-sentences, then we would expect a similar RT-difference between the two conditions. In contrast, if no such difference is involved, we expect no RT-contrast, and an interaction with the results for *stop*. The latter prediction was borne out. RTs for the ALWAYS PICTURE (2,383 ms) and the NEVER PICTURE (2,321 ms) did not differ significantly from one another. Comparing the results statistically to those for *stop* reported above (analyzed as a between-subjects, within-items design with a maximal random effects structure for the latter) yielded a significant interaction ( $\beta = 743.1$ ,  $SE = 224.5$ ,  $t = -3.31$ ,  $\chi^2 = 9.12$ ,  $p < 0.01$ ).

A potential concern about this first follow-up is that it involved empty calendar slots. In particular, one might worry that the NEVER PICTURE version, which conceptually corresponded to the more difficult *stop*-condition with an unmet presupposition, might lend itself to a relatively easy task-strategy of rejection based on the completely empty calendar strip, thus hiding potential delay effects. A second follow-up addressed this issue by filling the relevant calendar slots with another image type instead (see right side of Figure 12). While there was a small numerical difference between the ALWAYS PICTURE (5,505 ms) and the NEVER PICTURE (5,735 ms) in the results of this experiment, the difference was not statistically significant<sup>22</sup>. Comparing these results to the data obtained for *stop* from above, we again find a statistical interaction ( $\beta = 156.13$ ,  $SE = 72.93$ ,  $t = -2.14$ ,  $\chi^2 = 4.48$ ,  $p < 0.05$ ).

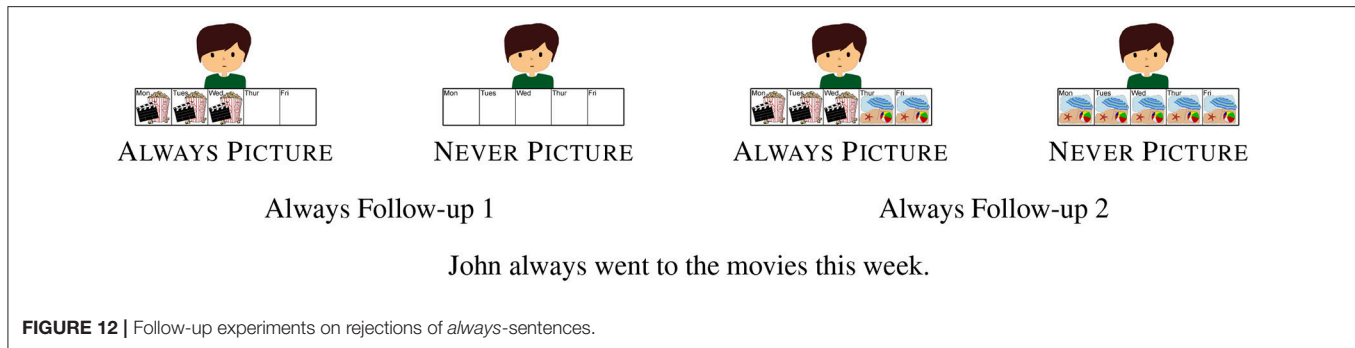
What both of these follow-ups suggest, then, is that while there is an asymmetry in the role of the two inferences in question in the case of *stop*, this is not the case for the different aspects of the entailments of *always*. While this of course does not conclusively show that all entailments have the same processing status, it further suggests that in the case of *stop*, we are not dealing with two aspects of the overall entailment, as posited by the SI approach to Ps. In contrast, these results are consistent with the traditional perspective that the relevant inferences associated with affirmative *stop* sentences (i.e., (45)) have different statuses (i.e., simply entailed vs. entailed and presupposed).

## 5. GENERAL DISCUSSION

We set out to investigate the SI approach to Ps by trying to answer the main question outlined in (48). The predictions of the SI approach to Ps in regards to this question are repeated in (49-a) and (49-b). Experiment Ia, Ib and II set out to test prediction in (49-b). Experiments IIIa-c tested the prediction in (49-a).

- (48) **Main question:** Do behavior patterns yield evidence for a distinction between Ps and entailments in affirmative contexts and between Ps and SIs in other contexts?

<sup>22</sup>Note that the overall longer RTs here are due to a slight variation in task, where a context sentence was included and the events in the calendar were revealed in two steps. Since the main measures of interest are a comparison between the two *always*-conditions and the interaction, this main effect of the task does not affect the interpretation of the results for our purposes.



- (49) **Predictions:** All else being equal,
- In affirmative contexts, Ps and entailments should behave uniformly.
  - In all other contexts, Ps and SIs should behave uniformly.

First, we will focus on Experiments Ia and Ib, as these produced results that were consistent with the prediction in (49-b). Following this, we will consider the other experiments, which produced results that were not in line with the predictions in (49-a) and (49-b), and discuss the challenge they pose for the SI approach to Ps.

### 5.1. What Doesn't Challenge the SI Approach to Ps

To briefly recap the situation in the literature, the classic finding since Bott and Noveck (2004) is that rejecting a sentence when its SI is false takes more time than accepting it. The same paradigm was then applied to Ps by Chemla and Bott (2013) and they found the opposite result: rejecting a negated sentence whose presupposition is not globally met takes *less* time than accepting it. On the basis of this result, Chemla and Bott (2013) concluded that Ps, unlike SIs, are not associated with a delay and that the answer to the question in (48) is positive: the processing of Ps and SIs is different, which in turn is a challenge for unified accounts like the SI approach to Ps. On the other hand, Romoli and Schwarz (2015) found that accepting negated sentences with a true presupposition is faster than accepting it when its P is not satisfied in the context, and they found parallel results for SIs, with faster acceptance of inference interpretations than no-inference interpretations. On the basis of this result, these authors concluded that there is no clear overall evidence for either SIs or Ps being associated with a delay or for the two inferences being different. On the face of it, the results from these two studies appear in conflict and they seem to give us opposite answers to the question of whether Ps and SIs differ. However, there is an obvious difference between these studies, which could account for the different results produced. Specifically, the two studies looked at different comparisons across acceptance and rejection responses; while Chemla and Bott (2013) compared acceptance vs. rejection responses of the same item, Romoli and Schwarz (2015) compared acceptance vs. acceptance responses across different items. Gaining a

comprehensive comparative perspective required looking at both acceptance and rejection responses systematically, and this constituted the main motivation for Experiment Ia and Ib.

In Experiment Ia, we compared direct and indirect SIs using the paradigm from Romoli and Schwarz (2015), to test whether their finding was specific to indirect SIs. Moreover, we extended their approach by comparing both acceptance vs. acceptance responses as well as rejection vs. rejection responses across items. Both direct and indirect SIs yielded faster responses in the inference condition than in the no-inference condition when we considered acceptance responses, thus replicating Romoli and Schwarz (2015) on indirect SIs and extending their results to direct ones. On the other hand, looking at rejections yielded the opposite pattern, as rejections in the inference condition were slower than in the no-inference condition. Thus, we find uniformity between direct and indirect SIs and we also reconcile the findings of Chemla and Bott (2013) and Romoli and Schwarz (2015) to some extent<sup>23</sup>. In Experiment Ib, we extended the same paradigm to Ps, by looking at sentences with *stop* under negation. The RT pattern was parallel to that for SIs, with a cross-over interaction reflecting opposite patterns for acceptance and rejection responses<sup>24</sup>.

The uniformity in the overall shape of the RT patterns of direct SIs, indirect SIs and Ps in these experiments is in line with the prediction in (49-b) and thus provides no evidence against the SIs approach to Ps. Moreover, we found no evidence for either Ps or SIs being associated with a delay in RTs, a point that we will return to in a moment.

### 5.2. What Does Challenge the SI Approach to Ps

In Experiment II, we investigated the effect of prosody on the availability of inference interpretations for SIs and Ps. In contrast to the results from Experiment Ia and Ib, the results

<sup>23</sup>Note that, while as far as RTs are concerned our results are comparable for ISIs and DSIs, the rate of implicature interpretations is significantly higher for DSIs. It's possible that this is simply due to complexities introduced by negation, but a more detailed explanation will have to be fleshed out in future work.

<sup>24</sup>Note that these results touch on an issue that has been investigated in detail elsewhere; namely, the effect of accepting/rejecting positive/negative sentences. In general, the work in this area seems to be consistent with our results, in that, judging sentences as true has been found to take longer than judging them as false (Wason, 1959). For a recent summary of the relevant literature see Dale and Duran (2011).

of Experiment II went against the prediction in (49-b). That is, Experiment II found directly opposite effects of placing prosodic stress on the inference-triggering expressions for SIs and Ps: inference rates decreased for SIs, relative to written stimuli, but increased for Ps. These results run against the SI approach to Ps' prediction of uniformity of behavior across these inferences.

With regards to the first prediction of the SI approach to Ps' (49-a), namely that in affirmative contexts, elements of meaning that have traditionally been thought of as Ps and entailments should behave uniformly. This prediction stems from the fact that the SI approach to Ps analyses the relevant inferences as simple entailments, and was addressed by Experiments IIIa-c. Experiment IIIa tested prediction (49-a) by comparing the entailment and the presupposition of "stop" in affirmative sentences. Specifically, it compared the behavior (measured as RTs) of participants who were rejecting a picture based on the notions that something was happening before or that it is not happening any longer, respectively. As the SI approach to Ps treats both of these elements of meaning as simple entailments, it did not predict a difference in RT behavior between these conditions. On the other hand, the traditional approach makes no specific predictions in regard to this comparison, but is perfectly compatible with there being a difference between the two. Experiment IIIa found a difference in the RTs associated with these different rejection responses, with slower responses for presupposition-based rejections, in line with previous findings (Kim, 2007; Schwarz, 2016b). This result is consistent with the traditional approach to Ps, but is a challenge for the SI approach to Ps. One way the SI approach to Ps could overcome this challenge would be to argue that not all simple entailments are on a par with one another with regard to RT behavior patterns, and so, Experiment IIIa's result should not be taken as indicative of a difference in their nature (i.e., they could still both be simple entailments of "stop"). Experiment IIIb and IIIc set out to explore this proposal by comparing the RTs associated with rejections based on two elements of meaning that have both been traditionally analyzed as simple entailments of "always." These experiments found no difference in the RT behavior of rejections based on these two different simple entailments. These results make the possible explanation of Experiment IIIa's results (that different simple entailments have differing RT patterns) by the SI approach to Ps less plausible. As this approach would now need to also explain why the RT behavior of the simple entailments of "stop" differed, while those of "always" did not.

It is worth considering these results in light of other recent experimental work which has also challenged the predictions of the SI approach to Ps. In particular, two other recent studies investigated the prediction in (49-b) by looking at how different populations interacted with these elements of meaning, using a Covered Picture selection task parallel to the one employed in the experiments reported here. Bill et al. (2016) and Kennedy et al. (2014) find that healthy adults, children (ranging from 4–7), and individuals with Broca's Aphasia (BAs) relate to Ps and SIs differently. Healthy adults and BAs tend to respond based on an inference reading when

responding to sentences associated with SIs, while children are more likely to access a no-inference reading. In contrast, for presuppositions, children and BAs pattern together and are more likely than healthy adults to respond based on an inference interpretation. Regardless of the exact explanation for each population's behavior in the respective cases, the fact that we get a dissociation in the patterns across populations, in particular with the BAs patterning with different groups for Ps and SIs, goes against the prediction in (49-b). Therefore, these results, combined with our present results provide strong evidence against treating SIs and Ps in an entirely uniform manner.

### 5.3. Are SIs (and Ps) Associated With RT Delays?

Results such as those found by Bott and Noveck (2004) are commonly interpreted to indicate that implicatures require a costly computation that lead to delays in processing (Bott and Noveck, 2004; Huang and Snedeker, 2009; Bott et al., 2012). Our results, on the other hand, did not involve a general delay in the inference conditions, for either SIs or Ps. In particular, when comparing acceptance judgments in Experiment Ia and Ib, cases where the Target picture was compatible with the inference interpretation were faster than ones where it was only compatible with the no-inference interpretation. This is incompatible with an account that simply posits two stages—an initial stage where only the literal meaning is available, and a later stage, where the inference interpretation is available—and maps these onto response time results. Both of the visible pictures involved in the acceptance comparison are compatible with the literal meaning, and thus should yield equivalent response patterns (or, if anything, a delay in the inference condition). In contrast with the acceptance comparison, the comparison of rejection responses yielded a pattern where responses based on an inference interpretation were slower. On their own, these might be seen as compatible with an account based on processing delays for inference interpretation. But given the cross-over interaction in our results, an alternative explanation of the effects is called for.

In the following, we sketch how the RT patterns in our data can be captured in terms of a conflict between pragmatic principles. To begin with, the relatively rapid acceptances based on inference interpretations suggests that the inferences are readily available. But why should the acceptance of pictures that are only compatible with a no-inference interpretation be slower? It cannot be due to a delay in availability of the no-inference interpretation since a), the inference interpretation entails the no-inference interpretation and b) rejections of pictures based on the no-inference reading are fast. An alternative explanation of the overall pattern in our data starts from the observation that delays arise precisely in those circumstances where the inference and no-inference interpretations conflict with one another. For example, we find relatively slow Target picture acceptances when the target is compatible with the no-inference interpretation but incompatible with the inference interpretation (**Figure 3A** for DSIs, **Figure 3C** for ISIs, and **Figure 3B** for



Ps). Similarly, Covered Picture selections are also slow in the very same circumstances. One possibility then, is that there are opposing pressures favoring the respective interpretations, and that delays arise precisely when there is a conflict between these factors. More specifically, we assume that comprehenders follow a general principle of charity, i.e., they generally try to construe utterances in such a way that they are true of the circumstances at hand. In our case, charity can plausibly be seen as corresponding to selecting the Target picture, as that is the obvious and salient option at hand. On the other hand, it is intuitively plausible that inference interpretations are generally preferred. For SIs, this is in line with naive speakers' intuitions about the meaning of *some*<sup>25</sup>. For Ps, a preference for an inference interpretation is in line with the common claim in the literature that interpretations including presuppositions seem to be the clear default, whereas no-inference interpretations are often thought to only be marginally available.

In sum, we assume the following two principles at work:

- (50) **Charity:** Construe sentences as true if possible<sup>26</sup>.
- (51) **Inference preference:** Inference interpretations are preferred (for both SIs and Ps)

The pressures of selecting the Target picture and the preference for inference interpretations oppose one another in precisely those conditions where we find a RT delay in our data. In the +LIT/−INF conditions, the principle of charity favors the Target picture, and the preference for inference interpretations favors the Covered Picture. Whether participants end up choosing the Target or the Covered Picture, their responses are delayed in these cases, compared to Covered Picture and Target picture selections in the relevant control conditions<sup>27</sup>. It is interesting to relate this account to an idea presented by Katsos and Bishop (2011), who explain acquisition data in terms of pragmatic tolerance: from our perspective, one could see this in terms of the charity principle being stronger in children than the preference for inference interpretations.

## 6. CONCLUSION

Recent proposals in the theoretical literature have put forth a unified view of a variety of inferences that traditionally have been seen as falling into different classes, under the umbrella of SIs. A simple and powerful approach to investigating these unified proposals experimentally is to compare the inferences

in question directly to one another, using behavioral measures. Everything else being equal, unified accounts predict uniform behavior. This approach has been applied fruitfully to the case of free choice inferences (Chemla and Bott, 2014; Tieu et al., 2016) and multiplicity inferences (Tieu et al., 2014), among others. We applied it to the comparison between classical SIs and Ps to investigate the uniformity prediction of recent SI approaches to Ps (Chemla, 2009; Romoli, 2015 among others). Previous results from the literature (Chemla and Bott, 2013; Romoli and Schwarz, 2015) bearing on this issue have yielded conflicting results. We proposed that the different results were due to differences in terms of what types of responses (in terms of acceptances vs. rejection responses) were compared. Our first few experiments (Ia & Ib) show that, once the acceptance vs. rejection pattern is factored in, then, in regards to the processing patterns, there is no longer any clear evidence for differences between the inference types. Furthermore, these results challenge the common interpretation of previous RT findings that implicatures are associated with an RT-delay due to the cost of computing these inferences online, and we sketched an alternative perspective based on our results. However, when we turned to Experiment II, we found that, counter to the predictions of the SI approach to Ps, there was a difference in the way these inferences were affected by prosody. In Experiment IIIa, we tested another prediction of SI approaches to Ps, namely that the relevant inferences of sentences including triggers like *stop* are simple entailments in affirmative contexts, which (again, everything else being equal) predicts uniform behavior with other simply entailed content. The results of this experiment showed that participants were slower to select the Covered Picture based on content that is traditionally thought to be entailed and presupposed compared with content traditionally thought to be simply/only entailed. These results are not consistent with the expectations of the SI approaches to Ps. In Experiments IIIb and c we investigated the plausibility of a possible explanation that SI approach to Ps could use to account for the differences in Experiment IIIa; that different simple entailments might show differing RT behavior. We investigated this possible claim by comparing the RT behavior associated with two simple entailments of “always,” and found no difference between them. These results reduce the plausibility of Experiment IIIa's results being accounted for with such an explanation. So, going back to the question of whether there is evidence from processing for a difference between SIs and Ps, we can now give it a positive answer: there is evidence for a difference between Ps and SIs. The first piece of evidence being the difference in the way Ps and SIs interact with prosody, and the second being the difference in how Ps and simple entailments are treated in affirmative sentences. Finally, our results link up quite nicely with recent evidence from the study of language acquisition (Bill et al., 2016) and Broca's Aphasia (Kennedy et al., 2014), which also produced results differentiating SIs and Ps in terms of responses patterns across populations. Considering these past findings, as well as our current results, it would appear that the SI approach to Ps is faced with a genuine challenge.

<sup>25</sup>Indeed, as anyone that has taught introductory logic can confirm, it takes substantial effort to convince students that *some*-statements are in principle compatible with universal scenarios, i.e., that *some* does not literally mean *some but not all*.

<sup>26</sup>In our set-up, this plays out as a pressure to select the Target picture, if possible.

<sup>27</sup>Note that, as RT-measurements are a relatively late and global measure of linguistic processing, our results do not preclude the possibility of there also being an initial delay associated with SI derivation, as found in studies measuring online processing more directly, such as Huang and Snedeker (2009) and others. Thanks to Jesse Snedeker for discussion on this point.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the National Statement on Ethical Conduct in Human Research (2007), National Health and Medical Research Council, Australian Government. The protocol was approved by the Macquarie University Human Research Ethics Committee. Data collection at the University of Pennsylvania took place with approval of the university's Institutional Review Board under protocol # 811457. All subjects gave written

informed consent in accordance with the Declaration of Helsinki.

## AUTHOR CONTRIBUTIONS

CB, JR, and FS equally contributed to designing and implementing all the reported experiments, as well as to writing this paper. CB and JR oversaw data collection for Experiment Ia, and FS for Experiments Ib, II, and IIIa-c. FS handled the statistical analyses of the data.

## FUNDING

We gratefully acknowledge support from NSF-grant BCS-1349009 to FS and the Australian Research Council Centre of Excellence in Cognition and its Disorders (CE110001021) to JR and CB.

## REFERENCES

- Abrusán, M. (2011). "Triggering verbal presuppositions," in *Semantics and Linguistic Theory (SALT) 20*, eds L. Nan and D. Lutz (Vancouver, BC), 684–701.
- Abrusán, M. (2016). Presupposition cancellation: explaining the 'soft-hard' trigger distinction. *Nat. Lang. Semantics* 24, 165–202. doi: 10.1007/s11050-016-9122-7
- Abusch, D. (2002). "Lexical alternatives as a source of pragmatic presupposition," in *Semantics and Linguistic Theory (SALT) 12*, ed B. Jackson (Ithaca, NY: CLC Publications), 1–19.
- Abusch, D. (2010). Presupposition triggering from alternatives. *J. Semant.* 27, 1–44. doi: 10.1093/jos/ffp009
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001
- Bates, D. M. (2005). Fitting linear mixed models in R. *R News* 5, 27–30. doi: 10.1016/j.cogpsych.2008.09.001
- Beaver, D. (2001). *Presupposition and Assertion in Dynamic Semantics*. Stanford, CA: CSLI Publications, Stanford University.
- Beaver, D. (2010). "Have you noticed that your belly button lint colour is related to the colour of your clothing?," in *Presuppositions and Discourse: Essays Offered to Hans Kamp*, eds R. Bauerle, U. Reyle, and T. E. Zimmerman (Bingley: Emerald group Publishing Limited), 65–100.
- Beaver, D., and Geurts, B. (2012). "Presuppositions," in *Semantics: An International Handbook of Natural Language Meaning, Vol. 3*, eds C. Maienborn, K. von Stechow, and P. Portner (Berlin: Mouton de Gruyter), 2432–2460.
- Bill, C., Romoli, J., Schwarz, F., and Crain, S. (2016). Scalar implicatures versus presuppositions: the view from acquisition. *Topoi* 35, 57–71. doi: 10.1007/s11245-014-9276-1
- Bott, L., Bailey, T. M., and Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *J. Mem. Lang.* 66, 123–142. doi: 10.1016/j.jml.2011.09.005
- Bott, L., and Noveck, I. (2004). Some utterances are underinformative. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006
- Breheny, R., Ferguson, H. J., and Katsos, N. (2013). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Lang. Cogn. Process.* 28, 443–467. doi: 10.1080/01690965.2011.649040
- Breheny, R., Katsos, N., and Williams, J. (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. doi: 10.1016/j.cognition.2005.07.003
- Chemla, E. (2009). Presuppositions of quantified sentences: experimental data. *Nat. Lang. Semant.* 17, 299–340. doi: 10.1007/s11050-009-9043-9
- Chemla, E. (2010). Similarity: towards a unified account of scalar implicatures, free choice permission and presupposition projection. Unpublished manuscript ENS.
- Chemla, E., and Bott, L. (2013). Processing presuppositions: dynamic semantics vs pragmatic enrichment. *Lang. Cogn. Process.* 38, 241–260. doi: 10.1080/01690965.2011.615221
- Chemla, E., and Bott, L. (2014). Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition* 130, 380–396. doi: 10.1016/j.cognition.2013.11.013
- Chemla, E., and Singh, R. (2014). Remarks on the experimental turn in the study of scalar implicature, Part II. *Lang. Linguist. Compass.* 8, 387–399. doi: 10.1111/lnc3.12080
- Chevallier, C., Noveck, I., Nazir, T., Bott, L., Lanzetti, V., and Sperber, D. (2008). Making disjunctions exclusive. *Q. J. Exp. Psychol.* 61, 1750–1761. doi: 10.1080/17470210701712960
- Chierchia, G. (2004). "Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface," in *Structures and Beyond: The Cartography of Syntactic Structures*, Vol. 3, ed A. Belletti (Oxford: Oxford University Press), 39–103.
- Cremers, A., and Chemla, E. (2014). "Direct and indirect scalar implicatures share the same processing signature," in *Pragmatics, Semantics and the Case of Scalar Implicatures*, Language and Cognition, ed S. Pistoia Reda (Basingstoke: Palgrave Macmillan), 201–240.
- Dale, R., and Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cogn. Sci.* 35, 983–996. doi: 10.1111/j.1551-6709.2010.01164.x
- Esipova, M. (2018). "Focus on what's not at issue: gestures, presuppositions, appositives under contrastive focus," in *Proceedings of Sinn und Bedeutung (SUB) 22* (Berlin: ZAS Papers in Linguistics), 367–384.
- Foppolo, F., and Marelli, M. (2017). No delay for some inferences. *J. Semant.* 34, 659–681. doi: 10.1093/jos/ffx013
- Fox, D., and Spector, B. (2018). Economy and embedded exhaustification. *Nat. Lang. Semantics* 26, 1–50. doi: 10.1007/s11050-017-9139-6
- Gamut (1991). *Logic, Language and Meaning*. Chicago, IL: University of Chicago Press.
- Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition, and Logical Form*. New York, NY: Academic Press.
- Grice, P. (1975). "Logic and conversation," in *The Logic of Grammar*, eds D. Davidson and G. Harman (Encino, CA: Dickenson), 64–75.
- Grodner, D. J., Klein, N. M., Carbary, K. M., and Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition* 116, 42–55. doi: 10.1016/j.cognition.2010.03.014
- Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts, Amherst.

- Heim, I. (1983). "On the projection problem for presuppositions," in *Proceedings of WCCFL 2*, ed D. P. Flickinger (Stanford, CA: Stanford University, CSLI Publications), 114–125.
- Heim, I., and Kratzer, A. (1998). *Semantics in Generative Grammar*. Malden, MA: Blackwell.
- Horn, L. (1972). *On the Semantic Properties of Logical Operators in English*. Ph.D. thesis, UCLA.
- Horn, L. (1989). *A Natural History of Negation*. Chicago, IL: University of Chicago Press.
- Huang, Y., Spelke, E., and Snedeker, J. (2013). What exactly do number words mean? *Lang. Learn. Dev.* 9, 105–129. doi: 10.1080/15475441.2012.658731
- Huang, Y. T., and Snedeker, J. (2009). Online interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cogn. Psychol.* 58, 376–415.
- Karttunen, L. (1974). Presupposition and linguistic context. *Theor. Linguist.* 1, 181–194. doi: 10.1515/thli.1974.1.1-3.181
- Katsos, N., and Bishop, D. V. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015
- Kennedy, L., Bill, C., Schwarz, F., Crain, S., Folli, R., and Romoli, J. (2014). "Scalar implicatures vs presuppositions: the view from Broca's aphasia," in *Proceedings of NELS 40* (Amherst, MA: GLSA).
- Kim, C. (2007). "Processing presupposition: verifying sentences with 'only,'" in *Papers from 31st Penn Linguistics Colloquium*, eds J. Tauberer, A. Eliam, and L. MacKenzie (Philadelphia, PA: Penn Working Papers in Linguistics), 213–226.
- Magri, G. (2010). *A Theory of Individual-Level Predicates Based on Blind Mandatory Scalar Implicatures*. Ph.D. thesis, Massachusetts Institute of Technology.
- Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicatures. *Cognition* 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1
- Papafragou, A., and Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition* 86, 253–282. doi: 10.1016/S0010-0277(02)00179-8
- Romoli, J. (2012). *Soft But Strong: Neg-Raising, Soft Triggers, and Exhaustification*. Ph.D. thesis, Harvard University.
- Romoli, J. (2015). The presuppositions of soft triggers are obligatory scalar implicatures. *J. Semant.* 32, 173–219. doi: 10.1093/jos/fft017
- Romoli, J., and Sauerland, U. (2017). "Presupposition and accommodation," in *The Routledge Handbook of Pragmatics*, eds A. Barron, G. Yueguo, and G. Steen (London: Routledge), 257–276.
- Romoli, J., and Schwarz, F. (2015). "An experimental comparison between presuppositions and indirect scalar implicatures," in *Experimental Perspectives on Presuppositions*, Studies in Theoretical Psycholinguistics, ed F. Schwarz (Dordrecht: Springer), 215–240.
- Schlenker, P. (2008). Be articulate: a pragmatic theory of presupposition projection. *Theor. Linguist.* 34, 157–212. doi: 10.1515/THLI.2008.013
- Schwarz, F. (2015). "Introduction: aspects of meaning in context - theoretical issues and experimental perspectives," in *Experimental Perspectives on Presuppositions*, ed F. Schwarz (New York, NY: Springer), 1–38.
- Schwarz, F. (2016a). Experimental work in presupposition and presupposition projection. *Annu. Rev. Linguist.* 2, 273–292. doi: 10.1146/annurev-linguistics-011415-040809
- Schwarz, F. (2016b). False but slow: evaluating statements with non-referring definites. *J. Semant.* 33, 177–214. doi: 10.1093/jos/ffu019
- Schwarz, F., Romoli, J., and Bill, C. (2015). Scalar implicatures processing: slowly accepting the truth (literally). in *Proceedings of Sinn und Bedeutung*, ed E. C. Zeijlstra (Göttingen), 553–570.
- Simons, M. (2001). "On the conversational basis of some presuppositions," in *Semantics and Linguistic Theory (SALT) 11*, eds R. Hastings, B. Jackson, and Z. Zvolenszky (Ithaca, NY: Cornell University), 431–448.
- Simons, M., Roberts, C., Beaver, D., and Tonhauser, J. (2017). The best question: explaining the projection behavior of factives. *Discourse Process.* 54, 187–206. doi: 10.1080/0163853X.2016.1150660
- Stalnaker, R. (1974). "Pragmatic presuppositions," in *Semantics and Philosophy*, eds M. Munitz and D. Unger (New York, NY: New York University Press), 197–213.
- Sudo, Y. (2012). *On the Semantics of Phi Features on Pronouns*. Ph.D. thesis, MIT.
- Tieu, L., Bill, C., Romoli, J., and Stephen, C. (2014). "Plurality inferences are scalar implicatures: Evidence from acquisition," in *Proceedings of the 24th Semantics and Linguistic Theory Conference*, eds T. Snider, S. D'Antonio, and M. Weigand (New York, NY: New York University), 122–136.
- Tieu, L., Romoli, J., Peng, Z., and Crain, S. (2016). Children's knowledge of free choice inferences and scalar implicatures. *J. Semant.* 33, 269–298. doi: 10.1093/jos/ffv001
- van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *J. Semant.* 9, 333–377. doi: 10.1093/jos/9.4.333
- Van Tiel, B., Van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *J. Semant.* 33, 137–175. doi: 10.1093/jos/ffu017
- Von Stechow, K. (2008). What is presupposition accommodation, again? *Philos. Perspect.* 22, 137–170. doi: 10.1111/j.1520-8583.2008.00144.x
- Wason, P. C. (1959). The processing of positive and negative information. *Q. J. Exp. Psychol.* 11, 92–107. doi: 10.1080/17470215908416296
- Wilson, D. (1975). Presupposition, assertion, and lexical items. *Linguist. Inq.* 6, 95–114.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Bill, Romoli and Schwarz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Context-Sensitivity and Individual Differences in the Derivation of Scalar Implicature

Xiao Yang<sup>1,2\*</sup>, Utako Minai<sup>2</sup> and Robert Fiorentino<sup>1</sup>

<sup>1</sup> Neurolinguistics and Language Processing Laboratory, Department of Linguistics, University of Kansas, Lawrence, KS, United States, <sup>2</sup> Developmental Psycholinguistics Laboratory, Department of Linguistics, University of Kansas, Lawrence, KS, United States

## OPEN ACCESS

### Edited by:

Anne Colette Reboul,  
Claude Bernard University Lyon 1,  
France

### Reviewed by:

Jacques Moeschler,  
Université de Genève, Switzerland  
Cristina Grisot,  
Université de Genève, Switzerland

### \*Correspondence:

Xiao Yang  
xiaoyang@ku.edu

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 01 June 2018

**Accepted:** 24 August 2018

**Published:** 20 September 2018

### Citation:

Yang X, Minai U and Fiorentino R  
(2018) Context-Sensitivity  
and Individual Differences  
in the Derivation of Scalar Implicature.  
Front. Psychol. 9:1720.  
doi: 10.3389/fpsyg.2018.01720

The derivation of scalar implicatures for the quantifier *some* has been widely studied to investigate the computation of pragmatically enriched meanings. For example, the sentence “I found some books” carries the semantic interpretation that at least one book was found, but its interpretation is often enriched to include the implicature that not all the books were found. The implicature is argued to be more likely to arise when it is relevant for addressing a question under discussion (QUD) in the context, e.g., when “I found some books” is uttered in response to “Did you find all the books?” as opposed to “Did you find any books?”. However, most experimental studies have not examined the influence of context on *some*, instead testing *some* sentences in isolation. Moreover, no study to our knowledge has examined individual differences in the ability to utilize context in interpreting *some*, whereas individual variation in deriving implicatures for *some* sentences in isolation is widely attested, with alternative proposals attributing this variation to individual differences in cognitive resources (e.g., working memory) or personality-based pragmatic abilities (e.g., as assessed by the Autism-Spectrum Quotient). The current study examined how context influences the interpretation of *some* in a story-sentence matching task, where participants rated *some* statements (“I cut some steaks”) uttered by one character, in response to another character’s question (QUD) that established the implicature as relevant (“Did you cut all the steaks?”) or irrelevant (“Did you cut any steaks?”). We also examined to what extent individuals’ sensitivity to QUD is modulated by individual differences via a battery of measures assessing cognitive resources, personality-based pragmatic abilities, and language abilities (which have been argued to modulate comprehension in other domains). Our results demonstrate that QUD affects the interpretation of *some*, and reveal that individual differences in sensitivity to QUD are modulated by both cognitive resources and personality-based pragmatic abilities. While previous studies have argued alternatively for cognitive resources or personality-based pragmatic abilities as important for deriving implicatures for *some* in isolation, we argue that arriving at a context-sensitive interpretation for *some* depends on both cognitive and personality-based properties of the individual.

**Keywords:** scalar implicature, question under discussion (QUD), individual differences, working memory, attentional control, Autism-Spectrum Quotient (AQ), pragmatic abilities



## INTRODUCTION

In conversational exchanges, interlocutors commonly convey meanings which go beyond the literal semantic content of the utterance and require the generation of pragmatic inferences on the part of the comprehender. A widely researched phenomenon which is argued to involve pragmatic inferencing is the interpretation of the quantifier *some*. For instance, the utterance in (1) semantically entails that *at least one, and possibly all* of the students is hardworking, yet pragmatically the interpretation is often enriched with the implicature that *not all* of the students are hardworking (Noveck and Sperber, 2007; Katsos and Cummins, 2010).

- (1) Some of the students are hardworking.  
 Semantic entailment: At least one, and possibly all of the students is hardworking.  
 Pragmatic implicature: Not all the students are hardworking.

The two readings differ in whether *all* is negated, since the semantic reading does not exclude the possibility that *all* may hold. In addition, the pragmatic implicature differs from the inherent semantic meaning, in that *not all* is cancellable but the semantic entailment *at least one* is not (Grice, 1989; see also Geurts, 2010), as shown in example (2) below:

- (2) Non-cancellable semantic entailment:  
 Some of the students are hardworking. #In fact, none of them are.  
 Cancellable pragmatic implicature:  
 Some of the students are hardworking. In fact, all of them are.

It has been argued that *some* is on a scale of quantifiers varying in informativity (i.e., how specific a quantifier is), ranging from the least to the most informative, e.g., *<some, many, all>* (Horn, 1972). Scalar implicature thus refers to the common intuition that a less informative item implies the negation of a more informative item on the scale, with *some* taken to imply *not all*. This meaning is often argued to arise due to interlocutors' expectation that utterances shall be optimally informative, as formalized by Gricean maxims (Grice, 1989); thus, the comprehender can infer that the speaker must mean that the more informative term *all* does not apply if they opted to use the less informative term *some*.

It is important to point out, however, that *some* is not interpreted with the *not all* implicature in all cases. For example, linguistic analyses of *some* highlight that the likelihood of interpreting *some* with the implicature is heavily influenced by the broader context in which the *some* sentence appears (Roberts, 2004; see also Chierchia et al., 2012). One specific context-level factor that has been argued to play an important role in determining whether *some* is interpreted with the implicature is the question under discussion (QUD). QUD refers to the crucial issue in the discourse that is expected to be addressed by a relevant answer. The extent to which the *not all* implicature is realized in the interpretation of *some* is argued to depend on whether it is relevant under the current QUD (Roberts, 2004, 2012). Consider the following conversational exchanges, where

the *some* utterances made by Speaker B in (3) and (4) are in response to different questions asked by Speaker A (examples adapted from Levinson, 2000; Politzer-Ahles and Fiorentino, 2013):

- (3) Upper-bound QUD  
 Speaker A: "Are all the students in this lab hardworking?"  
 Speaker B: "Some of them are."  
 (4) Lower-bound QUD  
 Speaker A: "Is there any evidence against them?"  
 Speaker B: "Some of their documents are forgeries."

The reading that *not all* the students are hardworking strongly arises in B's reply in (3). However, B's reply in (4) can be felicitously interpreted without the *not all* implicature, as *at least one and possibly all* of their documents was a forgery. This is due to the difference in QUD in the two conversations, established by A's questions. The QUD in (3) involves *all* the students, thus *some* in the reply should address A's question and consequently be interpreted with the *not all* implicature. Conversely, the QUD in (4) involves whether there is *at least one* piece of evidence; thus *some* can be simply interpreted as *at least one* without the *not all* implicature, as *all* is irrelevant under A's question. A QUD that highlights *all* and thus encourages the *not all* implicature is often termed as *upper-bound* (as in 3), while a QUD that does not encourage the implicature is termed as *lower-bound* (as in 4).

As is illustrated in examples (3) and (4) above, the QUD is often established in discourse through the linguistic utterances of the interlocutors in a conversation. These utterances may establish an issue that needs to be addressed, and thus indicate what is expected from an appropriate answer in the current discourse. In examples (3) and (4) above, for example, Speaker A establishes a QUD through a linguistic utterance whose properties (e.g., the choice of "all" versus "any") set the stage for interpreting subsequent utterances containing *some* with or without the implicature. Thus, even though an answer sentence with *some* may be ambiguous by itself, contextual cues such as QUD are often provided by interlocutors which can be used to disambiguate the optimal reading of *some* in the discourse.

We would like to note that there are a range of theories that aim to account for how the *not all* implicature is generated, ranging from those positing the implicature is only generated when relevant (e.g., Relevance Theory approaches; Sperber and Wilson, 2002), to those in which the implicature is always generated, but may be canceled when not relevant (e.g., the Default view; Levinson, 2000), as well as views in which the *not all* interpretation is due to the presence/absence of a silent exhaustive focus operator (e.g., the Grammatical view; Chierchia, 2004; Chierchia et al., 2012) rather than to pragmatic inferencing. While these approaches differ with respect to the specific mechanism by which the *not all* meaning comes into consideration as part of the interpretation of *some*, they share the common assumption that context is important in determining whether *some* is ultimately interpreted with this meaning<sup>1</sup>. However, despite the crucial role of context

<sup>1</sup>For example, under the Default view (Levinson, 2000), the implicature can be canceled when it is irrelevant in the context, while under Relevance Theory



highlighted in linguistic analyses of scalar implicature, only a handful of experimental studies have examined the extent to which comprehenders are indeed sensitive to context in interpreting *some*, with the majority of the literature instead testing *some* sentences in isolation, as we discuss below. Thus, the primary aim of the current study is to determine experimentally the extent to which comprehenders are sensitive to contextual information such as QUD and to characterize and account for the variability that individuals may show in sensitivity to context in interpreting *some*.

## EXPERIMENTAL STUDIES ON SCALAR IMPLICATURE

Studies on scalar implicature that do not manipulate context often test *some* sentences that are underinformative. These sentences are semantically true but pragmatically infelicitous based on, for example, world knowledge, such as the sentence in (5a) (examples from Noveck and Posada, 2003).

- (5a) Underinformative:                Some dogs have paws.  
 (5b) True and informative:        Some people have pets.

In (5a), if *some* is interpreted as *some but not all*, the sentence is pragmatically infelicitous, as it would be more informative to use *all* instead (since all dogs have paws), although the sentence is semantically true (at least one dog has paws). Underinformative sentences can thus be used to test whether or not a scalar implicature has been generated: if the *not all* implicature is realized, the infelicity should lead to increased rates of rejection as compared to true and informative sentences such as (5b), in judgment tasks, and to evidence of processing disruption in online studies. Using this paradigm, studies have found that adult native speakers, when analyzed as a group, generally show sensitivity to pragmatic infelicity (e.g., Noveck and Posada, 2003; Bott and Noveck, 2004; De Neys and Schaeken, 2007; Huang and Snedeker, 2009; Hunt et al., 2013; Tomlinson et al., 2013).

However, one major drawback in studies establishing underinformativity based on world knowledge (as in example 5a) is that they require that participants draw on their world knowledge and verify if counterexamples exist to evaluate the sentences (e.g., dogs that have no paws). Judgments may thus depend on participants' ability to consult the relevant world knowledge, and may vary based on participants' beliefs about how typical the world under discussion is when they are presented with these odd utterances (see Degen et al., 2015). Another drawback is that effects of underinformativity can be confounded with lexical differences across conditions. Notice that (5a) for

example contains the lexical items *dogs* and *paws*, while (5b) contains *people* and *pets* (see, e.g., Nieuwland et al., 2010 for discussion regarding this concern). Issues regarding world knowledge do not arise in variants of the underinformativity approach that provide the information needed to determine the felicity of the *some* sentences. This is often done by presenting a visual display with a number of objects and then asking participants to evaluate a *some* sentence about these objects (e.g., Huang and Snedeker, 2009; Hunt et al., 2013; Antoniou et al., 2016; among others). However, a general limitation of studies testing *some* sentences presented in isolation is that they do not directly target the comprehension of *some* within a broader context that provides information indicating whether the implicature is relevant, which might better approximate how comprehenders typically must interpret *some* sentences during everyday language use.

Another finding from this line of research on *some* in isolation is that individual native speakers have been shown to vary greatly from one another in terms of whether the *not all* implicature is derived. Many studies have revealed that native speakers generally fall into two groups: one group of speakers that consistently rejects underinformative sentences, suggesting that they interpret *some* pragmatically, and another group that consistently accepts underinformative sentences, suggesting that they interpret *some* semantically, without realizing the implicature (Noveck and Posada, 2003; Bott and Noveck, 2004; Hunt et al., 2013; Heyman and Schaeken, 2015; Antoniou et al., 2016). In a picture-sentence verification study on *some* by Hunt et al. (2013), for example, among the 24 adult native speakers of English tested, 11 rejected over 80% of the underinformative sentences, while 10 accepted over 80% of the underinformative sentences and only 3 showed no strong preference. Researchers have begun to investigate what underlies this individual variation, examining which properties of the individual may modulate the extent to which they are likely to derive scalar implicatures during the processing of *some* in isolation (e.g., Nieuwland et al., 2010; Dieussaert et al., 2011; Marty and Chemla, 2013; Tomlinson et al., 2013; Heyman and Schaeken, 2015; Antoniou et al., 2016). We review this literature in Section "Two Accounts for Individual Differences in Scalar Implicature" below.

A handful of studies have investigated the effect of context on the comprehension of *some* by manipulating QUD within a discourse (e.g., Breheny et al., 2006; Zondervan et al., 2008; Politzer-Ahles and Fiorentino, 2013; Degen and Goodman, 2014; Dupuy et al., 2016; Politzer-Ahles and Husband, 2018). In a recent study on scalar implicature in French by Dupuy et al. (2016), participants were presented with visual stories in which a character acted upon all objects (e.g., a boy hiding five out of five car toys), and a question-answer dialog about the story between two puppets (Dupuy et al., 2016, experiment 3). The first puppet's question was either "Did the boy hide all the cars?" or "Did the boy hide cars?", representing either an upper-bound QUD or a lower-bound QUD, to which the second puppet answered, "The boy has hidden some cars." Participants were asked to judge if the second puppet's answer was right by selecting "yes" or "no." Dupuy et al. (2016) found that participants were more likely to select "no" under the upper-bound QUD compared to the

view (Sperber and Wilson, 2002), context may affect whether the implicature is generated in the first place (see also the constraint-based account recently proposed by Degen and Tanenhaus, 2015). Under the Grammatical view (Chierchia, 2004), one may argue that context determines whether the silent operator is present or not. As such, the context manipulation in the current study is not intended to adjudicate among these alternative accounts of scalar implicature derivation, but instead to examine to what extent comprehenders are indeed sensitive to contextual information when interpreting *some*, and to probe the extent and origins of individual differences in sensitivity to context in the interpretation of *some*.

lower-bound QUD, suggesting that contextual information such as QUD does influence the comprehension of *some*. Although these studies have yielded evidence for the influence of context, none of them has systematically addressed individual differences. This leaves open the question of to what extent individuals differ in comprehending *some* utterances as dictated by the demands of context, and what abilities may make one better able to compute context-dependent interpretations for *some*. We address this question in the current study.

## Two Accounts for Individual Differences in Scalar Implicature

As discussed by Antoniou et al. (2016), the literature has yielded two main accounts which offer qualitatively different explanations for the individual variation observed in the comprehension of *some*: the “personality-based” account and the “cognitive resources” account.

The personality-based account posits that an individual’s likelihood of interpreting *some* with the *not all* implicature depends on personality traits, such as one’s awareness of the pragmatic use of language in everyday life (Nieuwland et al., 2010; Katsos and Bishop, 2011; Feeney and Bonnefon, 2013). Nieuwland et al. (2010), for example, examined the relationship between unimpaired adult individuals’ interpretation of *some* and their scores on the Autism-Spectrum Quotient (AQ, Baron-Cohen et al., 2001), a questionnaire assessing individuals’ autistic traits in a range of domains, including the everyday use of language in social communication (the Communication subscale of the AQ, “AQ-Comm subscale”). To examine individuals’ derivation of scalar implicatures, they compared brain responses to the object word in underinformative sentences (e.g., *lungs* in 6a) and in true and felicitous control sentences (e.g., *pets* in 6b).

- (6a) Underinformative: Some people have lungs, which require good care.
- (6b) True and informative: Some people have pets, which require good care.

Nieuwland et al. (2010) found a larger N400 EEG response for the object in underinformative sentences (6a) as compared to the control sentences (6b). However, this effect was limited to a subgroup of participants with better sensitivity to the pragmatic use of language in social communication as measured by AQ-Comm. In contrast, a subgroup with less sensitivity to the pragmatic use of language in social communication showed an effect in the opposite direction. Nieuwland et al. (2010) thus suggested that the ability to realize the *not all* implicature online depends on an individual’s awareness of the pragmatic aspects of language use in everyday life (see also Feeney and Bonnefon, 2013 for similar findings regarding the relation between self-perceived honesty and the scalar item *or*). However, the effects of personality-based factors have not been consistently found across studies; for example, Heyman and Schaeken (2015) did not find a robust relation between the interpretation of *some* and their personality-based measures, which included AQ and the Big-Five Personality Test, thus leaving open the question of to what extent personality traits modulate scalar implicature derivation.

In contrast to the personality-based account, the cognitive resources account proposes that variation in cognitive resources may affect the extent to which an individual is able to interpret *some* with the *not all* implicature. Scalar implicature has been characterized by some researchers as a costly process potentially involving multiple processing steps (De Neys and Schaeken, 2007; Huang and Snedeker, 2009; Dieussaert et al., 2011; Marty and Chemla, 2013; Tomlinson et al., 2013; see also Barbet and Thierry, 2016). For example, the generation of implicatures itself may be costly, which is particularly emphasized in psycholinguistic accounts of scalar implicature that reference processing cost in arguing that implicatures may be generated only when relevant to the context rather than by default. The interpretation of *some* also arguably involves processes such as the encoding and maintenance of information, including information regarding the context and the interlocutors in the context, in order to determine whether the interpretation of *some* is more optimal with or without the implicature. It may also require switching between generated interpretations of *some* that do or do not have implicature (or, under the Grammatical view, representations that do or do not include a silent operator), all of which may rely on an individual’s cognitive resources such as working memory and attentional control (see also Antoniou et al., 2016 for a recent discussion of how cognitive resources may come into play under alternative conceptions of how implicatures are generated).

Studies examining the influence of cognitive resources have typically adopted dual-task paradigms where participants respond to underinformative *some* statements while simultaneously attending to a secondary task to which cognitive resources must be allocated (De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013; Heyman and Schaeken, 2015). Other studies have included independent measures of individuals’ cognitive abilities in order to test the relationship between these cognitive resources and the processing of *some* sentences (Antoniou et al., 2016). For example, Dieussaert et al. (2011) elicited participants’ true/false judgments for underinformative *some* sentences while simultaneously memorizing dot patterns. Participants were asked to first memorize a dot pattern. Next, they judged an underinformative *some* sentence. Finally, they were prompted to reproduce the dot pattern. Differing in complexity, the dot patterns were intended to engender either a high cognitive load or a low cognitive load. Individual participants’ working memory capacity was also assessed via the Operation Span Task. Dieussaert et al. (2011) found that participants were overall more likely to accept underinformative sentences when they tried to memorize high-load patterns than low-load patterns. This effect was only observed among the participants with low working memory capacity, while those with higher working memory capacity showed similar judgments regardless of high or low cognitive load. Dieussaert et al. (2011) thus suggested that realizing the *not all* implicature requires sufficient cognitive resources.

While the literature on scalar implicature has typically focused either on cognitive resources or on personality traits, to our knowledge there have only been a few studies on the derivation of

scalar implicature for *some* that examined the role of both types of factor in the same study (Heyman and Schaeken, 2015; Antoniou et al., 2016; for an examination of scalar terms other than quantifiers, see Husband, 2014). Heyman and Schaeken (2015) examined the effect of a range of factors, including cognitive abilities and personality traits, on Dutch speakers' judgments for underinformative statements based on world knowledge, such as *Some oaks are trees*. They found that neither cognitive nor personality-based factors robustly predicted individual variation in speakers' judgments.

However, a recent study by Antoniou et al. (2016) revisited this issue, testing the effects of both types of individual differences on the interpretation of underinformative *some* sentences when underinformativity was established within the experiment, rather than based on world knowledge. In a picture-sentence verification task, participants were asked to judge underinformative statements like "There are hearts on some of the cards" as true or false, based on a visual display showing hearts on all five cards. Participants were also assessed on a battery of measures targeting cognitive resources and personality-based factors, including working memory (Backward Digit Span Task and Reading Span Task), attentional control (Stroop Task and the Simon task), cognitive flexibility (the Number-letter Task), autistic traits (Autism-Spectrum Quotient), personality traits (Big Five Inventory and Honesty/Integrity/Authenticity scale), and verbal and non-verbal IQ. Antoniou et al. (2016) found that interpreting *some* with the *not all* implicature was robustly predicted by working memory and age; individuals with larger working memory capacity were more likely to consistently derive the implicature (rejecting at least 4 out of 6 underinformative sentences in their study) than those with smaller working memory capacity, as were younger individuals. Other individual difference measures, including those assessing autistic and personality traits (e.g., Autism-Spectrum Quotient, Big-Five Inventory, and Honesty/Integrity/Authenticity Scale), did not turn out to be significant predictors. Antoniou et al. (2016) interpret the results as lending support for the cognitive resources account. They posit that the processes involved in computing scalar implicature may demand sufficient working memory, but also note that their stimuli, which included a high proportion of unambiguous fillers (e.g., when the picture shows 3/5 cards with stars), could also place a burden on working memory resources and thus hinder implicature generation. Regarding their finding that personality-based factors did not modulate scalar implicature derivation, Antoniou et al. (2016) suggest that personality-based factors might not account for variability in interpreting *some* robustly when working memory is included in the analysis, which previous studies like Nieuwland et al. (2010) did not test. However, they also discuss the possibility that personality-based factors may be more important for interpreting *some* sentences in richer discourse contexts. They speculate that, when tested in more naturalistic/communicative discourse contexts, it is possible that both cognitive and personality-based factors may modulate an individual's likelihood of deriving the implicature for *some* (see Marty and Chemla, 2013 for similar discussion).

## CURRENT STUDY

In the current study, we examine the role of individual differences in the derivation of scalar implicatures for *some* when presented in a communicative discourse context involving two interlocutors. Our primary aim is to directly test whether individual differences in scalar implicature derivation for *some* in context are better accounted for by cognitive resources or by personality-based pragmatic abilities, or whether both cognitive and personality-based abilities may play a role, as speculated by Antoniou et al. (2016). We address the following two main research questions in this study: First, we examine whether individuals' interpretation of *some* is influenced by the QUD, which is established via a brief discourse context involving utterances by two interlocutors. If comprehenders are able to utilize QUD in interpreting *some* in context, then the *not all* implicature should be more likely to arise when the QUD makes it relevant (upper-bound) than when the QUD does not (lower-bound). Second, we examine which properties of the individual modulate one's ability to interpret *some* based on the QUD, by including a battery of measures targeting abilities that are potentially important for computing a context-dependent interpretation of *some*, including measures targeting both cognitive resources and personality-based pragmatic abilities. We test the prediction that not only cognitive resources but also personality-based pragmatic abilities may affect individuals' ability to utilize QUD in interpreting *some*, given that the current study examines scalar implicature derivation in a discourse context involving communication between two interlocutors. As we discuss below, we also include assessments of individuals' language abilities to address to what extent language abilities may account for variation in scalar implicature derivation in context.

To address these questions, we tested participants using a story-sentence matching task in which one character utters a sentence containing *some*, such as "I folded some of the sweaters," following a question from another character that either establishes the relevance of the *not all* implicature (e.g., "Did you fold all the sweaters?") or that the implicature is irrelevant (e.g., "Did you fold any sweaters?"). How many objects were acted upon is illustrated in a visual display (showing, e.g., either 0, 2, or 4 folded sweaters). If an individual is sensitive to the context (whether the implicature has been established as relevant or as irrelevant), then they should be less likely to accept the target sentence (e.g., "I folded some of the sweaters," when 4 of 4 sweaters had been folded) when the implicature is relevant than when it is irrelevant.

Participants also completed a battery of individual differences assessments targeting cognitive resources and personality-based pragmatic abilities, allowing us to directly test proposals that individual differences in scalar implicature derivation have their origin in either cognitive resources or personality-based pragmatic abilities, or instead may make recourse to both types of ability. In addition to examining these two commonly tested potential sources of variation (cognitive resources and personality traits), we also examine individual differences in language skills as a third possible source of variability.



Relationships between language skills and pragmatic abilities have been explored in studies in the literature on children (e.g., Katsos et al., 2011) and on adults with Autism and Asperger Syndrome (e.g., Pijnacker et al., 2009). Yet in studies examining unimpaired adults, language skills have rarely been tested in the scalar implicature literature as a possible source of individual variability. Therefore, the current study included language skills as a third measure of individual differences, in order to examine whether language skills may be among the sources of variability contributing to individual differences in scalar implicature derivation among unimpaired adult native speakers.

## Participants

Sixty-four native English speakers (19 male, mean age = 21.9, age range = 18–53) who were naïve about the purpose of the study were recruited from the University of Kansas community. All participants provided informed consent before participating and received a cash payment or course credit upon completing their visit.

## Main Task: Story-Sentence Matching Task

The current study utilized a story-sentence matching task to probe the interpretation of *some* in a context in which the interpretation of the target *some* sentences depends on the QUD. We constructed 32 target trials, each of which consisted of a short story presented in text and pictures on a series of slides, about two characters carrying out an action involving a set of four objects (e.g., cutting four pieces of steak). The first slide introduced the characters and the objects. The second slide always showed that 4/4 objects were changed (e.g., all four steaks were shown as cut); following Hunt et al. (2013), all acted-upon objects were also highlighted by a red square to remove any ambiguity regarding the number of objects acted-upon, and the text showed that the objects ended up as shown in the picture (e.g., “In the end, the steaks look like this.”). On the third slide, a brief conversation between the characters was presented, in which one character asked, “Have you cut all the steaks?” or “Have you cut any steaks?” and the other responded “I cut some steaks.” Then a rating question appeared, asking the participant to rate how well the response matched with what happened in the story on a 7-point Likert scale. After they responded by clicking on their chosen rating on the scale, a button showing “click here for the next story” appeared at the bottom center of the screen, which triggered the next story once clicked on (see **Figure 1** for the depiction of an example trial). The slides were automatically presented at a comfortable reading speed (first slide 8000 ms, second slide 3000 ms, third slide 4000 ms), and the rating question and the scale were untimed. Participants were asked to read the stories carefully and answer the question at the end of each story.

In the target trials, we established an upper-bound or lower-bound QUD by including “all” or “any” in the first character’s question sentence, following Politzer-Ahles and Fiorentino (2013). When the question includes “all,” as in “Have you cut

all the steaks?”, the QUD is established as upper-bound, as the character is asking about whether each and every steak in the set has been cut. Therefore, the response sentence with *some* is expected to be interpreted as *at least one but not all*, such that the added *not all* implicature addresses the upper-bound QUD. The response should thus be underinformative since all four objects were acted upon as shown by the picture. In contrast, when the question includes “any,” as in “Have you cut any steaks?”, the QUD is established as lower-bound, as the character is asking about whether at least one steak has been cut. In this scenario, *some* is expected to be interpreted as *at least one* without the *not all* implicature. Thus, the response sentence with *some* should be felicitous given that at least one object has been acted upon. Therefore, if comprehenders are sensitive to the QUD in interpreting *some*, they should rate the target *some* sentences lower when the question sentence includes “all” (the upper-bound QUD; *all* condition, henceforth), compared to when the question sentence includes “any” (the lower-bound QUD; *any* condition, henceforth).

In order to mask the purpose of the study and to elicit participants’ interpretation of *some* in unambiguously true/felicitous or false sentences, we also included 32 filler trials that have the same story structure, presentation format, and QUD manipulation as the target trials. However, the filler trials differed from the target trials in two ways: (1) the character’s answer sentence included *only some* instead of *some* (e.g., “I cut only some steaks”), which should be unambiguously interpreted as *some but not all*; and (2) the number of acted-upon objects in the picture was 0/4, 2/4 or 4/4, while there were always 4/4 acted-upon objects in the target trials. These configurations thus made the filler trials patently true or patently false depending on the number of acted-upon objects, regardless of the QUD being upper-bound or lower-bound; the fillers were true when 2/4 objects were acted upon, and false when either 0/4 or 4/4 objects were acted upon.

From the two target conditions (*all* condition and *any* condition), we generated two lists of targets by alternating the QUD for each story, such that each participant would see all 32 target stories, but no participant would encounter the same story in both conditions. Each list included a total of 64 unique stories presented in random order (32 targets and 32 fillers), with half of the targets (16) in the *all* condition and the other half (16) in the *any* condition. Participants were randomly assigned to complete only one list. The truth value of the response sentences, number of acted-upon objects, and the QUD were balanced across all the trials (see **Supplementary Table S1**, for a summary of the properties of the target and filler stimuli).

## Measures of Individual Differences

The current study examined three potential sources of individual variation as discussed above (cognitive resources, personality-based socio-pragmatic abilities, and language skills), testing participants on a battery of measures targeting these three domains. In the following section, we describe how these sources of variations were assessed.

John and his coworker were working in a restaurant to develop new steak recipes. Here are the steaks they were going to use.

In the end, the steaks looked like this.

John's coworker asked him, "Have you cut all the/any steaks?"  
John quickly replied, "I cut some steaks."

John's coworker asked him, "Have you cut all the/any steaks?"  
John quickly replied, "I cut some steaks."  
How well did John's response match with what happened in the story?  
1 2 3 4 5 6 7

**FIGURE 1 |** Sample display of a target trial in the main story-sentence matching task. Each trial is depicted on four consecutive slides, numbered 1–4 here.

## Measures of Cognitive Resources

### Working memory capacity: count span task

Working memory capacity has been widely suggested as a factor that may account for individual differences in deriving scalar implicatures (De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013). When interpreting *some* under a specific QUD in the discourse, as is required in the current study, sufficient working memory may also be required to encode and maintain the QUD throughout the task, and to accurately represent what happens in an event which spans across multiple sentences and visual representations of the story scene. Thus, interpreting *some* in context and making distinctions between QUDs may require sufficient working memory capacity.

In the current study, we assessed individual working memory capacity via Count Span (Conway et al., 2005), which measures non-verbal working memory in a counting and recalling task. In the Count Span task, the participant was asked to count out loud the number of appearances of a specific shape when they viewed an array of shapes on the computer screen. The experimenter recorded the numbers that the participant had counted on each screen, after which a screen with a new array of shapes appeared. After between 2 and 6 screens, the participant would be prompted to recall the numbers they counted on the previous set of screens in their order of occurrence by entering the digits on the keyboard. Following Conway et al. (2005), we calculated an accuracy score for this task by comparing their total number

of correctly recalled digits versus the total numbers of counted digits.

### Domain-general context maintenance: dot pattern expectancy task

Although the previous literature has tended to focus on working memory as a factor that may modulate an individual's ability to process scalar implicatures, two additional processing-related factors which may be particularly important in processing *some* in context are domain-general context-maintenance ability and attentional control. Domain-general context maintenance involves holding prior information in working memory and utilizing it to subsequently determine task-relevant responses (Cohen et al., 1999). This ability has been examined as a potentially important source of variability in studies on the influence of context on ambiguity resolution in language tasks (e.g., Cohen et al., 1999, lexical ambiguity; Boudewyn et al., 2015, referential ambiguity). The ability is arguably also relevant when processing *some* under different QUDs, which involves maintaining the prior cues to the QUD and using them to determine whether an upper-bound or lower-bound meaning is supported under the current context. Therefore, distinguishing between QUDs and utilizing them in the interpretation of *some* sentences may require sufficient context maintenance ability.

The current study assessed domain-general context maintenance ability via the Dot Pattern Expectancy Task (DPX), a version of the Continuous Performance Test in which



participants respond to visually presented cue-target pairs, and must make a designated response only when both the cue and the target come in a specified form (Rosvold et al., 1956; Cohen et al., 1999). In the DPX, each trial includes a pair of dot patterns, with a “cue” pattern in white and a “probe” pattern in blue. The trials were comprised of four types: AX, AY, BX, and BY. AX trials are the target trials, while all other types are non-targets in that they either include a non-target probe pattern (AY), or include a non-target cue pattern followed by a target or non-target probe pattern (BX and BY, respectively). Therefore, the identity of the cue determines whether the following probe constitutes a target trial, as a non-target cue directly indicates a non-target trial regardless of the identity of the probe. For an individual with a high level of context maintenance ability, they should be able to not only correctly recognize the target pattern for AX trials, but also correctly detect the non-target cue (context) and refrain from making a target response for BX trials despite the target probe. In contrast, an individual with lower context maintenance ability should make more errors in BX trials by ignoring the context and incorrectly making a target response only based on the probe.

When completing this task, participants were instructed to press either a “no” or a “yes” button on a keyboard upon seeing each dot pattern; they should only press “yes” after seeing the target blue pattern following the target white pattern, and press “no” for all the other patterns. After each key press, they heard either a “bing” or a “buzz” sound indicating whether they had pressed the correct key. Four practice sessions were administered at the beginning of the task along with instructions. Participants practiced until they had reached 80% accuracy and had responded correctly to at least 1 BX trial, before they began the main session. The task was administered using Paradigm (Tagliaferri, 2005), with a cue duration of 1000 ms, an interstimulus interval of 2000 ms, target presentation for 500 ms, a response window of 1500 ms, and an intertrial interval of 1200 ms. There were 144 trials in total (104 AX, 16 AY, 16 BX, and 8 BY), with each trial type evenly distributed across four blocks.

A *d*-prime score was computed for the DPX task, following Cohen et al. (1999). The *d*-prime indexes the sensitivity to context in this task, which accounts for accuracy including both hit rate on target trials (AX trials) and false-alarms (BX trials, which included non-target cue pattern followed by a target). For each participant, their *d*-prime was calculated by  $z(\text{the accuracy on AX trials}) - z(\text{the error rate on BX trials})$ . Following standard procedure, hit rates of 1 were corrected to  $(1 - 1/160)$ , and false alarm rates of 0 were corrected to  $1/16$ . Higher *d*-prime scores represent greater sensitivity to domain-general context cues.

#### Attentional control: number Stroop task

Attentional control involves the ability to attend to crucial information in the presence of distractions and to inhibit the information that is irrelevant to the current task (e.g., Kane and Engle, 2002). Higher levels of attentional control ability have been found to facilitate performance in cognitive and language tasks involving selective attention and suppression of irrelevant information (e.g., Hutchison, 2007; Bialystok and Martin, 2004; Boudewyn et al., 2012; Abutalebi and Green, 2016).

In the literature on scalar implicature, a handful of studies have included measures of attentional control or inhibition ability as a submeasure of cognitive resources involved in interpreting *some* in isolation, although they have not commonly found it to have a significant effect (e.g., Heyman and Schaeken, 2015; Antoniou et al., 2016). We included an attentional control measure in the current study since sufficient attentional control may be important for generating QUD-dependent interpretations for *some* in context, where the comprehender needs to suppress one interpretation and pursue the other one that is relevant under the current QUD, while processing a relatively large amount of linguistic and visual material compared to a typical study on *some* in isolation.

We assessed attentional control via the number Stroop task, following the procedure outlined in Bush et al. (2006). In each trial, the participant was asked to count the number of words presented on the computer screen, which could be any number between 1 and 4, and to press the corresponding number key as quickly and as accurately as possible on a button pad. Trials were presented in 8 blocks of 20 trials; 4 blocks included congruent trials and 4 included incongruent trials, the order of which was counterbalanced across blocks. In congruent trials, the words were common animal words (e.g., “dog dog”; correct response is 2), while in incongruent trials the words were number words that do not match with the quantity of words on the screen (e.g., “one one one”; correct response is 3). Thus, for incongruent trials, participants must maintain their attention toward the quantity of words on the screen while avoiding distraction from the meanings of the words, in order to achieve the correct answer. We computed a Stroop interference score for each participant by subtracting the percent accuracy for congruent trials from that of incongruent trials, such that higher values reflect better attentional control ability<sup>2</sup>.

#### Personality-Based Socio-Pragmatic Abilities

A number of studies have suggested that individual variation in the derivation of scalar implicature has its origin in personality traits (Nieuwland et al., 2010; Katsos and Bishop, 2011; Feeney and Bonnefon, 2013). As Antoniou et al. (2016) speculate, personality-based factors such as sensitivity to the pragmatic use of language in everyday life may be particularly important when processing language in more conversational settings, as opposed to when interpreting *some* sentences outside of any discourse. Thus, an individual's ability to utilize QUD in order to arrive at a pragmatically felicitous interpretation of *some* in context in the current study may depend at least in part on their sensitivity to the pragmatic use of language in everyday life (which we refer to below as their socio-pragmatic abilities).

In the current study, socio-pragmatic abilities were assessed via Autism-Spectrum Quotient questionnaire (AQ, Baron-Cohen et al., 2001), which assesses individuals' general social and communicative skills based on the level of autistic-like traits that their responses demonstrate. The questionnaire includes

<sup>2</sup>We also computed Stroop interference scores based on reaction time (by subtracting the reaction time of incongruent trials from congruent trials); since this score was correlated with the accuracy-based interference score ( $r = 0.316$ ), only the accuracy-based interference score is used in the analysis.

50 statements about self-perceived characteristics, with 10 statements from each of the five subscales examining traits that vary across the autism spectrum (social skills, attention switch, communication, attention to detail, and imagination). Participants were asked to read each statement and answer to what degree each statement truly reflects themselves, by choosing from four levels: definitely agree, slightly agree, slightly disagree, or definitely disagree. Following Baron-Cohen et al. (2001), the answers were scored by assigning 1 to “definitely agree” or “slightly agree,” and 0 to “definitely disagree” or “slightly disagree” for the statements that indicate strong autistic traits, and assigning 0 to “definitely agree” or “slightly agree,” and 1 to “definitely disagree” or “slightly disagree” for the statements that do not indicate strong autistic traits. The total score and the scores for each of the five subscales were calculated by adding up the scores for the corresponding items. Thus, higher AQ scores are taken to reflect weaker socio-pragmatic abilities.

### Language Skills

A third source of variation that may modulate an individual's derivation of scalar implicatures is language skills. Although native speakers have been assumed to share a native grammar and thus have similar language abilities, recent studies have revealed that adult monolingual native speakers do show variability in native language processing (e.g., Kemper and Sumner, 2001; Pakulak and Neville, 2010; Borovsky et al., 2012; Dąbrowska, 2012; Van Dyke et al., 2014). Language skills have been shown to play an independent role in accounting for variability in native language processing in a range of domains, even when examined together with assessments of non-linguistic cognitive resources such as working memory (Van Dyke et al., 2014), which recommends the inclusion of language skills in studies examining individual variation in studies on language comprehension.

The measures of language skills in the current study targeted vocabulary, assessed by the Peabody Picture Vocabulary Test 4th edition (PPVT-4, Dunn and Dunn, 2007), and exposure to print materials, assessed by Author and Magazine Recognition task (Acheson et al., 2008). Vocabulary skills have been shown to predict comprehension success at the word level and the sentence level (e.g., Braze et al., 2007; Perfetti, 2007; Boudewyn, 2015). It has been argued that having strong, detailed lexical representations leads to efficient and successful comprehension in a number of ways, such as by reducing interference between lexical representations during comprehension (see e.g., Van Dyke et al., 2014 for recent evidence). It has also been suggested that those with stronger lexical representations may be better able to process the meanings of words and integrate words in context in order to derive meanings and generate inferences during passage comprehension (see e.g., Hamilton et al., 2013). When interpreting *some* in context, those with better vocabulary skills may be better able to recognize the two possible readings of *some* and to select an optimal reading according to the current context.

Print exposure has been shown to account for individual differences in performance across several linguistic domains, ranging from orthographic and phonological processing through sentence and discourse comprehension (Acheson et al., 2008;

Arnold et al., 2018). It has been argued that increased print exposure may lead to increased sensitivity to linguistic cues that facilitate comprehension, including as regards the resolution of ambiguity in discourse (e.g., Arnold et al., 2018). Arnold et al. (2018) found that individuals with greater print exposure as assessed by the Author and Magazine Recognition Task were better at resolving ambiguous pronoun reference using discourse cues. Increased sensitivity to discourse cues and patterns as a result of greater print exposure may also impact the interpretation of *some* in context; individuals with greater print exposure may be better able to recognize and utilize QUD in order to arrive at a coherent interpretation of the ambiguous term *some* in the discourse.

### Vocabulary: Peabody Picture Vocabulary Test

The PPVT-4 is a standardized test of receptive vocabulary that spans several subject fields. In each trial, the participant heard an English word pronounced by an experimenter and was asked to select from among four pictures the one that corresponds to the word. The trials were numbered and organized into sets of 12, with increasing level of difficulty. A starting set was initially picked based on the participant's chronological age, following the PPVT-4 manual. If the participant made 2 or more errors in this set, then the experimenter would go back to the previous set and test that as the new starting set, until the participant made 1 or 0 errors in a set. As they responded to the trials, the participant's answers were manually recorded and the total number of errors within each set was tracked by the experimenter on the PPVT-4 testing booklet. The task came to an end either when the participant made 8 or more errors within a set, or when they have completed the last set of the entire test. For each participant, a raw score was first computed by subtracting the total number of errors from the number of completed items; this raw score was then standardized based on the participant's chronological age, using the standardization chart provided in the PPVT-4 manual.

### Exposure to print materials: Author and Magazine Recognition Task

We measured exposure to print materials via the Author and Magazine Recognition Task (ART and MRT, Acheson et al., 2008). The ART consists of a list of 130 author names and the MRT a list of 130 magazine titles. Half of the names in the ART are real authors' names and the other half are foils that look like author names; similarly, half the titles in the MRT are real magazine titles and the other half are foils that appear to be magazine titles. The real items in both tasks are from popular reading materials covering various topics and genres. Participants were asked to select real authors' names in the ART and real magazine titles in the MRT without guessing, by entering an “X” beside the items in an Excel spreadsheet. Following Acheson et al. (2008), answers were scored by assigning 1 point for a correctly identified real item, and −1 point for a foil item incorrectly identified as real, generating a total score for the ART and for the MRT for each participant. The ART and MRT scores were then averaged into one single score, with higher values reflecting more extensive print exposure.

## Composite Scores for Measures of Individual Differences

We computed composite scores for cognitive resources and for language skills, based on the conceptual relatedness of the specific measures and the correlations among the scores within each domain (see **Supplementary Table S2**, for summary statistics for each of the individual difference measures, and **Supplementary Table S3** for pairwise correlations between the measures). That is, the composite score for Cognitive Resources was calculated by summing the Count Span score, Dot-pattern Expectancy *d*-prime score, and Stroop interference score. The Language Skills composite was calculated by summing the standardized PPVT-4 score and the Author and Magazine Recognition task score. Total AQ score was used to quantify individuals' Socio-pragmatic Abilities. We used Total AQ rather than the AQ-Communication Subscale (used, e.g., in Nieuwland et al., 2010), since all the subscale scores strongly correlated with the total AQ score. However, we note that the pattern of results reported below does not change if the AQ-Comm score rather than Total AQ is used in the analysis. This generates three individual difference scores that were included as predictors in the model-fitting: Cognitive Resources, Socio-pragmatic Abilities (Total AQ score), and Language Skills<sup>3</sup>. Before model fitting, the three scores were standardized using *z*-transformation, so that they are on similar numerical scales as required by mixed effect models (Jaeger, 2008).

## Procedure

Participants completed all the experimental tasks in the Neurolinguistics and Language Processing Laboratory at the University of Kansas. Tasks were administered in the following order, with break times in between each task: Peabody Picture Vocabulary Test 4th edition, Autism-Spectrum Quotient questionnaire, Author and Magazine Recognition Task, the Story-sentence Matching Task, Count Span task, Dot Pattern Expectancy task, and Number Stroop task. The entire session took about 1 h 30 min to complete in one visit to the laboratory.

## Summary of Predictions

Our first research question concerns the extent to which participants are sensitive to QUD in interpreting *some*. If participants are able to utilize QUD in interpreting *some*, then a main effect of QUD is expected to emerge, such that the ratings for the target sentences should be lower in the *all* condition than those in the *any* condition. Our second research question concerns the role of individual differences in the context-sensitive interpretation of *some*. If the interpretation of *some* in context is impacted by individual differences in both Cognitive Resources and Socio-pragmatic Abilities, then we expect interactions to emerge between QUD and the Cognitive Resources composite measure, as well as between QUD and the Socio-pragmatic Abilities measure.

<sup>3</sup>We note that the composite score for Cognitive Resources, the composite score for Language Skills, and the Socio-Pragmatic Abilities score are not correlated (see **Supplementary Table S4** for pairwise correlations between the composite scores).

If individual differences in Language Skills also impact the interpretation of *some* in context, an interaction between QUD and the Language Skills composite measure is expected to emerge.

## RESULTS

### Data Pre-processing and Modeling

The ratings in the main experiment were statistically analyzed by fitting a cumulative link mixed model (the *clmm* function from the package *ordinal*) with a *probit* link function (Christensen, 2010) in the R programming environment. We chose to use the cumulative link mixed model as it can analyze categorical outcomes while incorporating subject-level and item-level random effect structures, which is an advantage over traditional regression models (Jaeger, 2008; Cummings, 2012). The *probit* link function allows us to analyze rating responses by underlyingly modeling the log-transformed odds ratio of increasing the rating by 1 on the Likert scale (e.g., rating an utterance as 5 over 4, or as 6 over 5, etc., on the 7-point scale).

Model fitting began by including the following predictors of interest: the fixed factors QUD (*all*, *any*), and interactions between QUD and each of the individual difference scores: QUD  $\times$  Cognitive Resources, QUD  $\times$  Socio-pragmatic Abilities, and QUD  $\times$  Language Skills. Participant and Item were included as random intercepts. The initial model was then optimized by backward-fitting via log-likelihood ratio tests: if removing a predictor from the initial model did not reduce the model fit, then a simpler model without that predictor was built; on the contrary, if removing a predictor led to worse fit, then the predictor was retained. Following this procedure, the final model included the fixed effect of QUD and two interaction terms: QUD  $\times$  Cognitive Resources, QUD  $\times$  Socio-pragmatic Abilities, as well as Participant and Item as random intercepts.

### Effects of QUD and Individual Difference Measures

The two main research questions in the current study concern whether QUD modulates the rating of *some* sentences, which should be reflected by lower ratings in the *all* condition than in the *any* condition, and to what extent sensitivity to QUD is subject to individual differences in cognitive resources, personality-based pragmatic abilities, and language abilities, which would be reflected in a significant interaction between the QUD and a given measure of individual differences. Although all the variables of interest for both research questions were incorporated into one model, we report the results separately for each research question below.

A few things should be kept in mind when interpreting the effects in the final model, which is summarized in **Table 1**. Because of the *probit* link function, the coefficients represent the effect of predictors on the odds ratio of increasing the ratings, not directly on the ratings *per se*. Regarding the QUD effect, as the *all* condition was dummy-coded as the baseline condition, the effect of QUD appeared as the effect of the *any* condition,



**TABLE 1** | Summary of the final model analyzing  $N = 64$  participants' ratings as a function of QUD and individual difference measures.

	$\beta$	$SE$	$z$	$p$
QUD	0.5963	0.0487	12.25	<0.001
QUD $\times$ Cognitive Resources	0.13935	0.04896	2.846	<0.01
QUD $\times$ Socio-pragmatic Abilities	-0.19605	0.04774	-4.107	<0.001

QUD reflects the difference between the ratings in the any condition compared to the all condition, which is coded as the baseline.

as compared to the *all* condition. Finally, since the individual difference scores have been standardized to fit in the same model, the effects involving these scores should be interpreted based on standardized units.

To address the role of QUD, we examined the main effect of QUD in the model. The main effect of QUD is indeed significant, indicating that overall participants were more likely to provide higher ratings for target utterances in the *any* condition compared to the *all* condition ( $\beta = 0.5963$ ,  $SE = 0.0487$ ,  $z = 12.25$ ,  $p < 0.001$ ). In short, this finding suggests that the derivation of the scalar implicature for *some* was affected by the QUD as established in the discourse context (see **Supplementary Figure S1** for a visualization of the differences in mean raw ratings between the *all* condition and the *any* condition). To confirm that this effect of QUD does not just reflect an overall preference for the *any* versus the *all* stimuli regardless of whether the stimuli contained *some* (the targets, where QUD matters) or *only some* (the fillers, where QUD does not matter), we examined responses to the fillers, which also had *all* versus *any* QUDs but had a target sentence with *only some*, where ratings should not be sensitive to QUD. As expected, QUD did not modulate ratings in the fillers ( $\beta = 0.1334$ ,  $SE = 0.8479$ ,  $z = 0.157$ ,  $p = 0.875$ ).

To address whether sensitivity to QUD in interpreting *some* is impacted by individual differences in cognitive resources, personality-based pragmatic abilities, and language skills, we examined interactions between the QUD effect and individual difference scores in each of these three domains. Among the individual difference measures, QUD significantly interacted with both Cognitive Resources ( $\beta = 0.1394$ ,  $SE = 0.0489$ ,  $z = 2.846$ ,  $p < 0.05$ ) and Socio-pragmatic Abilities ( $\beta = -0.1961$ ,  $SE = 0.0477$ ,  $z = -4.11$ ,  $p < 0.001$ ), indicating that the QUD effect is modulated by individual differences in both domains. Sensitivity to QUD increased with greater cognitive resources, and with better socio-pragmatic abilities (note that since better socio-pragmatic abilities are indexed by lower AQ scores, the coefficient for the interaction term QUD  $\times$  Socio-pragmatic Abilities is negative). Regarding the role of Language Skills, the fact that QUD  $\times$  Language Skills was excluded during the model fitting indicated that Language Skills was not a significant predictor of individual sensitivity to QUD in our study.

## DISCUSSION

The current study investigated the interpretation of the scalar quantifier *some* in contexts which establish the *not all* scalar implicature as relevant (upper-bound contexts) or irrelevant

(lower-bound contexts). We examined to what extent native speakers are sensitive to context in interpreting *some* and which individual differences may best account for variability across individuals in the ability to utilize contextual information to interpret *some*. Overall, we found that native speakers as a group do distinguish the meaning of *some* based on the QUD, such that the *not all* implicature is more likely to arise under an upper-bound QUD than a lower-bound QUD. While the interpretation of *some* is typically described as context-sensitive in linguistic analyses, the findings of the current study converge with those of a still relatively limited number of experimental studies in demonstrating sensitivity to QUD in the interpretation of *some* during language comprehension (Politzer-Ahles and Fiorentino, 2013; Degen and Goodman, 2014; Dupuy et al., 2016; Politzer-Ahles and Husband, 2018). However, the findings of the current study also revealed individual differences in the extent to which QUD affects the interpretation of *some*, which depended both on an individual's cognitive resources and on their personality-based pragmatic abilities. While previous studies on the processing of *some* in isolation have alternatively argued that the derivation of scalar implicatures depends on cognitive resources or on personality traits, our findings are unique in demonstrating that the derivation of scalar implicatures, when tested in a discourse context, indeed makes recourse to both types of abilities.

## The Role of Cognitive Resources in Context Sensitivity

Our finding that individuals with greater cognitive resources show greater sensitivity to the context in interpreting *some*, as evidenced by the significant interaction of QUD  $\times$  Cognitive Resources, converges with the those of a number of studies arguing that sufficient cognitive resources are required for an individual to derive scalar implicatures (e.g., De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013). In our study, there are a number of possible ways that greater cognitive resources may have led to increased sensitivity to QUD. The interpretation of *some* with respect to a given QUD requires successfully attending to the contextual cues that establish QUD, as well as the encoding and maintenance of that information throughout the discourse. Upon encountering *some*, previously encountered information needs to be recalled and utilized to compute a context-sensitive interpretation for *some*, and the selected meaning for *some* must be maintained while possibly inhibiting the other meaning. All of these processes would arguably make recourse to the kinds of cognitive resources assessed in the current study (working memory, attentional control, and ability to maintain contextual information during processing), which regard an individual's ability to encode and maintain information and direct attention while also processing bottom-up input. Individuals with greater cognitive resources may also be better at consistently attending to and utilizing contextual information in order to interpret the target utterances over the course of an experiment that involved a relatively large number of target and filler trials, which itself may incur some amount of processing burden.

## The Role of Personality-Based Factors in Context Sensitivity

The current study also revealed that personality-based factors such as socio-pragmatic abilities (as measured by the AQ) also modulated sensitivity to QUD; those with greater socio-pragmatic abilities made a larger distinction between QUDs in their ratings, thus lending support to accounts arguing that personality traits modulate an individual's likelihood of deriving scalar implicatures (e.g., Nieuwland et al., 2010; Feeney and Bonnefon, 2013). In the current study, those individuals with greater awareness of the pragmatic aspects of communication in daily life, as assessed by AQ, were more sensitive to whether the *not all* implicature for *some* had been established as relevant within the conversational context.

Our findings converge with a number of previous studies demonstrating relationships between scalar implicature derivation and cognitive resources on one hand, and with a number of studies demonstrating relationships between scalar implicature derivation and personality traits on the other hand. Interestingly, in one previous study on the derivation of scalar implicature for *some* in isolation which did assess both potential sources, personality-based factors were not found to be a significant predictor of realizing the *not all* implicature (Antoniou et al., 2016). Recall that Antoniou et al. (2016) examined the interpretation of *some* without discourse context, where the acceptability of *some* sentences should only depend on a visual depiction that either made them felicitous or infelicitous. As Antoniou et al. (2016) acknowledged, socio-pragmatic abilities may not robustly modulate the interpretation of *some* in this type of task as it does not establish any kind of conversational exchange or discourse involving more than one interlocutor, and thus may not prompt the participant to make use of their understanding of the pragmatics of conversation in deciding how to interpret *some*.

The fact that Antoniou et al. (2016) did not observe an effect of personality-based pragmatic abilities while the current study did find such an effect is consistent with the claim that these abilities may be particularly important for taking contextual information into account when interpreting *some*, in particular that from communicative discourse contexts. This is exactly the kind of context provided in our story-sentence matching task, where a conversation between two interlocutors established the QUD determining the relevance of the implicature. Our findings thus strongly argue that individuals rely on both types of ability in the interpretation of *some* under conversational discourse contexts.

## The Role of Language Skills

Among our individual difference measures, language skills (measured via a composite of vocabulary size and exposure to print materials) did not prove to modulate individual sensitivity to QUD in interpreting *some*. It is worth noting that in Antoniou et al. (2016), their measure of verbal IQ (a sentence repetition task) also did not significantly predict individuals' derivation of the *not all* implicature. Although neither the current study nor Antoniou et al. (2016) found evidence of a relationship between

language skills and implicature derivation for *some*, a question to be examined in future research is whether language skills may become increasingly important when the relevance of the implicature is established in more linguistically rich contexts, perhaps with less visual information, which may place greater demands on the comprehender to construct and process the discourse through careful comprehension of a larger amount of text or speech input. Future studies could also examine whether different measures of language abilities might better account for individual variability in the derivation of scalar implicatures in context, such as passage comprehension measures which more directly target the processing of discourse.

More broadly, it may also be interesting for future research to examine to what extent language skills as well as cognitive resources and personality-based factors may influence the derivation of implicatures for scalar terms other than the quantifier *some*. Moreover, future research examining individual differences in sensitivity to context in the interpretation of scalar terms using online measures such as self-paced reading (e.g., Breheny et al., 2006; Politzer-Ahles and Fiorentino, 2013), eye-tracking (e.g., Politzer-Ahles and Husband, 2018), or neurolinguistic methods (e.g., Hartshorne et al., 2015; Politzer-Ahles and Gwilliams, 2015), may also provide new insights regarding how individual differences in the domains examined in the current study impact the derivation of scalar implicatures during the dynamics of language processing.

## CONCLUSION

This study demonstrates that comprehenders vary in their ability to utilize context cues in interpreting *some* in context. Moreover, this variability is associated with individual differences in both cognitive resources and personality-based pragmatic abilities. While previous studies on the processing of *some* without manipulating context have argued for one or the other of these sources in order to account for individual variability in deriving scalar implicatures, the current study establishes for the first time that computing pragmatically enriched meanings based on the broader discourse indeed draws upon both kinds of skills.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Human Research Protection Program (HRPP) at the University of Kansas. The protocol was approved by the Institutional Research Board at the University of Kansas (Study #00004409). All subjects gave written informed consent in accordance with the Declaration of Helsinki.

## DATA AVAILABILITY

The de-identified raw data supporting the conclusion of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher upon written request.



## AUTHOR CONTRIBUTIONS

XY, UM, and RF contributed the conception and design of the study. XY administered the experiment, organized the data, and performed the statistical analyses. XY, UM, and RF contributed to manuscript writing and editing.

## ACKNOWLEDGMENTS

We are grateful to Dr. Alison Gabriele and Dr. Steven Politzer-Ahles for their helpful discussion and feedback regarding this project, and to our research assistant Vann Hassell for help with data collection. We are also thankful to Jonah

Bates, Dr. Kate Coughlin, John-Patrick Doherty, Dr. Andrew McKenzie, Dr. Annie Tremblay, and Dr. Jie Zhang for their help with recruiting participants, and the members of Research in Acquisition and Processing Seminar at the University of Kansas and the audience at the 30th *CUNY Conference on Human Sentence Processing* for their constructive feedback.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01720/full#supplementary-material>

## REFERENCES

- Abutalebi, J., and Green, D. (2016). Neuroimaging of language control in bilinguals: neural adaptation and reserve. *Bilingualism* 19, 689–698. doi: 10.1017/S1366728916000225
- Acheson, D. J., Wells, J. B., and MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behav. Res. Methods* 40, 278–289. doi: 10.3758/BRM.40.1.278
- Antoniou, K., Cummins, C., and Katsos, N. (2016). Why only some adults reject under-informative utterances. *J. Pragmatics* 99, 78–95. doi: 10.1016/j.pragma.2016.05.001
- Arnold, J. E., Strangmann, I. M., Hwang, H., Zerkle, S., and Nappa, R. (2018). Linguistic experience affects pronoun interpretation. *J. Mem. Lang.* 102, 41–54. doi: 10.1016/j.jml.2018.05.002
- Barbet, C., and Thierry, G. (2016). Some alternatives? Event-related potential investigation of literal and pragmatic interpretations of some presented in isolation. *Front. Psychol.* 7:1479. doi: 10.3389/fpsyg.2016.01479
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J. Autism. Dev. Disord.* 31, 5–17. doi: 10.1023/A:1005653411471
- Bialystok, E., and Martin, M. M. (2004). Attention and inhibition in bilingual children: evidence from the dimensional change card sort task. *Dev. Sci.* 7, 325–339. doi: 10.1111/j.1467-7687.2004.00351.x
- Borovsky, A., Elman, J. L., and Fernald, A. (2012). Knowing a lot for one's age: vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *J. Exp. Child Psychol.* 112, 417–436. doi: 10.1016/j.jecp.2012.01.005
- Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006
- Boudewyn, M. A. (2015). Individual differences in language processing: electrophysiological approaches. *Lang. Linguistics Comp.* 9, 406–419. doi: 10.1111/lnc3.12167
- Boudewyn, M. A., Long, D. L., and Swaab, T. Y. (2012). Cognitive control influences the use of meaning relations during spoken sentence comprehension. *Neuropsychologia* 50, 2659–2668. doi: 10.1016/j.neuropsychologia.2012.07.019
- Boudewyn, M. A., Long, D. L., Traxler, M. J., Lesh, T. A., Dave, S., Mangun, G. R., et al. (2015). Sensitivity to referential ambiguity in discourse: the role of attention, working memory, and verbal ability. *J. Cogn. Neurosci.* 27, 2309–2323. doi: 10.1162/jocn\_a\_00837
- Braze, D., Tabor, W., Shankweiler, D. P., and Mencl, W. E. (2007). Speaking up for vocabulary: reading skill differences in young adults. *J. Learn. Disabil.* 40, 226–243. doi: 10.1177/00222194070400030401
- Breheny, R., Katsos, N., and Williams, J. (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. doi: 10.1016/j.cognition.2005.07.003
- Bush, G., Whalen, P. J., Shin, L. M., and Rauch, S. L. (2006). The counting Stroop: a cognitive interference task. *Nat. Protoc.* 1, 230–233. doi: 10.1038/nprot.2006.35
- Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Struct. Beyond* 3, 39–103.
- Chierchia, G., Fox, D., and Spector, B. (2012). “Scalar implicature as a grammatical phenomenon,” in *Semantics: An international handbook of natural language meaning*, Vol. 3, eds K. von Stechow, C. Maienborn, and P. Portner (Berlin: Mouton de Gruyter), 2297–2331.
- Christensen, R. H. B. (2010). *Ordinal - regression models for ordinal data*. R package version, 22.
- Cohen, J. D., Barch, D. M., Carter, C., and Servan-Schreiber, D. (1999). Context-processing deficits in schizophrenia: converging evidence from three theoretically motivated cognitive tasks. *J. Abnorm. Psychol.* 108:120. doi: 10.1037/0021-843X.108.1.120
- Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., and Engle, R. W. (2005). Working memory span tasks: a methodological review and user's guide. *Psychon. Bull. Rev.* 12, 769–786. doi: 10.3758/BF03196772
- Cummings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Lang. Res.* 28, 369–382. doi: 10.1177/0267658312443651
- Dąbrowska, E. (2012). Different speakers, different grammars: individual differences in native language attainment. *Linguistic Approaches Biling.* 2, 219–253. doi: 10.1075/lab.2.3.01dab
- De Neys, W., and Schaeken, W. (2007). When people are more logical under cognitive load: dual task impact on scalar implicature. *Exp. Psychol.* 54, 128–133. doi: 10.1027/1618-3169.54.2.128
- Degen, J., and Goodman, N. D. (2014). “Lost your marbles? The puzzle of dependent measures in experimental pragmatics,” in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, eds P. Bello, M. Guarini, M. McShane, and B. Scassellati (Austin, TX: Cognitive Science Society).
- Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171
- Degen, J., Tessler, M. H., and Goodman, N. D. (2015). “Wonky worlds: Listeners revise world knowledge when utterances are odd,” in *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, (Austin, TX: Cognitive Science Society).
- Dieussaert, K., Verkerk, S., Gillard, E., and Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *Q. J. Exp. Psychol.* 64, 2352–2367. doi: 10.1080/17470218.2011.588799
- Dunn, L. M., and Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. Minneapolis, MN: NCS Pearson.
- Dupuy, L. E., der Henst, V., Cheylus, A., and Reboul, A. C. (2016). Context in generalized conversational implicatures: the case of some. *Front. Psychol.* 7:381. doi: 10.3389/fpsyg.2016.00381
- Feeeny, A., and Bonnefon, J. F. (2013). Politeness and honesty contribute additively to the interpretation of scalar expressions. *J. Lang. Soc. Psychol.* 32, 181–190. doi: 10.1177/0261927X12456840
- Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511975158

- Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Hamilton, S. T., Freed, E. M., and Long, D. L. (2013). Modeling reader and text interactions during narrative comprehension: a test of the lexical quality hypothesis. *Discourse Processes* 50, 139–163. doi: 10.1080/0163853X.2012.742001
- Hartshorne, J. K., Snedeker, J., Liem Azar, S. Y. M., and Kim, A. E. (2015). The neural computation of scalar implicature. *Lang. Cogn. Neurosci.* 30, 620–634. doi: 10.1080/23273798.2014.981195
- Heyman, T., and Schaeken, W. (2015). Some differences in some: examining variability in the interpretation of scalars using latent class analysis. *Psychol. Belgica* 55, 1–18. doi: 10.5334/pb.bc
- Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English*. Doctoral dissertation, University of California, Los Angeles, CA.
- Huang, Y. T., and Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: evidence from real-time spoken language comprehension. *Dev. Psychol.* 45, 1723–1739. doi: 10.1037/a0016704
- Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., and Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: evidence from picture-sentence verification. *Neurosci. Lett.* 534, 246–251. doi: 10.1016/j.neulet.2012.11.044
- Husband, E. M. (2014). “A subclinical study of the cognitive resources underlying scalar implicature: A focus on scalar adjectives,” in *UCLA Working Papers in Linguistics*, Vol. 18, eds C. T. Schütze and L. Stockall (Los Angeles, CA: UCLA Department of Linguistics), 189–211.
- Hutchison, K. A. (2007). Attentional control and the relatedness proportion effect in semantic priming. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 645–662. doi: 10.1037/0278-7393.33.4.645
- Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007
- Kane, M. J., and Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: an individual-differences perspective. *Psychon. Bull. Rev.* 9, 637–671. doi: 10.3758/BF03196323
- Katsos, N., and Bishop, D. V. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015
- Katsos, N., and Cummins, C. (2010). Pragmatics: from theory to experiment and back again. *Lang. Linguistics Compass* 4, 282–295. doi: 10.1111/j.1749-818X.2010.00203.x
- Katsos, N., Roqueta, C. A., Estevan, R. A. C., and Cummins, C. (2011). Are children with specific language impairment competent with the pragmatics and logic of quantification? *Cognition* 119, 43–57. doi: 10.1016/j.cognition.2010.12.004
- Kemper, S., and Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychol. Aging* 16, 312–322. doi: 10.1037/0882-7974.16.2.312
- Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: The MIT press.
- Marty, P. P., and Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Front. Psychol.* 4:403. doi: 10.3389/fpsyg.2013.00403
- Nieuwland, M. S., Ditman, T., and Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: an ERP investigation of informativeness and pragmatic abilities. *J. Mem. Lang.* 63, 324–346. doi: 10.1016/j.jml.2010.06.005
- Noveck, I., and Sperber, D. (2007). “The why and how of experimental pragmatics: the case of ‘scalar inferences,’” in *Advances in Pragmatics*, ed. N. Burton-Roberts (Basingstoke: Palgrave), 184–212.
- Noveck, I. A., and Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain Lang.* 85, 203–210. doi: 10.1016/S0093-934X(03)00053-1
- Pakulak, E., and Neville, H. J. (2010). Proficiency differences in syntactic processing of monolingual native speakers indexed by event-related potentials. *J. Cogn. Neurosci.* 22, 2728–2744. doi: 10.1162/jocn.2009.21393
- Perfetti, C. (2007). Reading ability: lexical quality to comprehension. *Sci. Stud. Read.* 11, 357–383. doi: 10.1080/10888430701530730
- Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J. P., and Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *J. Autism. Dev. Disord.* 39, 607–618. doi: 10.1007/s10803-008-0661-8
- Politzer-Ahles, S., and Fiorentino, R. (2013). The realization of scalar inferences: context sensitivity without processing cost. *PLoS One* 8:e63943. doi: 10.1371/journal.pone.0063943
- Politzer-Ahles, S., and Gwilliams, L. (2015). Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography. *Lang. Cogn. Neurosci.* 30, 853–866. doi: 10.1080/23273798.2015.1027235
- Politzer-Ahles, S., and Husband, E. M. (2018). Eye movement evidence for context-sensitive derivation of scalar inferences. *Collabra Psychol.* 4:3. doi: 10.1525/collabra.100
- Roberts, C. (2004). “Context in dynamic interpretation,” in *The Handbook of Pragmatics*, eds L. R. Horn and G. Ward (Oxford: Blackwell Publishing), 197–200.
- Roberts, C. (2012). Information structure: towards an integrated formal theory of pragmatics. *Semant. Prag.* 5, 1–69. doi: 10.3765/sp.5.6
- Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D. Jr., and Beck, L. H. (1956). A continuous performance test of brain damage. *J. Consult. Psychol.* 20, 343–350. doi: 10.1037/h0043220
- Sperber, D., and Wilson, D. (2002). Pragmatics, Modularity and Mind-reading. *Mind Lang.* 17, 3–23. doi: 10.1111/1468-0017.00186
- Tagliaferri, B. (2005). *Paradigm. Perception Research Systems, Inc. Version 1*.
- Tomlinson, J. M. Jr., Bailey, T. M., and Bott, L. (2013). Possibly all of that and then some: scalar implicatures are understood in two steps. *J. Mem. Lang.* 69, 18–35. doi: 10.1016/j.jml.2013.02.003
- Van Dyke, J. A., Johns, C. L., and Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition* 131, 373–403. doi: 10.1016/j.cognition.2014.01.007
- Zondervan, A., Meroni, L., and Gualmini, A. (2008). “Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults,” in *Semantics and Linguistic Theory XVIII*, Vol. 18, eds T. Friedman and S. Ito (Ithaca, NY: Cornell University), 765–777.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Yang, Minai and Fiorentino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Linking Hypothesis and Number of Response Options Modulate Inferred Scalar Implicature Rate

Masoud Jasbi<sup>1\*</sup>, Brandon Waldon<sup>2</sup> and Judith Degen<sup>2</sup>

<sup>1</sup> Department of Linguistics, Harvard University, Cambridge, MA, United States, <sup>2</sup> Department of Linguistics, Stanford University, Stanford, CA, United States

## OPEN ACCESS

### Edited by:

Anne Colette Reboul,  
Claude Bernard University Lyon 1,  
France

### Reviewed by:

Daniel Grodner,  
Swarthmore College, United States  
Andrea Beltrama,  
Universität Konstanz, Germany

### \*Correspondence:

Masoud Jasbi  
masoud\_jasbi@fas.harvard.edu

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 11 May 2018

**Accepted:** 21 January 2019

**Published:** 12 February 2019

### Citation:

Jasbi M, Waldon B and Degen J  
(2019) Linking Hypothesis and  
Number of Response Options  
Modulate Inferred Scalar Implicature  
Rate. *Front. Psychol.* 10:189.  
doi: 10.3389/fpsyg.2019.00189

The past 15 years have seen increasing experimental investigations of core pragmatic questions in the ever more active and lively field of experimental pragmatics. Within experimental pragmatics, many of the core questions have relied on the operationalization of the theoretical notion of “implicature rate.” Implicature rate based results have informed the work on acquisition, online processing, and scalar diversity, inter alia. Implicature rate has typically been quantified as the proportion of “pragmatic” judgments in two-alternative forced choice truth value judgment tasks. Despite its theoretical importance, this linking hypothesis from implicature rate to behavioral responses has never been extensively tested. Here we show that two factors dramatically affect the “implicature rate” inferred from truth value judgment tasks: (a) the number of responses provided to participants; and (b) the linking hypothesis about what constitutes a “pragmatic” judgment. We argue that it is time for the field of experimental pragmatics to engage more seriously with its foundational assumptions about how theoretical notions map onto behaviorally measurable quantities, and present a sketch of an alternative linking hypothesis that derives behavior in truth value judgment tasks from probabilistic utterance expectations.

**Keywords:** scalar implicature, methodology, linking hypothesis, experimental pragmatics, truth value judgment task

## 1. INTRODUCTION

The past 15 years have seen the rise and development of a bustling and exciting new field at the intersection of linguistics, psychology, and philosophy: *experimental pragmatics* (Chierchia et al., 2001; Noveck and Posada, 2003; Bott and Noveck, 2004; Papafragou and Tantalou, 2004; Breheny et al., 2006, 2013; De Neys and Schaeken, 2007; Noveck and Reboul, 2008; Bonnefon et al., 2009; Geurts and Poussoulous, 2009; Huang and Snedeker, 2009; Grodner et al., 2010; Barner et al., 2011; Katsos and Bishop, 2011; Tomlinson et al., 2013; Degen and Tanenhaus, 2015, 2016; Bott and Chemla, 2016; van Tiel et al., 2016). Experimental pragmatics is devoted to experimentally testing theories of how language is used in context. How do listeners draw inferences about the – often underspecified – linguistic signal they receive from speakers? How do speakers choose between the many utterance alternatives they have at their disposal?

The most prominently studied phenomenon in experimental pragmatics is undoubtedly *scalar implicature*. Scalar implicatures arise as a result of a speaker producing the weaker of two ordered scalemates (Horn, 1972; Grice, 1975; Hirschberg, 1985; Geurts, 2010). Examples are provided in (1-2).

- (1) Some of her pets are cats.

*Implicature:* Some, but not all, of her pets are cats.

*Scale:* <all, some>

- (2) She owns a cat or a dog.

*Implicature:* She owns a cat or a dog, but not both.

*Scale:* <and, or>

A listener, upon observing the utterances in (1-2) typically infers that the speaker intended to convey the meanings listed as *Implicatures*, respectively. Since Grice (1975), the agreed-upon abstract rationalization the listener could give for their inference goes something like this: the speaker could have made a more informative statement by producing the stronger alternative (e.g., *All of her pets are cats* in (1)). If the stronger alternative is true, they should have produced it to comply with the Cooperative Principle. They chose not to. Assuming the speaker knows whether the stronger alternative is true, it must not be true. The derivation procedure for *ad hoc* exhaustivity inferences such as in (3) is assumed to be calculable in the same way as for scalar implicatures, though the scale is assumed to be contextually driven.

- (3) She owns a cat.

*Implicature:* She owns only a cat.

*Scale:* <cat and dog, cat>

Because the basic reconstruction of the inference is much more easily characterized for scalar implicatures than for other implicatures, scalar implicatures have served as a test bed for many questions in experimental pragmatics, including, but not limited to:

1. Are scalar inferences default inferences, in the sense that they arise unless blocked by (marked) contexts (Horn, 1984; Levinson, 2000; Degen, 2015)?
2. Are scalar inferences default inferences, in the sense that they are computed automatically in online processing and only canceled in a second effortful step if required by context (Bott and Noveck, 2004; Breheny et al., 2006; Huang and Snedeker, 2009; Grodner et al., 2010; Politzer-Ahles and Fiorentino, 2013; Tomlinson et al., 2013; Degen and Tanenhaus, 2016)?
3. What are the (linguistic and extra-linguistic) factors that affect whether a scalar implicature is derived (Breheny et al., 2006, 2013; De Neys and Schaeken, 2007; Bonnefon et al., 2009; Zondervan, 2010; Chemla and Spector, 2011; Bergen and Grodner, 2012; Degen and Goodman, 2014; Degen, 2015; Degen and Tanenhaus, 2015, 2016; Potts et al., 2015; de Marneffe and Tonhauser, in press)?
4. How much diversity is there across implicature types, and within scalar implicatures across scale types, in whether or not an implicature is computed (Doran et al., 2012; van Tiel et al., 2016)?
5. At what age do children acquire the ability to compute implicatures (Noveck, 2001; Musolino, 2004; Papafragou and Tantalou, 2004; Barner et al., 2011; Katsos and Bishop, 2011; Stiller et al., 2015; Horowitz et al., 2017)?

In addressing all of these questions, it has been important to obtain estimates of *implicature rates*. For 1., implicature rates

from experimental tasks can be taken to inform whether scalar implicatures should be considered default inferences. For 2., processing measures on responses that indicate implicatures can be compared to processing measures on responses that indicate literal interpretations. For 3., contextual effects can be examined by comparing implicature rates across contexts. For 4., implicature rates can be compared across scales (or across implicature types). For 5., implicature rates can be compared across age groups.

A standard measure that has stood as a proxy for implicature rate across many studies is the proportion of “pragmatic” judgments in truth value judgment paradigms (Noveck, 2001; Noveck and Posada, 2003; Bott and Noveck, 2004; De Neys and Schaeken, 2007; Geurts and Poussoulous, 2009; Chemla and Spector, 2011; Degen and Goodman, 2014; Degen and Tanenhaus, 2015). In these kinds of tasks, participants are provided a set of facts, either presented visually or via their own knowledge of the world. They are then asked to judge whether a sentence intended to describe those facts is true or false (or alternatively, whether it is right or wrong, or they are asked whether they agree or disagree with the sentence). The crucial condition for assessing implicature rates in these kinds of studies typically consists of a case where the facts are such that the stronger alternative is true and the target utterance is thus also true but underinformative. For instance, Bott and Noveck (2004) asked participants to judge sentences like “Some elephants are mammals”, when world knowledge dictates that all elephants are mammals. Similarly, Degen and Tanenhaus (2015) asked participants to judge sentences like “You got some of the gumballs” in situations where the visual evidence indicated that the participant received all the gumballs from a gumball machine. In these kinds of scenarios, the story goes, if a participant responds “FALSE”, that indicates that they computed a scalar implicature, e.g., to the effect of “Not all elephants are mammals” or “You didn’t get all of the gumballs”, which is (globally or contextually) false. If instead a participant responds “TRUE”, that is taken to indicate that they interpreted the utterance literally as “Some, and possibly all, elephants are mammals” or “You got some, and possibly all, of the gumballs”.

Using the proportion of “FALSE” responses on true but underinformative trials as a proxy for implicature rate is common in experimental pragmatics. For example, in one of the first studies to investigate scalar implicatures experimentally, Noveck (2001) tested adults’ and children’s interpretations of the scalar items *might* and *some*. The dependent measure in Noveck (2001) was the rate of “logically correct responses,” i.e., responding “yes” to statements such as *Some giraffes have long necks* or *There might be a parrot [in the box]* when there had to be a parrot in the box. He found that children responded “yes” more frequently than adults, and concluded that children interpret scalar items *some* and *might* more logically (i.e., literally). Similarly in another landmark study, Papafragou and Musolino (2003) tested children and adults interpretation of the following set of scalar items: <two, three>, <some, all>, and <finish, start>. The dependent measure in this study was the proportion of “No” responses to a puppet’s underinformative statement. The study concluded that “while adults overwhelmingly rejected infelicitous descriptions,



children almost never did so.” Furthermore, the study compared implicature rates across scales and concluded that “children also differed from adults in that their rejection rate on the numerical scale was reliably higher than on the two other scales.” In their final experiment, Papafragou and Musolino (2003) modified their task to invite scalar inferences more easily. They reported that this manipulation resulted in a significantly higher rejection rates. Based on these results, they concluded that children’s ability to compute implicatures is affected by the type of scalar item as well as children’s awareness of the task’s goals. Since these early pioneering studies, the rate of “FALSE” (or “No,” “Wrong,” “Disagree”) responses on underinformative trials in truth-value judgment tasks has become a commonplace dependent measure (Geurts and Poussoulous, 2009; Doran et al., 2012; Potts et al., 2015, *inter alia*.)

Given the centrality of the theoretical notion of “implicature rate” to much of experimental pragmatics, there is to date a surprising lack of discussion of the basic assumption that it is adequately captured by the proportion of “FALSE” responses in truth value judgment tasks [but see Geurts and Poussoulous, 2009; Katsos and Bishop, 2011; Benz and Gotzner, 2014; Degen and Goodman, 2014; Sikos et al., 2019]. Indeed, the scalar implicature acquisition literature was shaken up when Katsos and Bishop (2011) showed that simply by introducing an additional response option, children started looking much more pragmatic than had been previously observed in a binary judgment paradigm. Katsos and Bishop (2011) allowed children to distribute a small, a big, or a huge strawberry to a puppet depending on “how good the puppet said it.” The result was that children gave on average smaller strawberries to the puppet when he produced underinformative utterances compared to when he produced literally true and pragmatically felicitous utterances, suggesting that children do, in fact, display pragmatic ability even at ages when they had previously appeared not to.

But this raises an important question: in truth value judgment tasks, how does the researcher know whether an interpretation is literal or the result of an implicature computation? The binary choice task typically used is appealing in part because it allows for a direct mapping from response options—“TRUE” and “FALSE”—to interpretations—literal and pragmatic. That the seeming simplicity of this mapping is illusory becomes apparent once a third response option is introduced, as in the Katsos and Bishop (2011) case. How is the researcher to interpret the intermediate option? Katsos and Bishop (2011) grouped the intermediate option with the negative endpoint of the scale for the purpose of categorizing judgments as literal vs. pragmatic, i.e., they interpreted the intermediate option as pragmatic. But it seems just as plausible that they could have grouped it with the positive endpoint of the scale and taken the hard line that only truly “FALSE” responses constitute evidence of a full-fledged implicature. The point here is that there has been remarkably little consideration of *linking hypotheses* between behavioral measures and theoretical constructs in experimental pragmatics, a problem in many subfields of psycholinguistics (Tanenhaus, 2004). We argue that it is time to engage more seriously with these issues.

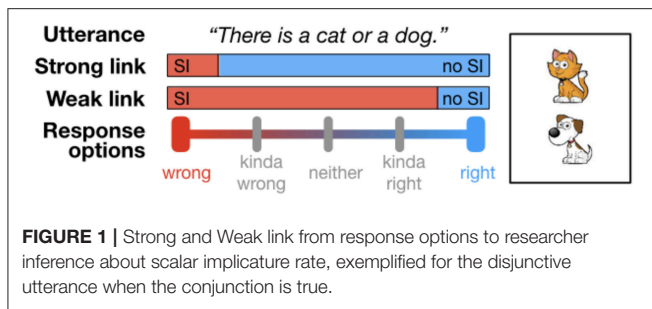
We begin by reporting an experiment that addresses the following question: do the number of response options provided in a truth value judgment task and the way that responses are grouped into pragmatic (“SI”) and literal (“no SI”) change inferences about scalar implicature rates? Note that this way of asking the question assumes two things: first, that whatever participants are doing in a truth value judgment task, the behavioral measure can be interpreted as providing a measure of interpretation; and second, that listeners either do or do not compute an implicature on any given occasion. In the General Discussion we will discuss both of these issues. Following Degen and Goodman (2014), we will offer some remarks on why truth value judgment tasks are better thought of as measuring participants’ estimates of speakers’ *production* probabilities. This will suggest a completely different class of linking hypotheses. We then discuss an alternative conception of scalar implicature as a probabilistic phenomenon, a view that has recently rose to prominence in the subfield of probabilistic pragmatics (Franke and Jäger, 2016; Goodman and Frank, 2016). This alternative conception of scalar implicature, we argue, affords developing and testing quantitative linking hypotheses in a rigorous and motivated way.

Consider a setup in which a listener is presented a card with a depiction of either one or two animals (see **Figure 1** for an example). As in a standard truth value judgment task, the listener then observes an underinformative utterance about this card (e.g., “There is a cat or a dog on the card”) and is asked to provide a judgment on a scale with 2, 3, 4, or 5 response options, with endpoints “wrong” and “right.”<sup>1</sup> In the binary case, this reproduces the standard truth value judgment task. **Figure 1** exemplifies (some of) the researcher’s options for grouping responses. Under what we will call the “Strong link” assumption, only the negative endpoint of the scale is interpreted as evidence for a scalar implicature having been computed. Under the “Weak link” assumption, in contrast, any response that does not correspond to the positive endpoint of the scale is interpreted as evidence for a scalar implicature having been computed. Intermediate grouping schemes are also possible, but these are the ones we will consider here. Note that for the binary case, the Weak and Strong link return the same categorization scheme, but for any number of response options greater than 2, the Weak and Strong link can in principle lead to differences in inferences about implicature rate.

Let’s examine an example. Assume three response options (wrong, neither, right). Assume further that each of the three responses was selected by a third of participants, i.e., the distributions of responses is 1/3, 1/3, and 1/3. Under the Strong link, we infer that this task yielded an implicature rate of 2/3. Under the Weak link, we infer that this task yielded an implicature rate of 1/3. This is quite a drastic difference if we are, for instance, interested in whether scalar implicatures are inference defaults and we would like to interpret an implicature rate of above an arbitrary threshold (e.g., 50%) as evidence for

<sup>1</sup> An open question concerns the extent to which the labeling of points on the scale affects judgments (e.g., “wrong”–“right” vs. “false”–“true” vs. “disagree”–“agree”). Studies vary in the labeling of scale points.





such a claim. Under the Strong link, we would conclude that scalar implicatures are not defaults. Under the Weak link, we would conclude that they are. In the experiment reported in the following section, we presented participants with exactly this setup. We manipulated the number of response options between participants and analyzed the results under different linking hypothesis<sup>2</sup>.

## 2. EXPERIMENT

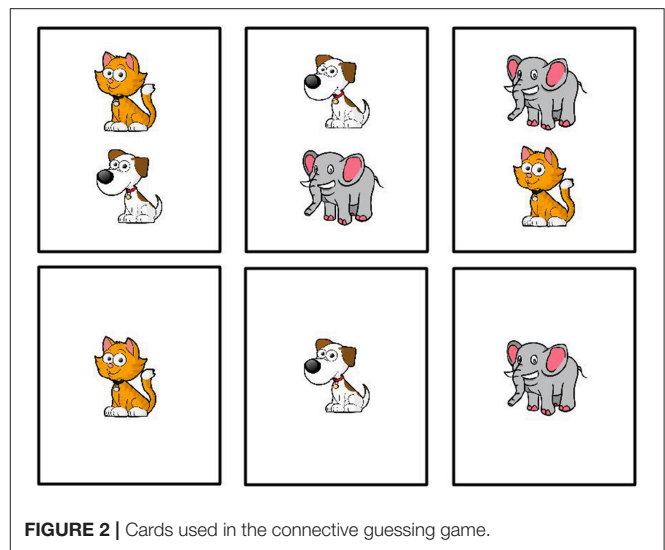
Participants played an online card game in which they were asked to judge descriptions of the contents of cards. Different groups of participants were presented with different numbers of response options. On critical trials, participants were presented with descriptions for the cards that typically result in exhaustivity implicatures ("There is a cat on the card" when there was a cat and a dog) or scalar implicatures ("There is a cat or a dog on the card" when there was a cat and a dog). We categorized their responses on such trials according to the Weak and the Strong link introduced above, and tested whether the number of response options and the linking hypothesis led to different conclusions about the rate of computed implicatures in the experimental task.

### 2.1. Methods

#### 2.1.1. Participants

Two hundred participants were recruited via Amazon Mechanical Turk. They optionally provided demographic information at the end of the study. Participants' mean age was 35. We also asked participants if they had any prior training in logic. 40 participants reported that they did, while 160 had no

<sup>2</sup>Researchers may vary with respect to which linking hypothesis (Weak vs. Strong, or others) they consider most plausible. Supporters of the Weak link may argue that there are three theoretically motivated categories of judgments: false, true but infelicitous, true and felicitous. Under such an account, it is plausible that false and true+felicitous responses occupy the ends of the false/true scale while true but infelicitous responses occupy the mid portion. On critical trials, participants judge underinformative statements that are true but infelicitous and therefore the mid portion of the scale can provide evidence for implicature computation. However, supporters of the Strong link may argue that if a participant computes an implicature, their response in the task should show the commitment to that interpretation by judging the underinformative utterance as false. Any other response shows that they have not truly computed an implicature. So far, these discussions have remained largely informal. In this paper we stay neutral with respect to the plausibility of each link and only aim to demonstrate the consequences of assuming them.



prior training in logic. All participants' data was included in the final analysis<sup>3</sup>.



#### 2.1.2. Materials and Procedure

The study was administered online through Amazon Mechanical Turk<sup>4</sup>. Participants were first introduced to the set of cards we used in the study (Figure 2). Each card depicted one or two animals, where an animal could be either a cat, a dog, or an elephant. Then participants were introduced to a blindfolded fictional character called Bob. Bob was blindfolded to avoid violations of ignorance expectations associated with the use of disjunction (Chierchia et al., 2001; Sauerland, 2004). Participants were told that Bob would guess the contents of the cards and their task was to indicate whether Bob's guess was wrong or right. On each trial, participants saw a card and a sentence representing Bob's guess. For example, they saw a card with a cat and read the sentence "There is a cat on the card." They then provided an assessment of Bob's guess. The study ended after 24 trials.

Two factors were manipulated within participants: card type and guess type. There were two types of cards, cards with only one animal on them and cards with two animals. There were three types of guesses: simple (e.g., *There is a cat*), conjunctive (e.g., *There is a cat and a dog*), and disjunctive (e.g., *There is a cat or a dog*). Crossing card type and guess type yielded trials of varying theoretical interest (see Figure 3): critical underinformative trials that were likely to elicit pragmatic inferences (either scalar or exhaustive) and control trials that were either unambiguously true or false. Each trial type occurred three times with randomly

<sup>3</sup>This study was carried out in accordance with the recommendations of the Common Rule, Federal Office for Human Research Protections. The protocol was approved by the Stanford University IRB 2 (non-medical research). All subjects gave Informed consent, documentation was waived by the IRB.

<sup>4</sup>The experiment can be viewed here [https://cdn.rawgit.com/thegricean/si-paradigms/94a590f0/experiments/main/1\\_methods/online\\_experiment/connective\\_game.html](https://cdn.rawgit.com/thegricean/si-paradigms/94a590f0/experiments/main/1_methods/online_experiment/connective_game.html)

elephant	cat	cat and dog	cat or dog	
control: unambiguously false	control: unambiguously true	control: unambiguously false	control: unambiguously true	
control: unambiguously false	<b>critical: exhaustivity implicature</b>	control: unambiguously true	<b>critical: scalar implicature</b>	

**FIGURE 3 |** Trial types (critical and control). Headers indicate utterance types. Rows indicate card types. Critical trials are marked in bold.

sampled animals and utterances that satisfied the constraint of the trial type. Trial order was randomized.

On critical trials, participants could derive implicatures in two ways. First, on trials on which two animals were present on the card (e.g., cat and dog) but Bob guessed only one of them (e.g., “There is a cat on the card”), the utterance could have a literal interpretation (“There is a cat and possibly another animal on the card”) or an exhaustive interpretation (“There is only a cat on the card”). We refer to these trials as “exhaustive”. Second, on trials on which two animals were on the card (e.g., a cat and a dog) and Bob used a disjuncton (e.g., “There is a cat or a dog on the card”), the utterance could have the literal, inclusive, interpretation, or a pragmatic, exclusive interpretation. We refer to these trials as “scalar.”

In order to assess the effect of the number of response options on implicature rate, we manipulated number of response options in the forced choice task between participants. We refer to the choice conditions as “binary” (options: *wrong*, *right*), “ternary” (options: *wrong*, *neither*, *right*), “quaternary” (options: *wrong*, *kinda wrong*, *kinda right*, *right*), and “quinary” (*wrong*, *kinda wrong*, *neither*, *kinda right*, *right*). Thus, the endpoint labels always remained the same. If there was an uneven number of response options, the central option was *neither*. Participants were randomly assigned to one of the four task conditions.

## 2.2. Results and Discussion

The collected dataset contains 50 participants in the binary task, 53 in the ternary task, 43 in the quaternary task, and 54 in the quinary task. **Figures 4–7** show the proportions of response choices in each of the 8 trial types on each of the four response tasks, respectively. We report the relevant patterns of results qualitatively before turning to the quantitative analysis of interest.

### 2.2.1. Qualitative Analysis

In the binary task, participants were at or close to ceiling in responding “right” and “wrong” on unambiguously true and false trials, respectively (see **Figure 4**). However, on underinformative trials (i.e., a “cat” or “cat or dog” description for a card with both a cat and a dog), we observe pragmatic behavior: on exhaustive trials, participants judged the utterance “wrong” 14% of the time; on scalar trials, participants judged the utterance “wrong” 38%

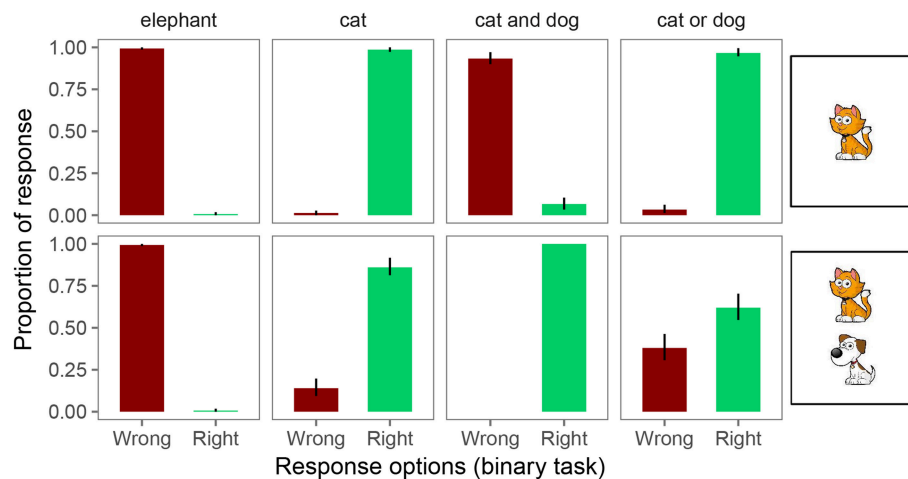
of the time. That is, both under the Weak and Strong link assumptions introduced in the Introduction, inferred implicature rate on exhaustive trials is 14% and on scalar trials 38%.

In the ternary task, participants were also at or close to ceiling in responding “right” and “wrong” on unambiguously true and false trials, respectively (see **Figure 5**). And again, on underinformative trials (a “cat” and “cat or dog” description for a card with both a cat and a dog), we observed pragmatic behavior: on exhaustive trials, participants considered the guess “wrong” 8% of the time and neither wrong nor right 12% of the time. On scalar trials, participants judged the guess “wrong” 23% of the time and “neither” 11% of the time. This means that the Weak and Strong link lead to different conclusions about implicature rates on the ternary task. Under the Weak link, inferred implicature rate on exhaustive trials is 20%; under the Strong link it is only 8%. Similarly, under the Weak link, inferred implicature rate on scalar trials is 34%; under the Strong link it is only 23%.

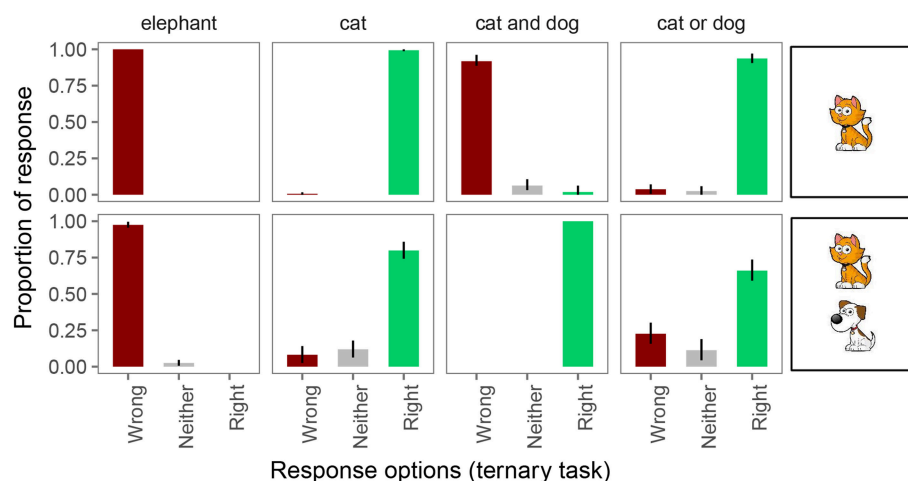
In the quaternary task (**Figure 6**), participants were again at or close to ceiling in responding “right” and “wrong” on 4 of the 6 unambiguously true and false trials. However, with four response options, two of the control conditions appear to be showing signs of pragmatic infelicity: when a conjunction was used and only one of the animals was on the card, participants considered the guess “wrong” most of the time (46%), but they often considered it “kinda wrong” (32%) or even “kinda right” (19%). This suggests that perhaps participants considered the notion of a partially true or correct statement in our experimental setting. Disjunctive descriptions of cards with only one animal, while previously at ceiling for “right” responses, were downgraded to only “kinda right” 26% of the time, presumably because these utterances are also underinformative, though the degree of underinformativeness may be less egregious than on scalar trials.

On underinformative exhaustive trials, we observed pragmatic behavior as before: participants judged the guess “wrong” 2% of the time, “kinda wrong” 5% of the time, and “kinda right” 66% of the time. On scalar trials, participants judged the guess “wrong” 6% of the time, “kinda wrong” 12% of the time, and “kinda right” 43% of the times.

Thus, we are again forced to draw different conclusions about implicature rates depending on whether we assume the



**FIGURE 4 |** Proportion of responses for the binary forced choice judgments. Error bars indicate 95% binomial confidence intervals (Sison and Glaz, 1995).



**FIGURE 5 |** Proportion of responses for the ternary forced choice judgments. Error bars indicate 95% multinomial confidence intervals (Sison and Glaz, 1995).

Weak link or the Strong link. Under the Weak link, inferred implicature rate on exhaustive trials is 73%; under the Strong link it is only 2%. Similarly, under the Weak link, inferred implicature rate on scalar trials is 61%; under the Strong link it is only 6%.

Finally, **Figure 7** shows the proportion of responses in the quinary task. Performance on the 4 pragmatically felicitous control trials was again at floor and ceiling, respectively. The 2 control conditions in which the quaternary task had revealed pragmatic infelicity again displayed that pragmatic infelicity in the quinary task, suggesting that this is a robust type of pragmatic infelicity that, nonetheless, requires fine-grained enough response options to be detected experimentally.

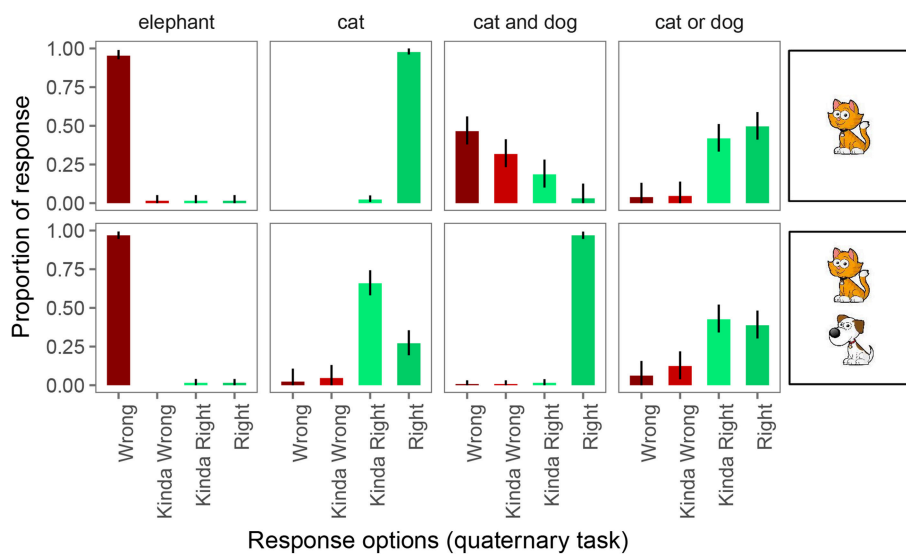
On underinformative exhaustive trials, we observed pragmatic behavior as before: participants judged the guess “wrong” 2% of the time, “kinda wrong” 1% of the time, “neither” 1% of the time, and “kinda right” 72% of the time. On scalar trials, participants judged the guess “wrong” 6% of the time, “kinda wrong” 4%

of the time, “neither” 1% of the time, and “kinda right” 52% of the time.

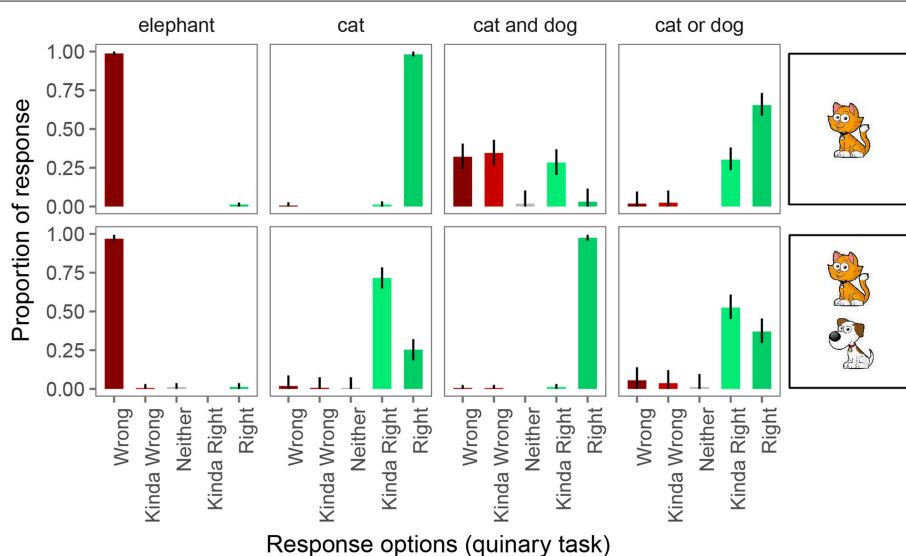
Thus, we would again draw different conclusions about implicature rates depending on whether we assume the Weak link or the Strong link. Under the Weak link, inferred implicature rate on exhaustive trials is 76%; under the Strong link it is only 2%. Similarly, under the Weak link, inferred implicature rate on scalar trials is 63%; under the Strong link it is only 6%.

### 2.2.2. Quantitative Analysis

Our primary goal in this study was to test whether the estimated implicature rate in the experimental task is affected by the linking hypothesis and the number of response options available to participants. To this end, we only analyzed the critical trials (exhaustive and scalar). In particular, we classified each data point from critical trials as constituting an implicature (1) or not (0) under the Strong and Weak link. **Figure 8** shows the resulting implicature rates by condition and link. It is immediately



**FIGURE 6** | Proportion of responses for the quaternary forced choice judgments. Error bars indicate 95% multinomial confidence intervals (Sison and Glaz, 1995).



**FIGURE 7** | Proportion of responses for the quinary forced choice judgments. Error bars indicate 95% multinomial confidence intervals (Sison and Glaz, 1995).

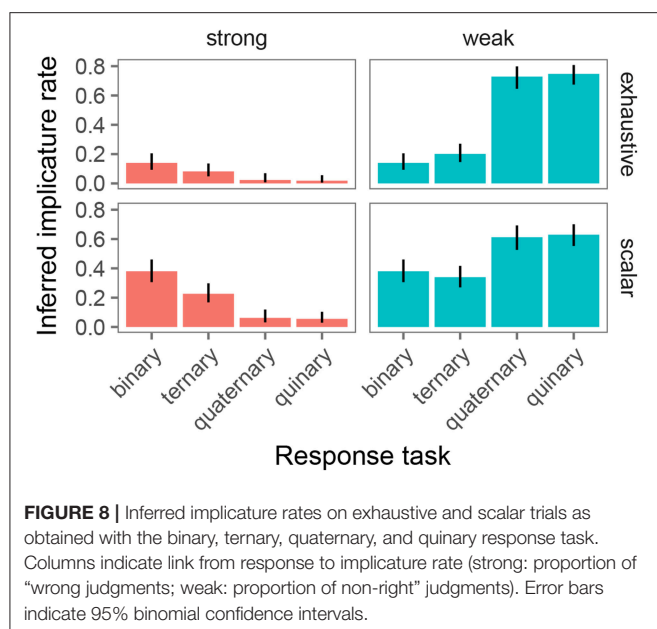
apparent that there is variability in inferred implicature rate. In particular, the Weak link appears to result in greater estimates of implicature rates in tasks with four or five response options, compared to the Strong link. For the binary and ternary task, the assumed link appears to play a much smaller role.

To analyze the effect of link and response options on inferred implicature rate, we used a Bayesian binomial mixed effects model using the R package “brms” (Bürkner, 2016) with weakly informative priors.<sup>5</sup> The model predicted the log odds of implicature over no implicature from fixed effects of *response*

*type* (binary, ternary, quaternary, quinary—dummy-coded with binary as reference level), *link* (strong vs. weak—dummy-coded with strong as reference level), and *trial type* (exhaustive vs. scalar—dummy-coded, with exhaustive as reference level), as well as their two-way and three-way interactions. Following Barr et al. (2013), we included the maximal random effects structure justified by the design: random intercepts for items (cards) and participants, random by-participant slopes for link, trial type, and their interaction, and random by-item slopes for link, trial type, response type, and their interactions. Since the number of response options was a between-participant variable we did not include random slopes of response options for participants. Four chains converged after 2,000 iterations each (warmup = 1,000).

<sup>5</sup>For more information about the default priors of the “brms” package, see the brms package manual.





**TABLE 1 |** Model parameter estimates and their credible intervals.

Predictors	Estimate	2.5%	97.5%	Evidence
Intercept	−8.60	−13.98	−4.53	*
Link = Weak	−0.15	−4.86	4.77	
Task = Quaternary	−1.83	−8.08	4.20	
Task = Quinary	−4.05	−10.90	2.38	
Task = Ternary	−1.45	−7.31	4.56	
Implicature = Scalar	6.09	1.00	12.29	*
Link = Weak : Task = Quaternary	14.03	7.24	21.88	*
Link = Weak : Task = Quinary	17.28	10.64	25.80	*
Link = Weak : Task = Ternary	3.81	−1.49	9.22	
Link = Weak : Implicature = Scalar	0.90	−4.01	6.43	
Task = Quaternary : Implicature = Scalar	−5.67	−13.66	1.54	
Task = Quinary : Implicature = Scalar	−2.31	−9.30	4.61	
Task = Ternary : Implicature = Scalar	−1.31	−7.70	4.65	
Link=Weak : Task=Quaternary : Implicature=Scalar	−3.29	−12.07	4.55	
Link=Weak : Task=Quinary : Implicature=Scalar	−7.74	−16.59	−0.16	*
Link=Weak : Task=Ternary : Implicature=Scalar	−1.44	−7.00	4.22	

Rows marked with an asterisk in the evidence column do not contain 0 in the credible interval, thereby providing evidence for an effect.

**Table 1** summarizes the mean parameter estimates and their 95% credible intervals.  $\hat{R} = 1$  for all estimated parameters. All the analytical decisions described here were pre-registered<sup>6</sup>.

The model provided evidence for the following effects: First, there was a main effect of trial type such that scalar trials resulted in greater implicature rates than exhaustive trials (Mean Estimate = 6.09, 95% Credible Interval=[1, 12.29]). Second, there was an interaction between link and number of response options such that the quaternary task (Mean Estimate = 14.03,

95% Credible Interval = [7.24, 21.88]) and the quinary task (Mean Estimate = 17.28, 95% Credible Interval = [10.64, 25.80]) resulted in greater implicature rates with a weak link than with a strong link, but there was no evidence of a link-dependent difference in inferred implicature rate for the binary and ternary task. Finally, there was a three-way interaction between link, trial type, and number of response options, driven by the binary/quinary contrast (Mean Estimate = −7.74, 95% Credible Interval=[−16.59, −0.16]). Simple effects analysis on only the binary and quinary trials, separately for the exhaustive and scalar subset of the data, revealed that the three-way interaction is driven by a different effect of number of response options under the Weak vs. Strong link for the two inference types. Specifically, on exhaustive trials, number of response options (2 vs. 5) only resulted in greater implicature rates under the Weak ( $\beta = .2$ ,  $p < 0.0001$ ), but not the Strong link ( $\beta = -0.8$ ,  $p < .82$ ). In contrast, on scalar trials, number of response options (2 vs. 5) resulted in greater implicature rates under the Weak ( $\beta = 3.6$ ,  $p < 0.005$ ) link, and in lower implicature rates under the Strong link ( $\beta = -4.0$ ,  $p < 0.0007$ ).

In sum, both number of response options and link affected the inferred implicature rate, as did the type of inference (exhaustive vs. scalar).

### 3. GENERAL DISCUSSION

#### 3.1. Summary and Methodological Discussion

In this paper we asked whether linking hypothesis and number of response options available to participants in truth value judgment tasks affects inferred implicature rates. The results presented here suggest they do. A linking assumption that considered the highest point on the scale literal and any lower point pragmatic (Weak link) resulted in higher implicature rates in tasks with 4 or 5 response options compared to the standard two options. A linking hypothesis that considered the lowest point on the scale pragmatic and any higher point literal (Strong link) reported lower implicature rates in tasks with 4 or 5 options compared to the standard two options. The results suggest that the choice of linking hypothesis is a crucial analytical step that can significantly impact the conclusions drawn from truth value judgment tasks. In particular, there is danger for pragmatic ability to be both under- and overestimated.

While the binary truth value judgement task avoids the analytic decision between Strong and Weak linking hypothesis, the results reported here suggest that binary tasks can also underestimate participants' pragmatic competence. In binary tasks, participants are often given the lowest and highest points on a scale (“wrong” vs. “right”) and are asked to report pragmatic infelicities using the lowest point (e.g., “wrong”). The study reported here showed that on trials with true but pragmatically infelicitous descriptions, participants often avoided the lowest point on the scale if they were given more intermediate options. Even though the option “wrong” was available to participants in all tasks, participants in tasks with intermediate options chose it

<sup>6</sup>Our preregistration can be accessed at <https://aspredicted.org/tq3sz.pdf>

less often. In computing implicature rate, this pattern manifested itself as a decrease in implicature rate under the Strong link when more response options were provided, and an increase in implicature rate under the Weak link when more response options were provided. These observations are in line with Katsos and Bishop (2011)'s argument that pragmatic violations are not as severe as semantic violations and participants do not penalize them as much. Providing participants with only the extreme ends of the scale (e.g., wrong/right, false/true) when pragmatic violations are considered to be of an intermediate nature risks misrepresentation of participants' pragmatic competence. It further suggests that in studies that use binary tasks to investigate response-contingent processing, proportions of "literal" responses may be a composite of both literal and pragmatic underlying interpretations that just happen to get mapped differently onto different response options by participants.

This study did not investigate the effect of response labels on the inferred implicature rate. However, the results provided suggestive evidence that some options better capture participant intuitions of pragmatic infelicities than others. Among the intermediate options, "kinda right" was chosen most often to report pragmatic infelicities. The option "neither" was rarely used in the ternary and quinary tasks (where it was used as a midpoint), suggesting that participants interpreted pragmatic infelicities as different degrees of being "right" and not "neither right nor wrong." Therefore, options that capture degrees of being "right" like "kinda right" may prove most suitable for capturing infelicity in the long run. We leave this as a methodological issue for future research.

The study had three further design features worth investigating in future work. First, the utterances were ostensibly produced by a blindfolded character. This was an intentional decision to control for violation of ignorance expectations with disjunction. A disjunction such as "A or B" often carries an implication or expectation that the speaker is not certain which alternative actually holds. Future work should investigate how the violation of the ignorance expectation interacts with link and number of response options in inferred implicature rate. Second, in this study we considered exhaustive and scalar implicatures with *or*. If the observed effects of link and number of response options hold in general, they should be observable using other scales, e.g., on implicatures with *some*. Finally, our experiment was designed as a guessing game and the exact goal or task-relevant Question Under Discussion of the game was left implicit. Given the past literature on QUD effects on scalar implicature, we expect that different goals—e.g., to help the character win more points vs. to help the character be more accurate—would affect how strict or lenient participants are with their judgments and ultimately affect implicature rate in the task (Zondervan, 2010; Degen and Goodman, 2014). Future work should systematically vary the goal of the game and explore its effects on the inferred implicature rate. But crucially, it's unlikely that the observed effects of number of response options and linking hypothesis on inferred implicature rate are dependent on any of the discussed design choices.

### 3.2. Revisiting the Linking Hypothesis

On the traditional view of the link between implicature and behavior in sentence verification tasks, scalar implicature is conceptualized as a binary, categorical affair – that is, an implicature is either "calculated" or it isn't, and the behavioral reflexes of this categorical interpretation process should be straightforwardly observed in experimental paradigms. This assumption raises concerns for analyzing variation in behavior on a truth value judgment task; for example, why did the majority of respondents in the binary condition of our experiment answer "right" to an utterance of the underinformative "There is a cat or dog" when the card had both a cat and a dog on it? And why did a sizeable minority nonetheless choose "wrong" in this same condition?

To explain these data on the traditional view, we are forced to say that a) not all participants calculated the implicature; or that b) some participants who calculated the implicature did not choose the anticipated (i.e., "wrong") response due to some other cognitive process which overrode the "correct" implicature behavior; or some mixture of (a) and (b). We might similarly posit that one or both of these factors underlie the variation in the ternary, quaternary, and quinary conditions. However, without an understanding of how to quantitatively specify the link between implicature calculation and its behavioral expression, the best we can hope for on this approach is an analysis which predicts general qualitative patterns in the data (e.g., a prediction of relatively more "right" responses than "wrong" responses in a given trial of our binary truth value judgment task, or a prediction of a rise in the rate of "right"/"wrong" responses between two experimental conditions, given some contextual manipulation). However, we should stress that to the best of our knowledge, even a qualitative analysis of this kind of variation in behavior on sentence verification tasks – much less the effect of the number of response choices on that behavior – is largely underdeveloped in the scalar implicature literature.

We contrast the above view of implicature and its behavioral reflexes with an alternative linking hypothesis. Recent developments in the field of probabilistic pragmatics have demonstrated that pragmatic production and comprehension can be captured within the Rational Speech Act (RSA) framework (Frank and Goodman, 2012; Degen et al., 2013, 2015; Goodman and Stuhlmüller, 2013; Kao et al., 2014; Qing and Franke, 2015; Bergen et al., 2016; Franke and Jäger, 2016; Goodman and Frank, 2016). Much in the spirit of Gricean approaches to pragmatic competence, the RSA framework takes as its point of departure the idea that individuals are rational, goal-oriented communicative agents, who in turn assume that their interlocutors similarly behave according to general principles of cooperativity in communication. Just as in more traditional Gricean pragmatics, pragmatic inference and pragmatically-cooperative language production in the RSA framework are, at their core, the product of counterfactual reasoning about alternative utterances that one might produce (but does not, in the interest of cooperativity). However, the RSA framework explicitly and quantitatively models cooperative interlocutors as agents whose language production and comprehension is

a function of Bayesian probabilistic inference regarding other interlocutors' expected behavior in a discourse context.

Specifically, in the RSA framework we model pragmatically competent listeners as continuous probabilistic distributions over possible meanings (states of the world) given an utterance which that listener observes. The probability with which this listener  $L_1$  ascribes a meaning  $s$  to an utterance  $u$  depends upon a prior probability distribution of potential states of the world  $P_w$ , and upon reasoning about the communicative behavior of a speaker  $S_1$ .  $S_1$  in turn is modeled as a continuous probabilistic distribution over possible utterances given an intended state of the world the speaker intends to communicate. This distribution is sensitive to a rationality parameter  $\alpha$ , the production cost  $C$  of potential utterances, and the informativeness of the utterance, quantified via a representation of a literal listener  $L_0$  whose interpretation of an utterance is in turn a function of that utterance's truth conditional content  $[[u]](s)$  and her prior beliefs about the state of the world  $P_w(s)$ .

$$\begin{aligned} P_{L_1}(s|u) &\propto P_{S_1}(u|s) * P_w(s) \\ P_{S_1}(u|s) &\propto \exp(\alpha(\log(P_{L_0}(s|u)) - C(u))) \\ P_{L_0}(s|u) &\propto [[u]](s) * P_w(s) \end{aligned}$$

This view contrasts with the traditional view in that it is rooted in a quantitative formalization of pragmatic competence which provides us a continuous measure of pragmatic reasoning. In the RSA framework, individuals never categorically draw (or fail to draw) pragmatic inferences about the utterances they hear. For example, exclusivity readings of disjunction are represented in RSA as relatively lower posterior conditional probability of a conjunctive meaning on the  $P_L$  distribution given an utterance of "or", compared to the prior probability of that meaning. Thus, absent auxiliary assumptions about what exactly would constitute "implicature," it is not even possible to talk about rate of implicature calculation in the RSA framework. The upshot, as we show below, is that this view of pragmatic competence does allow us to talk explicitly and quantitatively about rates of observed behavior in sentence verification tasks.

We take inspiration from the RSA approach and treat participants' behavior in our experimental tasks as the result of a soft-optimal pragmatic speaker in the RSA framework. That is, following Degen and Goodman (2014), we proceed on the assumption that behavior on sentence verification tasks such as truth value judgment tasks, is best modeled as a function of an individual's mental representation of a cooperative speaker ( $S_1$  in the language of RSA) rather than of a pragmatic listener who interprets utterances ( $P_{L_1}$ )<sup>7</sup>. In their paper, Degen and Goodman show that sentence verification tasks are relatively more sensitive to contextual features like the Question Under

Discussion than are sentence interpretation tasks, and that this follows if sentence interpretation tasks—but not sentence verification tasks—require an additional layer of counterfactual reasoning about the intentions of a cooperative speaker.

A main desideratum of a behavioral linking hypothesis given the RSA view of pragmatic competence is to transform continuous probability distributions into categorical outputs (e.g., responses of "right"/"wrong" in the case of the binary condition of our experiment). For a given utterance  $u$  and an intended communicated meaning  $s$ ,  $S_1(u | s)$  outputs a conditional probability of  $u$  given  $s$ . For example, in the binary condition of our experiment where a participant evaluated "There is a cat or a dog" when there were both animals on the card, the participant has access to the mental representation of  $S_1$  and hence to the  $S_1$  conditional probability of producing the utterance "cat or dog" given a cat and dog card:  $S_1(\text{"cat or dog"} | \text{cat and dog})$ . According to the linking hypothesis advanced here, the participant provides a particular response to  $u$  if the RSA speaker probability of  $u$  lies within a particular probability interval. We model a responder,  $R$ , who in the binary condition responds "right" to an utterance  $u$  in world  $s$  just in case  $S_1(u|s)$  meets or exceeds some probability threshold  $\theta$ :

$$\begin{aligned} R(u, w, \theta) \\ = \text{"right"} \text{ iff } S_1(u|s) \geq \theta \\ = \text{"wrong"} \text{ otherwise} \end{aligned}$$

The model of a responder in the binary condition is extended intuitively to the condition where participants had three response options. In this case, we allow for two probability thresholds:  $\theta_1$ , the minimum standard for an utterance in a given world state to count as "right", and  $\theta_2$ , the minimum standard for "neither". Thus, in the ternary condition,  $R(u, s, \theta_1, \theta_2)$  is "right" iff  $S_1(u | s) \geq \theta_1$  and "neither" iff  $\theta_1 > S_1(u | s) \geq \theta_2$ . To fully generalize the model to our five experimental conditions, we say that  $R$  takes as its input an utterance  $u$ , a world state  $s$ , and a number of threshold variables dependent on a variable  $c$ , corresponding to the experimental condition in which the participant finds themselves (e.g., the range of possible responses available to  $R$ ).

Given  $c = \text{"ternary"}$

$$\begin{aligned} R(u, w, \theta_1, \theta_2) \\ = \text{"right"} \text{ iff } S_1(u | s) \geq \theta_1 \\ = \text{"neither"} \text{ iff } \theta_1 > S_1(u | s) \geq \theta_2 \\ = \text{"wrong"} \text{ otherwise} \end{aligned}$$

Given  $c = \text{"quaternary"}$

$$\begin{aligned} R(u, w, \theta_1, \theta_2, \theta_3) \\ = \text{"right"} \text{ iff } S_1(u | s) \geq \theta_1 \\ = \text{"kinda right"} \text{ iff } \theta_1 > S_1(u | s) \geq \theta_2 \\ = \text{"kinda wrong"} \text{ iff } \theta_2 > S_1(u | s) \geq \theta_3 \\ = \text{"wrong"} \text{ otherwise} \end{aligned}$$

Given  $c = \text{"quinary"}$

$$\begin{aligned} R(u, w, \theta_1, \theta_2, \theta_3, \theta_4) \\ = \text{"right"} \text{ iff } S_1(u | s) \geq \theta_1 \\ = \text{"kinda right"} \text{ iff } \theta_1 > S_1(u | s) \geq \theta_2 \\ = \text{"neither"} \text{ iff } \theta_2 > S_1(u | s) \geq \theta_3 \\ = \text{"kinda wrong"} \text{ iff } \theta_3 > S_1(u | s) \geq \theta_4 \\ = \text{"wrong"} \text{ otherwise} \end{aligned}$$

<sup>7</sup>Degen and Goodman (2014) argue that sentence verification is more plausibly construed as a production task rather than as an interpretation task because participants, unlike in natural language comprehension, are provided with the ground truth about the state of the world that a speaker is describing. Thus, participants are in essence being asked to assess the quality of a speaker's utterance. In contrast, Degen and Goodman argue, true interpretation tasks are characterized by the listener inferring what the state of the world is that the speaker is describing, for instance by selecting from one of multiple interpretation options.

In an RSA model,  $S_1(u | s)$  will be defined for any possible combination of possible utterance and possible world state. One consequence of this is that for the purposes of our linking hypothesis, participants are modeled as employing the same decision criterion – does  $S_1(u | s)$  exceed the threshold? – in both “implicature” and “non-implicature” conditions of a truth value judgment task experiment. That is, participants never evaluate utterances directly on the basis of logical truth or falsity: for example, our blindfolded character Bob’s guess of “cat and dog” on a cat and dog card trial is “right” to the vast majority of participants not because the guess is logically true but because  $S_1(\text{“cat and dog”} | \text{cat and dog})$  is exceedingly high.

For further illustration, we use our definition of a pragmatically-competent speaker  $S_1$  (as defined above) to calculate the speaker probabilities of utterances in states of the world corresponding to our experimental conditions (i.e., for “cat,” “dog,” “cat and dog,” and “elephant,” given either a cat on the card, or both a cat and a dog on the card). In calculating these probabilities, we assume that the space of possible utterances is the set of utterances made by Bob in our experiment (i.e., any possible single, disjunctive, or conjunctive guess involving “cat,” “dog,” or “elephant”). For the purposes of our model, we assume a uniform cost term on all utterances. We furthermore assume that the space of possible meanings corresponds to the set of possible card configurations that a participant may have seen in our experiment, and that the prior probability distribution over these world states is uniform. Lastly, we set  $\alpha$ —the speaker rationality parameter—to 1. The resulting speaker probabilities are shown in **Figure 9**.<sup>8</sup>

The linking hypothesis under discussion assumes that speaker probabilities of utterance given meaning are invariant across a) our four different experimental conditions, b) across participants, and c) within participants (that is, participants do not update their  $S_1$  distribution in a local discourse context). We note that the assumption (b) may conceivably be relaxed by allowing one or more of the parameters in the model – including the prior probability over world states  $P_w$ , the cost function on utterances  $C$ , or the rationality parameter  $\alpha$ —to vary across participants. We also note that assumption (c) in particular is in tension with a growing body of empirical evidence that semantic and pragmatic interpretation is modulated by rapid adaptation to the linguistic and social features of one’s interlocutors (Fine et al., 2013; Kleinschmidt and Jaeger, 2015; Yildirim et al., 2016).

However, if we should like to keep the above simplifying assumptions in place, then this linking hypothesis commits us to explaining variation in the data in terms of the threshold parameters of our responder model  $R$ . Consider first the variation in response across different experimental conditions on a given trial, e.g., evaluation of a guess of “cat and dog” when the card contains both a cat and a dog. The variation in the proportion of responses of “right” on this trial between the binary, ternary, quaternary, and quinary conditions indicates that the threshold value for “right” responses must vary across conditions; that is,

we predict that the  $\theta$  of the binary condition will differ from, e.g., the  $\theta_1$  of the ternary condition as well as the  $\theta_1$  of the quaternary condition. We also observed variation in response on this trial within a single condition (for example, a sizeable minority of participants responded “wrong” to this trial in the binary condition). Thus, this linking hypothesis is committed to the notion that threshold values may vary across participants, such that a speaker probability of utterance  $S_1(u | s)$  can fall below  $\theta$  for some subset of participants while  $S_1(u | s)$  itself remains constant across participants.

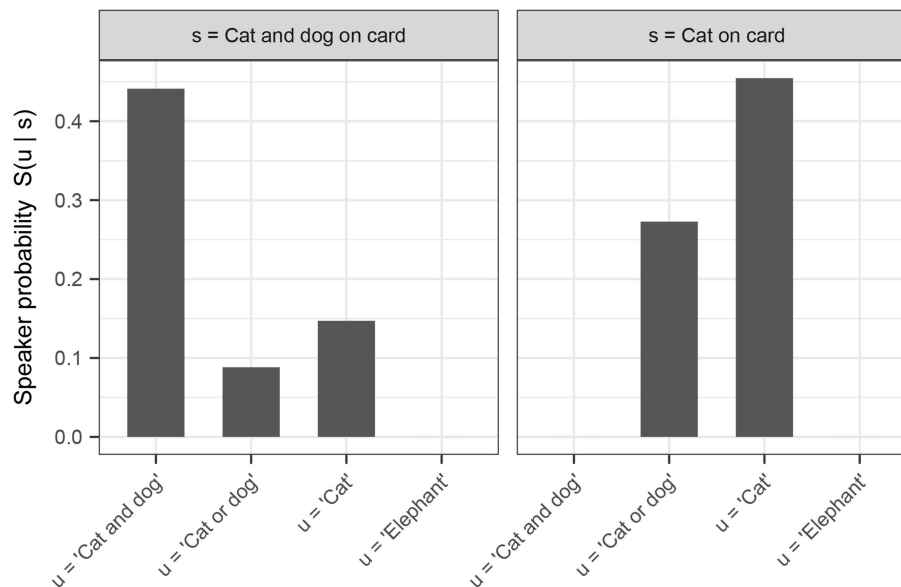
Lastly, for two utterances of the same conditional probability and in the same experimental condition, participants in our experiment sometimes provided a judgment of “right” to one utterance but “wrong” to the other. That is, there was within-subject variation in this experiment. One way to represent such variation would be to posit that the parameterization of threshold values proceeds stochastically and that threshold values are recalibrated for every individual sentence verification task. Rather than representing a threshold as a discrete value  $N$  between 0 and 1, we can represent that threshold as a distribution over possible threshold values – with mass centered around  $N$ . Whenever an individual encounters a single trial of our truth value judgment task experiment, a threshold value is sampled from this distribution. By allowing values of  $\theta$  to vary stochastically in this way, we can capture that  $S_1(u | s)$  can fall both above and below  $\theta$  for a given participant.

The model in its present form already captures an interesting asymmetry in inferred implicature rates between exhaustive and scalar trials of the experiment: note specifically (c.f. **Figure 8**) that inferred implicature rates are greater in the binary and ternary conditions for scalar trials over exhaustive trials. This is expected given the model’s inferred speaker probabilities: the speaker probability of producing “There is a cat on the card” in the context of there being a cat and dog on the card (an exhaustive implicature-inducing trial) is greater than the speaker probability of producing “There is a cat or a dog on the card” in that same context (a scalar implicature-inducing trial). Assuming noisy  $\theta$  values centered around  $N$ , participants are expected to respond “Right” more frequently on exhaustive than on scalar trials, which is precisely what is observed. Recall that these probabilities were derived via the simplifying assumption of uniform cost on utterances; in fact, adding cost to relatively complex disjunctive sentences over simple declarative sentences only predicts a more pronounced asymmetry in the experimentally-observed direction.

As suggested above, the quantitative predictions of our model will depend crucially on the values assigned to its free parameters – including (but not limited to) the probability thresholds and speaker costs of possible utterances. However, the values of these parameters can be estimated in a principled and informed manner through Bayesian statistical analysis of our experimental data. Samples from prior distributions over possible parameter values yield predicted patterns of response, which are then compared against empirically-observed response patterns in order to determine the a posteriori probability that these values are in fact the “real” latent parameter values. The resulting

<sup>8</sup>Note that the probabilities in each facet don’t sum to 1 because the model considers all possible disjunctive, conjunctive, and simple utterances, while we are only visualizing the ones corresponding to the experimental conditions.





**FIGURE 9 |** Speaker probabilities of utterances on the exhaustive and scalar trials, as obtained using the model described in this section.

posterior distributions are sampled from in turn, in order to parameterize the model and assess overall quantitative fit given the data. Though we leave a quantitative assessment of our model to future work, we sketch the general procedure here to emphasize that the model is amenable to rigorous and data-driven evaluation.

One empirical problem is the pattern of responses we observed for “cat and dog” on trials where there was only a cat on the card. Because this utterance is strictly false in this world state, it is surprising—on both the traditional view as well as on the account developed here—that participants assigned this utterance ratings above “wrong” with any systematicity. However, this is what we observed, particularly in the quaternary and quinary conditions of the experiment, where a sizeable minority of participants considered this utterance “kinda right”. As **Figure 9** demonstrates, the conditional speaker probability of this utterance in this world state is 0; thus, there is no conceivable threshold value that would allow this utterance to ever be rated above “wrong” (on the reasonable assumption that the thresholds in our responder model  $R$  should be non-zero). Any linking hypothesis will have to engage with this data point, and we leave to future work an analysis which captures participants’ behavior in this condition.

For the time being, however, we present the above analysis as a proof of concept for the following idea: by relaxing the assumptions of the traditional view of scalar implicature—namely, that scalar implicatures either are or are not calculated, and that behavior on sentence verification tasks directly reflects this binary interpretation process—we can propose quantitative models of the variation in behavior that is observed in experimental settings. We note that the linking hypothesis proposed here is just one in the space of possible hypotheses. For example, one might reject this threshold-based analysis

in favor of one whereby responses are the outcomes of sampling on the (pragmatic speaker or pragmatic listener) probability distributions provided by an RSA model. We leave this systematic, quantitative investigation to future work. For now we emphasize that explicit computational modeling of behavioral responses is a tool that is available to researchers in experimental pragmatics. While using the RSA framework as the modeling tool requires revising traditional assumptions about the nature of scalar implicature by relaxing the crisp notion of scalar implicature as something that is or is not “calculated” in interpretation, it provides new flexibility to explicitly discuss behavior in experimental settings. One need not adopt the RSA framework as the tool for hypothesizing and testing the link between theoretical constructs and behavior in pragmatic experiments. However, the empirical findings we have reported here—that the inferences researchers draw about “implicature rate” are volatile and depend on various features of the paradigm and the linking hypothesis employed—strongly suggest that experimental pragmatics as a field must engage more seriously with the foundational questions of what we are measuring in the experiments we run.

Concluding, we have shown in this paper that inferred “implicature rate”—a ubiquitous notion in theoretical and experimental pragmatics—as estimated in truth value judgment tasks, depends on both the number of responses participants are provided with as well as on the linking hypothesis from proportion of behavioral responses to “implicature rate”. We further sketched an alternate linking hypothesis that treats behavioral responses as the result of probabilistic reasoning about speakers’ likely productions. While a thorough model comparison is still outstanding, this kind of linking hypothesis opens a door toward more systematic and rigorous formulation and testing of linking hypotheses between theoretical

notions of interest in pragmatics and behavioral responses in experimental paradigms.

## AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the study. MJ conducted the online survey studies;

reported the results and performed the statistical analysis. BW conducted the modeling and wrote the discussion section. JD wrote the theoretical introduction, and contributed to the experimental section and the discussion section and the modeling sections. All authors contributed to manuscript revision, read, and approved the submitted version.

## REFERENCES

- Barner, D., Brooks, N., and Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition* 118, 84–93. doi: 10.1016/j.cognition.2010.10.010
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001
- Benz, A., and Gotzner, N. (2014). "Embedded implicatures revisited: issues with the truth-value judgment paradigm," in *Proceedings of the Formal & Experimental Pragmatics Workshop, European Summer School for Language, Logic and Information (ESSLLI)*, eds J. Degen, M. Franke, and N. D. Goodman (Tübingen), 1–6.
- Bergen, L., and Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 1450–1460. doi: 10.1037/a0027850
- Bergen, L., Levy, R., and Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semant. Pragmat.* 9, 1–46. doi: 10.3765/sp.9.20
- Bonnefon, J.-F., Feeney, A., and Villejoubert, G. (2009). When some is actually all: scalar inferences in face-threatening contexts. *Cognition* 112, 249–258. doi: 10.1016/j.cognition.2009.05.005
- Bott, L., and Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *J. Mem. Lang.* 91, 117–140. doi: 10.1016/j.jml.2016.04.004
- Bott, L., and Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006
- Breheny, R., Ferguson, H. J., and Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition* 126, 423–440. doi: 10.1016/j.cognition.2012.11.012
- Breheny, R., Katsos, N., and Williams, J. (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. doi: 10.1016/j.cognition.2005.07.003
- Bürkner, P.-C. (2016). brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01
- Chemla, E., and Spector, B. (2011). Experimental evidence for embedded scalar implicatures. *J. Semant.* 28, 359–400. doi: 10.1093/jos/ffq023
- Chierchia, G., Crain, S., Teresa, M., Guasti, M. T., Gualmini, A., and Meroni, L. (2001). "The acquisition of disjunction: evidence for a grammatical view of scalar implicatures," in *Proceedings of the 25th Annual Boston University Conference on Language Development*, A. H.-J. Do, L. Domínguez and A. Johansen (Somerville, MA: Cascadia Press), 157–168.
- de Marneffe, M.-C., and Tonhauser, J. (in press). "Inferring meaning from indirect answers to polar questions: the contribution of the rise-fall-rise contour," in *Questions in Discourse*, eds E. Onea, M. Zimmermann, and K. von Heusinger (Leiden: Brill Publishing).
- De Neys, W., and Schaeken, W. (2007). When people are more logical under cognitive load - dual task impact on scalar implicature. *Exp. Psychol.* 54, 128–133. doi: 10.1027/1618-3169.54.2.128
- Degen, J. (2015). Investigating the distribution of "some" (but not "all") implicatures using corpora and web-based methods. *Semant. Pragmat.* 8, 1–55. doi: 10.3765/sp.8.11
- Degen, J., Franke, M., and Jäger, G. (2013). "Cost-based pragmatic inference about referential expressions," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 376–281.
- Degen, J., and Goodman, N. D. (2014). "Lost your marbles? The puzzle of dependent measures in experimental pragmatics," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, eds P. Bello, M. Guarini, M. McShane, and B. Scassellati (Austin, TX: Cognitive Science Society), 397–402.
- Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature A constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171
- Degen, J., and Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: a visual world eye-tracking study. *Cogn. Sci.* 40, 172–201. doi: 10.1111/cogs.12227
- Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. in *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 548–553.
- Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88, 124–154. doi: 10.1353/lan.2012.0008
- Fine, A. B., Jaeger, T. F., Farmer, T. F., and Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE* 8:e77661. doi: 10.1371/journal.pone.0077661
- Frank, M. C., and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science* 336:998. doi: 10.1126/science.1218633
- Franke, M., and Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Z. Sprachwissenschaft*, 35, 3–44. doi: 10.1515/zfs-2016-0002
- Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press.
- Geurts, B., and Pouscoulous, N. (2009). Embedded implicatures?? *Semant. Pragmat.* 2, 1–34. doi: 10.3765/sp.2.4
- Goodman, N. D., and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* 20, 818–829. doi: 10.1016/j.tics.2016.08.005
- Goodman, N. D., and Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Top. Cogn. Sci.* 5, 173–184. doi: 10.1111/tops.12007
- Grice, H. P. (1975). Logic and conversation. *Syntax Semant.* 3, 41–58.
- Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition* 116, 42–55. doi: 10.1016/j.cognition.2010.03.014
- Hirschberg, J. (1985). *A Theory of Scalar Implicature*. Ph.D. thesis, University of Pennsylvania; Garland Publishing Company.
- Horn, L. (1972). *On the Semantic Properties of the Logical Operators in English*. Ph.D. thesis, UCLA.
- Horn, L. (1984). "Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature," in *Meaning, Form, and Use in Context: Linguistic Applications*, ed D. Schiffrin (Washington, DC: Georgetown University Press), 11–42.
- Horowitz, A. C., Schneider, R. M., and Frank, M. C. (2017). The trouble with quantifiers: exploring children's deficits in scalar implicature. *Child Dev.* 89, e572–e593. doi: 10.1111/cdev.13014
- Huang, Y. T., and Snedeker, J. (2009). On-line interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cogn. Psychol.* 58, 376–415. doi: 10.1016/j.cogpsych.2008.09.001
- Kao, J., Wu, J., Bergen, L., and Goodman, N. D. (2014). Nonliteral understanding of number words. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12002–12007. doi: 10.1073/pnas.1407479111
- Katsos, N., and Bishop, D. V. M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and

- implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015
- Kleinschmidt, D. F., and Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev.* 122, 148–203. doi: 10.1037/a0038695
- Levinson, S. C. (2000). *Presumptive Meanings - The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.
- Musolino, J. (2004). The semantics and acquisition of number words: integrating linguistic and developmental perspectives. *Cognition* 93, 1–41. doi: 10.1016/j.cognition.2003.10.002
- Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1
- Noveck, I., and Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain Lang.* 85, 203–210. doi: 10.1016/S0093-934X(03)00053-1
- Noveck, I. A., and Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends Cogn. Sci.* 12, 425–431. doi: 10.1016/j.tics.2008.07.009
- Papafragou, A., and Musolino, J. (2003). Scalar implicature: experiments at the semantics-pragmatics interface. *Cognition* 86, 253–282. doi: 10.1016/S0010-0277(02)00179-8
- Papafragou, A., and Tantalou, N. (2004). Children's computation of implicatures. *Lang. Acquisit.* 12, 71–82. Available online at: <https://www.jstor.org/stable/20011567>
- Politzer-Ahles, S., and Fiorentino, R. (2013). The realization of scalar inferences: context sensitivity without processing cost. *PLoS ONE* 8:e63943. doi: 10.1371/journal.pone.0063943
- Potts, C., Lassiter, D., Levy, R., and Frank, M. C. (2015). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *J. Semant.* 33, 755–802. doi: 10.1093/jos/ffv012
- Qing, C., and Franke, M. (2015). "Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning," in *Bayesian Natural Language Semantics and Pragmatics*, Vol. 2, eds H. Zeevat and H.-C. Schmitz (Cham: Springer International Publishing), 201–220. doi: 10.1007/978-3-319-17064-0\_9
- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguist. Philos.* 27, 367–391. doi: 10.1023/B:LING.0000023378.71748.db
- Sikos, L., Kim, M., and Grodner, D. J. (2019). Social context modulates tolerance for pragmatic violations in binary but not graded judgments. *Front. Psychol.*
- Sison, C. P., and Glaz, J. (1995) Simultaneous confidence intervals and sample size determination for multinomial proportions. *J. Am. Stat. Assoc.* 90, 366–369.
- Stillier, A. J., Goodman, N. D., and Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Lang. Learn. Dev.* 11, 176–190. doi: 10.1080/15475441.2014.927328
- Tanenhaus, M. K. (2004). "On-line sentence processing: past, present and future," in *On-line Sentence Processing: ERPS, Eye Movements and Beyond*, eds M. Carreiras and C. Clifton (London: Psychology Press), 371–392.
- Tomlinson, J. M., Bailey, T. M., and Bott, L. (2013). Possibly all of that and then some: scalar implicatures are understood in two steps. *J. Mem. Lang.* 69, 18–35. doi: 10.1016/j.jml.2013.02.003
- van Tiel, B., van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *J. Semant.* doi: 10.1093/jos/ffu017
- Yildirim, I., Degen, J., Tanenhaus, M.K., and Jaeger, T.F. (2016). Talker-specificity and adaptation in quantifier interpretation. *J. Mem. Lang.* 87, 128–143. doi: 10.1016/j.jml.2015.08.003
- Zondervan, A. (2010). *Scalar Implicatures or Focus: An Experimental Approach*. Ph.D. thesis, Universiteit Utrecht, Amsterdam.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Jasbi, Waldon and Degen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Social Context Modulates Tolerance for Pragmatic Violations in Binary but Not Graded Judgments

Les Sikos<sup>1\*</sup>, Minjae Kim<sup>2</sup> and Daniel J. Grodner<sup>3</sup>

<sup>1</sup>Department of Language Science and Technology, Saarland University, Saarbrücken, Germany, <sup>2</sup>Department of Psychology, Boston College, Boston, MA, United States, <sup>3</sup>Department of Psychology, Swarthmore College, Swarthmore, PA, United States

## OPEN ACCESS

### Edited by:

Anne Colette Reboul,  
Claude Bernard University Lyon 1,  
France

### Reviewed by:

Masoud Jasbi,  
Harvard University,  
United States  
Greta Mazzaggio,  
Università degli Studi di Firenze, Italy

### \*Correspondence:

Les Sikos  
sikos@coli.uni-saarland.de

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 16 May 2018

**Accepted:** 20 February 2019

**Published:** 20 March 2019

### Citation:

Sikos L, Kim M and Grodner DJ  
(2019) Social Context Modulates  
Tolerance for Pragmatic Violations in  
Binary but Not Graded Judgments.  
Front. Psychol. 10:510.  
doi: 10.3389/fpsyg.2019.00510

A common method for investigating pragmatic processing and its development in children is to have participants make binary judgments of underinformative (UI) statements such as *Some elephants are mammals*. Rejection of such statements indicates that a (not-all) scalar implicature has been computed. Acceptance of UI statements is typically taken as evidence that the perceiver has not computed an implicature. Under this assumption, the results of binary judgment studies in children and adults suggest that computing an implicature may be cognitively costly. For instance, children under 7 years of age are systematically more likely to accept UI statements compared to adults. This makes sense if children have fewer processing resources than adults. However, Katsos and Bishop (2011) found that young children are able to detect violations of informativeness when given graded rather than binary response options. They propose that children simply have a greater tolerance for pragmatic violations than do adults. The present work examines whether this pragmatic tolerance plays a role in adult binary judgment tasks. We manipulated social attributes of a speaker in an attempt to influence how accepting a perceiver might be of the speaker's utterances. This manipulation affected acceptability rates for binary judgments (Experiment 1) but not for graded judgments (Experiment 2). These results raise concerns about the widespread use of binary choice tasks for investigating pragmatic processing and undermine the existing evidence suggesting that computing scalar implicatures is costly.

**Keywords:** language, pragmatics, inference, pragmatic tolerance, scalar implicature, truth value judgment, social cognition

## INTRODUCTION

Much of what we communicate in conversation is implicit. For example, if a speaker says, "Some of the students passed the test," comprehenders often infer that *not all* of the students passed. This is a pragmatic inference. It arises because communication is typically cooperative. Cooperative speakers should, among other things, make the strongest statement compatible with their knowledge. This follows from the maxim of quantity (Grice, 1975). The speaker chose a relatively vague expression (*some*) rather than a more specific one (*all*). The comprehender can thus infer that the speaker was not in a position to use the more informative expression.



This frequently leads to the inference that a stronger statement (*All of the students passed the test*) is false.

This is an example of a *scalar implicature* (Horn, 1972). In recent years, scalar implicatures have become a central testing ground for investigating how implicit meanings are computed and how pragmatic communication abilities develop. To explore these issues, researchers frequently ask participants to judge underinformative (UI) statements such as *Some elephants are mammals* (see Katsos and Cummins, 2012 for a review). These utterances are literally true, but their implicit *not-all* meaning is false. The rejection of a UI statement in a binary sentence acceptability judgment task is thought to indicate that a pragmatic inference has been computed. Acceptance is taken as an indication that only a literal interpretation has been computed.

There is considerable variation across individuals and situations in judgments for UI statements. Studies report that anywhere from 23 to 83% of adult respondents judge such sentences false depending on various factors (see Dieussaert et al., 2011 for review). One important factor appears to be cognitive processing resources. Participants take longer to judge UI statements as false rather than true. This is consistent with the notion that participants initially compute the literal meaning of UIs before engaging in an effortful second stage process of computing the pragmatic meaning. In support, when participants are given less time to respond (Bott and Noveck, 2004; Bott et al., 2012) or are asked to do a secondary memory task (De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013) the acceptance rate of UI statements increases, but not the acceptance rates for patently true or false statements (e.g., *All elephants are mammals*, *Some elephants are reptiles*). Further, individuals with smaller working memory capacity exhibit greater acceptance of UI sentences (Feeney et al., 2004; Dieussaert et al., 2011). Acceptance rates also decrease when a larger proportion of stimuli are UI statements or when alternative utterances are made more salient (Foppolo et al., 2012). Both of these manipulations should make it easier to make the comparisons necessary to generate the inference. These results are anticipated if computing scalar inferences requires time and cognitive resources.

In contrast to adult response patterns, developmental studies on the acquisition of scalar inference report that children under 7-years-old reliably accept UI statements.<sup>1</sup> This has led many researchers to conclude that young children lack the cognitive resources or the pragmatic competence to derive conversational inferences at adult-like levels (see Noveck and Reboul, 2008). However, studies that do not use judgment tasks generally indicate that young children can generate scalar implicatures. Pouscoulous et al. (2007) asked children to perform an act out task to make a display of boxes accurately conform to a statement. In a situation where five of five boxes contained a token, the experimenter said, "I would like some of the boxes to contain a token." Nearly 70% of 4-year-olds removed a coin from at least one of the boxes. This strongly suggests that they generated a *not all* implicature. Similar evidence was found

by (Horowitz et al., 2018; Experiment 2) using a referential identification task. The experimenter said, "On the cover of my book, some of the pictures are cats." Children as young as 4.5 years old reliably selected a book for which two of four pictures were cats more often than a book for which four of four pictures were cats.

Katsos and Bishop (2011; see also Veenstra et al., 2018) propose that the acceptance of pragmatically infelicitous statements in binary judgment tasks may reflect a greater tolerance of pragmatic violations rather than a lack of pragmatic competence *per se*. They found that when participants were given a ternary rather than binary judgment task (awarding a cartoon speaker a "small," "big," or "huge" strawberry reward), 5- to 6-year-old children and adults were both sensitive to informativity (i.e., they gave UI statements a smaller reward than optimally informative statements such as *Some mammals are elephants*) and tolerant of pragmatic violations (i.e., they gave UI statements a bigger reward than false statements). In a separate study, they replicated the typical pattern whereby children at this age systematically accept UI statements in a binary judgment task. Katsos and Bishop concluded that children do in fact detect violations of informativity for UI statements, but do not consider these pragmatic violations grave enough to warrant outright rejection in a binary choice task. In other words, children may in general be more charitable and forgiving in binary judgment tasks than adults.

Note that recognizing UI statements as underinformative requires computing alternative statements that might have been made (such as *All elephants are mammals*) and determining whether any of these alternatives are more optimally informative than what was actually said. These are precisely the steps involved in generating a scalar implicature. Indeed, the computation of alternatives has been proposed as the primary cognitive bottleneck in scalar implicature computation in adults and children (Barner et al., 2011; Marty and Chemla, 2013; Tiel and Schaeken, 2017). Katsos and Bishop's pattern of results indicates that children do generate scalar inferences and that this is observable when provided with an appropriate response scale. This result is thus problematic for the view that children lack the cognitive resources or pragmatic skills necessary to generate scalar implicatures. It also calls into question the use of binary choice scales for investigating scalar implicatures in children. The primary goal of the current studies is to examine whether pragmatic tolerance might also play a role in binary judgment tasks for adults.

A potential issue with binary response options is that they artificially constrain the perceivers' choices. In natural conversation, there are many moves available to an interlocutor who is asked to judge the validity of a statement. For instance, a UI statement might elicit an explanatory qualification (*True, but incomplete or inappropriate*; *Not quite*) or a request for clarification (*Do you mean not all?*). Indeed in most circumstances, it would be uncooperative to merely tell the speaker that they were right or wrong without providing some additional feedback. This is because there are multiple ways that a statement can be infelicitous. It may be false, off topic, vague, suffer from presupposition failure (e.g., *The current king*

<sup>1</sup>This is true regardless of whether the task is a statement evaluation task or a truth value judgment (see Foppolo et al., 2012).

of France is bald), or otherwise inapt. A UI statement is neither completely true nor false but pragmatically odd. Thus, even when an individual computes the scalar inference, making a binary judgment compels the perceiver to make a complex metalinguistic judgment about where to place the threshold for acceptability. This raises the possibility that variability in binary response tasks reflects differences in response selection processes when faced with two poor options rather than, or in addition to, differences in computing a pragmatic inference. On this view, we would anticipate that determining where to set the threshold in a binary choice task could be influenced by factors that affect how forgiving the addressee might be toward the speaker's utterance. This would be true even in cases where these factors are not directly relevant to whether an implicature has been generated.

In contrast, a ternary judgment task provides an intermediate response option that allows respondents an explicit way to signal that UI statements are worse than patently true statements, but better than patently false ones. If so, in situations where participants are provided with three response options rather than two, the intermediate response should be favored (ala Katsos and Bishop, 2011) regardless of the social context or cognitive task demands.

Most previous studies of adult UI sentence processing have asked participants to make judgments on isolated, context-free sentences as stimuli. However, computing a pragmatic inference requires that the comprehender recover the communicative intentions of a cooperative speaker. With context-free sentences, it is unclear what the communicative intentions of the speaker might be: some participants may not attempt to compute a pragmatic interpretation at all given the lack of social context, while others might attempt to attribute particular characteristics and intentions to the speaker in order to judge their pragmatic felicity. As a result, variability in response judgments could be at least partially due to differences in the social attributions that comprehenders covertly ascribe to the disembodied speaker. In an attempt to control this potential aspect of variability, the studies below provide rich communicative contexts with clear goals within which participants are asked to make their judgments.

Furthermore, we hypothesized that social attributes of the speaker might influence how tolerant the perceiver is of the speaker's utterance. For example, people may be more tolerant of pragmatic violations from speakers they consider to be more likeable. While such attributes do not change the fundamental communicative task and hence should not affect whether an implicature has been drawn, they may make the participant more or less accepting of the speaker's utterances. The experiments below directly test this hypothesis by manipulating the social attributes of the speaker. If variability in binary choice tasks reflects response selection processes rather than different rates of implicature computation, then this social manipulation will have a greater effect on judgments of UI statements when using a binary scale (Experiment 1) than when using a ternary scale (Experiment 2). In sum, we are interested in whether pragmatic tolerance is affected by social attributes of the speaker, a manipulation that should not directly affect implicature computation *per se*.

## EXPERIMENT 1

The goal of Experiment 1 was to test whether attributes of the speaker that are not directly related to the communicative task can affect adult comprehenders' tolerance for pragmatic violations in a binary judgment task. Participants were provided with a specific social context. They were assigned to tutor an 8-year-old boy on a biology exam on which he was asked to create quantified statements involving animal species and classes. This task provides a plausible cover story for why the speaker might make UI and patently false statements. It also makes clear the purpose of his utterances and the perceiver's role in the communication. Participants were given a brief description of the student as a Sympathetic, Unsympathetic, or Non-native English-speaking child. The Sympathetic speaker was described as kind and adorable. The Unsympathetic speaker was depicted as cruel and obnoxious. The Non-native speaker was described as speaking English as a foreign language. Importantly, his native language was described as lacking quantifiers.

The aim of this speaker manipulation was to create differing social contexts that might influence adults into being more or less charitable with their judgments of the speaker's pragmatic violations. For instance, previous work has shown that individuals who are perceived as more likeable receive higher scores on performance assessments in various situations (e.g., Sonnentag, 1998). It was expected that the Sympathetic speaker condition would elicit greater charity from participants. This in turn might engender increased tolerance for pragmatic infelicity relative to the Unsympathetic condition. The Non-native speaker was included to potentially increase the rate of rejections by providing social motivation to focus specifically on the appropriate use of quantifiers. Since participants were told that Bobby's native tongue lacks words for specifying quantities, they may have elected to pay special attention to his use of quantifiers in order to help him. This could have led to decreased tolerance for using *some* when *all* would have been more informative compared with the other speaker conditions.

Though speaker type was manipulated between subjects by altering the introductory text, the stimuli, feedback options, and core judgment task were identical for all participants. UI statements in this test-taking context are less optimally informative than a potential alternative statement for all three speaker types. Thus, we should anticipate that implicature rates are similar across the different speakers. If the rate of rejections is different across speakers, this would be evidence that binary judgments are driven by processes other than implicature calculation *per se*.

## Materials and Methods

### Participants

A total of 102 English-speaking adults were recruited to participate in an online questionnaire through Amazon's Mechanical Turk in exchange for \$0.60. Participants were restricted to those living in the United States, who had completed at least 100 Human Intelligence Tasks (HITs), and who had an excellent performance record on previous HITs (minimum 97% approval rating). The survey was implemented and hosted

on Qualtrics. Four participants failed to submit their data at the end of the survey.

### Stimuli

A total of 120 categorical statements were constructed in 6 sentence types, with 20 statements per type (Table 1). All statements contained a quantifier (*all* or *some*) followed by a subset-superset relationship that paired an animal exemplar (subset) with an animal category (superset). Critical items (UI) were literally true but pragmatically false. Thus, acceptability judgments for such items had no correct or incorrect answer. The remaining sentence types were fillers that described either patently true or patently false subset-superset relations. Ten counterbalanced lists were constructed from these materials such that each list contained ten UI items and ten filler items (two items each of sentence types F1–F5), and no exemplar from a category was used more than once per list. Thus, each list contained 50% UI statements. This proportion has been shown to elicit a high percentage of pragmatic responses in adults (Dieussaert et al., 2011).

### Instructions

Three parallel sets of instructions were created. They differed only in their characterization of the speaker. All participants saw the following: “Imagine that you have been assigned as a tutor to a young student named Bobby. Bobby is currently studying basic biology. He has just taken a test in which he had to make true sentences out of animal names, animal traits and amount words (‘some,’ ‘all,’ ‘none’). While he has a solid understanding of the animals he studied in class, he has trouble forming appropriate sentences to communicate his knowledge. Your task is to go over each item of the test with Bobby, tell him how he did, and to provide additional feedback to help him create better sentences.” Participants then read one of the following descriptions:

1. *Sympathetic speaker*. “Bobby’s teacher has told you that Bobby is an adorable, funny, outgoing, 8-year-old boy with an unfortunate developmental disorder. Like most children with this disorder, Bobby is eager to interact socially with the people around him but he is hindered with significant speech and language delays. Although Bobby is now a reasonably good communicator, he still lags significantly behind his age-matched peers.”
2. *Unsympathetic speaker*. “Bobby’s teacher has told you that Bobby is a very difficult and obnoxious 8-year-old boy who is often suspended from school because of his repeated violent outbursts. For example, he recently broke a 5-year-old girl’s arm and then laughed at her while she cried. His teachers have told you that Bobby learns best when given clear and direct feedback on tests and assignments.”
3. *Non-native speaker*. “Bobby’s teacher has told you that Bobby is a bright, friendly, 8-year-old boy from Brazil who speaks Gazuungu, an Amazonian language that is known for a number of unusual features. In particular, Gazuungu has no ‘amount words’ for generic quantities less than 10, so

**TABLE 1** | Examples of sentence types.

Type	Example	Correct response
F1	All birds are parrots	“Not quite”
F2	All cats are birds	“Not quite”
F3	All parrots are birds	“That’s right”
F4	Some birds are parrots	“That’s right”
F5	Some cats are birds	“Not quite”
UI	<b>Some parrots are birds</b>	<b>?</b>

it has no equivalents for English words like ‘some.’ Instead, quantities less than 10 must be described using exact numbers. Bobby already knows quite a bit of English but he would like to learn to speak it perfectly. Bobby is patient and does not mind being corrected because it means he is learning.”

### Procedure

Participants were randomly assigned to a speaker condition. After the instructions, participants completed two practice items (not UI statements). Participants were then randomly assigned to one of the 10 stimulus lists. All 20 experimental items were presented on a single screen with the order of items randomized for each participant. Participants responded to each item by selecting between two radio buttons labeled “That’s right” and “Not quite,” and then provided any additional explanation they thought might be useful for Bobby in a text entry field (e.g., “That’s right. Tigers, like other mammals, have fur”). The survey took approximately 10–15 min to complete.

### Exit Survey

Following the experimental task, participants were given three 3-option multiple choice questions designed to assess attentiveness to the speaker characteristics: (1) *How old is Bobby?* Options: 6, 8, 12; (2) *How was this student described?* Options: Kind, Amazonian, Obnoxious; and (3) *What subject is he studying?* Options: Biology, Mathematics, Geography. Participants were then asked to judge how likeable Bobby was on a 7-point Likert scale followed by eight demographic questions.

## Results

### Statistical Methods and Exclusion Criteria

Response data were modeled with logistic mixed effect regression using the *glmer* function in the *lme4* package within the statistical language R (Bates et al., 2014b) and all models consisted of the maximal participant and item random effects structure justified by the data and design (Barr et al., 2013; Bates et al., 2014a). To render model coefficients more interpretable, continuous independent variables were centered around their mean and categorically manipulated predictors were sum coded. Reported coefficients are in logit units.

Two participants were eliminated for reporting that their age of English acquisition was in adulthood (Each learned at 24 or older, all other participants learned at age 6 or younger). The mean accuracy for responses to filler items (statements type

F1-F5) was used as a proxy for attentiveness to the task. Three participants were excluded for accuracy rates below 70%. The remaining 93 participants were relatively evenly distributed across speaker conditions ( $N_{\text{Non-native}} = 31$ ;  $N_{\text{Sympathetic}} = 28$ ;  $N_{\text{Unsympathetic}} = 34$ ). For these participants, mean accuracy rates to filler items were high ( $M = 95\%$ ,  $SE = 8.3\%$ ) and did not differ across conditions ( $z_s \ll 1$ ). Responses are depicted in **Figure 1**.

### Judgments of UI Statements

For UI sentences, the rate of rejections was reliably affected by speaker type: A maximum likelihood ratio test revealed that a model containing speaker type as a fixed effect provided a better fit to the data than one without ( $\chi^2[2] = 5.15$ ,  $p = 0.076$ ). Pairwise comparisons indicated that Non-native Bobby was reliably more likely to be rejected than Unsympathetic Bobby ( $\beta = 1.93$ ,  $SE = 0.98$ ,  $z = 1.97$ ,  $p < 0.05$ ) and marginally more than Sympathetic Bobby ( $\beta = 1.89$ ,  $SE = 1.01$ ,  $z = 1.88$ ,  $p = 0.06$ ). There were no differences in rejections for Sympathetic and Unsympathetic Bobby ( $z = 0.09$ ). There were no effects of speaker condition for any of the filler sentence categories (all  $z_s < 1$ ).

### Exit Survey Results

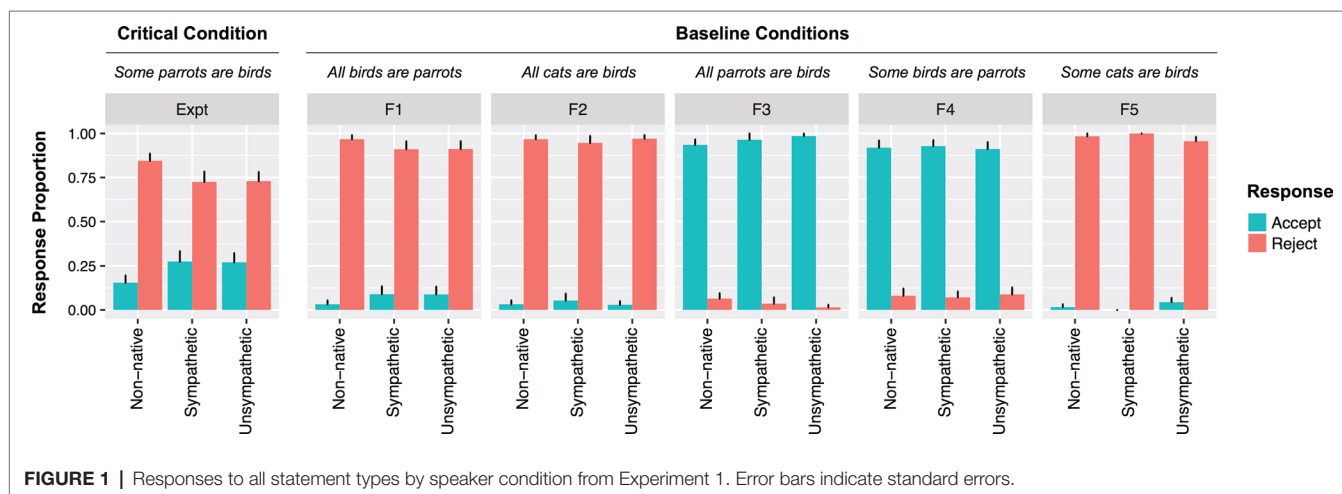
Participants were extremely accurate at providing Bobby's age (93.9%), and academic subject (98%). Performance was not significantly different across conditions ( $t_s < 1$ ). However, performance was less impressive for remembering the critical description of Bobby (79.8%). Only 50% of participants in the Unsympathetic condition selected "obnoxious" as the best description of Bobby, while the remaining 50% selected "kind." In contrast, 100 and 97% of participants in the Sympathetic and Non-native speaker conditions respectively selected the appropriate descriptor. It was important to establish that the effect of speaker type on UI judgments was driven by participants who paid attention to the description. To this end, analyses were repeated excluding individuals who provided the wrong description for Bobby. When only responders who were attending to the key manipulation were considered, the trends observed

for the whole data set strengthened. A model containing speaker type as a fixed effect provided a reliably better fit to the data than one without ( $\chi^2[2] = 6.6$ ,  $p < 0.05$ ). Pairwise comparisons indicated that Non-native Bobby was significantly more likely to be rejected than either Sympathetic Bobby ( $\beta = 2.83$ ,  $SE = 1.42$ ,  $z = 1.99$ ,  $p < 0.05$ ) or Unsympathetic Bobby ( $\beta = 2.31$ ,  $SE = 1.02$ ,  $z = 2.26$ ,  $p < 0.05$ ). There were no differences in rejections for Sympathetic and Unsympathetic Bobby ( $z < 1$ ). There were no effects of speaker condition for any of the filler sentence categories (all  $z_s < 1$ ).

### Likeability

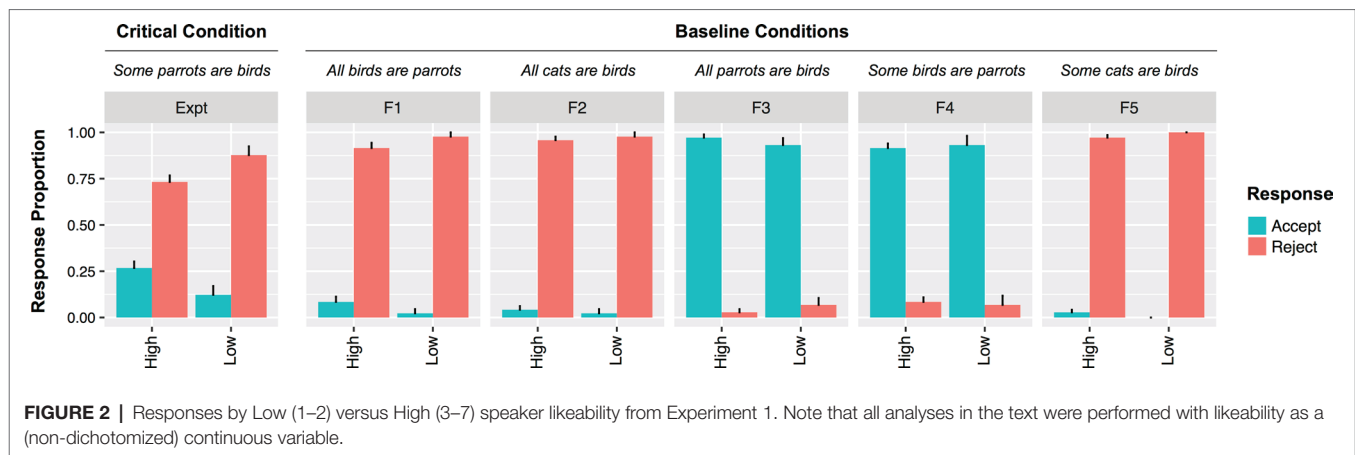
Surprisingly, participants in the Non-native speaker condition rated Bobby significantly less likeable than those in either the Unsympathetic ( $F(1,65) = 265$ ,  $p < 0.001$ ) or Sympathetic ( $F(1,57) = 249$ ,  $p < 0.001$ ) speaker conditions (Non-native:  $M = 2.24$ ,  $SE = 0.19$ ; Unsympathetic:  $M = 6.0$ ,  $SE = 0.16$ ; Sympathetic:  $M = 6.21$ ,  $SE = 0.17$ ). These differences persisted when only participants who correctly recalled the speaker description were included in the analysis ( $p_s < 0.001$ ; Non-native:  $M = 2.16$ ,  $SE = 0.17$ ; Unsympathetic:  $M = 5.83$ ,  $SE = 0.19$ ; Sympathetic:  $M = 6.21$ ,  $SE = 0.17$ ). It was unexpected to find that Non-native Bobby was perceived to be the least likeable and that Unsympathetic Bobby was rated nearly as likeable as Sympathetic Bobby. We discuss possible explanations for this below.

A mixed effects model with likeability as a predictor of rejections fared reliably better than one without ( $\chi^2[2] = 4.3$ ;  $p < 0.05$ ). The more likeable participants rated Bobby, the less likely they were to reject UI statements ( $\beta = 0.37$ ,  $SE = 0.18$ ,  $z = 2$ ,  $p < 0.05$ ). When only participants who accurately recalled the description of Bobby were included, the relationship between likeability and rejection rate was still present ( $\chi^2[2] = 6.2$ ;  $\beta = 0.53$ ,  $SE = 0.23$ ,  $z = 2.33$ ,  $p < 0.05$ ). In order to establish whether the effect of likeability was unique to UI statements, a model including sentence type (filler vs. UI), likeability, and their interaction was fit to the data. A model containing the interaction term fared reliably better than one without ( $\chi^2[1] = 4.7$ ;  $\beta = 0.3$ ,  $SE = 0.14$ ,  $z = 2.1$ ,



**FIGURE 1 |** Responses to all statement types by speaker condition from Experiment 1. Error bars indicate standard errors.





$p < 0.05$ ). This was because there were differential effects of likeability for different sentence types. Though Bobby's likeability reliably predicted rejections to UI statements, it did not predict rejections to any other sentence type ( $z_s < 0.1$ ). **Figure 2** depicts the different patterns for participants who rated Bobby highly unlikely (rated 1 or 2) versus those who rated Bobby as more likeable.

## Discussion

The results from Experiment 1 demonstrate that social context can modulate adult comprehenders' tolerance for pragmatic violations in a binary judgment task. Findings revealed that participants in the Non-native speaker condition rated Bobby significantly less likeable than did participants in either the Unsympathetic or Sympathetic speaker condition. Moreover, participants in the Non-native speaker condition were also significantly less likely to accept UI utterances than participants in the Unsympathetic or Sympathetic speaker conditions. Finally, when collapsing across speaker conditions, results showed that participants who strongly disliked Bobby were less likely to accept critical UI items than participants who gave Bobby a higher likeability rating.

The current design does not allow us to tease apart exactly which specific social factors underlie the greater rejection rate for UI utterances in the Non-native speaker condition. It could be that participants demanded a higher threshold for correctness for non-native Bobby because he was less likeable. It could also be that they focused more on the use of quantifiers because the instructions highlighted that Bobby's native language differs from English in this dimension. Because likeability was inversely correlated with the Non-native speaker condition, we cannot assess the independent contributions of these factors. Regardless, the results indicate that social aspects of the task influenced binary judgments for UI statements, but this was not observed for statements that were patently true or false. This pattern of results indicates that binary judgments of UI sentences are sensitive to social factors that are not directly relevant to the implicature calculation. We return to possible explanations for the surprising likeability results in the Non-native speaker condition in the General Discussion.

An unresolved question is how to interpret acceptances. Rejections of UI statements putatively indicate that an implicature was drawn, but it is not clear whether acceptances entail that no implicature was drawn. To investigate this question, we conducted an unplanned exploratory analysis of the text responses provided by participants to UI statements. If participants generated an implicature, then it would be reasonable to correct Bobby by providing a more optimally informative statement, thereby cancelling the implicature. For instance, for a UI sentence of the form "Some subsets are supersets" a participant might have provided the stronger alternative "All subsets are supersets." Responses were coded with respect to whether they contained the stronger alternative either explicitly or using an elided form (e.g., "All of them are"). Consistent with expectations, when participants rejected UI statements, they overwhelmingly provided the stronger alternative ( $M = 85.7\%$  of trials,  $SE = 3.1\%$ ). There were no reliable differences among speaker conditions (Sympathetic:  $M = 90.4\%$ ,  $SE = 4.3\%$ ; Unsympathetic:  $M = 82.4\%$ ,  $SE = 5.9\%$ ; Non-native:  $M = 85.2\%$ ,  $SE = 5.6\%$ ;  $\chi^2[2] = 0.9$ ,  $p = 0.9$ ). For acceptances, there were fewer strong alternatives provided but still a substantial number ( $M = 21.1\%$ ,  $SE = 6.1\%$ ). There was no reliable effect of speaker condition (Sympathetic:  $M = 0.8\%$ ,  $SE = 0.8\%$ ; Unsympathetic:  $M = 21.2\%$ ,  $SE = 9.0\%$ ; Non-native:  $M = 47.5\%$ ,  $SE = 16\%$ ;  $\chi^2[2] = 2.35$ ,  $p = 0.31$ ). It is possible that participants generated implicatures on these trials, though we cannot be certain. They may have provided the stronger statement for reasons unrelated to cancelling an unwarranted implicature. At a minimum, we can conclude that in these cases participants did not lack the cognitive resources to compute the strong alternative or to recognize its relevance to the weaker UI utterance. This indicates that participants can accept UI statements even in cases where they recognize that there are other more optimally informative utterances available.

## EXPERIMENT 2

The goal of Experiment 2 was to test whether speaker likeability continues to modulate pragmatic tolerance when participants are

given a ternary rather than binary judgment task. Based on the results of Katsos and Bishop (2011), we predicted that any differences in pragmatic tolerance due to the differences in perceived speaker likeability would be reduced or eliminated. This is because the intermediate response option provides participants with an explicit way to convey that UI statements are less than optimal but are better than patently false statements. Thus most participants on most trials should choose the intermediate response option.

## Materials and Methods

### Participants

A total of 102 English-speaking adults were recruited *via* Mechanical Turk. Eight failed to submit their data at the end of the survey, leaving data from 94 participants for analysis.

### Materials and Procedure

The stimuli, instructions, procedure, and exit survey were identical to those in Experiment 1, with the following exception: participants were given three response options instead of two (“That’s right,” “Not quite,” “That’s wrong”).

## Results and Discussion

### Exclusion Criteria

Three participants were removed for indicating that they were adults when they learned English (30 or older. All other participants were 6 or younger). Filler items were judged incorrect if participants responded “That’s Right” to a patently false item (F1, F2, F5) or if they failed to respond “That’s Right” to a patently true item (F3, F4). Four participants were excluded for accuracy below 70%. The remaining participants were relatively evenly distributed across the three speaker conditions ( $N_{Non-native} = 30$ ;  $N_{Sympathetic} = 32$ ;  $N_{Unsympathetic} = 25$ ) and had high mean accuracy ( $M = 94.9\%$ ;  $SE = 0.8\%$ ) (see **Figure 3**).

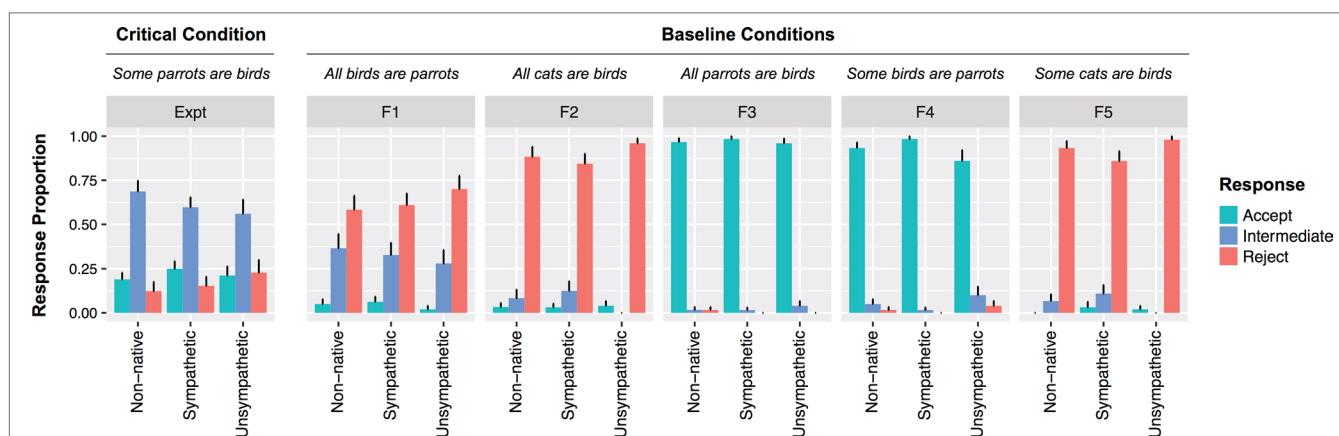
### Judgments of UI Statements

As predicted, the inclusion of an intermediate judgment option had clear effects on participant responses to UI items: in all

speaker conditions, participants had a strong preference for the intermediate response option (**Figure 3**). For no other sentence type was the intermediate response the preferred option. A mixed effect model including speaker type was not reliably better at explaining the rate of rejections than one without ( $\chi^2[2] = 0.4$ ,  $p = 0.82$ ). Speaker type was also not related to the rates of acceptances ( $\chi^2[2] = 1$ ,  $p = 0.61$ ).

To establish whether speaker type had a reliably smaller effect on UI judgments in the ternary task relative to the binary task, the rejection data from both Experiments 1 and 2 were combined and fit to a model crossing experiment and speaker type. A model without the interaction of these factors fared worse than a model including the interaction ( $\chi^2[2] = 4.9$ ,  $p = 0.087$ ). Thus speaker type had a stronger effect for the binary judgment task relative to the ternary judgment task on rejection rates. To investigate this interaction further, models were fit to subsets of the data consisting of each pair of the three speaker conditions. The difference between rejection rates in the Non-native and Unsympathetic speaker conditions was reliably different across experiments ( $\beta = 2.05$ ,  $SE = 0.95$ ,  $z = 2.17$ ,  $p < 0.05$ ). For the Non-native and Sympathetic conditions, this difference was marginally reliable across experiments ( $\beta = 1.5$ ,  $SE = 0.84$ ,  $z = 1.83$ ,  $p = 0.067$ ). In contrast, there was no interaction between speaker type and experiment in predicting rejection rates for the Sympathetic and Unsympathetic speaker conditions. ( $z = 0.59$ ).

For acceptances, a model containing the interaction of experiment and speaker type was numerically, but not reliably, better at explaining the data than one without ( $\chi^2[2] = 2.8$ ,  $p = 0.25$ ). When considering just the Non-native and Unsympathetic speaker conditions, there was a marginal interaction between speaker type and experiment ( $\beta = 0.92$ ,  $SE = 0.54$ ,  $z = 1.7$ ,  $p = 0.09$ ). This arose because speaker type had a stronger effect on acceptances for the binary judgment task than for the ternary judgment task. There was no interaction in acceptances between the Non-native and Sympathetic speaker conditions across experiments ( $z = 0.93$ ). Nor was there an interaction in acceptance rates for the



**FIGURE 3 |** Responses to all statement types by speaker condition from Experiment 2. Error bars indicate standard errors.

Unsympathetic and Sympathetic speaker conditions across experiments ( $z = 0.74$ ).

### Exit Survey

Accuracy patterns in Experiment 2 were similar to those from Experiment 1. Participants were extremely accurate at providing Bobby's age (91.2%), and academic subject (97.8%). Performance did not differ across conditions ( $t_s < 1$ ). Performance was again worse for remembering the critical description of Bobby (85.7%). Participants in the Sympathetic and Non-native conditions were highly accurate (94.1 and 100% respectively), but participants in the Unsympathetic Bobby condition were much less accurate (57.7%). When only data from participants who described Bobby correctly were included in the analysis of rejection rates, the pattern was similar to results from all participants. Speaker condition did not reliably predict rejections ( $\chi^2[2] = -0.74$ ,  $p = 1$ ), acceptances ( $\chi^2[2] = 0.57$ ,  $p = 0.75$ ), or intermediate responses ( $\chi^2[2] = 1.11$ ,  $p = 0.57$ ).

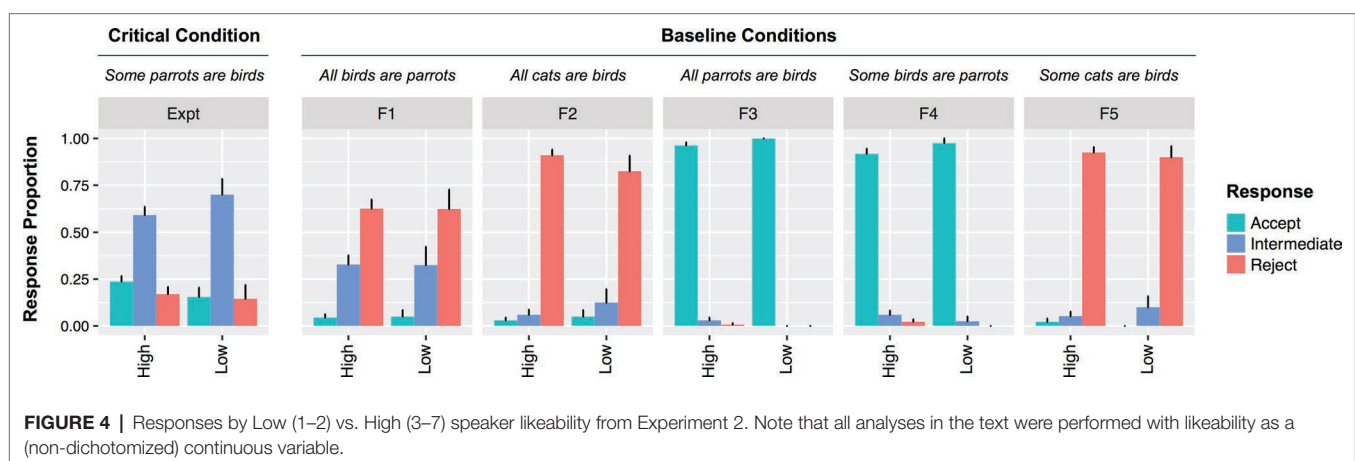
When including just those participants who correctly recalled the speaker description, the interactions across experiments in rejection and acceptance rates became more apparent. A model containing the interaction of experiment and speaker type on rejections performed marginally better than a model without this term ( $\chi^2[2] = 4.75$ ,  $p = 0.09$ ). There was an interaction between speaker type and experiment rejection rates reliable for the Non-native and Sympathetic speaker conditions ( $\beta = 1.83$ ,  $SE = 0.89$ ,  $z = 2.07$ ,  $p < 0.05$ ) and marginal for the Non-native and Unsympathetic conditions ( $\beta = 2.16$ ,  $SE = 1.26$ ,  $z = 1.7$ ,  $p = 0.08$ ). There was no such interaction for the Sympathetic and Unsympathetic speaker conditions ( $z = 0.01$ ). Thus, the effect of speaker type on rejection rates was reliably larger for Experiment 1 with binary response options compared to Experiment 2 with ternary response options.

Parallel analyses were performed on acceptances using only data from participants who described Bobby correctly. A model containing the interaction between speaker type and experiment was marginally better than a model that did not contain this term ( $\chi^2[2] = 4.9$ ,  $p = 0.08$ ). There was an interaction between speaker type and experiment when considering just the Non-native

and Unsympathetic speaker conditions ( $\beta = 1.32$ ,  $SE = 0.63$ ,  $z = 2.11$ ,  $p < 0.05$ ). There was a trend toward an interaction for the Non-native and Sympathetic speaker conditions across experiments ( $\beta = 0.72$ ,  $SE = 0.53$ ,  $z = 1.35$ ,  $p = 0.18$ ). There was no interaction across experiments for the Unsympathetic and Sympathetic speaker conditions ( $z = 0.81$ ). Thus, just as with rejections, the effect of speaker type on acceptances was larger with binary response options than with ternary response options.

Experiment 2 also replicated the surprising speaker-likeability finding from Experiment 1: participants in the Non-native speaker condition rated Bobby significantly less likeable than did participants in the Unsympathetic or Sympathetic conditions (Non-native:  $M = 2.10$ ,  $SE = 0.23$ ; Unsympathetic:  $M = 5.92$ ,  $SE = 0.23$ ; Sympathetic:  $M = 5.76$ ,  $SE = 0.20$ ). Exit survey results also revealed that only 58% of participants in the Unsympathetic condition selected "obnoxious" as the best description of Bobby, while the remaining 42% selected "kind." In contrast, 95 and 100% of participants in the Sympathetic and Non-native speaker conditions respectively selected the appropriate descriptor (see General Discussion for possible explanations for this finding).

However, in contrast to the results found with the binary judgment task in Experiment 1, likeability had no effect on rejections ( $\chi^2[2] = 1.08$ ,  $p = 0.3$ ). There was also no relationship between likeability and acceptances ( $\chi^2[2] = 0.17$ ,  $p = 0.68$ ). When only participants who accurately recalled the description of Bobby were included, these patterns were unchanged (rejections:  $\chi^2[2] = 0.58$ ,  $p = 0.44$ ; acceptances:  $\chi^2[2] = 0.9$ ,  $p = 0.34$ ). To investigate whether the effect of likeability was different for UI and other sentence types (filler vs. UI) an interactional analysis was performed. There was a main effect of sentence type whereby fillers were rejected more often than UI statements ( $\beta = 0.99$ ,  $SE = 0.24$ ,  $z = 4.06$ ,  $p < 0.001$ ). There was no effect of likeability ( $z = 0.88$ ,  $p = 0.38$ ). Importantly, there was no interaction between likeability and sentence type in predicting rejections ( $z = 0.91$ ,  $p = 0.36$ ) nor acceptances ( $z = 0.92$ ,  $p = 0.36$ ). Thus, unlike Experiment 1 where responses to UI items were specifically affected by likeability for binary judgments, there was no difference in the (null) effects of likeability for ternary judgments (See **Figure 4**).



## Discussion

The results from Experiment 2 indicate that social context did not modulate participants' tolerance for pragmatic violations when participants were given an intermediate option in a ternary judgment task. In contrast to Experiment 1, the positive correlation between speaker likeability and acceptance of critical items is eliminated when participants have an intermediate response option. This indicates that the locus of social context effects in Experiment 1 was in selecting a response (i.e., determining what the threshold for rejection is), rather than being related to computing the inference.

Experiment 2 also addresses a potential concern with the speaker manipulation in Experiment 1. Though the task itself is unchanged across speaker conditions, it is still logically possible that the manipulation of speaker description somehow affected implicature calculation indirectly. For instance, if the speaker descriptions fundamentally changed the communicative goals of the task in disparate ways. If so, then it is conceivable that the results from Experiment 1 reflect differences in implicature rates across conditions rather than differences in response selection. The results from Experiment 2 rebut this interpretation. The rates of implicatures in Experiment 2 (inferred from either rejections or acceptances) were not affected by speaker condition nor likeability as they were in Experiment 1. Since the only difference between Experiments 1 and 2 is the response options available to participants, this difference strongly indicates that implicature processes were unaffected.

Similar exploratory analyses to those in Experiment 1 were performed on participants' text feedback. The rate of strong alternative statements provided for trials in which the participant did not accept the UI statement (both intermediate responses and rejections) was similar to Experiment 1 ( $M = 85.0\%$ ,  $SE = 3.1\%$ ). There were no reliable differences among speaker conditions (Sympathetic:  $M = 82.8\%$ ,  $SE = 5.5\%$ ; Unsympathetic:  $M = 80.5\%$ ,  $SE = 7.6\%$ ; Non-native:  $M = 91.1\%$ ,  $SE = 4.3\%$ ;  $\chi^2[2] = 2.3$ ,  $p = 0.32$ ). For trials on which the participant accepted the UI statement, the rate of feedback containing strong alternative statements ( $M = 3.8\%$ ,  $SE = 2.8\%$ ) was numerically lower than that for acceptances in Experiment 1 ( $M = 21.1\%$ ). There were no reliable differences for different speaker conditions (Sympathetic:  $M = 3.4\%$ ,  $SE = 2.6\%$ ; Unsympathetic:  $M = 0\%$ , Non-native:  $M = 7.1\%$ ,  $SE = 7.1\%$ ; model unidentifiable). One possible explanation for the reduction from Experiment 1 to 2 is that participants who generated an implicature and who wanted to provide corrective feedback for Bobby without rejecting his statement could avail themselves of the intermediate response in Experiment 2. In Experiment 1, they would have had to accept the statement.

## GENERAL DISCUSSION

We set out to test whether manipulating social context can modulate adult acceptability judgments of UI utterances. We manipulated the perceived likeability of the speaker by providing participants with a specific social context and a

detailed description of their interlocutor against which they were asked to make their judgments. In Experiment 1, participants rejected UI utterances from the Non-native speaker more frequently than from either the Unsympathetic or Sympathetic speakers when given only a binary response option. At the same time, participants disliked the Non-native speaker relative to the other speakers. This pattern of effects indicates that social context can influence pragmatic judgments when participants are forced to choose between rejection and acceptance. Note that the cognitive task was identical in all conditions and participants were randomly assigned to speaker conditions. Thus, it is unlikely that participants in the Non-native speaker condition had more cognitive resources than those in the other conditions. Moreover, participants were equally accurate on filler items across conditions. Social factors only influenced judgments on the UI items, where the pragmatic and literal meanings diverged.

In Experiment 2, the same materials were employed, but participants had three response options and could therefore give more graded feedback. In this case, the acceptance rate was not affected by our social context manipulation. Thus, the positive correlation between speaker likeability and acceptance of critical items is eliminated when participants have an intermediate response option. In this case, participants did not have to deliberate over where to place the boundary of acceptability—the intermediate response option provided participants with an explicit way to signal that UI statements are less than optimal but are better than patently false statements.

The relative likeability of the speakers is somewhat surprising. We had predicted that the Unsympathetic speaker condition would engender the least amount of charity from participants. However, both experiments found that likeability ratings were lowest in the Non-native speaker condition. One possible explanation for this unexpected result is that participants were displaying ethnocentric tendencies (were prejudiced against non-native speakers and/or immigrants). An alternative explanation may be related to the high rate of patently false statements (30% of the items) in the experimental design. Participants may have been able to rationalize such “poor performance” from both the Sympathetic and Unsympathetic speakers: Sympathetic Bobby was described as having a developmental disorder and Unsympathetic Bobby was described as “very difficult.” Non-native Bobby, on the other hand, was described as “bright.” This may have led participants in the Non-native speaker condition to become more irritated with his poor performance. A related finding was also surprising. Likeability ratings for Unsympathetic Bobby were not reliably different than ratings for Sympathetic Bobby (even among participants who correctly remembered unsympathetic Bobby being labeled “obnoxious” by his teachers). One possible explanation for this finding is that Unsympathetic Bobby may have garnered compassion rather than aversion; participants may have attributed his poor behavior to external causes (e.g., poor parenting) rather than to the child himself. Importantly, these issues are tangential to the critical finding. Because these manipulations should not directly influence the actual computation of a scalar inference, any difference in responses between binary



and ternary judgments is better explained by differences in response selection processes than by different rates of implicature computation. Therefore, we take the current findings as clear evidence that social factors unrelated to generating the implicature itself can modulate adult comprehenders' tolerance for pragmatic violations in a binary judgment task.

An open question is how to interpret acceptances in the present studies. One possibility is that participants in Experiment 1 recognized that Bobby's utterance was not optimally informative, but decided that this violation was not sufficient to assign it the same rating as patently false statements. If so, many of these individuals would have likely preferred an intermediate option. On this view, we should have seen a reduction in the rate of acceptances in Experiment 2. There were indeed small numerical reductions for the Unsympathetic speaker (30.6 vs. 22.9%) and for the Sympathetic speaker (27.5 vs. 24%) who were both deemed likeable, but the rate of acceptances increased slightly for the Non-native speakers (13.3 vs. 19%) who were deemed unlikeable. However, the overall rate of acceptances did not fall dramatically when provided with an intermediate option. There are at least two plausible accounts for this. One is that there were, by chance, fewer genuine implicatures drawn in Experiment 2 than Experiment 1. On this view, non-acceptances in the ternary task might more accurately reflect implicature generation than rejections in the binary task. If so, then the small reduction in acceptances from Experiment 1 to Experiment 2 would have been larger if the two groups of participants generated implicatures at the same rate. A second possibility is that the intermediate responses were still too harsh for some individuals who generated implicatures. As a result, they elected to accept UI statements even with an intermediate option available. In this case, an additional intermediate option (e.g., "mostly right") might have revealed still more individuals who are sensitive to underinformativity (see Jasbi, Waldon, and Degen, submitted). Either of these possibilities, either singly or in combination, could have led to the pattern observed.

## CONCLUSION

The present studies demonstrate that pragmatic tolerance can contribute to the variability found in adult responses to UI utterances in binary judgment tasks. Many studies take the non-acceptance of a UI statement to be evidence that the comprehender has computed a scalar inference and the acceptance of a UI statement as evidence that they have not. The results above call these assumptions into question. We have shown that adult comprehenders, like children (Katsos and Bishop, 2011), will accept a UI statement even in the same situations where they recognize it as non-optimal. Unlike patently false or true statements, UIs are neither completely wrong nor completely correct. When forced to select between two inapt options in a binary choice task, social factors can tip the balance so that participants choose to reject UI statements more often for certain speakers. In contrast, a ternary judgment task allows

participants to clearly indicate that UI utterances are intermediately acceptable between patently true and false statements. With a more apt intermediate response option, participants are not as affected by social aspects of the speaker. More work is needed to establish what aspects of the social context are most influential for binary judgments, and to determine why children are less likely to reject pragmatically infelicitous statements than adults.

What we do have evidence for is that binary judgments are affected by selection processes, which are unrelated to implicature computation, in a way that graded judgments are not. Binary judgments are perhaps the most widespread method for investigating implicature processing and development. The present work thus demonstrates that results garnered from binary judgment tasks must be interpreted with caution.

## DATA AVAILABILITY

The data sets analyzed for this study can be found at <https://drive.google.com/drive/folders/15qSxN7dXPP7GKKJA8Ks9dr3nInJf4OL6?usp=sharing>

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations set forth in the Belmont Report, federal regulations (e.g., DHHS regulations 45 CFR Part 46), and Swarthmore College policies. The protocol was approved by the Swarthmore College Internal Review Board. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Internal Review Board at Swarthmore College. Participants were provided with and completed an electronic consent form before beginning the study.

## AUTHOR CONTRIBUTIONS

DG, LS, and MK all made contributions to the conception and design, study implementation, data acquisition an analysis, interpretation, and write up of this work.

## FUNDING

Funding for the present work was provided by research grants from Swarthmore College to author DG.

## ACKNOWLEDGMENTS

The authors wish to acknowledge Howard Lam and Raul Anchirai for assistance in developing the surveys to administer these studies. We are also grateful to the audience at the 2016 CUNY Conference on human sentence processing for helpful feedback on an earlier presentation of this work.

## REFERENCES

- Barner, D., Brooks, N., and Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition* 118, 84–93.
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001
- Bates, D., Kliegl, R., Vasishth, S., and Baayen, R. H. (2014a). Parsimonious mixed models. Available from: arXiv: 1506.04967.
- Bates, D., Maechler, M., Bolker, B., Walker, S. (2014b). Data from: lme4: linear mixed-effects models using Eigen and S4. R Package. Version 1.1–7. <http://CRAN.R-project.org/package=lme4>
- Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006
- Bott, L., Bailey, T. M., and Grodner, D. (2012). Distinguishing speed from accuracy in scalar inferences. *J. Mem. Lang.* 66, 123–142. doi: 10.1016/j.jml.2011.09.005
- De Neys, W., and Schaeken, W. (2007). When people are more logical under cognitive load. *Exp. Psychol.* 54, 128–133. doi: 10.1027/1618-3169.54.2.128
- Dieussaert, K., Verkerk, S., Gillard, E., and Schaeken, W. (2011). Some effort for some: further evidence that scalar inferences are effortful. *Q. J. Exp. Psychol.* 64, 2352–2367. doi: 10.1080/17470218.2011.588799
- Feeney, A., Scafton, S., Duckworth, A., and Handley, S. J. (2004). The story of some: everyday pragmatic inference by children and adults. *Can. J. Exp. Psychol.* 58, 121–132. doi: 10.1037/h0085792
- Foppolo, F., Guasti, M. T., and Chierchia, G. (2012). Scalar implicatures in child language: give children a chance. *Lang. Learn. Dev.* 8, 365–394. doi: 10.1080/15475441.2011.626386
- Grice, H. P. (1975). "Logic and conversation" in *Syntax and semantics vol. 3: Speech acts*. eds. P. Cole and J. L. Morgan (New York: Academic Press), 41–58.
- Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis. Los Angeles: UCLA.
- Horowitz, A. C., Schneider, R. M., and Frank, M. C. (2018). The trouble with quantifiers: exploring children's deficits in scalar implicature. *Child development* 89, e572–e593.
- Katsos, N., and Bishop, D. V. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and inference. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015
- Katsos, N., and Cummins, C. (2012). Scalar inference: theory, processing and acquisition. *Nouveaux cahiers de linguistique française* 30, 39–52.
- Marty, P. P., and Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Front. psychol.* 4, 403.
- Noveck, I. A., and Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends Cogn. Sci.* 12, 425–431. doi: 10.1016/j.tics.2008.07.009
- Pouscoulous, N., Noveck, I. A., Politzer, G., and Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language acquisition* 14, 347–375.
- Sonnentag, S. (1998). Identifying high performers: do peer nominations suffer from a likeability bias? *Eur. J. Work Organ. Psy.* 7, 501–515. doi: 10.1080/135943298398547
- Tiel, B., and Schaeken, W. (2017). Processing conversational implicatures: alternatives and counterfactual reasoning. *Cogn. Sci.* 41, 1119–1154. doi: 10.1111/cogs.12362
- Veenstra, A., Hollebrandse, B., and Katsos, N. (2018). Why some children accept under-informative utterances. *Pragmat. Cogn.* 24, 297–313. doi: 10.1075/pc.00003.vee

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Sikos, Kim and Grodner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Believing What You're Told: Politeness and Scalar Inferences

Diana Mazzarella<sup>1\*</sup>, Emmanuel Trouche<sup>2</sup>, Hugo Mercier<sup>3</sup> and Ira Noveck<sup>4</sup>

<sup>1</sup> Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), Berlin, Germany, <sup>2</sup> Cognition and Development Lab, Department of Psychology, Yale University, New Haven, CT, United States, <sup>3</sup> Institut Jean Nicod, Centre National de la Recherche Scientifique, Paris, France, <sup>4</sup> Institut des Sciences Cognitives Marc Jeannerod, Centre National de la Recherche Scientifique, Bron, France

The experimental pragmatics literature has extensively investigated the ways in which distinct contextual factors affect the computation of scalar inferences, whose most studied example is the one that allows “Some X-ed” to mean *Not all X-ed*. Recent studies from Bonnefon et al. (2009, 2011) investigate the effect of politeness on the interpretation of scalar utterances. They argue that when the scalar utterance is face-threatening (“Some people hated your speech”) (i) the scalar inference is less likely to be derived, and (ii) the semantic interpretation of “some” (*at least some*) is arrived at slowly and effortfully. This paper re-evaluates the role of politeness in the computation of scalar inferences by drawing on the distinction between “comprehension” and “epistemic assessment” of communicated information. In two experiments, we test the hypothesis that, in these face-threatening contexts, scalar inferences are largely *derived* but are less likely to be *accepted as true*. In line with our predictions, we find that slowdowns in the face-threatening condition are attributable to longer reaction times at the (latter) epistemic assessment stage, but not at the comprehension stage.

**Keywords:** experimental pragmatics, scalar inference, some, face, politeness, epistemic vigilance

## OPEN ACCESS

### Edited by:

Penka Stateva,  
University of Nova Gorica, Slovenia

### Reviewed by:

Francesca Foppolo,  
Università degli studi di Milano  
Bicocca, Italy  
Francesca Marina Bosco,  
Università degli Studi di Torino, Italy

### \*Correspondence:

Diana Mazzarella  
mazzarella@leibniz-zas.de

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 30 January 2018

**Accepted:** 18 May 2018

**Published:** 13 June 2018

### Citation:

Mazzarella D, Trouche E, Mercier H  
and Noveck I (2018) Believing What  
You're Told: Politeness and Scalar  
Inferences. *Front. Psychol.* 9:908.  
doi: 10.3389/fpsyg.2018.00908

## INTRODUCTION

Scalar inferences are classically described as pragmatic enrichments (made by a listener) when a speaker uses a weaker term (e.g., “some”) to communicate a narrowed, more informative, meaning that excludes a stronger term (e.g., “all”). Consider the following example:

- (1) a. Some students failed the exam.
- b. Not all of the students failed the exam.

While the semantic (encoded) meaning of “some” is compatible with *all*, (1a) is frequently interpreted as communicating (1b). The inference from (1a) to (1b) is called *scalar* because, based on earlier accounts (e.g., Horn, 1984), the enrichment exploits an implicit scale of informativeness that ranges from *some* to *all*. The explanation goes as follows: the addressee assumes that the speaker would have said “all” if she thought the statement with “all” was true; the choice of “some” thus implies either that she does not know whether *all* is the case or that she believes *all* is not the case. If it is reasonable to assume that the speaker knows whether the stronger alternative holds or not, the use of a relatively weak expression is taken to indicate that the speaker believes the stronger alternative to be false.

Scalar inferences have become the drosophila of experimental pragmatics as they have the means to provide a clear test case to investigate the interaction of semantics with contextual information in sentence processing. The way the scalar enrichment is carried out has been vigorously debated in this experimental literature (see Chemla and Singh (2014a,b) and Noveck (2018) for recent

reviews). On one side are those who argue that scalar inferences occur routinely and independently of context, i.e., by default (Levinson, 2000). On the other side are those who defend context-sensitivity and argue that such enrichments occur as a function of particular features of a task or conversational context and that these particular instances need not arise with the mere expression of a weak scalar term (see Sperber and Wilson, 1986/1995; Carston, 1998, but also Chierchia (2013) for a discussion of context-sensitivity within a grammatical framework). Much evidence supports the latter characterisation. As has been reported and summarized elsewhere (Noveck and Sperber, 2007; Noveck and Reboul, 2008; Noveck and Spotorno, 2014), linguistically encoded readings are often sufficient for making on-line interpretations with utterances containing weak scalar terms.

Recent research has looked into the time course of the derivation of scalar inferences in real-time language comprehension. Many studies report that enriched readings (e.g., *some but not all*) are linked with the availability or application of supplementary processing (e.g., see Bott and Noveck, 2004; Breheny et al., 2006; De Neys and Schaeken, 2007; Huang and Snedeker, 2009; Bott et al., 2012). However, there is disagreement in the literature concerning the source of these slowdowns (see Foppolo and Marelli, 2017). Moreover, it appears that the speed at which a scalar inference is computed may depend on features of the context of utterance, such as the naturalness and availability of alternatives (Breheny et al., 2013; Degen and Tanenhaus, 2015).

The debate concerning the kinds of contextual factors that affects the computation of scalar inferences has been recently enriched by the work of Bonnefon and colleagues. This work investigates the effect of politeness on the derivation of scalar inferences and it presents a case for the following two claims. First, politeness is likely to block the computation of the scalar inference. Second, the unenriched interpretation of the scalar utterance in politeness contexts requires supplementary processing costs (Bonnefon et al., 2009, 2011; Feeney and Bonnefon, 2012). The aim of this paper is to address these two claims by (i) assessing the robustness of Bonnefon and colleagues' results, and by (ii) providing a finer-grained analysis of the processing of scalar utterances in these experiments. This analysis will be based on the distinction between "comprehension" and "epistemic assessment" (see, Mazzarella, 2015), and will ultimately describe Bonnefon and colleagues' results as linked to the process of "epistemic assessment" (with no direct bearing on the comprehension of the scalar utterance).

## FACE-THREATENING CONTEXTS AND SCALAR INFERENCE

In a series of studies, Bonnefon and colleagues investigated the derivation of scalar inferences in face-threatening contexts. These are situations in which the public image or positive identity of the addressee is threatened (Brown and Levinson, 1987). For instance, consider the following example:

- (2) a. Some people hated your speech.  
b. Not all the people hated your speech.

The scalar utterance (2a) carries a threat toward the public image of the addressee (represented by lack of public approval and support). Because of this, Bonnefon and colleagues argue, the addressee would be less likely to derive the scalar inference (2b) and would take the use of "some" as a polite device adopted by the speaker to sugar-coat the information conveyed. Specifically, they claim that "face-threatening contexts make the narrowed interpretation of "some" less appropriate" and report that, in line with this, their "result suggests that people's tendency to draw the scalar inference from "some X-ed" to "not all X-ed" decreases when X threatens the face of the listener" (Bonnefon et al., 2009, p. 250–251).

The empirical support for these claims comes from a series of similar studies where the authors investigate the interpretation of face-threatening and face-boosting scalar utterances. While the former carry a threat toward the "face" of the addressee, the latter reinforce his positive identity. They present participants with short vignettes in which they are asked to imagine to have carried out a publicly observed act (such as giving a speech in front of a small group of people). Critically, participants are provided with feedback in the form of a scalar utterance. The feedback is negative in the face-threat condition ("Some people hated your speech") and positive in the face-boost one ("Some people loved your speech"). This is followed up with a meta-linguistic question about the feedback. The entire task can be broken down into three parts: the task's background information (3), the scalar utterance (4), and what we call the *semantic compatibility question* (5), which prompts the critical "Yes" or "No" responses:

- (3) Imagine you gave a speech at a small political rally. You are discussing your speech with Denise, who was in the audience. There were 6 other people in the audience. You are considering whether to give this same speech to another audience.
- (4) Hearing this, Denise tells you that "Some people hated your speech."
- (5) Given what Denise told you, do you think that it is possible that everybody hated your speech?

Importantly, the semantic compatibility question is the measure used to determine whether or not a scalar inference has been derived. A "No" answer is taken to suggest that participants have derived the scalar inference (hence the perceived incompatibility between what Denise said and the state of affairs in which *everybody* hated/loved the speech). On the contrary, a "Yes" answer is taken to reveal that participants have adopted the semantic interpretation of "some," *at least some and possibly all*, which is consistent with the possibility that *everybody* hated/loved the speech.

Bonnefon et al. (2009, 2011) have consistently found that the percentage of participants answering "Yes" is significantly higher in the face-threat condition than it is in the face-boost condition (see **Table 1**). That is, participants are more likely to think that it is possible that everybody *hated* the speech—after being told that "some" did—than to think that everybody *loved* the speech when similarly told that "some" did.

Bonnefon et al. (2011) went further by measuring response times, which covered the scalar utterance (4) and the time to answer the semantic compatibility question (5). They reported



**TABLE 1 |** Percentage of “Yes” responses to the semantic compatibility question.

	Face-threat (%)	Face-boost (%)
Bonnefon et al. (2009)	42	17
Bonnefon et al. (2011)	55	27

For each cell, the complement corresponds to “No” responses.

that response times were longer when participants answered “Yes” as opposed to “No,” but only in the face-threat condition (the interaction was significant at the 0.5 level using one-tailed tests). The authors concluded that in face-threatening contexts the semantic interpretation of “some” is associated with extra-processing effort. In light of this, Bonnefon et al. (2011, p. 3393) argue that politeness “appears to add a layer of complexity to the usual processes involved in the interpretation of “some.”” This layer of complexity would make the semantic or “broad” interpretation of “some” the more effortful one.

To summarize, Bonnefon and colleagues put forward the following two claims: (i) scalar inferences are less likely to be derived in face-threatening contexts; (ii) the semantic interpretation of face-threatening scalar utterances involves an extra cognitive cost. Taken together, these two claims are presented as an interesting challenge to current cognitive models of scalar inference. Specifically, they target the assumption that semantic interpretations are always less effortful to arrive at than pragmatic interpretations: “Showing that face-threatening contexts encourage broad interpretations whilst making them harder would require to revisit this basic assumption.” (Bonnefon et al., 2011, p. 3390).

In what follows, we take the following three steps. First, we analyse the structure of Bonnefon et al.’s task and describe a confound that undermines their main claim. Once this confound is exposed, it becomes arguable that the two interpretations linked to the existential quantifier “some” are not themselves the source of the exceptional results. Second, we consider a relatively new line of research that distinguishes between “comprehension” and “epistemic assessment” of the communicated content (Sperber et al., 2010) and discuss its implications for the role of politeness on the processing of scalar utterances. Finally, we introduce our experiments which aim at disentangling the interpretation of the scalar utterance from the participant’s evaluative task.

## COMPREHENSION AND EPISTEMIC ASSESSMENT: A METHODOLOGICAL CONFOUND AND A THEORETICAL CONFLATION

The starting point of our critical discussion involves a closer analysis of Bonnefon and colleagues’ original paradigm, and especially the (2011) follow-up paper that further presented reaction times. We focus on (i) the way reaction times were collected, and, (ii) the nature of the test question (i.e., the semantic compatibility question).

To start, it is crucial to notice that, in Bonnefon et al.’s (2011) study, the scalar utterance and the semantic compatibility question, (4) and (5) above, were displayed together on the screen as a block of text. The critical reaction time thus measured a (long) interval in which participants read the block *and* answered the question: the response time measure began with the advancing of the visual display of the text to “Hearing this, Denise tells [...]” and ended when the answer key was pressed. Despite the length of the block and the fact that, arguably, two tasks are involved—reading the utterance and providing a Yes/No answer to the semantic compatibility question—the interaction effect is described as depending on the interpretation of the scalar utterance alone. Bonnefon et al. conclude that the semantic interpretation of “some” is derived more slowly than the pragmatic one, and thus that politeness increases the processing effort required to arrive at the semantic interpretation.

Furthermore, Bonnefon et al. see a direct link between responses to the semantic compatibility question (e.g., [...] *do you think that it is possible that everybody hated your speech?*) and the interpretation of the scalar utterance, with “Yes” answers corresponding to a semantic interpretation of “some,” and “No” answers to a pragmatic one. However, as discussed by Mazzarella (2015), the semantic compatibility question relates to the participant’s *belief* that a certain state of affairs is likely to hold. It is *not* inherently a question about the speaker’s informative intention (about the *interpretation* of the scalar utterance). In Bonnefon et al.’s task, the participant is asked to evaluate the likelihood concerning a state of affairs, e.g., that everybody hated the speech, when told earlier that “some” did. Note that the answer to the semantic compatibility question is not entirely dependent on the *interpretation* of the scalar utterance: participants could interpret it as communicating the scalar inference *Not all the people hated your speech*, and yet end up believing that it is possible that everyone indeed hated the speech. This is because the comprehension of a speaker’s intended meaning and one’s acceptance of it do not always go hand in hand. That is, an audience can *understand* a speaker’s communicated content, e.g., that *not everyone hated the speech*, without *believing* it. Whether or not a listener accepts the incoming information depends on the plausibility of this information as well as on the trust the listener grants the communicator (Sperber et al., 2010).

The above discussion points toward the distinction between “comprehension” and “epistemic assessment”: while the former process relates to the interpretation of a piece of communicated information, the latter determines whether we believe it. This distinction has a long tradition in philosophy of language at least since the work of Austin (1962/98). Austin (1962/98) distinguishes between “securing the uptake” of an utterance, that is, comprehending its meaning and illocutionary force, from the utterance’s perlocutionary effects. The latter comprise a range of cognitive and behavioral effects, “effects upon the feelings, thoughts, or actions of the audience” (Austin, 1962/98, p. 11), which go beyond uptake. These include the beliefs the audience forms with respect to what is communicated. Crucially, the latter may differ from the beliefs the communicator intended to induce in the audience.

This distinction between comprehension and epistemic assessment is overlooked in Bonnefon and colleagues' work, as it is in the experimental pragmatics literature more generally. Instead of evaluating a participant's answer as a measure of accepting/rejecting the speaker's implied meaning, Bonnefon and colleagues directly map "Yes" answers from the semantic compatibility question to a semantic *interpretation* of "some" and "No" answers from the same question to a pragmatic *interpretation* of "some." Furthermore, they explain the reaction time data as linked to the processes involved in the interpretation of the scalar utterance ("comprehension"), with no role assigned to epistemic assessment.

In light of the cognitive distinction between comprehension and epistemic assessment, Mazzarella (2015) proposes an alternative explanation of the data. She suggests that face-threatening contexts reduce the perceived honesty of the speaker and, as a result, decrease the likelihood of accepting the scalar inference as true. Mazzarella's (2015) hypothesis accounts for the higher percentage of "Yes" answers in the face-threat condition by suggesting that these answers reflect a rejection of the scalar inference. This possibility is not only theoretically plausible, but also empirically grounded. From a theoretical perspective, it is plausible to assume that the attribution of politeness concerns to the speaker might negatively affect the likelihood of her sharing some face-threatening information. If the speaker cared about saving the face of the addressee, she would try to minimize his face loss. In doing this, she could decide to withhold some relevant information, or even lie to the addressee<sup>1</sup>. If the addressee believes that this is the case, he might consider a polite speaker as less reliable from an epistemic point of view. Crucially, this hypothesis receives some support from two studies ran by Bonnefon et al. (2009) themselves. In their Experiment 2, they showed that, when presented with situations in which the speaker is described as knowing that *all* (e.g., that all the people hated/loved the speech), participants judge it as more likely that the speaker would use the word "some" in face-threatening contexts ("Some people hated your speech") than in face-boosting ones ("Some people loved your speech"). Furthermore, in a second rating study, they asked participants to rate how "accurate," "considerate," "honest," and "nice" it was of the speaker to use the word "some" in a context in which the speaker knew that *some but not all* and in a context in which she knew that *all*. Crucially, in the *all* condition, the use of "some" was rated as inaccurate and dishonest in both face-threatening and face-boosting contexts, but nice and considerate only in face-threatening contexts. If Mazzarella's (2015) account is on the right track, longer reaction times associated with "Yes" answers would be better explained as linked to the process of epistemically *evaluating* the scalar inference, which is triggered by the presence of the test question in (5).

<sup>1</sup>Mazzarella (2015) suggests that the speaker would go as far in her face-saving work as "plausible deniability" allows (Lee and Pinker, 2010). That is, the speaker might not want to commit to a blatantly false statement, but rather communicate a falsity only implicitly (as in the case of the scalar inference *Not all people hated your speech* in a situation in which the speaker knows that everybody hated the speech).

## THE CURRENT EXPERIMENTS

The aim of the current experiments is two-fold. First, we aim to confirm the robustness of Bonnefon and colleagues' categorical results, irrespective of the confound described in the previous section. Second, we will collect reaction time measures while addressing this confound in order to better determine how participants interpret the scalar utterance and the semantic compatibility question. That is, we adopt the same experimental paradigm but separate the presentation of the scalar utterance, (4), from that of the semantic compatibility question, (5). This way we can distinguish between what we call "the comprehension stage" and the "the epistemic assessment stage" and measure reaction times from each part (RT<sub>UTTERANCE</sub> and RT<sub>QUESTION</sub>) and combine them. By separating the scalar utterance from the test question, we can better isolate the source of reaction time effects. Study 1 relies on Bonnefon et al.'s (2011) original materials, while Study 2 introduces some motivated changes to the materials in order to increase the likelihood that the scalar inference would be effectively derived.

### Study 1

In Study 1, we adopted Bonnefon et al.'s task but made two modifications. First, to better study the processes of deriving the scalar inference and of epistemically evaluating its factual plausibility, we separated the presentation of the scalar utterance and the semantic compatibility question. Second, we introduced a new question, which we refer to as the *conversational implicature question*: "Given what Denise tells you, do you think that she means that you should give the speech again to another group?" This question was presented after the semantic compatibility question, in order to preserve the integrity of Bonnefon et al.'s original task. See Appendix A for a thorough comparison between our materials and Bonnefon et al.'s (2011).

As a reminder, the first aim of Study 1 is to confirm Bonnefon et al.'s (2011) results. This would mean more "Yes" responses to the semantic compatibility question in the face-threat condition, as well as longer reaction times overall (for RT<sub>UTTERANCE</sub> + RT<sub>QUESTION</sub>) for those responses when compared to the face-boost condition. Once this is accomplished, we will more carefully inspect reaction times to each part (RT<sub>UTTERANCE</sub> and RT<sub>QUESTION</sub>).

As Bonnefon et al. (2011) maintain that slowdowns in the face-threat condition are linked directly to the way in which "some" is interpreted, they predict that the source of the expected interaction would be an observable difference between the face-threat and the face-boost conditions at the presentation of the scalar utterance. Specifically, they predict longer reaction time at the scalar utterance phase (i.e., RT<sub>UTTERANCE</sub>) for "Yes" answers than for "No" answers in the face-threat condition (with the latter being comparable with "Yes" and "No" answers in the face-boost condition). Following Mazzarella (2015), we predict that slowdowns will be observed for "Yes" answers uniquely at the presentation of the semantic compatibility question (for RT<sub>QUESTION</sub>), and not at the presentation of the scalar utterance

(for  $RT_{\text{UTTERANCE}}$ ). Specifically, we predict longer reaction time at the question phase for “Yes” answers than “No” answers in the face-threat condition.

The aim of introducing the *conversational implicature question* is to determine the extent to which answers to the semantic compatibility question depend on the participants’ understanding of the speaker’s intention in the vignette. To understand why this is relevant, note that the speaker’s utterance should be considered an indirect answer to the vignette’s tacit question—should the addressee give the speech again? In the face-threatening version of this story, it is plausible that the addressee would take the utterance (*Some... hated*) as an indirect negative answer, which licenses the genuinely conversational implicature that the addressee should not give the speech again. Interestingly, this kind of implicature is warranted in the face-threat condition regardless of whether the scalar utterance is given a semantic interpretation—*At least some (and possibly all)* of the people hated your speech—or a pragmatic one—*Some but not all* of the people hated your speech. In both cases, the word “hated” (the face-threat condition) largely suffices for answering the indirect question regardless of one’s concern for processing the word “Some”. Turning to the face-boosting context, the semantic and the pragmatic readings of the scalar utterance (*Some people loved your speech*) do interact differentially with the task’s indirect question. If the utterance in the face-boosting context is interpreted as *You should not give the speech again* (which is plausible given Denise’s tepid utterance) this judgment is consistent with an enriched reading of the scalar utterance (i.e., *only some people loved your speech so it is not clear that you should give the speech again*). On the other hand, if the scalar utterance is taken to implicate that *You should give the speech again*, it is more likely that the listener interpreted the existential quantifier as *Some and perhaps all*. Crucially, the interpretative path of the face-boosting utterance is bound up with the way it is interpreted.

It is worth noting that the negative valence of the verb “hate” may be stronger than the positive valence of the verb “love.” As a result, answers to the conversational implicature question may be more subject to individual variability in the face-boost condition than in the face-threat condition. As a result of this asymmetry between the face-boosting and the face-threatening stories, it is not clear whether they are equally likely to lead to the pragmatic enrichment of “Some”. By explicitly introducing the *conversational implicature question* in our task, we thus investigate this potential asymmetry between the face-threatening and the face-boosting context.

## Material and Methods

We recruited 399 participants through Amazon Mechanical Turk<sup>2</sup> (228 men, 171 women, mean age 32.3,  $SD = 10.1$ ). Each participant read the Speech vignette either in the face-boosting

version or in the face-threatening one. The face-threatening version of the Speech vignette read as follows:

- (6a) Imagine you gave a speech at a small political meeting. You are discussing your speech with Denise, who was also there. There were 6 other people in the audience that day. You tell Denise that you are thinking about giving the same speech to another group.
- (6b) Hearing this, Denise tells you that “Some people hated your speech.”

We made some minor adjustments to Bonnefon et al.’s task in order to make it clearer (e.g., it is more appropriate to call a gathering of six people a “meeting” rather than a “rally”). Texts corresponding to (6a) and (6b) were displayed in two separate screens. In the face-boost version, *Some people hated your speech* was replaced with *Some people loved your speech*. After reading the story, participants were asked the following questions (always in this order):

- (6c) Given what Denise told you, do you think that it is possible that everybody hated [loved] your speech?
- (6d) Given what Denise tells you, do you think that she means that you should give the speech again to another group?

These questions were followed by two options, “Yes” and “No.” Participants were required to click on one of them.

This study was carried out in accordance with the Décret n° 2017-884, whose article R. 1121- 1. -II indicates that such research does not have to receive IRB approval in France. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

## Analysis

In order to retain the cleanest data possible, we first removed from our analysis participants who (i) unnecessarily clicked on the relevant screen more than once<sup>3</sup> (91 participants), as well as participants who (ii) showed a clear lack of attention during the task by exceeding the following reaction times:  $RT_{\text{UTTERANCE}} > 20$  s,  $RT_{\text{QUESTION}} > 30$  s (3 participants). Finally, we log transformed the data and we removed 13 participants as outliers using the criteria of 2.5 SD away from the mean for  $RT_{\text{UTTERANCE}}$  and for  $RT_{\text{QUESTION}}$ . Our final sample included 292 participants.

## Results

### The semantic compatibility question

In the face-threat condition, 45% of participants answered “Yes” to the question “...do you think it is possible that everybody hated your speech?” This percentage dropped to 32% in the face-boost condition (Fisher exact test,  $p = 0.02$ ,  $OR = 1.7$ ). The difference across conditions, though slightly narrower than those reported by Bonnefon et al. (2009, 2011), replicates their

<sup>2</sup>On the general reliability of MTurk data see, e.g., Capaldi (2017) and Zwaan et al. (2017). For evidence concerning the reliability of on-line reaction time data, see Crump et al. (2013).

<sup>3</sup>This “clicking criterion” was introduced in order to make sure that reaction times would reflect *spontaneous* interpretation or epistemic assessment of the scalar utterance. By eliminating errant or superfluous clicks we could be assured that slowdowns (or lack of thereof) would reflect an immediate response and not, for example, an inability to find the relevant answer box, a temporary interruption of the task, or a second thought (with no way to distinguish among these possibilities).

original findings and confirms that participants are significantly more likely to respond positively to the semantic compatibility question in the face-threat condition than they are in the face-boost one (Table 2).

### The conversational implicature question

When asked whether the speaker meant that the participant should give the speech again, the participants in the face threat condition were practically unanimous in saying “No”—only 7% responded with “Yes.” In the face-boost condition, participants were more divided: 36% said “No” and 64% answered “Yes.” This result confirms our prediction that the conversational implicature would vary across the two conditions.

In order to see how the semantic compatibility question is influenced by the conversational implicature (Table 2), we split participants according to their answers. Preserving the order of the questions, there are four possibilities: Yes-Yes, Yes-No, No-Yes, No-No. Tables 3, 4 provide the distribution of the participants in the face-threat and face-boost conditions respectively.

**TABLE 2 |** Percentage of “Yes” responses to the semantic compatibility question and the conversational implicature question in Study 1.

	Hearing this, Denise tells you that	
	“Some people hated your speech” Face-threat condition	“Some people loved your speech” Face-boost condition
<b>SEMANTIC COMPATIBILITY QUESTION</b>		
do you think that it is possible that everybody hated [loved] your speech?	45%	32%
<b>CONVERSATIONAL IMPLICATURE QUESTION</b>		
do you think that she means that you should give the speech again to another group?	7%	64%

For each cell, the complement corresponds to “No” responses.

**TABLE 3 |** Percentage of “Yes”/“No” answers to the semantic compatibility question and the conversational implicature question in the face-threat condition (Study 1).

Semantic compatibility question	Conversational implicature question	Proportion of total
Yes: 45%	Yes: 3%	Yes-Yes: 1%
	No: 97%	Yes-No: 44%
Face-threat		
No: 55%	Yes: 10%	No-Yes: 6%
	No: 90%	No-No: 49%

These data are particularly relevant with regard to the face-boost condition as they provide insight into how to interpret participants’ answers to the semantic-compatibility question. As discussed above, a “No” answer to the conversational implicature question provides us with an indication about the scalar inference. That is, it is plausible to assume that those who carry out and adopt the scalar inference in the face-boost condition (to infer and commit to *Some but not all the people loved your speech* by saying “No” to the semantic compatibility question) are more likely to conclude *You should not give the speech again*. Arguably, one might not want to give a speech again if only a subset of the group loved the speech. This allows us to speculate about the pattern of answers for the group of participants who answered “Yes” to the conversational implicature question (*You should give the speech again*). Most likely, those who answer “Yes” to both questions in the face-boost condition represent a group of participants who have arguably not derived the scalar inference at all (28% of the total).

### Reaction times

In the top of Table 5, we present the results by combining  $RT_{\text{UTTERANCE}}$  and  $RT_{\text{QUESTION}}$  in order to make them comparable to Bonnefon et al.’s (2011). A two-way ANOVA was conducted that examined the effect of Response Type and Face Condition (Face-boost/Face-threat) on the combined response times. The results do not strongly replicate Bonnefon et al. (2011). While our results are in line with their data, we only find a tendency toward an interaction,  $[F_{(1, 288)} = 2.75, p = 0.10]$ , reflecting the fact that participants took longer to answer “Yes” but only in the face-threat condition. We did not find a main effect for Response Type  $[F_{(1, 288)} = 0.32, p = 0.57]$  nor for Face Condition  $[F_{(1, 288)} = 0.53, p = 0.46]$ .

We then analyse these results in more detail by breaking them up into  $RT_{\text{UTTERANCE}}$  and  $RT_{\text{QUESTION}}$ . We ran the same ANOVA analysis, first using  $RT_{\text{UTTERANCE}}$  as the dependent variable and then using  $RT_{\text{QUESTION}}$  as the dependent variable. With regard to  $RT_{\text{UTTERANCE}}$ , it showed no main effect of Face Condition  $[F_{(1, 288)} = 1.09, p = 0.30]$ , no main effect of Response Type  $[F_{(1, 288)} = 0.00, p = 0.99]$ , and no interaction

**TABLE 4 |** Percentage of “Yes”/“No” answers to the semantic compatibility question and the conversational implicature question in the face-boost condition (Study 1).

Semantic compatibility question	Conversational implicature question	Proportion of Total
Yes: 32%	Yes: 87.5%	Yes-Yes: 28%
	No: 12.5%	Yes-No: 4%
Face-boost		
No: 68%	Yes: 53%	No-Yes: 36%
	No: 47%	No-No: 32%



**TABLE 5 |** Mean response time (in seconds) for “Yes” and “No” answers to the semantic compatibility question, in the face-boost and in the face-threat conditions (Study 1).

	Face-threat	Face-boost
<b>SCALAR UTTERANCE + SEMANTIC COMPATIBILITY QUESTION</b>		
“Yes”	8.2 s (2.6)	7.5 s (2.9)
“No”	7.6 s (2.7)	7.9 s (3.1)
<b>SCALAR UTTERANCE</b>		
“Yes”	3.0 s (1.0)	3.1 s (1.5)
“No”	2.9 s (1.3)	3.3 s (1.6)
<b>SEMANTIC COMPATIBILITY QUESTION</b>		
“Yes”	5.2 s (1.9)	4.4 s (2.1)
“No”	4.6 s (1.9)	4.6 s (2.1)

Standard deviations are included in parentheses.

effect [ $F_{(1, 288)} = 1.12, p = 0.29$ ]. On the other hand, with regard to  $RT_{QUESTION}$ , the ANOVA revealed a main effect of Face Condition [ $F_{(1, 288)} = 4.21, p = 0.04$ ] but no main effect for Response Type [ $F_{(1, 288)} = 0.58, p = 0.45$ ]. There was a tendency toward an interaction [ $F_{(1, 288)} = 3.51, p = 0.06$ ].

Finally, a logistic regression analysis was conducted to predict Response Type (“Yes”/“No”) using as predictors Face Condition,  $RT_{UTTERANCE}$ ,  $RT_{QUESTION}$ , and the interaction between Face Condition and both RT measures. A test of the full model against a constant only model failed to reach statistical significance, indicating that the predictors as a set did not distinguish between “Yes” and “No” answers (chi square = 9.16,  $p = 0.10$  with  $df = 5$ ).

## Discussion

In line with Bonnefon and colleagues, the results suggest that in face-threatening contexts people are more likely to answer positively to the semantic compatibility question, i.e., indicating that some participants ultimately believed that *Everybody hated your speech* when told that *Some people hated your speech*. This replicates the results of a series of studies by Bonnefon and colleagues (Bonnefon et al., 2009, 2011; Feeney and Bonnefon, 2012) and confirms the robustness of this effect.

In contrast, Study 1 did not provide robust support to Bonnefon et al.’s (2011) original reaction time claim. Study 1 failed to replicate their significant interaction when considering the utterance and question together, though it did reveal a tendency in the expected direction ( $p = 0.10$ ). That is, participants who answered “Yes” tended to be slower overall, and only in the face-threat condition.

However, thanks to our design, we could further investigate potential reaction time differences at the presentation of the scalar utterance ( $RT_{UTTERANCE}$ ) and of the semantic compatibility question ( $RT_{QUESTION}$ ). Contrary to predictions based on Bonnefon et al.’s (2011), we did not find any significant reaction time difference for  $RT_{UTTERANCE}$ . We did find a significant effect of condition for  $RT_{QUESTION}$ , with slower responses in the face-threat condition than in the face-boost condition. Furthermore, in line with predictions based on Mazzarella (2015), the interaction effects suggest that “Yes”

answers (5.2 s) tend to take longer than “No” answers (4.6 s) in the face-threat condition (the latter being comparable with “Yes” and “No” answers in the face-boost condition—4.4 and 4.6 s respectively).

In sum, Study 1 provides no evidence that the scalar utterance is interpreted at different speeds across the two conditions. On the other hand, the data do suggest that the process of epistemic evaluation, which operates when answering the semantic compatibility question, is the source of reaction time differences. These results confirm our hypothesis that the process of epistemically evaluating the piece of incoming information plays a crucial role in Bonnefon and colleagues’ task, and that this needs to be distinguished from the process of interpreting the scalar utterance. Epistemic assessment may lead to the rejection of the scalar inference, particularly in these face-threatening contexts (because of politeness considerations). Given that the rejection of the executed scalar inference would ultimately provide a “Yes” answer, this would explain why the percentage of “Yes” answers to the semantic compatibility question increases in the face-threat condition, and why they tend to be longer than “No” answers.

## Study 2

Study 2 aimed at increasing the likelihood that participants would derive the scalar inference. To achieve this, we manipulated the background scenario in the following two ways: (i) we increased the relevance of the scalar inference by introducing a slightly different implicit question that puts the focus on the delivery of the speech (*You tell Denise that you would like to know the audience’s reaction* instead of *You tell Denise that you are thinking about giving the same speech to another group*) and; (ii) we explicitly characterize the speaker (Denise) as knowledgeable with regard to the question at issue (i.e., the so-called “competence assumption” is made clear, see, e.g., Sauerland, 2004; Breheny et al., 2013). See Appendix A for a thorough comparison with Bonnefon et al.’s (2011) material.

As discussed in the literature, the scalar inference is facilitated when the speaker is assumed to be in a position to know the entire situation, i.e., that she is in a position to know that the stronger alternative is false. That is, an utterance of *Some people loved your speech* would be taken to license the scalar inference *Not everybody loved your speech* if one could assume that the speaker knows everybody’s opinion about the speech and is in a position to rule out the possibility that everybody loved it. Bonnefon et al.’s scenarios do not clearly attribute such knowledge to the speaker. In fact, it is not clear whether in the original paradigm Denise is at all aware of the opinion of all the other members of the audience. This leaves open the possibility that some participants might have assumed that Denise did not know everyone’s opinion, and so the listener is arguably not in a position to know whether a scalar inference is called for. Our manipulation (ii) should overcome this limitation. It follows that if the manipulations in (i) and (ii) are effective, the percentage of “Yes” answers to the semantic-compatibility question in the face-boost condition should drop from Study 1 to Study 2 because the scalar inference is called for with greater confidence. Assuming that our manipulations do facilitate the derivation of the scalar inference, we will be

in a better position to analyze the role played by the scalar utterance and the semantic compatibility question with respect to the reaction time effects.

## Material and Methods

We recruited 398 participants through Amazon Mechanical Turk (230 men, 168 women, mean age 32.6,  $SD = 10.5$ ). Each participant read a version of the Speech vignette either in the face-boost version or in the face-threat version.

- (6a) Imagine you gave a speech at a small political meeting. You are discussing your speech with Denise, who was also there. There were 6 other people in the audience that day and you know that Denise spoke with all of them about it later. You tell Denise that you would like to know the audience's reaction.
- (6b) Hearing this, Denise tells you that "Some people hated your speech."

As in Study 1, the texts corresponding to (6a) and (6b) were displayed separately in two steps and the face-boost version presented the scalar utterance with *Some people loved your speech*. After reading the story, participants were presented the semantic compatibility question:

- (6c) Given what Denise told you, do you think that it is possible that everybody hated [loved] your speech?

There was no conversational implicature question.

This study was carried out in accordance with the Décret n° 2017-884, whose article R. 1121-1. -II indicates that such research does not have to receive IRB approval in France. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

## Analysis

Using the same criteria as Study 1, we excluded 87 participants because of the presence of unnecessary clicks, 2 participants because their RTs betrayed a clear lack of attention and 15 participants as outliers. Our final sample included 294 participants.

## Results

### Semantic compatibility question

Responses to the semantic compatibility question (see Table 6) reveal that few participants in the face-boost condition thought it was possible that everyone loved the speech, while in the face-threat condition, participants remained split between the two answers (Fisher exact test,  $p < 0.001$ ,  $OR = 5.7$ ). The minor modifications in Study 2—rendering the speaker omniscient and the utterance more relevant—prompted the anticipated result. The result is that the face-threat/face-boost distinction is much clearer here than in Study 1.

### Reaction times

We first determined whether the combined RTs prompted effects reminiscent of Bonnefon et al. (2011). Table 7 displays the RTs collected. A two-way ANOVA was conducted that examined the effect of Response Type ("Yes"/"No") and Face

Condition (Face-boost/Face-threat) on the combined response times ( $RT_{UTTERANCE} + RT_{QUESTION}$ ). It revealed no significant effects (main effect of Face Condition [ $F_{(1, 290)} = 2.71$ ,  $p = 0.10$ ], main effect of Response Type [ $F_{(1, 290)} = 1.14$ ,  $p = 0.29$ ], interaction effect [ $F_{(1, 290)} = 1.02$ ,  $p = 0.32$ ].

We then performed the same ANOVA using  $RT_{UTTERANCE}$  as a dependent variable and found no significant effects (main effect of Face Condition [ $F_{(1, 290)} = 0.05$ ,  $p = 0.83$ ], main effect of Response Type [ $F_{(1, 290)} = 1.17$ ,  $p = 0.28$ ], interaction effect [ $F_{(1, 290)} = 0.62$ ,  $p = 0.43$ ]. However, when we used  $RT_{QUESTION}$  as dependent variable, the ANOVA revealed a main effect of Face Condition [ $F_{(1, 290)} = 4.91$ ,  $p = 0.03$ ], a main effect of Response Type [ $F_{(1, 290)} = 5.06$ ,  $p = 0.03$ ], and a tendency toward an interaction [ $F_{(1, 290)} = 3.18$ ,  $p = 0.08$ ].

Finally, a logistic regression analysis was conducted to predict Response Type ("Yes"/"No") using as predictors Face Condition,  $RT_{UTTERANCE}$ ,  $RT_{QUESTION}$ , and the interaction between Face Condition and both RT measures (Face Condition\* $RT_{UTTERANCE}$  and Face Condition\* $RT_{QUESTION}$ ). A test of the full model against a constant only model was statistically significant, indicating that the predictors as a set reliably distinguish between "Yes" and "No" answers (chi square = 53.46,  $p < 0.001$  with  $df = 5$ ). A Wald test showed

**TABLE 6 |** Percentage of "Yes" responses to the semantic compatibility question in Study 2.

	Hearing this, Denise tells you that	
	"Some people hated your speech" Face-threat condition	"Some people loved your speech" Face-boost condition
<b>SEMANTIC COMPATIBILITY QUESTION</b>		
do you think that it is possible that everybody hated [loved] your speech?	45%	12.5%

For each cell, the complement corresponds to "No" responses.

**TABLE 7 |** Mean response time (in seconds) for "Yes" and "No" answers to the semantic compatibility question, in the face-boost and in the face-threat conditions (Study 2).

	Face-threat	Face-boost
<b>SCALAR UTTERANCE + SEMANTIC COMPATIBILITY QUESTION</b>		
"Yes"	8.4 s (3.2)	7.2 s (2.5)
"No"	7.7 s (3.2)	7.5 s (3.2)
<b>SCALAR UTTERANCE (COMPREHENSION ASSESSMENT STAGE)</b>		
"Yes"	2.8 s (1.2)	2.8 s (0.9)
"No"	3.1 s (1.3)	2.9 s (1.3)
<b>SEMANTIC COMPATIBILITY QUESTION</b>		
"Yes"	5.6 s (2.4)	4.4 s (1.9)
"No"	4.6 s (2.3)	4.6 s (2.5)

Standard deviations are included in parentheses.

that the interaction Face Condition\*RT<sub>QUESTION</sub> significantly predicts response type [Wald (1) = 4.7,  $p = 0.03$ ]. An increase of 1 s in response time for answering the semantic compatibility question and being in the face-threat condition makes participants 1.4 times more likely to answer “Yes.” By contrast, neither the interaction Face Condition\*RT<sub>UTTERANCE</sub> nor any of the factors individually (Face Condition, RT<sub>UTTERANCE</sub> and RT<sub>QUESTION</sub>) turned out to be significant predictors [Face Condition\*RT<sub>UTTERANCE</sub>: Wald (1) = 2.1,  $p = 0.15$ , Condition: Wald (1) = 2.3,  $p = 0.13$ , RT<sub>UTTERANCE</sub>: Wald (1) = 0.05,  $p = 0.82$ , RT<sub>QUESTION</sub>: Wald (1) = 0.04,  $p = 0.84$ ]. It is important to note that, while the interaction Face Condition\*RT<sub>UTTERANCE</sub> is not statistically significant, the tendency goes in a direction opposite to the Face Condition\*RT<sub>QUESTION</sub> interaction effect. That is, an increase of 1 s in reading time for the scalar utterance and being in the face threat condition makes participants 1.4 times less likely to answer “Yes” to the semantic compatibility question [ $OR = 0.7$ ]. In other words, those subjects who spend more time reading the scalar utterance tend to respond negatively later to the semantic compatibility question.

## Discussion

In line with our expectations, modifying the task so that it maximizes the coherence between the background story and the utterance while also presenting Denise as omniscient about the relevant group of people affected the percentage of “Yes” and “No” answers to the semantic compatibility question. Compared to the task in Study 1, Study 2 presents higher rates of negative responses to the semantic compatibility question in the face-boosting condition. That is, by ensuring the relevance of the scalar inference, we replicated and sharpened the results of Study 1 (in line with Bonnefon and colleagues’ work).

As in Study 1, we found no evidence of reaction time differences with respect to the presentation of the scalar utterance (RT<sub>UTTERANCE</sub>), as should be predicted by Bonnefon et al. (2011). In fact, our results suggest that, if anything, being in the face-threat condition and displaying longer reading times for the scalar utterance is less likely to produce “Yes” answers to the semantic compatibility question. This is in direct contrast with Bonnefon et al.’s claim that longer reaction times for “Yes” responses in the face-threat condition are linked to the extra costs imposed by politeness considerations on the processing of the scalar utterance. Our results show that, on the contrary, slow “Yes” responses are due to the process of epistemic assessment, which takes exceptionally longer in the face-threat condition when participants answer the semantic compatibility question. This interaction was indeed the only significant predictor of Response Type to the semantic compatibility question.

## GENERAL DISCUSSION

Bonnefon and colleagues investigate the effect of politeness in the computation of scalar inferences. Based on their findings, they suggest that politeness “appears to add a layer of complexity to the usual processes involved in the interpretation of *some*” (Bonnefon et al., 2011, p. 3393). Specifically, they claim that the effect of politeness is two-fold: on the one hand, it blocks the

derivation of the scalar inference in face-threatening contexts and, on the other, it makes the semantic interpretation of “some” (*at least some and possibly all*) appear to be an effortful step. While their (2011) findings—which reveal slowdowns when giving positive responses to the semantic compatibility question—are consistent with their account, their analysis is based on a task whose dependent measure does not isolate the scalar utterance. Their task involves reading a scalar utterance that (a) serves as an indirect response to a more pressing question and that; (b) is then re-assessed meta-linguistically in the task’s semantic compatibility question, which is its real dependent measure.

The main aim of our studies was to reevaluate Bonnefon et al.’s findings and to reanalyse the task through the lens of *epistemic vigilance*. While comprehension involves the pragmatic ability to infer the speaker’s meaning from linguistic and contextual cues, epistemic assessment involves what Sperber et al. (2010) call a capacity for epistemic vigilance, which enables hearers to avoid being accidentally or intentionally misinformed. Sperber et al. (2010) have suggested that there are two main factors affecting the believability of a piece of communicated information: the reliability of its source and the believability of its content. Our hypothesis—based on Mazzarella (2015)—was that epistemic vigilance toward the source may affect the believability of scalar inferences in face-threatening contexts. In such contexts, participants recognize that the speaker is trying to be nice and polite (by allowing the listener to generate a reading that can be glossed as *Some but not all the people hated your speech*); however, participants also recognize that it is probable that the speaker’s comment is perhaps well meaning but that it is not entirely honest and, consequently, they judge part of what she communicates as not true, so they do not accept it. While participants are likely to conclude that *everyone hated the speech*, participants answer affirmatively to the semantic compatibility question because they are rejecting the speaker’s communicated information. This shows how there is a distinction to be made between what is communicated in a comprehension stage and what is believed (or not) in an epistemic assessment stage. We have argued that responses in this task are due to reactions at the epistemic assessment stage; according to Bonnefon et al., the task’s question is merely a measure of scalar inference-making.

In order to adjudicate between the two competing claims, we experimentally separated the task’s scalar utterance from the dependent measure—responses to the semantic compatibility question. We thus (1) separated the reading of the scalar utterance from the reading/responding to the test question and we (2) made the utterance more relevant to the participant’s task wherever possible. With respect to (1) we took separate reading time measures of the scalar utterance and the response to the semantic compatibility question (while also adding the two together) in both of our studies. Overall, we find no evidence that the scalar utterance is interpreted at different speeds across the two (face-threat vs. face-boost) conditions. Our data suggest, instead, that the process of epistemic evaluation, which operates when answering the semantic compatibility question, is the source of the reaction time differences. This undermines any claim that suggests that participants slow down while *drawing* a semantic reading of the utterance on line.

Overall, our results show that there is a range of responses to scalar utterances. As prior studies have shown, they are often not drawn at all. We see evidence of that here, through the large minority of participants in the face-boost condition of Study 1 who give Yes-Yes responses (in line with findings from Degen and Tanenhaus, 2015). As prior studies have also shown, including those generated by the current paradigm, participants can be encouraged to generate scalar enrichments (see Study 2) once the competence assumption can be more confidently endorsed. This can be seen through the high percentage of participants who respond negatively to the semantic compatibility question. The added value of the current work is that it shows that one can experimentally capture a third process. That is, work with the current paradigm shows that people often make the scalar enrichment because it is part of a speaker's communicated meaning but that eventually a listener can reassess that communicated information and reject it. In the current case, this is due to effects of politeness. We suggest that politeness affects the process of epistemically evaluating a piece of incoming information as presented by the speaker. When the speaker is perceived as motivated by politeness concerns, her reliability as a trustworthy informant becomes questionable. As a consequence, addressees are more likely to reject what the speaker is communicating to them (e.g., a pragmatically enriched scalar utterance). It is sensible to assume that taking these concerns into consideration requires additional processing effort (reflected in longer reaction times at the epistemic assessment stage).

## REFERENCES

- Austin, J. (1962/98). "How to do things with words," in *Pragmatics Critical Concepts*, Vol. 2, ed A. Kashe (London: Routledge), 7–64.
- Bonnefon, J.-F., De Neyes, W., and Feeney, A. (2011). "Processing scalar inferences in face-threatening contexts," in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, eds L. Carlson, C. Hölscher, and T. Shipley (Austin, TX: Cognitive Science Society), 3389–3395.
- Bonnefon, J.-F., Feeney, A., and Villejoubert, G. (2009). When some is actually all: scalar inferences in face-threatening contexts. *Cognition* 112, 249–258. doi: 10.1016/j.cognition.2009.05.005
- Bott, L., Bailey, T. M., and Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *J. Memory Lang.* 66, 123–142. doi: 10.1016/j.jml.2011.09.005
- Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *Cognition* 51, 437–457. doi: 10.1016/j.jml.2004.05.006
- Breheny, R., Ferguson, H. J., and Katsos, N. (2013). Taking the epistemic step: towards a model of on-line access to conversational implicatures. *Cognition* 126, 423–440. doi: 10.1016/j.cognition.2012.11.012
- Breheny, R., Katsos, N., and Williams, J. (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. doi: 10.1016/j.cognition.2005.07.003
- Brown, P., and Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press
- Capaldi, C. (2017). *Graduating From Undergrads: Are Mechanical Turk Workers More Attentive than Undergraduate Participants?* OSF. Available online at: <https://osf.io/d2zxw/>
- Carston, R. (1998). "Informativeness, relevance and scalar implicature," in *Relevance Theory: Applications and Implications*, eds R. Carston and S. Uchida (Amsterdam: John Benjamins), 179–236.
- Chemla, E., and Singh, R. (2014a). Remarks on the experimental turn in the study of scalar implicature, Part I. *Lang. Linguist. Compass* 8, 373–386. doi: 10.1111/lnc3.12081
- Chemla, E., and Singh, R. (2014b). Remarks on the experimental turn in the study of scalar implicature, Part II. *Lang. Linguist. Compass* 8, 387–399. doi: 10.1111/lnc3.12080
- Chierchia, G. (2013). *Logic in Grammar: Polarity, Free Choice, and Intervention*. Oxford: Oxford University Press.
- Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE* 8:e57410. doi: 10.1371/journal.pone.0057410
- Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171
- De Neyes, W., and Schaecken, W. (2007). When people are more logical under cognitive load: dual task impact on scalar implicature. *Exp. Psychol.* 54, 128–133. doi: 10.1027/1618-3169.54.2.128
- Feeney, A., and Bonnefon, J.-F. (2012). Politeness and honesty contribute additively to the interpretation of scalar expressions. *J. Lang. Soc. Psychol.* 32, 181–190. doi: 10.1177/0261927X12456840
- Foppolo, F., and Marelli, M. (2017). No delay for some inferences. *J. Semantics* 34, 659–681. doi: 10.1093/jos/ffx013
- Horn, L. (1984). "Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature," in *Meaning, Form, and Use in Context*, ed D. Shiffrin (Washington, DC: Georgetown University Press), 11–42.
- Huang, Y. T., and Snedeker, J. (2009). Online interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cogn. Psychol.* 58, 376–415. doi: 10.1016/j.cogpsych.2008.09.001

These data, all inspired by Bonnefon et al.'s paradigm, open up an interesting direction of research within the field of experimental pragmatics. Crucially, they highlight the importance of taking into consideration the cognitive distinction between comprehension and acceptance. This distinction, which has long been acknowledged in the philosophical literature thanks to the seminal work of Austin and Grice, have been neglected in the experimental pragmatics literature so far. The challenge for the future is to devise new paradigms to study comprehension and epistemic assessment as two distinct components in the process of forming beliefs via testimony.

## AUTHOR CONTRIBUTIONS

DM devised the project and the main conceptual ideas. DM and IN designed the studies. DM and ET implemented the studies and ET carried out the analysis of the results. DM wrote the manuscript with input from IN and HM. All authors helped shape the research and discussed the results.

## FUNDING

This work was supported by a Post-doctoral Study Grant awarded by the Fyssen Foundation to DM, an ANR grant [ANR-16-TC-0001-01] to HM and a Mercator Fellowship (under the auspices of Xprag.de) to IN. The publication of this article was funded by the Open Access Fund of the Leibniz Association.



- Lee, J. J., and Pinker, S. (2010). Rationales for indirect speech: the theory of the strategic speaker. *Psychol. Rev.* 117, 785–807. doi: 10.1037/a0019688
- Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.
- Mazzarella, D. (2015). Politeness, relevance and scalar inferences. *J. Pragmat.* 79, 93–10. doi: 10.1016/j.pragma.2015.01.016
- Noveck, I. (2018). *Experimental Pragmatics. The Making of a Cognitive Science*. Cambridge, MA: Cambridge University Press.
- Noveck, I. A., and Sperber, D. (2007). “The why and how of experimental pragmatics: the case of scalar inference,” in *Advances in Pragmatics*, ed N. Burton-Roberts (Basingstoke: Palgrave), 184–212.
- Noveck, I. A., and Spotorno, N. (2014). “Experimental pragmatics,” in *Handbook of Pragmatics*, ed J. Verschueren (Amsterdam: Benjamins Publishing Company), 1–30.
- Noveck, I. A., and Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends Cogn. Sci.* 12, 425–431. doi: 10.1016/j.tics.2008.07.009
- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguist. Philos.* 27, 367–391. doi: 10.1023/B:LING.0000023378.71748.db
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. (2010). Epistemic vigilance. *Mind Lang.* 24, 359–393. doi: 10.1111/j.1468-0017.2010.01394.x
- Sperber, D., and Wilson, D., (1986/1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.
- Zwaan, R., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., et al. (2017). *Some Psychological Effects Replicate Even Under Potentially Adverse Conditions*. Available online at: <https://osf.io/preprints/psyarxiv/rbz29/download?format=pdf>

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Mazzarella, Trouche, Mercier and Noveck. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX A

The table below displays the Speech story in the original version from Bonnefon et al. (2011) (translated from Dutch), as well as in the modified versions of Study 1 and Study 2.

Relevant changes to the original story are in bold. Horizontal lines indicate where participants were asked to advance the text.

Bonnefon et al., 2011	Study 1	Study 2	
Imagine you gave a speech at a small political rally. You are discussing your speech with Denise, who was in the audience. There were 6 other people in the audience. You are considering whether to give this same speech to another audience.	Imagine you gave a speech at a small political meeting. You are discussing your speech with Denise, who was also there. There were 6 other people in the audience that day. You tell Denise that you are thinking about giving the same speech to another group.	Imagine you gave a speech at a small political meeting. You are discussing your speech with Denise, who was also there. There were 6 other people in the audience that day and <b>you know that Denise spoke with all of them about it later. You tell Denise that you would like to know the audience's reaction.</b>	
Hearing this, Denise tells you that "Some people loved/hated your speech." Given what Denise told you, do you think that it is possible that everybody loved/hated your speech?	Hearing this, Denise tells you that "Some people loved/hated your speech."	Hearing this, Denise tells you that "Some people loved/hated your speech."	RT <sub>UTTERANCE</sub>
	Given what Denise told you, do you think that it is possible that everybody loved/hated your speech?	Given what Denise told you, do you think that it is possible that everybody loved/hated your speech?	RT <sub>QUESTION</sub>
	<b>Given what Denise tells you, do you think that she means that you should give the speech again to another group?</b>		



# Scalar Diversity, Negative Strengthening, and Adjectival Semantics

Nicole Gotzner<sup>1,2\*</sup>, Stephanie Solt<sup>1</sup> and Anton Benz<sup>1</sup>

<sup>1</sup> Leibniz-Centre General Linguistics, Berlin, Germany, <sup>2</sup> Department for German Language and Linguistics, Humboldt-University, Berlin, Germany

Previous research has demonstrated great variability in the rates of scalar inferences across different triggers (Doran et al., 2009; van Tiel et al., 2016). In the current study, we show that variation is more systematic than previously thought. In particular, we present experimental evidence suggesting that endorsements of scalar implicatures (i) are anti-correlated with the degree of negative strengthening of the stronger scale-mate (e.g., whether *John is not stunning* is interpreted as conveying that John is rather ugly) and (ii) are affected by the scale structure and the underlying scalar semantics of gradable adjectives (in particular boundedness, polarity, and adjectival extremeness). Overall, our research suggests that scale structure should be taken into account in theories of implicature.

## OPEN ACCESS

### Edited by:

Penka Stateva,  
University of Nova Gorica, Slovenia

### Reviewed by:

Richard Breheny,  
University College London,  
United Kingdom  
Mojmir Docekal,  
Masaryk University, Czechia

### \*Correspondence:

Nicole Gotzner  
gotzner@leibniz-zas.de

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 01 May 2018

**Accepted:** 17 August 2018

**Published:** 12 September 2018

### Citation:

Gotzner N, Solt S and Benz A (2018)  
Scalar Diversity, Negative  
Strengthening, and Adjectival  
Semantics. *Front. Psychol.* 9:1659.  
doi: 10.3389/fpsyg.2018.01659

**Keywords:** scalar implicature, scalar diversity, scale structure, gradable adjectives, negative strengthening, negation

## 1. INTRODUCTION

According to a tacit assumption in the theoretical and experimental literature, scalar implicature is based on a single mechanism, and the behavior of one scale generalizes to the whole family of scales (van Tiel et al., 2016). Contrary to this so-called uniformity assumption, experimental research has demonstrated great variability in the rates of scalar inferences across different triggers, in part being explained by factors such as grammatical category, boundedness, and semantic distance between scale-mates (Doran et al., 2009, 2012; van Tiel et al., 2016). These experimental studies have provided evidence that gradable adjectives in particular tend to yield low rates of scalar implicature (e.g., see the conclusions in Doran et al., 2012; Beltrama and Xiang, 2013).

In the current study, we focus on scalar implicatures and a specific kind of manner implicature triggered by negated adjectives, referred to as negative strengthening (Horn, 1989). Negative strengthening describes the phenomenon by which an utterance such as *John is not brilliant* receives a stronger interpretation than its semantic meaning, for example that John is “rather stupid” or less than intelligent. This interpretation is derived as a Manner or I implicature (Horn, 1989; Levinson, 2000) or explained as a blocking phenomenon in optimality theory (Blutner, 2000; Krifka, 2007). Theories agree that scalar implicature and negative strengthening are two different kinds of implicature, which arise from distinct conversational principles, the Q and R principles, respectively (Horn, 1989; Levinson, 2000). These principles are assumed to govern each other; therefore an interaction between the two kinds of pragmatic strengthening is expected (see Krifka, 2007)<sup>1</sup>.

<sup>1</sup> For example, Levinson (2000) assumes that Q and I implicatures are additive if an utterance triggers both kinds of inferences but if the two interpretations stand in conflict, the Q implicature wins over the I implicature.

In this paper, we present two experimental studies investigating the interaction of scalar implicature and negative strengthening in different types of gradable adjectives. We will show that for some adjectives the effect of scalar implicature may be masked by the presence of negative strengthening. Further, we provide evidence that the scale structure associated with the semantics of gradable adjectives affects the likelihood with which a scalar implicature and negative strengthening are derived.

In the following, we discuss how scalar implicature and negative strengthening affect the interpretation of gradable adjectives and we review previous studies on scalar diversity. Then, we present the results of two experiments and discuss the relevance of the findings to the phenomenon of scalar diversity.

## 1.1. Interaction Between Scalar Implicature and Negative Strengthening

In this section, we explore the interplay of different kinds of implicatures and the interpretations they lead to. To see the effect of semantic interpretation and pragmatic inference on statements involving weak and strong scalar terms, consider the scale of attractiveness depicted in **Figure 1**. The first line represents the semantic interpretation of an utterance like *John is attractive*, which is compatible with the stronger alternative statement *John is stunning*. The second line depicts the effect of scalar implicature, namely that the stronger statement is implicated not to obtain, such that the weaker term *attractive* is understood to apply only to the more restricted range of being “attractive but not stunning.” This implicature is based on the maxim of quantity, since the two scale-mates stand in an entailment relationship with each other and the stronger term is more informative than its weaker scale-mate.

Now consider the case where the strong scalar term appears under negation. As shown in the first line of **Figure 1**, *John is not stunning* entails on its semantic meaning merely that John is something less than stunning (i.e., it leaves open whether he might be attractive, merely average looking, or downright ugly). The third line of **Figure 1** shows the effect of negative strengthening on the stronger scalar item: If the statement *John is not stunning* is negatively strengthened, the resulting meaning is inconsistent with John being attractive (i.e., it is consistent with him being average looking or ugly). Horn (1989) posits that this interpretation is based on a conventionalized negative strengthening rule according to which negated adjectives receive a “rather un-adjective” interpretation. He assumes that for choosing the more complicated negated expression, the speaker must have had additional reasons such as politeness considerations. In a similar vein, an asymmetry between positive and negative expressions has been pointed out (Brown and Levinson, 1987; Horn, 1989; Ruytenbeek et al., 2017). For example, while an utterance like *John is not tall* is often interpreted as John being rather short, the statement with the antonym (*John is not short*) is unlikely to be strengthened in order to convey that John is rather tall. The assumption is that the positive adjective denotes a desirable property in contrast to its antonym, relating the negative asymmetry to euphemism and understatement (Brown and Levinson, 1987; Horn, 1989;

Krifka, 2007). However, it is easy to find counterexamples to this asymmetry and it is unclear which notion of polarity is the relevant one (e.g., emotional valency vs. negative morphology) or which adjective constitutes the positive form (see especially Ruytenbeek et al., 2017). We will return this issue below.

As the fourth line of **Figure 1** shows, when both apply, scalar implicature and negative strengthening divide up the range of possible interpretations categorically. However, as we will see this pattern may only hold in the case of certain types of gradable adjectives.

Note also that there is a possibility not reflected in **Figure 1**, namely that negated strong expressions receive a so-called scale-reversal or indirect implicature (e.g., see Horn, 1989; Chierchia, 2004; Romoli, 2012; Gotzner and Romoli, 2017). A classic example is that the utterance *John did not eat all of the cookies* implicates that he ate some of them. This scale reversal implicature occurs when the strong scale-mate appears under negation and it is assumed to arise by the same mechanism as (direct) scalar implicature. The crucial difference is that negation reverses entailment relationships and therefore the *not all* is replaced with the alternative *not some*. Thus, the negation of the stronger alternative *not some* leads to the inference that John ate some of the cookies.

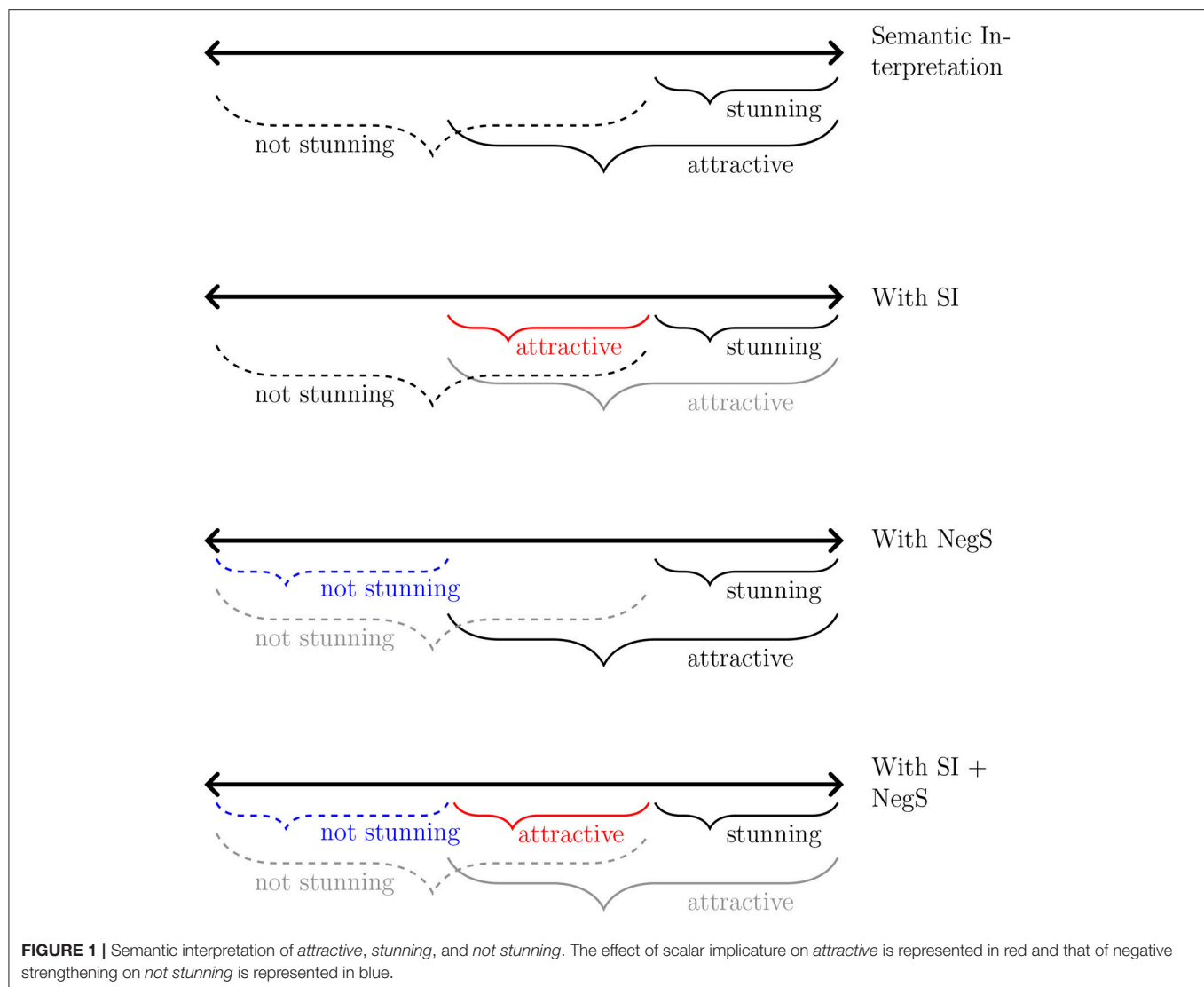
While scale reversal implicature and negative strengthening are based on different types of conversational principles, they stand in direct competition with each other. When a sentence contains a negated scalar term, scale reversal leads to the endorsement of the weaker scale-mate while negative strengthening excludes the weaker scale-mate. Thus, hearers may be inclined to take into account both considerations of informativeness and manner when deciding whether the weaker term applies (that is, whether the speaker wanted to convey that the weaker term applies or not).

## 1.2. Experimental Evidence for Scalar Diversity

There have been several experiments investigating the likelihood with which different scalar terms trigger a scalar implicature. Doran et al. (2009, 2012) investigated the availability of such inferences across a range of scale types, using a truth value judgment task in which participants were presented with a statement containing a weak scalar term and a fact which would support the use of a stronger term, and were asked to indicate whether a literally minded character “Literal Lucy” would say the sentence was true or false given that fact. The results showed that upper-bounding inferences were less likely to arise in the case of gradable adjectives than for quantifiers, cardinal numerals or rank orderings. Furthermore, only in the case of adjectives did the explicit mention of stronger scale-mate alternatives have the effect of increasing the rate of implicatures.

In an experiment employing a felicity-judgment task, Beltrama and Xiang (2013) similarly found evidence that adjectives behave differently from modal expressions with respect to the triggering of scalar implicatures, and furthermore that adjectives themselves differ in the extent to which they give rise to implicatures. Specifically, it was found that weak





positive adjectives (e.g., *decent*) tend to implicate the negation of the corresponding middle and extreme adjectives (e.g., *good*, *excellent*), but middle adjectives do not implicate the negation of the extreme adjective. No such difference was found for modal expressions. The authors suggest several possible explanations for their findings, including relative semantic distance between scale-mates, the particular semantic properties of extreme adjectives, and the unbounded nature of adjectival measurement scales as opposed to the bounded nature of modal scales (see also Simons and Warren, 2018 for further evidence on the role of boundedness and relevant discussion).

A more extensive and fine-grained investigation of potential variability in scalar implicature rates is that of van Tiel et al. (2016), who investigated 43 weak/strong scalar pairs from a variety of grammatical categories, including adjectives, determiners, verbs, and adverbs. In their experiment, participants were presented with statements involving a weak scalar term and

were asked whether they would infer the negation of a stronger scale-mate, for example whether the statement in (1) licenses the scalar inference in (2).

- (1) John is attractive
- (2) John is not stunning

The main finding of the van Tiel et al. (2016) study was a high variability in endorsement rates of the scalar inference across triggering expressions. For example, while few participants endorsed the potential scalar inference in (2) triggered by the weak term *attractive*, almost all participants endorsed the scalar inference associated with *some*. The authors also systematically investigated a range of factors that could account for scalar diversity. As potential predictors of variability in inference rates, van Tiel et al. (2016) probed the semantic distance between the weaker and stronger term, their association strength, the availability of the stronger term, its relative frequency, as well as the presence of an upper bound on the underlying measurement

scale. The only significant predictors were upper boundedness and semantic distance (as measured by a rating of the perceived difference in strength between the statements involving the weaker and stronger term). But a large proportion of the overall variance in inference rates remained unexplained by any of the potential predictors investigated. Overall, the study by van Tiel et al. (2016) has been taken as evidence refuting the uniformity hypothesis.

Benz et al. (2018a) revisited the methodology and findings of van Tiel et al. (2016), raising the possibility of a confound due to the presence of negative strengthening. As described above, the experimental materials in van Tiel et al.'s study included negated stronger scale-mates, which as discussed above may give rise to negative strengthening. Consider our example (2) from above. The utterance *John is not stunning* may be strengthened to convey that John is rather ugly, which is incompatible with the semantic meaning of *attractive*. This could have the effect of masking the presence of scalar implicature. That is, participants in van Tiel et al.'s task may have derived a scalar implicature for the weaker scale-mate but decided nonetheless to respond with *No* because the strengthened reading of the stronger scale-mate stood in conflict with their interpretation of the implicature-modified weaker term. If participants interpreted *John is attractive* as "attractive but not stunning" but *John is not stunning* in the conclusion sentence is interpreted as John being rather ugly, then the *No*-answer is simply based on the presence of negative strengthening—not the absence of scalar implicature for the weaker scale-mate. To be clear, this additional pragmatic strengthening comes into play because the conclusion sentence contains a negated strong scalar term and this affects the interpretation of the original statement of interest.

Benz et al. (2018a) carried out an experiment using the same set of materials used by van Tiel et al. in which participants saw a statement involving the negation of the stronger scale-mate and were asked whether the negation of the weaker term followed, as a measure of negative strengthening. For example, participants were asked whether an utterance like (2) *John is not stunning* suggests that John is not attractive. It was found that that endorsements of scalar implicature were anti-correlated with the degree of negative strengthening of the stronger scale-mate. The study thus provided evidence for the assumption that participants did not endorse the scalar implicature with certain triggers because they negatively strengthened the stronger term. Further, the authors presented additional analyses showing that the data by van Tiel et al. are consistent with a modified version of the uniformity assumption, once negative strengthening is taken into account.

The above study also found a potential explanatory role for factors including semantic distance and boundedness (i.e., the factors identified by van Tiel et al. as being significant predictors of scalar implicature rates). However, Benz et al. note that there was high degree of overlap between potential predictors in the stimulus material (e.g., between boundedness and grammatical category), making it difficult to draw firm conclusions as to the source of the observed effects.

### 1.3. Scale Structure, Adjective Meaning, and Implicature

The existing body of experimental research on scalar diversity has provided evidence that adjectives behave differently from other sorts of scalar items when it comes to the derivation of scalar implicatures. But it is less clear why this should be, or indeed the extent to which it is the case for all adjectival pairs or only certain salient subclasses.

The results of these previous studies also suggest that properties of the underlying measurement scales lexicalized by gradable adjectives (and perhaps items of other classes) play a role in determining the frequency at which they give rise to scalar implicatures<sup>2</sup>. Here too, there are a number of questions that remain to be explored.

As noted above, one factor found to be a significant predictor of scalar implicature rates is boundedness, namely whether or not the stronger member of a lexical scale denotes a scalar endpoint of some sort. The notion of boundedness is familiar from the literature on the semantics of gradable adjectives (see especially Kennedy and McNally, 2005; Kennedy, 2007), where it has been shown to explain a diverse range of combinatorial and interpretive phenomena. The central observation is that the measurement scales lexicalized by gradable adjectives may differ as to whether they have maximum and/or minimum points. This is claimed in particular to determine the interpretation of the adjective in its unmodified "positive" form. If the scale is lower closed, the corresponding adjective has an existential minimum standard (to be dirty is to have some amount of dirt); if it is upper closed, the adjective has a maximum standard (to be clean is to have a maximal degree of cleanness). Both of these are known as absolute interpretations. By contrast, if the scale is open on both ends, the adjective has a context-dependent relative standard (what counts as tall depends on the context and the sorts of entities under consideration).

Importantly, the bounded adjectival cases in van Tiel et al.'s study do not correspond to the class of maximum standard gradable adjectives from the adjectival literature, but rather involve a somewhat heterogeneous mix of measurement scale structures and adjective meanings. In some pairs, the weaker term is a relative gradable adjective, while the stronger term is a non-gradable adjective denoting the scalar endpoint; in the theory of Kennedy & McNally, this point is actually not part of the measurement scale lexicalized by the weaker adjective. Furthermore, in some such pairs (e.g., *cheap/free*) the stronger term denotes a scalar "zero" point, i.e., the complete absence of some property, whereas in others (e.g., *good/perfect*) it denotes some maximum point. Finally in other cases, the measurement scale itself is plausibly closed on both ends; the weaker term has a minimum standard existential interpretation while the stronger term is maximum-denoting (e.g., *allowed/obligatory, possible/certain*). In fact, van Tiel et al.'s experimental materials

<sup>2</sup>Note that the term "scale" is used in two distinct ways in this context, referring either to the scale of lexical alternatives involved in scalar implicature calculation or to the measurement scale that provides the semantic content of individual lexical items. When it is necessary to avoid confusion, we will use the terms "lexical scale" and "measurement scale" to distinguish these.

contain no “classic” examples of maximum-standard gradable adjectives. This gap makes it difficult to clearly diagnose the scope of the boundedness effect, and its source.

If the interpretation of the stronger scale-mate (namely whether or not it is endpoint-denoting) plays a role in determining the frequency at which scalar implicatures will arise, we might hypothesize that the interpretation of the weaker scale-mate will likewise play a role. And indeed, Benz et al. (2018a) find a difference between those lexical scales in which the weaker scale-mate has a greater-than-minimum or existential interpretation (L scales) and those where it invokes a mid-scale standard (M scales): only in the latter case does the negative correlation between scalar implicature and negative strengthening obtain. They note however that the set of M scales in the original materials from van Tiel et al. largely overlaps with the set of adjectival scales, while the L scales involve primarily items of other grammatical categories such as quantifiers and verbs; thus the potential role of this aspect of scale structure cannot be separated from that of grammatical category.

Put in different terms, the adjectival scales investigated to date in the scalar diversity literature largely involve relative gradable adjectives as the potential implicature trigger. Few minimum standard adjectives have been tested, and thus it is not yet known how this subclass will pattern with respect to the two types of implicature investigated here. One previous investigation by Leffel et al. (in press) showed that lower bounded adjectives like *late* and relative adjectives like *tall* are interpreted differently in the “not very” construction. In particular, Leffel et al. (in press) found that the utterance *John was not very late* yielded an inference to the positive form (that John was late) while the utterance *John is not very tall* was interpreted as meaning that John is not tall (with negative strengthening).

Finally, even among the relative gradable adjective pairs that make up the majority of the adjectival scales tested to date, there is diversity in the structures of the underlying measurement scales, and in how the individual members of the pairs relate to those scales. In particular, the items tested to date include both positive adjectives (e.g., *big/enormous*) and negative adjectives (e.g., *small/tiny*). As discussed above, positive vs. negative polarity has been argued to be relevant to the likelihood of negative strengthening, and we thus might expect it to play a role for scalar implicature too; but this has not yet been systematically investigated. Furthermore, in many of the pairs tested (e.g., *good/excellent*) the weaker term is a basic-level term while the stronger one is an extreme adjective (Morzycki, 2012); but in several cases (e.g., *adequate/good*), the weaker term describes something like a moderate degree of the property in question, while the stronger one is the basic-level term. Also as discussed earlier, Beltrama and Xiang (2013) found evidence for a lower level of scalar implicatures to the negation of an extreme adjective than to the negation of a mid-scale adjective; but the role of this factor as a potential predictor has not been taken into consideration in the more recent literature on scalar diversity.

In light of the issues discussed above, further research into the potential predictors of scalar diversity is needed, particularly as

it pertains to adjectival scales (see also a recent commentary by McNally, 2017).

## 1.4. Goals of the Current Study

The current study investigates the interplay of scalar implicature and negative strengthening for a broader and more balanced range of scalar adjectives. We have decided to focus on adjectival scales for several reasons. First, adjectives constituted the majority of items in van Tiel et al. (2016), and we wanted to further evaluate the claim that they generate low rates of implicature (Doran et al., 2012; Beltrama and Xiang, 2013). Second, the semantics of the class of gradable adjectives is well described and it is possible to tease apart factors related to the structure of the underlying measurement scales. Third, adjectives belong to the set of open class terms, thereby providing a rich set of items. In contrast to previous work, we include a much more varied set of adjectival scales, and code these on a fuller set of scalar properties that we hypothesize to be relevant to the availability of pragmatic inferences.

The first goal of our study is to determine whether the (anti-)correlation between scalar implicature and negative strengthening found by Benz et al. (2018a) for van Tiel et al.'s original items is also replicated for a wider range of adjectival scales. The second is to provide further insight into the predictors of variability in the rates of these inferences, with a focus on examining the role of factors relating to the underlying structure of the scales lexicalized by gradable adjectives.

## 2. EXPERIMENTS

### 2.1. Methods

#### 2.1.1. Participants

Participants with US IP addresses were recruited on Amazon's Mechanical Turk platform and were further screened for native language. In total, 220 native English speakers (mean age: 37.4, 95 female, 121 male, 4 gender information not given) took part in the study.

The experiments were conducted in accordance with the ethics policy of the Deutsche Forschungsgemeinschaft (DFG) under approval of grant Nr. BE 4348/4-1. Since the study involved a healthy adult population, no ethics consent was required according to institution's guidelines and national regulations. Participant's consent was obtained by virtue of survey completion and their data were fully anonymized.

#### 2.1.2. Materials

##### 2.1.2.1. Items

We created a set of 70 adjective pairs with weak and strong scale-mates<sup>3</sup>. We took all adjective pairs from the van Tiel et al. study (32) and added a further set of 38 adjective pairs to balance factors related to the scale structure of the adjectives. In particular, we added further absolute gradable adjectives (minimum standard and maximum standard), as well as more pairs where the

<sup>3</sup>The original list contained 71 pairs, but the pair *content/unhappy* was excluded from further analyses on the basis of diagnostics showing that the two terms are not on the same scale, but rather have opposite polarity.

Mary says:

*He is intelligent.*

Would you conclude from this that, according to Mary, he is not brilliant?

☐ Yes      ☐ No

**FIGURE 2** | Sample item of the scalar implicature task (based on van Tiel et al., 2016).

Mary says:

*He is not brilliant.*

Would you conclude from this that, according to Mary, he is not intelligent?

☐ Yes      ☐ No

**FIGURE 3** | Sample item of the negative strengthening task.

stronger scale-mate is non-extreme. **Tables A1, A2** in Appendix presents a list of all 70 adjective pairs. These were embedded in 7 separate tasks administered to 40 participants each (except for the politeness ratings which only involved 20 participants).

#### 2.1.2.2. Main tasks

The two main tasks employed the paradigm from van Tiel et al. (2016). Participants are presented with a scenario involving two characters, Mary and John, who make a series of statements. Their task is to decide whether a strengthened interpretation follows from a given statement. In the first task, participants were presented with the weaker term and had to indicate whether they endorse the negation of the stronger term, i.e., the scalar implicature. For example, Mary said: *John is intelligent* and participants were asked whether, according to Mary, John is not brilliant. **Figure 2** presents a sample display participants saw.

In the second main task, participants were asked whether the negation of the stronger term suggests the negation of the weaker term. For example, participants saw the statement *John is not brilliant* and were asked whether they conclude that John is not intelligent. The latter task is a measure of negative strengthening of the stronger scale-mate. **Figure 3** gives an example.

Two survey versions of the main tasks were created and administered to 20 participants each.

#### 2.1.2.3. Additional rating experiments

Additionally, we collected a variety of measures based on the methodology of van Tiel et al. (2016). First, we had participants rate the **semantic distance** between the statements involving the weaker and stronger scale-mate. In this task, participants were presented with a pair of two statements, one with the weak

term and one with the strong term. Participants were asked how much stronger the second statement is compared to the first one. They gave their answer on a 1–7 point Likert scale with 7 indicating that the second statement is much stronger and 1 that the statements are equally strong.

Second, we administered a **cloze task** to another set of participants, in order to measure association strength between the weaker and stronger terms. We used the open version of the task by van Tiel et al. in which participants had to mention three words that come into their mind upon seeing the statement with the weaker term. Participants' responses were then coded for the frequency of mentioning the stronger scale-mate. We employed a strict scheme for coding the responses, only taking into account exact mentions of the stronger scale-mate<sup>4</sup>.

Finally, participants also rated the kindness/politeness of statements involving the weaker term, the stronger term and the negated stronger term. Here we used the methodology of a previous study by Bonnefon et al. (2009). In each task, participants rated how nice the respective statement was on a 1–7 point scale with 7 indicating that the statement was very nice. This rating was included because negative strengthening has been discussed with respect to politeness considerations.

<sup>4</sup>We also computed the semantic similarity of the weak and strong term with latent semantic analysis (Landauer, 2006). For this analysis, we used the tool provided at <http://lsa.colorado.edu/> (pairwise comparison). Both the cloze task and the LSA analysis were intended as a measure of the association strength between the weaker and stronger alternative. However, since LSA values were not significant predictors for variability in scalar implicature and negative strengthening rates, we only kept the measure of cloze probability in the final statistical models reported in section 3.2.



**TABLE 1** | Overview of tasks.

Label	Task	Intended measure
Main task SI	Inference judgment (yes/no)	Scalar implicature
Main task NegS	Inference judgment (yes/no)	Negative strengthening
Semantic distance	Strength rating (1–7 scale)	Scale distinctness
Cloze task	Free word production	Association strength
Politeness weak	Kindness rating (1–7 scale)	Weak statement
Politeness strong	Kindness rating (1–7 scale)	Strong statement
Politeness “not” strong	Kindness rating (1–7 scale)	Negated strong statement

**Table 1** presents an overview of the different tasks we ran. All of these tasks, except for the politeness rating, were administered in two survey versions with different orders to 20 participants each.

#### 2.1.2.4. Annotation

In addition to the measures presented above, we annotated each pair on a range of scale-related properties, specifically the **boundedness** and **extremeness** of the stronger scale-mate, the **standard type** of the weaker scale-mate (minimum, relative, or maximum), and the **polarity** of the scale as a whole (positive or negative).

In making these annotations, the following diagnostics were used: A pair was coded as **upper bounded** if the stronger member of the pair denotes a scalar endpoint, as evidenced by compatibility with endpoint-oriented modifiers such as *almost*, *completely*, and *100 percent* (e.g., *completely clean* vs. ??*completely tall*). A pair was classified as **extreme** if the stronger member of the pair patterns as extreme using Morzycki’s [2012] test of compatibility with extreme adjectival modifiers such as *downright* and *flat-out* (e.g., *downright excellent* vs. ??*downright good*). Following the diagnostics of Kennedy and McNally (2005) and Kennedy (2007), a pair’s weaker member was classified as having a **minimum** standard if it is compatible with low-degree modifiers such as *slightly* and *a bit* (e.g., *slightly wet* vs. ??*slightly tall*), and if its negation entails a zero degree of the property in question; it was coded as having a **maximum** standard if it passes the tests for upper-boundedness described above, or shows other evidence of endpoint-orientation; and it was classified as **relative** otherwise. Note that pairs with a maximum-standard weaker scale-mate necessarily have a bounded (endpoint-denoting) stronger scale-mate, and represent cases of variation in the precision at which the standard is interpreted (e.g., *clean/spottless*). This will be relevant below.

Adjectival **polarity** proves to be the most complicated dimension to annotate. As discussed in Ruytenbeek et al. (2017), there are multiple notions of adjectival polarity, including morphological, dimensional, evaluative, and markedness-based ones, and individual tests do not apply equally well to all antonym pairs. We therefore followed those authors in implementing a step-wise classification, in which a series of tests were applied in sequence, as follows: (i) If the weaker member of the pair contains a negative morpheme, that pair was classified as negative. (ii) If

the pair is associated with a quantitatively measurable dimension, the adjective pairs associated with higher measurement values were classified as positive and those associated with lower measurement values were classified as negative, based on acceptability in the frame “something with a larger (smaller) number/amount of *x* is *Adj-er*”. Note that this test applies both to adjectives traditionally considered dimensional (e.g., *tall/short*: “something with a larger (smaller) number of inches of height is *taller (shorter)*”) as well as to those with more complicated relations to measurable dimensions (e.g., *dirty/clean*: “something with a larger (smaller) amount of dirt is *dirtier (cleaner)*”). (iii) For adjectives expressing value or taste judgments, an evaluative notion of polarity (“good” vs. “bad”) was applied; (iv) Tests (i)–(iii) left 9 pairs still unclassified (*damaged/broken*, *faulty/non-functional*, *sleepy/asleep*, *light/white*, *dark/black*, *special/unique*, *calm/unflappable*, *tired/exhausted*, *hungry/starving*). These were annotated for polarity based on the authors’ judgments. Note finally that this classification procedure identified some cases of conflict between dimensional and evaluative notions of polarity (e.g., *dirty* is dimensionally positive but arguably evaluatively negative). On account of our overall focus on the role of scale structural factors, we chose to prioritize the dimensional sense.

We also extracted the frequency of the weaker term, the stronger term and the negated stronger term from the Corpus of Contemporary American English (Davies, 2008). We calculated the relative frequency of the weaker and stronger term taking the logarithm of the frequency of the weaker divided by the stronger term (to make up for skewness of the distribution, see van Tiel et al., 2016). For negative strengthening, we took the logarithm of the frequency of the negated stronger term divided by that of the simple stronger term.

## 2.2. Results

### 2.2.1. Results of Main Tasks (SI and NegS)

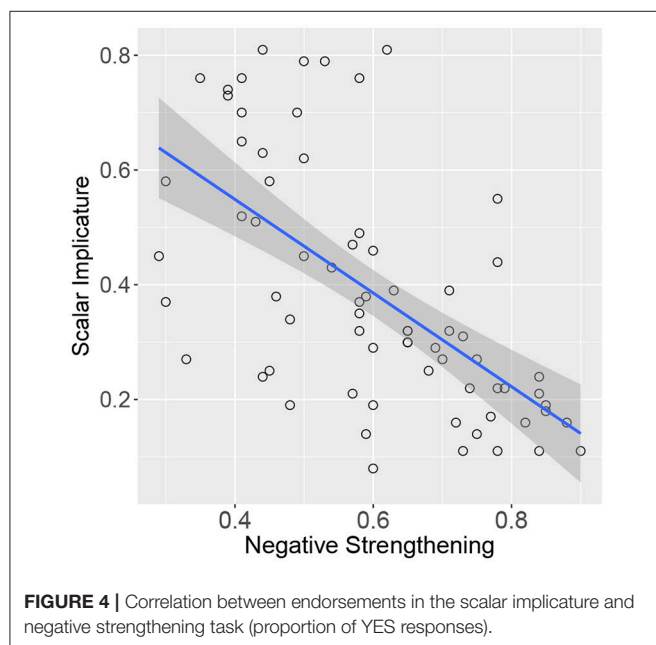
**Table 2** presents a sample of adjectives with different scale structures and their respective endorsement rates in the scalar implicature (SI) and negative strengthening (NegS) tasks. The results for all scales are presented in **Tables A1, A2** in the Appendix. A Pearson’s correlation test revealed that the two ratings were anti-correlated ( $r = -0.62$ ,  $p < 0.0001$ , see **Figure 4**). That is, the more likely participants applied negative strengthening to the stronger scale-mate, the less likely they were to endorse the scalar implicature.

### 2.2.2. Predicting Variability

We first calculated the mean values for each adjective pair in the seven different tasks. Then, we fit two linear regression models involving all predictors outlined above for the scalar implicature and the negative strengthening tasks respectively (see **Table 3**). The regression analysis showed that endorsements of the scalar implicature were higher for upper bounded scales ( $p < 0.01$ ), more distant scale-mates ( $p < 0.0001$ ), and higher for negative compared to positive scales ( $p < 0.05$ ). Conversely, extreme adjectives yielded lower endorsement rates compared to non-extreme ones ( $p < 0.0001$ ) and maximum standard weaker scale-mates lower rates than relative terms ( $p$

**TABLE 2 |** Example scales and their respective endorsement rates in the scalar implicature (SI) and negative strengthening (NegS) task.

Weak/strong term	Scale structure	SI	NegS
Cheap/free	Bounded rel neg non-extreme	0.76	0.41
Possible/certain	bounded min pos non-extreme	0.58	0.3
Clean/spotless	Bounded max neg extreme	0.27	0.75
Wet/soaked	Unbounded min pos extreme	0.24	0.44
Large/gigantic	Unbounded rel pos extreme	0.22	0.74
Scared/petrified	Unbounded rel neg extreme	0.14	0.75



<0.01)<sup>5</sup>. The multiple  $R^2$  of the SI model was 0.62 and the amount of explained variance for each predictor is listed in Table 3A.

The negative strengthening task showed the opposite pattern with lower endorsement rates for more distant scale-mates ( $p < 0.0001$ ) and higher rates for extreme adjectives ( $p < 0.01$ ). The negative strengthening rates were higher for maximum standard weaker scale-mates compared to relative ones ( $p < 0.05$ )<sup>6</sup>. The

<sup>5</sup>The initial model used relative weak terms as the reference level but we also computed a model with minimum standard adjectives as the reference level. Maximum standard adjectives yielded lower SI rates than minimum standard ones ( $p < 0.01$ ) while there was no difference between minimum and relative adjectives ( $p = 0.62$ ).

<sup>6</sup>Again we computed a second model with minimum standard adjectives as the reference level. Maximum standard adjectives yielded higher NegS rates than minimum standard ones ( $p < 0.05$ ) while there was no difference between minimum standard and relative adjectives ( $p = 0.36$ ).

**TABLE 3 |** Predictors of endorsements in (A) the scalar implicature and (B) negative strengthening task.

	Estimate	SE	t-value	p-value	$R^2$
<b>(A) SI</b>					
(Intercept)	-0.295	0.190	-1.547		
Weak min	-0.024	0.049	-0.495	0.623	
Weak max	-0.208	0.079	-2.652	0.010	0.060
Upper bounded	0.140	0.049	2.840	0.006	0.117
Semantic distance	0.132	0.028	4.763	0.000	0.136
Polarity neg	0.088	0.042	2.103	0.040	0.047
Extremeness	-0.206	0.052	-3.963	0.000	0.165
Politeness weak	0.017	0.034	0.513	0.610	0.004
Politeness strong	0.002	0.021	0.108	0.914	0.004
Cloze probability	-0.370	0.242	-1.526	0.132	0.069
Relative frequency	-0.024	0.019	-1.233	0.223	0.021
<b>(B) NegS</b>					
(Intercept)	1.276	0.316	4.038		
Weak min	-0.040	0.044	-0.905	0.370	
Weak max	0.146	0.069	2.121	0.038	0.081
Upper bounded	-0.073	0.044	-1.644	0.106	0.056
Semantic distance	-0.105	0.025	-4.151	0.000	0.184
Polarity neg	0.012	0.037	0.320	0.750	0.003
Extremeness	0.129	0.042	3.048	0.004	0.085
Politeness weak	-0.022	0.024	-0.930	0.357	0.008
Politeness not strong	-0.036	0.044	-0.833	0.408	0.011
Cloze probability	0.012	0.033	0.367	0.715	0.022
Relative frequency	0.263	0.216	1.219	0.228	0.071

multiple  $R^2$  of the NegS model was 0.52 and the amount of explained variance for each predictor is listed in Table 3B.

Finally, we assessed the effect of adding NegS rates as a predictor in the model for the SI task. The original model had an  $R^2$  of 0.62 and the new model with NegS as a predictor had a multiple  $R^2$  of 0.66; this improved fit was found to be significant (model comparison test with the anova function:  $p < 0.05$ ). The original factors extremeness, polarity and semantic distance remained as significant predictors in the new model but the difference between relative and maximum standard weaker scale-mates was marginal ( $p = 0.07$ ). The results of the model are presented in Table 4.

### 3. GENERAL DISCUSSION

#### 3.1. Summary of Main Findings

The current experiments showed that endorsements of scalar implicature are anti-correlated with the degree of negative strengthening of the stronger scale-mate. At the same time, we replicated the finding by van Tiel et al. (2016) that upper-bound denoting and semantically distant scale-mates yield higher endorsement rates in the scalar implicature task with our extended set of adjectival scales. Going beyond the latter study, we found that several additional factors related to the scale structure underlying the semantics of different adjective types

**TABLE 4 |** Model for endorsements in the scalar implicature with negative strengthening task as an additional predictor.

	Estimate	SE	t-value	p-value	R <sup>2</sup>
(Intercept)	0.091	0.247	0.369	0.713	
NegS	−0.339	0.145	−2.340	0.023	0.189
Weak min	−0.034	0.047	−0.719	0.475	
Weak max	−0.148	0.080	−1.856	0.068	0.043
Upper bounded	0.103	0.050	2.051	0.045	0.087
Semantic distance	0.097	0.031	3.170	0.002	0.086
Polarity neg	0.094	0.041	2.328	0.023	0.050
Extremeness	−0.171	0.052	−3.264	0.002	0.126
Politeness weak	0.007	0.033	0.223	0.825	0.003
Politeness strong	0.008	0.021	0.384	0.702	0.004
Cloze probability	−0.278	0.237	−1.172	0.246	0.050
Relative frequency	−0.019	0.019	−1.036	0.305	0.017

predict variability, in particular polarity, adjectival extremeness, and the nature of the standard invoked by the weaker scale-mate.

In our negative strengthening task, extremeness, and semantic distance also had an impact on endorsement rates but these effects went in the opposite direction. That is, negative strengthening rates were lower the more distant the scale-mates and, in turn, higher for extreme adjectives. We further found an effect for maximum standard weaker terms compared to relative and minimum standard terms in the negative strengthening task.

Finally, we computed a model for the scalar implicature task that took into account all factors (including negative strengthening) and with these factors we were able to account for 66 % of the observed variance.

### 3.2. Interaction Between Scalar Implicature and Negative Strengthening

At the beginning of this paper, we discussed the possibility that negative strengthening could mask the presence of scalar implicature in van Tiel et al. (2016)'s task. This hypothesis is supported by the finding of an anti-correlation between endorsement rates in the scalar implicature task and the negative strengthening task. In addition, negative strengthening rates were a significant predictor of endorsement rates in the scalar implicature task (and explained variance in addition to other significant predictors such as extremeness, polarity, and semantic distance). These findings provide evidence that, for some scales, participants did not endorse the scalar implicature due to the application of negative strengthening to the negated stronger scale-mate.

Looking at the endorsement rates in the two tasks in comparison, however, there are some scales which received high negative strengthening rates as well as high scalar implicature rates. We therefore take our findings to indicate that negative strengthening is one among many factors which determines whether a scalar implicature is derived. This is also evident in the fact that scale structure factors such as boundedness, polarity, and extremeness remained significant predictors even when negative

strengthening was taken into account. Further, for some scales, scalar implicature is robust and remains unaffected by negative strengthening while for other scales the propensity of triggering negative strengthening seems to be higher.

More generally, our findings corroborate the assumption that quantity and manner implicatures can both occur for the same pairs of lexical items. In other contexts, however the two might stand in competition with each other (see for example Levinson, 2000). Hence, our findings motivate further theoretical research into negative strengthening and how exactly different kinds of implicature are related to each other and how they interact in a specific context. In Gotzner et al. (in preparation), theoretical underpinnings of the attested interaction between scalar implicature and negative strengthening.

### 3.3. Scale Structure

We found that several factors related to scale structure had an effect on the rates at which the two kinds of inferences were generated. In what follows, we consider each of these in turn.

#### 3.3.1. Boundedness and the Absolute/Relative Distinction

In the present study, we found that participants were more likely to endorse a scalar inference if the lexical scale of alternatives was upper bounded, meaning that the stronger scale-mate denotes an endpoint on some underlying measurement scale. Thus we replicated van Tiel et al.'s finding that boundedness is a significant predictor of scalar implicature rates. However, we also found that it is not all upper bounded lexical scales that behave this way. Specifically, we observed low rates of scalar implicature endorsement when the weaker scale-mate is itself a maximum standard gradable adjective, while the stronger term denotes that standard interpreted at a higher level of precision (e.g., *clean/spotless*, *dry/parched*). Thus it is not upper boundedness *per se* that is associated with higher implicature rates, but rather those lexical scales in which an endpoint-denoting stronger term stands in opposition to a minimum standard or relative standard weaker term.

van Tiel et al. (2016) discussed the boundedness effect in terms of scale distinctness, a broader concept that encompasses also semantic distance (which also had a significant effect; see below). That is, if the stronger scale-mate denotes an upper bound it is more clearly distinguishable from the weaker term, and therefore participants may be more likely to derive a scalar implicature. While we think that this characterization is compatible with our findings, we would also like to entertain the possibility that scale boundedness plays a more fundamental role in implicature computation, though perhaps in different ways for different sorts of adjectival pairs.

As discussed above, the literature on adjectival semantics (Kennedy and McNally, 2005; Kennedy, 2007) draws a distinction between two types of gradable adjectives: absolute gradable adjectives, which lexicalize measurement scales that are closed on one or both ends, with those endpoints providing a fixed standard of comparison for the adjective; and relative gradable adjectives, which lexicalize open scales, and thus have

contextually determined standards. Psycholinguistic work has shown that listeners necessarily consider the standard value as part of the comprehension process of a sentence containing an absolute adjective (the “Obligatory Scale” hypothesis entertained by Frazier et al., 2008). It is plausible that similar factors might make scalar endpoints, and thus the adjectives that refer to them, particularly salient alternatives for the purposes of scalar implicature calculation.

Such an explanation holds potential in particular for adjectives such as *allowed/obligatory* and *possible/certain*, which arguably lexicalize totally closed scales. However, many of the upper-bounded adjectival pairs included in our study involved a relative gradable adjective as the weaker scale-mate (e.g., *cheap/free*, *scarce/unavailable*, *good/perfect*). In the adjectival literature, these are analyzed as lexicalizing totally open scales; the stronger term may then be analyzed as denoting a point that is actually not on the scale lexicalized by the weaker term. In these cases, we see it as possible that the use of the weaker scale-mate itself implies or even presupposes that the value described is on the non-endpoint portion of the scale, without any need for reference to a stronger potential alternative.

Factors relating to the lexical semantics of adjectives are also relevant in the case of unbounded scales of alternatives, that is, scales where the stronger scale-mate is not endpoint denoting. In most such pairs, the stronger term has a relative interpretation, according to which the standard is fixed contextually, with respect to the given comparison class (Solt, 2011; Solt and Gotzner, 2012). The values on the scale depend heavily on the noun that the adjective modifies and other contextual factors (see Rips and Turnbull, 1980 for psycholinguistic evidence that finding a standard for relative adjectives involves extra computation when the reference class is not mentioned).

For the computation of scalar implicature this may have the following consequences: (1) participants may not compute a scalar implicature to the negation of a stronger scale-mate with a relative interpretation because the stronger term does not stand in competition with the weaker term, i.e., because the stronger term might not come to mind in the same context or be relevant for the same comparison class.

Additionally, (2) relative adjectives may be less prone to implicature derivation because people have difficulty identifying the borderline for which the terms apply. Such a proposal has been made by Leffel et al. (in press), who formulated a constraint on implicatures such that they are not drawn if a borderline contradiction would be the result. Leffel et al. (in press) showed that lower bounded adjectives like *late* and relative adjectives like *tall* give rise to distinct inference patterns. For example, the utterance *John was not very late* yielded an inference to the positive form (that John was late) while the utterance *John is not very tall* was interpreted as meaning that John is not tall (with negative strengthening). Based on these data, Leffel and colleagues proposed a constraint according to which implicatures are not derived if they lead to a borderline contradiction. By the same token, we may hypothesize that participants in our study were reluctant to draw an inference from, say, *intelligent* to *not brilliant* or *big* to *not enormous* because they were uncertain as to where the scalar boundary for the stronger term lies.

Finally, while we found a difference in implicature rates between scales with endpoint-denoting and non-endpoint-denoting stronger scale-mates, and those with maximum standard vs. non-maximum-standard weaker scale-mates, there was perhaps surprisingly no difference between lexical scales with minimum-standard weaker terms and those with relative weaker terms (on either the scalar implicature task or the negative strengthening task). We see this as an issue requiring further investigation, in particular since the study by Leffel et al. (in press) found these two classes to behave differently with regards to a related variety of pragmatic inference.

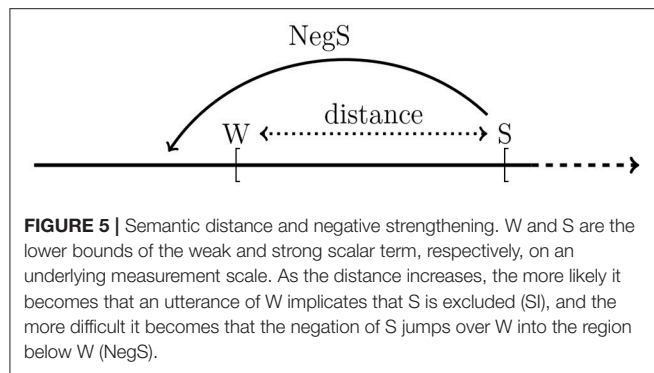
### 3.3.2. Extremeness

In our study, there was an additional effect related to scale structure that is relevant in this discussion. We found that extreme adjectives obtained lower implicature rates compared to non-extreme ones. If orientation toward the endpoint was the crucial factor in implicature computation, then extreme adjectives should yield higher implicature rates, contrary to what we have found. We assume that the effect of adjectival extremeness is of a different nature. Extreme adjectives have a particular semantics and they behave peculiarly in certain respects (see especially Morzycki, 2012 who entertains the view that extreme adjectives signal that the degree lies outside of the contextual range). Extreme adjectives may only be used in specific contexts and therefore again it might not arise as a competitor alternative out of the blue (see also Beltrama and Xiang, 2013). In turn, the use negated extreme adjectives may indicate that the situation is non-stereotypical thereby encouraging negative strengthening, as we have found in our negative strengthening experiment. This would be in line with the account of negative strengthening by Horn (1989) and Krifka (2007) according to which more complex expressions are used for less stereotypical instances. For example, the utterance *John is not tall* will tend to be used to describe cases that fall under the literal meaning of *short* (since *short* and *not tall* have the same literal meanings), but which are greater in height than the ones described by the utterance *John is short*.

### 3.3.3. Semantic Distance

Negative strengthening and scalar implicature are differentially affected by semantic distance: as semantic distance between weak and strong scale-mates increases the SI-rate increases and the NegS-rate decreases. There is a suggestive explanation of this behavior if semantic distance is considered as distance between the lower bounds of the weak and the strong scale-mate on an underlying measurement scale, see Figure 5. The semantics of the weak term (W) always includes that of the strong term (S), however, the most likely value on a measurement scale that the speaker had in mind when producing W and S may be some distance apart. We may think of semantic distance as the distance between the lower bounds of the intervals defined by W and S. As the distance between the lower bounds increases, the more likely it becomes that the speaker means by saying W that S is not the case, and, hence, it becomes more and more likely that subjects answer that saying W implies not-S (SI), i.e., that a scalar implicature occurs. Negative strengthening (NegS) is explained





as a *blocking* phenomenon (Horn, 1989; Levinson, 2000; Blutner, 2004; Krifka, 2007) which can be understood as a consequence of Horn's (1984) *principle of the division of pragmatic labor*, according to which a speaker who has a choice between a marked and an unmarked expression will prefer the unmarked one and, hence, signal by a choice of the marked expression that the unmarked one is not applicable. Hence, the existence of the unmarked expression *blocks* parts of the semantic meaning of the marked expression. In the case of scale mates (S,W) this means that if a speaker uses the marked expression not-S, then the existence of the unmarked W blocks not-S from having the meaning that could have been expressed by W. If the distance between W and S widens, W has to block a larger and larger interval on the underlying measurement scale, and it may become more and more improbable that W succeeds in doing this. As a result, the rate of negative strengthening will decrease with increasing semantic distance.

### 3.3.4. The Role of Scale Structure in Implicature Computation

Overall, we take our results to indicate that scale structure associated with the semantics of different adjectives systematically encourages or blocks certain inferences (see also Leffel et al., in press). We hypothesize that scale structure puts constraints on the range of potential values and thereby determines the alternatives used in implicature computation. Thus far, insights from the lexical semantics of scales have not been taken into account in the theory of implicature. Our investigation highlights the role of scale structure in pragmatic strengthening.

## 3.4. The Role of Polarity and Politeness

As mentioned in the introduction, an asymmetry between positive and negative adjectives has been taken as evidence that negative strengthening is related to politeness considerations. Evidence for such an asymmetry between positive and negative adjectives was found in the experimental studies by Ruytenbeek et al. (2017) but previous experimental studies cited therein provided mixed results.

In the current study, we did not find any evidence that politeness ratings predicted variability in scalar implicature or negative strengthening rates. We did, however, find that polarity

itself is an independent predictor of scalar implicature rates (though not of negative strengthening). Specifically, we found higher implicature rates with negative antonyms compared to positive ones.

Recall from section 3.1 that we chose to prioritize a dimensional notion of adjectival polarity, according to which the positive member of an antonym pair is the one that corresponds to a higher amount of some measurable property. As we noted above, this classification leads to some discrepancies with the evaluative notion of polarity. For example, according to the dimensional point of view the adjective *dirty* is the positive antonym (since it involves greater amounts of dirt), while *clean* is the negative one (involving lesser amounts of dirt). In contrast, the evaluative notion of polarity would result in exactly the opposite classification since typically *clean* seems to be considered a desirable property. We also ran some additional models in which we restricted our analysis to the clear cut cases of the dimensional view of polarity and this analysis replicated the main results for the effect of polarity in the scalar implicature task (while again no such effect was present in the negative strengthening task).

In fact, the negative adjectives that yielded the highest levels of scalar implicatures in our study included many for which the stronger scale-mate denotes the complete absence of some quantity or property (e.g., *inaudible*, *extinct*, *free*, *unavailable*). We hypothesize that there is something about this sort of meaning that is particularly likely to give rise to implicatures, and thus that our findings in this area are again primarily related to scale-based factors rather than socially or politeness-motivated ones.

Another way in which polarity may play a role (independent of politeness) in the derivation of scalar implicature is by introducing certain presuppositions. Cruse (1986) discusses differences in the scale structure between positive and negative members. He notes that for the interpretation of positive adjectives like *good* the whole scale is relevant while in the case of negative adjectives like *bad*, the underlying question is to put the predicate on the "badness scale." For this reason, it could be the case that when a speaker utters a sentence like *The movie was bad*, listeners are more likely to derive the inference that the movie was bad but not terrible. In effect, the presuppositions of the predicate may constrain the alternatives available for scalar implicature. Since positive members do not tend to introduce a presupposition it is less clear which alternatives are relevant and therefore hearers may be less likely to derive a scalar implicature.

We conclude that there is some evidence that polarity plays a role in implicature computation but the specific contributions to scalar implicature and negative strengthening need to be determined by further experimental research. It has to be kept in mind that our study was purely correlational (in contrast to other studies demonstrating politeness effects in scalar implicature computation such as Bonnefon et al., 2009). To discover effects of politeness, test sentences may have to be embedded within a rich conversational context in future studies and politeness may have to be manipulated directly in the experimental setup.

### 3.5. Methodological Issues

In a commentary, McNally (2017) argues that the methods used by van Tiel et al. were too crude to (i) detect certain implicatures and (ii) detect effects of the parameters they tested. Essentially, the problem McNally discusses is that adjectives are polysemous and in the absence of a context participants may construct the meaning on a fly and not think of the intended pair as scale-mates. This criticism also applies to the current study and it stresses the need to present test sentences within a conversational context. Our investigation particularly motivates further research into the impact of scale structure on implicature derivation. Yet investigating how a large variety of scales behave within an enriched communicative context has to be left to future research. One experimental paradigm which might be useful for this endeavor is the action-based task by Gotzner and Benz (2018), and its interactive version (Benz et al., 2018b), which has been implemented for the quantifier *some* and the connective *or* (Benz and Gotzner, 2017). The advantage of this paradigm is that utterances are embedded in a communicative situation and candidate readings are made relevant. In conclusion, it is vital to move to an experimental paradigm that introduces a context with respect to which statements involving scalar terms should be interpreted.

## 4. CONCLUSIONS

Our research revealed an interaction between scalar implicature and negative strengthening, which are based on distinct conversational principles, the Q and R principle, respectively (Horn, 1989; Levinson, 2000). Specifically, participants were less likely to endorse a scalar implicature when they applied negative strengthening to the stronger scale-mate. Importantly, we observed that gradable adjectives do not generally lead to low rates of scalar implicature. Rather, different factors determine which inferences arise with negative strengthening being one of them.

## REFERENCES

- Beltrama, A., and Xiang, M. (2013). "Is excellent better than good? Adjective scales and scalar implicatures," in *Sinn und Bedeutung*, Vol. 17 (Paris), 81–98.
- Benz, A., Bombi, C., and Gotzner, N. (2018a). "Scalar diversity and negative strengthening," in *Proceedings of Sinn und Bedeutung 22*, Vol. 1, eds U. Sauerland and S. Solt (Berlin: ZAS), 191–204.
- Benz, A., and Gotzner, N. (2017). "Embedded disjunctions and the best response paradigm," in *Sinn und Bedeutung 21* (Edinburgh).
- Benz, A., Gotzner, N., and Raithel, L. (2018b). "Embedded implicature in a new interactive paradigm," in *Proceedings of Sinn und Bedeutung 22*, Vol. 1, eds U. Sauerland and S. Solt (Berlin: ZAS), 205–221.
- Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *J. Semant.* 17, 189–216. doi: 10.1093/jos/17.3.189
- Blutner, R. (2004). "Pragmatics and the lexicon," in *The Handbook of Pragmatics*, eds L. Horn and G. Ward (Oxford: Blackwell Publishing), 488–514.
- Bonnefon, J. F., Feeney, A., and Villejoubert, G. (2009). When some is actually all: scalar inferences in face-threatening contexts. *Cognition* 112, 249–258. doi: 10.1016/j.cognition.2009.05.005
- Brown, P., and Levinson, S. C. (1987). *Politeness: Some Universal in Language Usage*, Vol. 4 *Studies in Interactional Sociolinguistics*. Cambridge: Cambridge University Press.
- Chierchia, G. (2004). "Scalar implicatures, polarity phenomena, and the syntax / pragmatics interface," in *Structures and Beyond*, ed A. Belletti (Oxford: Oxford University Press), 39–103.
- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.
- Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 560 Million Words, 1990–Present*. Provo, UT: Brigham Young University. Available Online at: <https://corpus.byu.edu/coca/>
- Doran, R., Baker, R. E., McNabb, Y., Larson, M., and Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *Int. Rev. Pragmat.* 1, 1–38. doi: 10.1163/187730909X12538045489854
- Doran, R., Ward, G., McNabb, Y., Larson, M., and Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88, 124–154. doi: 10.1353/lan.2012.0008
- Frazier, L., Clifton C., and Stolterfoht, B. (2008). Scale structure: processing minimum standard and maximum standard scalar adjectives. *Cognition* 106, 299–324. doi: 10.1016/j.cognition.2007.02.004
- We showed that the most important predictors explaining differences across triggers was the underlying scale structure of the adjectives we tested (in particular boundedness, semantic distance, extremeness, and polarity). Thus far, insights concerning the semantics of scales have not been well integrated into theories of scalar implicature and negative strengthening. Our findings highlight that adjectives with different scale structure give rise to distinct inference patterns. For this reason, we propose that the semantics of different scales should be a central aspect of study in theories of implicature.

## AUTHOR CONTRIBUTIONS

NG has carried out the experiments, analyzed the data, and written the first draft of the manuscript. NG, SS, and AB have all contributed to designing the experiments, annotating the factors and editing subsequent drafts of the manuscript.

## ACKNOWLEDGMENTS

We thank Richard Breheny, Eve Clark, Herbert Clark, Judith Degen, Napoleon Katsos, Manfred Krifka, Jacopo Romoli, Uli Sauerland, Chao Sun, Bob van Tiel as well as the audiences of SALT 28 at MIT for insightful discussion. We are also grateful to Henry Salfner for assistance with the experiments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the Xprag.de Initiative (Grant Nr. BE 4348/4-1), a grant awarded to SS (grant Nr. SO 1157/1-2), and the Bundesministerium für Bildung und Forschung (BMBF) (Grant Nr. 01UG1411). The publication of this article was funded by the Open Access Fund of the Leibniz Association.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01659/full#supplementary-material>

- Gotzner, N., and Benz, A. (2018). The best response paradigm: a new approach to test implicatures of complex sentences. *Front. Commun.* 2:21. doi: 10.3389/fcomm.2017.00021
- Gotzner, N., and Romoli, J. (2017). The scalar inferences of strong scalar terms under negative quantifiers and constraints on the theory of alternatives. *J. Semant.* 35, 95–126. doi: 10.1093/jos/ffx016
- Horn, L. R. (1984). “Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature,” in *Meaning, Form, and Use in Context: Linguistic Applications*, ed D. Schiffrin (Washington, DC: Georgetown University Press), 11–42.
- Horn, L. R. (1989). *A Natural History of Negation*. Chicago, IL: University of Chicago Press.
- Kennedy, C. (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguist. Philos.* 30, 1–45. doi: 10.1007/s10988-006-9008-0
- Kennedy, C., and McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81, 345–381. doi: 10.1353/lan.2005.0071
- Krifka, M. (2007). “Negated antonyms: creating and filling the gap,” in *Presupposition and Implicature in Compositional Semantics*, eds U. Sauerland and P. Stateva (Houndmills: Palgrave Macmillan), 163–177.
- Landauer, T. K. (2006). *Latent Semantic Analysis*. Wiley Online Library.
- Leffel, T., Cremers, A., Gotzner, N., and Romoli, J. (in press). Vagueness in implicature: the case of modified adjectives. *J. Semant.*
- Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicatures*. Cambridge, MA: MIT Press.
- McNally, L. (2017). “Scalar alternatives and scalar inference involving adjectives: a comment on van Tiel, et al. (2016),” in *Asking the Right Questions: Essays in Honor of Sandra Chung*, eds J. Ostrove, R. Kramer, and J. Sabbagh (Santa Cruz, CA: US Santa Cruz Previously Published Works), 17.
- Morzycki, M. (2012). Adjectival extremeness: degree modification and contextually restricted scales. *Nat. Lang. Linguist. Theory* 30, 567–609. doi: 10.1007/s11049-011-9162-0
- Rips, L. J., and Turnbull, W. (1980). How big is big? relative and absolute properties in memory. *Cognition* 8, 145–174.
- Romoli, J. (2012). *Soft But Strong. Neg-Raising, Soft Triggers, and Exhaustification*. Ph.D. thesis, Harvard University.
- Ruytenbeek, N., Verheyen, S., and Spector, B. (2017). Asymmetric inference towards the antonym: experiments into the polarity and morphology of negated adjectives. *Glossa* 2:92. doi: 10.5334/gjgl.151
- Simons, M., and Warren, T. (2018). A closer look at strengthened readings of scalars. *Q. J. Exp. Psychol.* 71, 272–279. doi: 10.1080/17470218.2017.1314516
- Solt, S. (2011). “Notes on the comparison class,” in *Vagueness in Communication (ViC2009), Revised Selected Papers (LNAI 6517)*, eds R. Nouwen, R. van Rooij, U. Sauerland, and H.C. Schmitz (Berlin; Heidelberg: Springer), 189–206.
- Solt, S., and Gotzner, N. (2012). “Experimenting with degree,” in *Proceedings of SALT, Vol. 22* (Chicago, IL), 353–364.
- van Tiel, B., van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *J. Semant.* 33, 107–135. doi: 10.1093/jos/ffu017

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Gotzner, Solt and Benz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# A Link Between Local Enrichment and Scalar Diversity

Chao Sun<sup>1,2</sup>, Ye Tian<sup>3</sup> and Richard Breheny<sup>2\*</sup>

<sup>1</sup> Leibniz-Centre General Linguistics, Berlin, Germany, <sup>2</sup> Division of Psychology and Language Sciences, University College London, London, United Kingdom, <sup>3</sup> Amazon Research, Cambridge, United Kingdom

## OPEN ACCESS

### Edited by:

Anne Colette Reboul,  
Claude Bernard University Lyon 1,  
France

### Reviewed by:

Emmanuel Chemla,  
UMR8554 Laboratoire de sciences  
cognitives et psycholinguistique  
(LSCP), France  
Thomas Castelain,  
Universidad de Costa Rica,  
Costa Rica

### \*Correspondence:

Richard Breheny  
r.breheny@ucl.ac.uk

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 27 April 2018

**Accepted:** 10 October 2018

**Published:** 01 November 2018

### Citation:

Sun C, Tian Y and Breheny R  
(2018) A Link Between Local  
Enrichment and Scalar Diversity.  
Front. Psychol. 9:2092.  
doi: 10.3389/fpsyg.2018.02092

Several recent studies have shown that different scalar terms are liable to give rise to scalar inferences at different rates (Doran et al., 2009, 2012; van Tiel et al., 2016). A number of potential factors have been explored to account for such *Scalar Diversity*. These factors can be seen as methodological in origin, or as motivated by widely discussed analyses of scalar inferences. Such factors allow us to explain some of the variation, but they leave much of it unexplained. In this paper, we explore two new potential factors. One is methodologically motivated, related to the choice of items in previous studies. The second is motivated by theoretical approaches which go beyond the standard Gricean approach to pragmatic effects. In particular, we consider *dual route* theories which allow for scalar inferences to be explained either using ‘global’ pragmatic derivations, like those set out in standard Gricean theory, or using local adjustments to interpretation. We focus on one such theory, based on the Bayesian Rational Speech Act approach (RSA-LU, Bergen et al., 2016). We show that RSA-LU predicts that a scalar term’s liability to certain kinds of local enrichment will explain some Scalar Diversity. In three experiments, we show that both proposed factors are active in the scalar diversity effect. We conclude with a discussion of the grammatical approach to local effects and show that our results provide better evidence for dual route approaches to scalar effects.

**Keywords:** scalar implicature, scalar diversity, scale homogeneity, local enrichment, lexical uncertainty

## INTRODUCTION

### The Scalar Diversity Phenomenon

Recent experimental studies investigated the rates at which scalar expressions of different lexical categories give rise to scalar inferences (SIs) (Doran et al., 2009, 2012; Beltrama and Xiang, 2012; van Tiel et al., 2016). It has been found in these studies that different scalar expressions give rise to SIs at different rates. van Tiel et al. (2016) employed an inference paradigm to test participants’ interpretation of statements containing scalar expressions. Several classes of scalar expressions were examined including quantifiers (e.g., <all, some>), modals (<certainly, possibly>), adjectives (<beautiful, pretty>) and verbs (<dislike, loathe>). **Figure 1** is an example of an item (van Tiel et al., 2016: Experiment 2). Participants read a statement uttered by a character. Then they were asked whether the speaker implied the negation of the stronger statement in which scalar expression was replaced by its stronger scale mate. For example, when the character states that the student is intelligent, participants are asked whether, according to the speaker, the student is not brilliant. A ‘Yes’ response indicates that participants drew the SI and a ‘No’ response indicates that the inference was unavailable.



van Tiel et al. (2016) found significant variation in the derivation rates of SIs across different scalar expressions, ranging from 4 to 100%. Quantifiers and modal expressions generated SIs more frequently than adjectives and verbs. Moreover, while quantifiers and modal expressions consistently gave rise to SIs, there was much greater variability among adjectives and verbs. These results were consistent with those reported in Doran et al. (2009). The scalar diversity effect has been replicated in several studies that have used different procedures and that also provided more context for the target utterance (see Experiment 1 below, also Simons and Warren, 2018; Sun and Breheny, 2018a).

Scalar inference is widely seen as a specific instance of conversational implicature (Horn, 1972; Grice, 1975; Gazdar, 1979; Geurts, 2010). Implicatures are contextual implications of what the speaker literally says, which are derived on the basis of expectations speakers and listeners have about each other. A scalar implicature for the experimental item, 'The student is intelligent,' mentioned above, would be that the student is not brilliant. It is widely agreed that the underlying meaning of 'intelligent' is such that someone counts as intelligent if their intellectual capacities place them anywhere at or above some standard of such capacities. Another scalar term whose meaning relates to the same scale may be anchored to a higher point. This would be the case with 'brilliant.' Thus the student being brilliant is consistent with a literal assertion of 'The student is intelligent.' The standard Gricean explanation for the SI, *the student is not brilliant* is based on the idea that interlocutors expect each other to be as informative, or specific as is relevant in context (see for example Geurts, 2010). From this expectation, one can reason to the conclusion that, according to a speaker who used 'intelligent,' they do not consider the student brilliant. According to the design of van Tiel et al.'s (2016) study, all of the pairs of scalar terms have literal meanings with this scalar property. That is, the term that is not mentioned picks out a higher point on a scale than the one that is mentioned in the speaker's utterance. Thus, for all of the items used, the standard approach implies that a SI could be available. Although this standard account does not predict that there should be no scalar diversity, it does not predict that there will be diversity; at least not without further assumptions. For instance, there could be differences in terms of the relation between the weaker term uttered and the stronger alternative that needs to be evoked in order to derive the implication. Thus the interest in the scalar diversity phenomenon surrounds the question of what would explain this great variation in rates of 'Yes' response for different scalar terms.

In this paper, we will approach our discussion of factors responsible for scalar diversity in terms of their being either methodologically or theoretically motivated. Among theoretically motivated factors, we consider factors suggested by the standard Gricean theory and those that would follow from an augmented standard theory, which accommodates the widely acknowledge fact of local pragmatic effects.

In the following sub-section, we review empirical work so far presented that has accounted for some of the scalar diversity effect. Here, we introduce a new methodological factor, related to the polysemy of many of the scalar terms. We then introduce the idea that certain 'local' effects are not explained by the standard

John says:

*The student is intelligent.*

Would you conclude from this that,  
according to John, she is not brilliant?

Yes ☐

No ☐

**FIGURE 1 |** Sample item from van Tiel et al. (2016) – Experiment 2.

Gricean mechanism. We discuss an account of this phenomenon within the Bayesian, Rational Speech Act framework. We show how this RSA framework predicts scalar diversity to the extent that scalar terms are susceptible to local pragmatic processes.

In the second part of the paper, we present three experiments. The first is a replication of van Tiel et al.'s (2016) study. The second addresses the methodological problem of polysemy of scalar terms. The third tests the prediction concerning the relation between local enrichability and scalar diversity.

## Accounting for Scalar Diversity

If we approach the results of van Tiel et al.'s (2016) study from the standard Gricean perspective, one potential factor that may contribute to the scalar diversity effect is the lack of context in the experimental items. If we reconsider the item in **Figure 1** above, we can see that the utterance is presented without context. It is widely agreed that, from a Gricean perspective, stronger alternatives should only be considered for SI if that alternative is somehow relevant in context. Several experimental studies have shown that participants are able to infer implicit relevant context with the presentation of an experimental stimulus (Breheny et al., 2006; Bergen and Grodner, 2012). While all the scalar items are tested in van Tiel et al.'s (2016) study without context, it could be that items differ in the extent to which the relevant context can be inferred for different scalar terms. van Tiel et al. (2016) consider this possibility and dismiss it as likely to be an explanation for Scalar Diversity. Their case is supported to some extent by evidence from Doran et al. (2009). In that study, the sentence containing a scalar term is presented in explicit contexts that make the more informative alternative relevant and in explicit contexts that do not. Doran et al. (2009) report that rates of SI are affected by this contextual manipulation for their adjective scales but not for their quantifier scales (e.g., involving 'some'). But even in supportive context, Doran et al. (2009) found that rates for quantifiers were higher than for adjectives. Thus, the presence of explicit supportive context lessens the difference between scale types, it does not eliminate it. Further support for this conclusion comes from a corpus study reported in Sun and Breheny (2018a). Here participants read items selected from a corpus that had a wide variety of contexts. Again, the scalar diversity effect was lessened by the richer contexts associated with the items, but not eliminated.

van Tiel et al. (2016) explored a range of other explanations for the variability which they found. These explanations are

motivated by standard approaches to SI since they focus on the relation between the scalar term used and its alternative. van Tiel et al. (2016) hypothesized that the availability of the stronger alternative and the distinctness of the scale-mate may account for some of the variability in inference rates. The availability of the stronger alternative was measured in four parameters including association strength, grammatical class (open/closed), semantic relatedness and relative frequency of the scale-mate. One motivation for exploring availability might be that pairs of scalar terms may be more or less strongly associated with one another and this might be a factor in Scalar Diversity. However, in a regression analysis, van Tiel et al. (2016) found that none of the four parameters related to availability could independently explain scalar variability. This finding is corroborated in the study reported below, and in Sun and Breheny (2018a). A caveat should be entered at this point regarding measures of association. These have all been tested against the result of studies like van Tiel et al.'s (2016) inference task where the task stimuli mention the stronger scalar term as well as the weaker one. That is, in **Figure 1** above, 'brilliant' is mentioned as well as 'intelligent'; 'all' is mentioned as well as 'some'; and so forth. By mentioning the stronger scalar term ('brilliant', 'all', etc.) the task design may neutralize any difference in salience that might antecedently exist among scalar pairs. Thus it is possible that differences in association among scalar pairs could contribute to the scalar diversity effect, but that would be on top of other factors at play in the results reported to date.

The second kind of factor that van Tiel et al. (2016) consider is the distinctness of the scale-mate. Specifically, they sought a measure semantic distance (i.e., the difference in the perceived strengths between the pairs of scalar terms) and 'boundedness' (i.e., whether the underlying scale contained an endpoint). In contrast to measures of association, a regression analysis showed that semantic distance and boundedness did independently account for a significant amount of variance, where boundedness accounted for over three times more variance than did semantic distance.

Together, all of the measures explored by van Tiel et al. (2016) accounted for less than half of the variance, leaving a large amount of variation unexplained. Factors to do with the relation between scalar term and its alternative are the ones that are clearly suggested by the standard Gricean approach to SI. van Tiel et al. (2016) suggest that the availability of the stronger alternative and the distinctness of scale-mate are the only plausible factors that they could think of. Their conclusion is that the rest of the variation in inference rates among scalar terms must be unsystematic. In the rest of this section, we discuss two other kinds of factor motivated by considerations beyond standard Gricean theory.

## Methodological Factors

The first thing to consider about the scalar diversity effect is that there might be factors related to the methods used in these studies that contribute to the effect. One such factor is identified in Benz et al. (2017). This relates to the phenomenon known as negative strengthening. A negated scalar term might not simply denote the complement of its positive counterpart but may be understood

with a strengthened meaning. For example, 'not tall' is often understood not simply as denoting the set of things that are not at or above the contextual reference point in height, but somewhat below this standard. Negative strengthening is relevant to the methods used in van Tiel et al. (2016). Consider for example the item in **Figure 1**. The participant is asked to judge if the speaker thinks the student is not brilliant. To the extent that the scalar term under negation may undergo negative strengthening, the participant may respond negatively on the basis that 'not brilliant' is understood to mean somewhat less than brilliant, e.g., stupid. Benz et al. (2017) provide some evidence that adjective terms are more susceptible to negative strengthening and so this may have been a factor in van Tiel et al.'s (2016) results. However, it is not likely to be the sole remaining factor since other studies have probed for SIs without this kind of stimulus and still found the scalar diversity effect. For example, Sun and Breheny (2018a) employ the paraphrase task from Degen (2015). This task asked the participant whether 'intelligent but not brilliant' would be a good paraphrase for 'intelligent' in a given item. Here, there is no conflict with a negative strengthening inference.

In this paper, we wish to explore an issue related to items used in van Tiel et al.'s (2016) studies and others. This has to do with how homogeneous the senses of the scalar terms are. The relevant concept here is that scalar terms, such as 'brilliant' can be highly polysemous. 'Brilliant' can be understood as related to an underlying intelligence scale, but it can also be understood to be related to other scales to do with personality, such as kindness, or with other skills, as in a brilliant actor. Consider also the scale <unsolvable, hard> taken from van Tiel et al. (2016). 'hard' has a sense related to difficult. Under this sense, 'unsolvable' could be the hyponym of 'hard' with respect to problem-solving (e.g., 'this is a really hard question'), while 'unbearable' could be the hyponym of 'hard' with respect to suffering (e.g., 'times were hard at the end of the war'). Thus, it is sometimes the case that 'unsolvable' is not construed as being on the same entailment scale as 'hard', and the same happens with other scales such as <depleted, low>, <ridiculous, silly>, and <happy, content>.

When asked to judge whether 'hard' implies 'not unsolvable' or whether 'low' implies 'not depleted', participants in van Tiel et al.'s (2016) experiments may have evoked senses of these terms that are not on the same scale. By contrast, consider the scale <always, sometimes>, 'sometimes' and 'always' have fairly homogeneous senses across uses, relating to the frequency of an event. It would be difficult to construe these terms as not being in an entailment relation. Thus, when asked to judge whether 'sometimes' implies not always, participants were more likely to derive an implicature. We hypothesize that other things being equal, the more homogeneous the sense of the items in a pair, the higher the rate of scalar implicature derivation. We will test this hypothesis in Experiment 2.

## Theoretically Motivated Factors

Beyond methodological questions, we want to consider whether scalar diversity can be explained to some extent if we consider pragmatic theories that go beyond standard Gricean theory. In particular, standard Gricean theory has long been the target of criticism that the method of deriving conversational implicatures

cannot explain a large class of apparently pragmatic effects (Cohen, 1971; Wilson, 1975; Carston, 1988). This critical work shows that in some cases, the meaning of a sub-constituent of an utterance seems to be given a pragmatically augmented interpretation. Although early work on such ‘local enrichment’ did not focus on SI, recent research has (Noveck and Sperber, 2007; Chierchia et al., 2012; among others). An example of local enrichment involving SI is given in (1a) below, which could be glossed by imagining the constituent ‘hit some of the targets’ being given a reading, *hit some and not all of the targets*. This is indicated in (1b):

1. a. Exactly one player hit some of the targets.  
b. Exactly one player hit some but not all of the targets.

This example is based on materials in Chemla and Spector (2011) who discuss why the gloss in (1b) is not derivable using standard Gricean derivation. Potts et al. (2016) reports that participants in an experiment readily understood sentence (1a) according to the gloss in (1b). That local enrichment does occur in natural language is becoming a more widely accepted assumption.<sup>1</sup> Although very little experimental research has explored the conditions under which local processes occur, it is possible to incorporate the fact of local enrichment into a framework that also allows for ‘global’ implicature derivation, of the kind set out in the standard Gricean theory.<sup>2</sup> Such a dual-route framework is set out in Bergen et al. (2016) which augments a ‘standard’ Bayesian probabilistic approach to scalars, the Rational Speech Act (RSA) approach, with additional ‘lexical uncertainty’ (RSA-LU). This framework adopts a liberal stance toward (local) enrichment and posits a family of compositional semantic rules, each of which can represent different enrichments of a given constituent. This is coupled with a framework for reasoning with the uncertainty about which, if any, enrichment is being used. In order to see how such a dual-route approach might account for scalar diversity, it will be necessary to briefly outline some of the details of RSA-LU.<sup>3</sup>

The RSA approach aims to capture how speakers and listeners recursively model each other’s production and comprehension decisions. Like the standard Gricean approach, the standard RSA approach to SI assumes that a single literal interpretation could be assigned to a sentence containing a scalar term. A ‘literal listener’ uses Bayesian inference to model a speaker who chooses an utterance,  $u$ , on the assumption that (the speaker believes) it is true. If we assume that a literal interpretation of the sentence uttered determines the function  $\mathcal{L}$  from utterances and states of affairs to truth values, then the probability that the literal listener assigns to each state of affair after hearing the utterance,  $L_0$ , is

determined by the prior probability on the state of affairs and the truth value of utterance in that state of affairs as follows:

$$2. L_0(w|u) \propto P(w)\mathcal{L}(u, w)$$

A pragmatically sophisticated speaker who addresses  $L_0$  intending convey what is the case, is best served by choosing an utterance that is maximally specific, subject to preferences related to cost of the message. Putting aside some details, the distribution for the speaker’s choice of utterance is given as in (3-4) below:

$$3. S_1(u|w) \propto e^{\lambda U_1(u|w)}$$

$$4. U_1(u|w) = \log(L_0(w|u)) - \mathcal{C}(u)$$

Then a pragmatically sophisticated listener may make inferences about  $S_1$ ’s message according to Bayes’ rule:

$$5. L_1(w|u) \propto P(w)S_1(u|w)$$

Higher-order iterations,  $S_n$  and  $L_n$ , follow the same pattern.

This standard RSA model is capable of accounting for the fact that if the speaker says, ‘The nurse saw some of the signs,’ we are liable to infer that (according to the speaker) the nurse did not see all of the signs. In general, for scalar pair  $\langle S, W \rangle$ , where  $S$  is stronger than  $W$ , if the speaker utters  $W$ , we are liable to infer that she does not think  $S$  is true (see Bergen et al., 2016 for an illustration). Thus using only a single ‘literal’ semantic interpretation function, RSA shows that Bayesian reasoning among speaker and hearer can result in a SI. This in essence provides an account of SI in a broadly ‘Gricean’ way.

However, as mentioned, one can factor in the possibility of enrichments that cannot be explained using a ‘global’ Gricean inference which assume the literal semantics of the sentence. Thus, Bergen et al. (2016) allow that the speaker may use, and be understood to be using, an enriched interpretation of a certain clause type, or expression type. This can be done by supposing that each kind of enrichment for  $W$  constitutes a new semantic interpretation function  $L_i$ . Uncertainty about which, if any, enrichment is being employed in a given utterance can be captured at the level of the first pragmatically sophisticated listener,  $L_1$ , who marginalizes (takes the weighted average) over interpretation functions relative to the prior probabilities of each possible enrichment being used. This is indicated in a revised set of formulae in (6–9) below:

$$6. L_0(w|u, \mathcal{L}) \propto P(w)\mathcal{L}(u, w)$$

$$7. S_1(u|w, \mathcal{L}) \propto e^{\lambda U_1(u|w)}$$

$$8. U_1(u|w, \mathcal{L}) = \log(L_0(w|u, \mathcal{L})) - \mathcal{C}(u)$$

$$9. L_1(w|u) \propto P(w) \sum_{\mathcal{L} \in \Lambda} P(\mathcal{L})S_1(u|w, \mathcal{L})$$

The upshot of this move for simple cases containing unembedded scalar terms is that the strength of the SI (that the speaker does not believe that the stronger sentence is true) can be affected by the prior probability that the speaker intends the literal interpretation or one of the possible enrichments of  $W$ . If there is a high prior probability that the scalar term’s interpretation gets locally enriched to exclude states of affairs where  $S$  is true, then, overall, the strength of the SI that  $S$  is not true would be greater than it would be if no enrichment were used (i.e., if only

<sup>1</sup> In this theoretical introduction, we set aside the widely discussed idea that both local ‘scalar’ enrichments of the kind in (1) and simple unembedded scalar enrichments are both derived via linguistic means in the form of a syntactically represented exhaustification operator (Fox, 2007; Chierchia et al., 2012). Interpreting our results in relation to this theory is different to dual route theories being discussed here. We will return to this point in greater depth below.

<sup>2</sup> Some work on factors which impact on local enrichment include Chemla et al. (2017) and Sun and Breheny (2018c).

<sup>3</sup> More details can be found in Bergen et al. (2016). See also Potts et al. (2016).

the standard model were used). Thus, if the scalar term *W* is associated with a very low, or zero, prior probability that it is enriched this way, then the strength of the SI in a stimulus like that presented in van Tiel et al. (2016) will be lower than where it has a higher prior probability of such local enrichment.<sup>4</sup>

Let us refer to an enrichment of the interpretation of *W* so that it excludes cases where *S* is true as *upper-bound excluded local enrichment* (UBELE). It is in principle possible that scalar terms differ in the prior probabilities on this kind of enrichment. To the extent that these priors differ across scalar terms, we should see differences in rates of SIs in the task reported in van Tiel et al. (2016). Thus, RSA-LU predicts that variation in the strength of these priors could explain at least some of the scalar diversity effect. We explore this prediction in Experiment 3.

## THE CURRENT STUDIES

We tested three separate groups of people in Experiments 1–3. Experiment 1 is more or less a replication of Experiment 2 of van Tiel et al. (2016) using a different measurement scale. Our goal is to obtain a continuous measure of participants' judgment on the availability of SIs for each scalar pair. The remaining studies investigate whether scale homogeneity or liability of UBELE can account for some of the variation in the rates of SIs.

Scale homogeneity was operationalized in terms of a naturalness judgment on an 'X but not Y' construction where <X, Y> is a scalar pair and X can be understood as stronger than Y. In Experiment 2, a group of participants was asked to rate the naturalness of sentences of the form 'X but not Y' e.g., (10a–c):

10. a. The student is brilliant but not intelligent. <brilliant, intelligent>
- b. The water is hot but not warm. <hot, warm>
- c. The dancer finished but she did not start. <finish, start>

'But' has a *denial-of-expectation* conventional implicature. Thus, a sentence 'X but not Y' is felicitous to the extent that X can be construed to not strictly entail Y, but Y would normally be expected, given X. A scale with high homogeneity is one where the stronger term is interpreted to entail the weaker term. Entailment relations require that if X entails Y, whenever X holds, Y must hold. Therefore these 'X but not Y' sentences should be very unnatural if the contrasting predicates X and Y are on the same entailment scale. So if the naturalness rating for a 'but' sentence is low, it suggests a high degree of homogeneity for the given scale; whereas if the rating is high, then the degree of homogeneity is relatively low. Other things being equal, the more homogeneous the sense of the items in a pair, the higher the rate of scalar implicature derivation. We predict that the naturalness rating for scalar expressions in Experiment 2 should negatively correlate with the results of Experiment 1.

<sup>4</sup> Bergen et al. (2016) consider two kinds of enrichments for scalar term *W*. One which excludes states of affairs that also support *S* (this is UBELE, mentioned in the text above) and one which includes only states of affairs where *S* is true. In our experiments below, we consider only the predictions of RSA-LU based on variation in the priors on the first kind of enrichment. In other work (Sun and Breheny, 2018b), we consider also the second kind of enrichment.

Liability of UBELE is the degree to which a weak scalar term is liable to undergo local enrichment to exclude states of affairs where the stronger term is true. In Experiment 3, liability of UBELE is operationalized in terms of the naturalness judgment of an 'X so not Y' construction where <X, Y> is a scalar pair and X is stronger than Y. In Experiment 3, a separate group of participants rated the naturalness for sentences of the form, 'X so not Y' e.g., (11a–c).

11. a. The student is brilliant so not intelligent. <brilliant, intelligent>
- b. The water is hot so not warm. <hot, warm>
- c. The dancer finished so she did not start. <finish, start>

The discourse function of 'so' contrasts with that of 'but' in a number of ways (Blakemore, 2002). 'So' implies that the second segment follows in some way from the first. While 'X but not Y' suggest that one might expect Y, given X, 'X so not Y' suggests that one might expect not Y, given X. Thus, 'X so not Y' sentences should be more coherent to the extent that the weaker scalar expression can undergo UBELE. For example, to understand (11b) as felicitous, 'warm' must have its meaning locally enriched to be understood as 'warm but not hot.' Notice that this has to involve local enrichment rather than Gricean scalar-implicature reasoning because the weaker term is in the scope of negation.<sup>5</sup>

In Experiment 3, if the naturalness rating for 'so' sentences is low, it suggests that the scalar expression is less liable to be enriched to exclude the upper bound; whereas if the rating is high, then it is more liable to be so enriched. As mentioned above, RSA-LU predicts that greater liability for UBELE, the higher the ratings in an inference task of the kind presented in van Tiel et al. (2016). Thus, we predicted that the naturalness rating for scalar expressions in Experiment 3 should positively correlate with the results of Experiment 1.

## EXPERIMENT 1

### Methods

#### Participants

Thirty-six participants were recruited from University College London via an online psychological subject pool. All participants spoke English as a native language. Participants provided written informed consent, and this study was approved by the UCL Research Ethics Committee. Participants came into the lab to complete the testing on a laptop, in return for course credit or £2.5.

#### Materials and Procedure

We tested all 43 scale pairs from van Tiel et al. (2016) in an inference task to measure scalar implicature derivation. The only difference in procedure was that, instead of providing a yes/no response, participants were asked to rate on a 0–100 scale to indicate to what extent they could infer from

<sup>5</sup> Our 'so' task is not the only way to get a measure of liability for UBELE. Previous research has used corpus methods to get a measure of this effect. See Chemla (2013) and Potts and Levy (2015).



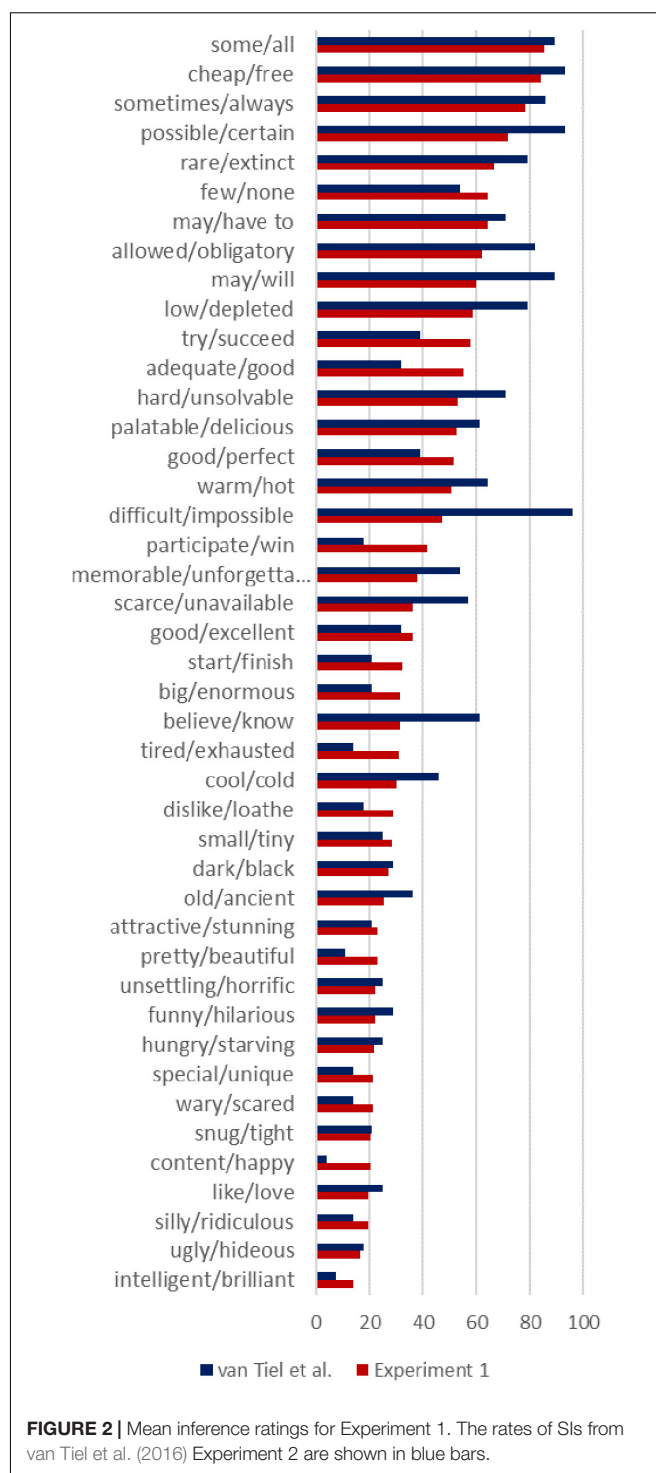
the speaker's statement that the speaker does not believe the stronger alternative. In van Tiel et al. (2016) Experiment 2, the statements were created based on the results of the sentence completion task, e.g., 'The \_\_\_ is attractive but she isn't stunning.' Three statements were selected for each scale, partially based on the completion frequency. Here, we selected the two more frequent statements for every scale (see **Appendix** for a list of items used). If the statements used in the original study had the same completion frequency, a random selection was made. We also used the exact same control items from van Tiel et al.'s (2016) experiment. Four lists were created, each participant judged either 21 or 22 experimental items and 7 control items. Thus, each experimental item was judged by 18 participants. Participants were randomly assigned to one of four lists. A randomized order of presentation of the items was created for each participant.

## Results

The mean ratings for entailments and non-coherent inferences were 86.97 ( $SD = 24.81$ ) and 8.3 ( $SD = 15.09$ ), respectively. Two participants were excluded from the analysis because their mean ratings for entailments or non-coherent inferences were two standard deviations away from the means. The mean ratings for all scalar items are shown in **Figure 2** (red bars). The rates of SIs from van Tiel et al. (2016, Experiment 2) are also included in that figure (blue bars).<sup>6</sup>

We carried out one-way ANOVAs with the ratings on the inference task as the dependent variable and lexical categories as the independent variable. The ratings were averaged by items (43 scales) before entering into the analysis. There was a statistically significant difference among lexical categories [ $F(3,39) = 9.52, p < 0.001$ ]. A Tukey *post hoc* test revealed that the ratings of SI for quantifiers ( $M = 76.03, SD = 10.89$ ) and modals ( $M = 64.35, SD = 5.24$ ) were significantly higher than for adjectives ( $M = 34.95, SD = 17.19$ ) and verbs ( $M = 35.30, SD = 13.17$ ), but there was no statistically significant differences between quantifiers and modals, and between adjectives and verbs. These results are in line with those seen in van Tiel et al. (2016). Inspecting the graph, one can see some differences among items, but the general pattern is the same.

To examine whether factors identified by van Tiel et al. (2016) explain some of the variation found in Experiment 1, we conducted a multiple linear regression analysis to predict the ratings of SIs in our Experiment 1 from all the potential factors reported in van Tiel et al. (2016) including association strength, grammatical class, word frequencies, semantic relatedness, semantic distance, and boundedness. The ratings of SIs in Experiment 1 were averaged by item (43 scales) before entering the analysis. The results of the linear regression are summarized in **Table 1**. The model explained 48.7% of the variance [ $R^2 = 0.56, F(6,35) = 7.48, p < 0.001$ ]. As in van Tiel et al. (2016) only semantic distance and boundedness were



**FIGURE 2 |** Mean inference ratings for Experiment 1. The rates of SIs from van Tiel et al. (2016) Experiment 2 are shown in blue bars.

significant predictors of the inference task results, whereas other factors did not make a significant contribution to the model.

## Discussion

Experiment 1 established that there is a considerable amount of variation among scalar terms in terms of how strongly

<sup>6</sup> We took data reported in van Tiel et al. (2016) P145 to build **Figure 2** and ran comparative analysis in Experiments 2 and 3.

**TABLE 1 |** Results of multiple linear regression for inference ratings of Experiment 1.

	Estimate	SE	t-Value	p-Value	R <sup>2</sup>
(Intercept)	4.651	21.135	0.22	0.827	
Association strength	0.024	0.108	0.22	0.827	0.007
Grammatical class	−13.575	9.429	−1.44	0.159	0.099
Word frequencies	−3.603	2.605	−1.38	0.175	0.016
Semantic relatedness	3.036	14.085	0.22	0.831	0.020
Semantic distance	7.234	3.203	2.26	0.030	0.106
Boundedness	20.802	4.897	4.25	0.000	0.315

they give rise to scalar implicatures. The general pattern found in van Tiel et al. (2016) was replicated, with a different measurement scale. Experiment 1 also replicated van Tiel et al.'s (2016) findings that semantic distance and boundedness only explain some of the variation.

## EXPERIMENT 2

### Methods

#### Participants

We invited Amazon Mechanical Turk workers located in the United States with a 95% approval rate on tasks previously performed for other requesters. Forty participants were recruited and were paid United States \$0.50 for their participation. The experiment was initiated by a consent statement approved by UCL Research Ethics Committee. Participants were asked to indicate their native language, but we paid them regardless of their answer to this question. Only participants with English as a native language were included in the analysis.

#### Materials and Procedure

**Figure 3** is an example item. We used the 43 scales investigated in Experiment 1 to construct experimental sentences for Experiment 2. The experimental sentences were of the form 'X but not Y' where, according to van Tiel et al. (2016), X and Y are a pair of scalar terms and X is stronger than Y. For example, 'The student is brilliant but not intelligent.' We constructed two experimental sentences for every scale (see **Appendix** for a list of items used). The nominal ('student') used in each experimental sentence was the same as for the corresponding statement in Experiment 1. For the verbs and auxiliary verbs like 'may,' experimental sentences were constructed differently to make sure that the weaker term was in the scope of negation (see **Appendix** for details); for instance, 'The lawyer will appear in person but it is not the case that he may appear in person.' In addition, we constructed seven filler sentences, which contained clearly felicitous (e.g., 'The banker is rich but not happy') and clearly infelicitous items (e.g., 'The man left the party but he never came'). Participants were asked to rate how natural these constructions are on a 1 (very unnatural) – 7 (very natural) scale. Each participant judged 43 experimental sentences and 7 fillers. Eight survey versions with pseudo-randomized order of items were

The student is brilliant but not intelligent.

very unnatural ○ ○ ○ ○ ○ ○ ○ very natural  
1 2 3 4 5 6 7

**FIGURE 3 |** Sample item in Experiment 2.

created. Participants were randomly assigned to one of eight surveys.

## Results

Two participants were excluded because their mean ratings for the infelicitous items were above 5. The mean ratings for the clearly felicitous and clearly infelicitous control items were 5.8 ( $SD = 1.82$ ) and 2.32 ( $SD = 1.91$ ). The mean rating for experimental items ranged from 1.33 ( $SD = 0.59$ ) (<will, may>) to 4.47 ( $SD = 2.11$ ) (<unique, special>). Critically, we found that the naturalness of the 'but' sentences correlated negatively with the ratings of SIs in Experiment 1 [ $r(41) = -0.31$ ,  $p = 0.04$ ] – see **Figure 4**. In addition, it correlated negatively with the results from van Tiel et al. (2016, Experiment 2) [ $r(41) = -0.36$ ,  $p = 0.02$ ]. These results confirmed the prediction outlined earlier. We defer discussion of these results until after the combined analysis.

## EXPERIMENT 3

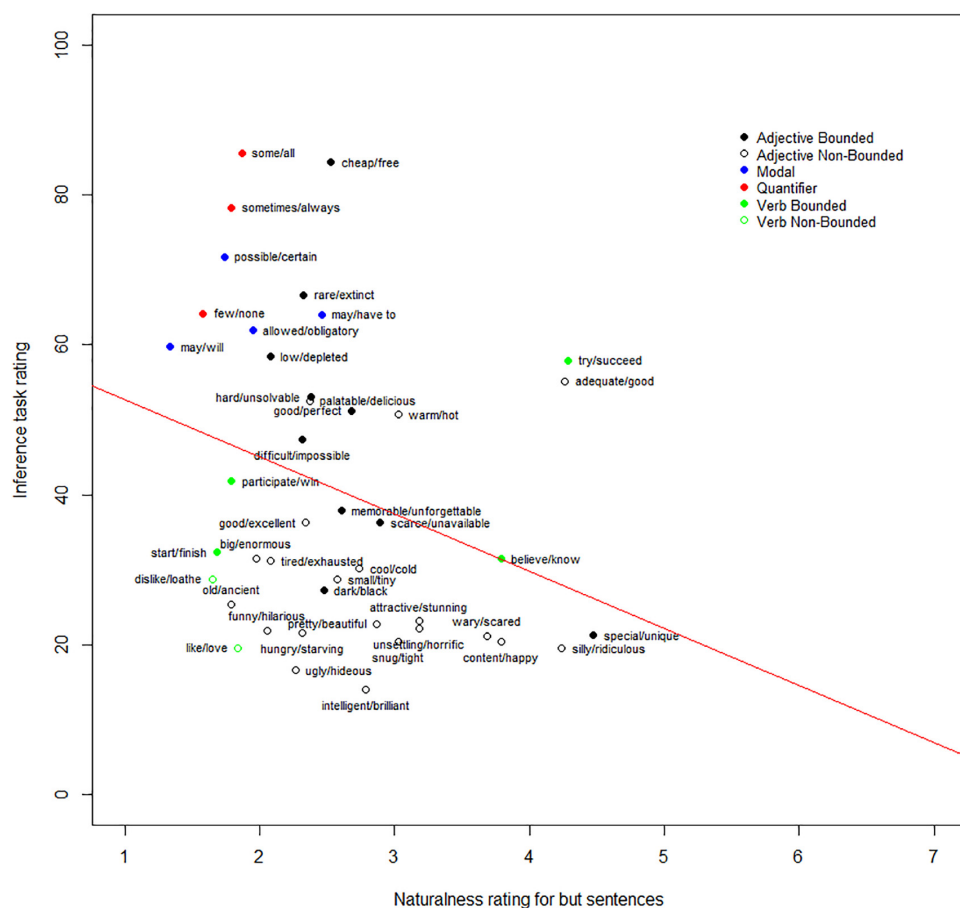
### Methods

#### Participants

Forty participants were recruited from University College London via an online psychological subject pool. All participants spoke English as a native language. Participants provided written informed consent, and this study was approved by the UCL Research Ethics Committee. Participants came into the lab to fill out a paper-based survey, in return for course credit or £1.

#### Materials and Procedure

**Figure 5** is an example item. We used 43 scales investigated in Experiment 1 to construct experimental sentences for Experiment 3. Two experimental sentences were constructed for each scale (see **Appendix** for a list of items used). The experimental sentences were of the form 'X so not Y,' where X is stronger than Y; for example, 'The student is brilliant so not intelligent.' As in Experiment 2, the nominal ('student') used in each experimental sentence was from statements used in Experiment 1. For the verbs and auxiliary verbs like 'may,' experimental sentences were constructed differently (see **Appendix** for details); for example, 'The lawyer will appear in person so it is not the case that he may appear in person.' Seven filler sentences were constructed, which contained clearly felicitous (e.g., 'The cup is red so not blue') and clearly infelicitous items (e.g., 'The banker is rich so not happy'). Participants were asked to indicate how natural these constructions are on a 1 (very unnatural) – 7 (very natural) point scale. Each participant judged 43 experimental sentences and 7 fillers. Eight paper-based



**FIGURE 4 |** Negative correlation between the absence of homogeneity and inference rate.

The student is brilliant so not intelligent.

very unnatural ○ ○ ○ ○ ○ ○ ○ very natural  
1 2 3 4 5 6 7

**FIGURE 5 |** Sample item in Experiment 3.

survey versions with pseudo-randomized order of items were created.

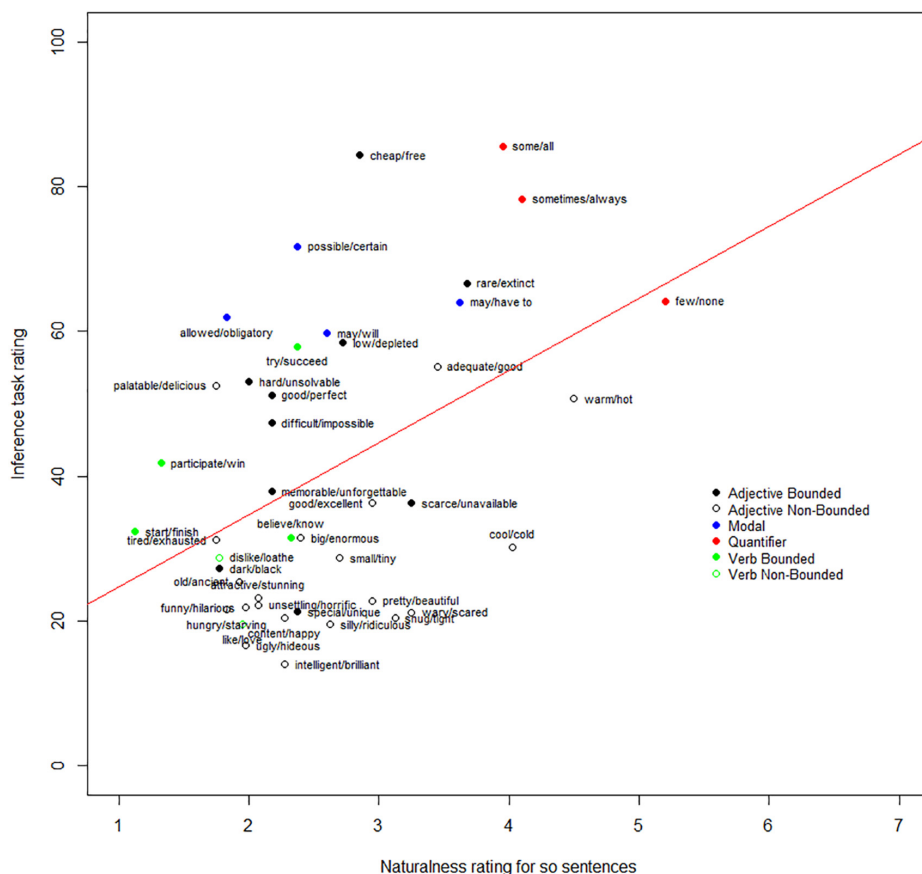
## Results

The mean rating for the clearly felicitous and clearly infelicitous control items were 5.89 ( $SD = 1.68$ ) and 1.53 ( $SD = 1.25$ ). The mean rating for experimental items ranged from 1.13 ( $SD = 0.33$ ) (<finish, start>) to 5.2 ( $SD = 1.99$ ) (<none, few>). We found that the naturalness of the 'so' sentences positively correlated with the ratings of SIs in Experiment 1 [ $r(41) = 0.44$ ,  $p = 0.004$ ] – see in **Figure 6**. In addition, the naturalness of the 'so' sentence also positively correlated with the results from van Tiel et al. (2016, Experiment 2) [ $r(41) = 0.35$ ,  $p = 0.02$ ].

## COMBINED ANALYSIS

To investigate the proportion of variance explained by all the potential factors, multiple linear regression analyses were conducted to predict the ratings of SIs in Experiment 1 from scale homogeneity degree, propensity for local enrichment, and all factors established in van Tiel et al. (2016). The rating of SIs in Experiment 1, and the naturalness rating from Experiments 2 and 3 were averaged by item (43 scales) before entering the analysis. The results of the linear regression are summarized in **Table 2**.

We found that the regression model accounted for 63% of the variance [ $R^2 = 0.70$ ,  $F(8,33) = 9.73$ ,  $p < 0.001$ ]. This contrasts with the 49% of variance explained without the ratings for the 'but' and 'so' tasks entered in the model. In this fuller model, the propensity for local enrichment, semantic distance and boundedness were substantial factors, with the propensity for local enrichment explaining 15%, semantic distance explaining 11%, and boundedness explaining 31%. None of the other factors significantly accounted for the variation in the rates of SIs. In this model, scale homogeneity did not significantly explain the variance. Scale homogeneity was highly correlated with semantic distance [ $r(41) = -0.53$ ,  $p < 0.001$ ]. Thus, the variance in inference ratings explained by scale homogeneity largely



**FIGURE 6 |** Positive correlation between the propensity of local enrichment and inference rate.

overlapped with the variance accounted for by semantic distance. When semantic distance was omitted from the model, scale homogeneity did explain a significant amount of the variance.

## GENERAL DISCUSSION

### Discussion of Experiments 2 and 3

We adapted the items from van Tiel et al. (2016, Experiment 2) for these two naturalness judgment tasks. Participants were asked to judge the felicity of sentences of the form ‘S but/so not W’ where ‘S’ is the stronger term from the SI judgment task (‘all,’ ‘hot,’ etc.) and ‘W’ is the weaker term (‘some,’ ‘warm,’ etc.). The respective sentences have different felicity conditions due to the function of ‘but’ and ‘so,’ respectively. We argue that the ‘but’ sentences probe scale homogeneity, while the ‘so’ sentences probe liability for UBELE.

Concerning scalar homogeneity, if participants find a way to read a sentence of the form ‘S but not W’ as felicitous, then it indicates that the items of this scalar pair could be constructed as not always being on the same scale. The results of Experiment 2 showed that the degree of homogeneity varied across different scales. That is, quantificational and modal scales, as well as most verb scales, are in clear entailment relation, but most adjective

**TABLE 2 |** Results of combined analysis.

	Estimate	SE	t-Value	p-Value	R <sup>2</sup>
(Intercept)	−31.274	24.967	−1.250	0.219	
Scale homogeneity	−3.142	3.008	−1.040	0.304	0.037
Local enrichment	10.442	2.668	3.910	< 0.001	0.149
Association strength	0.048	0.092	0.520	0.606	0.006
Grammatical class	1.506	9.032	0.170	0.869	0.067
Word frequencies	−2.926	2.262	−1.290	0.205	0.013
Semantic relatedness	−8.151	12.303	−0.660	0.512	0.015
Semantic distance	8.291	3.150	2.630	0.013	0.107
Boundedness	21.564	4.171	5.170	< 0.001	0.308

scales are not. We suggest that this variation in the degree of homogeneity is expected due to factors like underspecification or polysemy. We found that high homogeneity led to higher rates of SIs, compared to when homogeneity was low.

The results of Experiment 2 are related to the hypothesis discussed in Doran et al. (2009). They suggested that there are domain-general scalar expressions such as quantifiers and modals and domain-specific ones such as adjectives. The former are more likely to give rise to SIs in the absence of context, whereas the latter require more contexts in order to derive SIs.



Doran et al. (2009) found that only the derivation of adjective scales was affected by providing stronger scale mates in the context. This result might be due to the low homogeneity in adjective scales. That is, without restriction in the context, the use of scalar adjectives may evoke alternatives that are irrelevant in deriving scalar implicatures.

Since scalar homogeneity is strongly correlated with semantic distance, it raises the question of what the relation between the two concepts is. One possibility is that a low rating of semantic distance reflects the fact that scalar pairs are not uniformly on the same scale. That is, as it was measured in van Tiel et al. (2016), semantic distance may reflect, to some extent, both genuine semantic distance (a measure of distinctness) and scale homogeneity. For example, a high semantic distance rating for <all, some> may reflect genuine distinctness of the terms, while a lower rating for <unique, special> may reflect also a lack of scale homogeneity. Future research may seek an alternative means to measure distance that may de-confound these two dimensions.

Turning now to liability for UBELE, this is a new factor motivated by an extension of standard Gricean pragmatic theory. In Experiment 3, if participants find a way to read the sentences of the form ‘S so not W’ as felicitous, then it indicates that ‘W’ (e.g., a sentence containing ‘some’) has been locally enriched in the scope of negation to exclude situations where S is true (e.g., all). The results of Experiment 3 showed that the naturalness of ‘S so not W’ varied across different scales, suggesting that scalar terms differ in their propensity for being locally enriched in this way. The very strong positive correlation between the naturalness of ‘S so not W’ and the rates of SIs measured in the inference task suggested that liability of UBELE influences the judgment in van Tiel et al.’s (2016) original inference task. UBELE can give rise to what looks like a standard Gricean scalar implicature in the unembedded case and this could have inflated rates measured in the inference task.

## Theoretical Implications of Scalar Diversity

From the perspective of the standard Gricean approach to SIs, the existence of a scalar diversity effect among apparently good scalar pairs is not predicted without further assumptions. In previous research, a number of factors have been explored to account for scalar diversity. Apart from one methodologically motivated factor, these factors can all find motivation from the perspective of standard Gricean approaches to scales, relying on scalar alternatives. To date some variance has been explained by these theoretically motivated factors but much is left unexplained. Our contribution in this paper has been to add one more potential methodological factor (scale homogeneity) and one more theoretically motivated factor (liability to upper-bound excluding local enrichment).

As to scale homogeneity, we obtained the predicted negative correlation between ratings on our ‘but’ task and those on the inference task. However, these ratings were highly correlated with ratings for semantic distances and, in a full model that also includes semantic distance as a factor, ‘but’ task ratings did not emerge as a significant factor. We have indicated how future

research may explore to what extent it is lack of semantic distance and lack of scale homogeneity explain low rates of implicature, particularly for adjective items.

The results of the ‘so’ task clearly suggests a new factor unexplored in previous studies. This task operationalizes our idea that weak scalar terms differ in their propensity for being locally enriched to exclude the upper bound (UBELE). Our results provide confirmation for current pragmatic approaches to scalars that extend the standard Gricean approach to accommodate the fact of local enrichment. We focused in particular on RSA-LU (Bergen et al., 2016), to derive a prediction of a positive correlation between ratings on our ‘so’ task and the inference task. This is what we found. Moreover, we established that a model including this measure of liability for UBELE as a factor accounts for more variance than a model which includes only those factors motivated by the standard Gricean approach, explored in van Tiel et al. (2016).

We note, however, that ‘distinctiveness of alternatives’ factors, motivated by the standard Gricean model, remained significant in accounting for scalar diversity. This is expected in a dual-route pragmatic approach like RSA-LU. For in that approach, there are two routes to SI. One route is via so-called, ‘global’ inference about the speaker’s actions employing the literal semantics of the sentence and shared principles of conversation. This is akin to the standard Gricean derivation which relies on scalar alternatives. Thus distinctness of those alternatives, as well as their contextual relevance and availability, remain potential factors. The other is via a free enrichment process. That factors motivated by both routes contribute to accounting for Scalar Diversity is expected on the dual route account.

Until now we have not discussed grammatical theories of scalar implicature phenomena. According to widely cited versions of these theories (e.g., Fox, 2007; Chierchia et al., 2012), scalar implicatures of the kind tested in van Tiel et al.’s (2016) inference task are not derived using general pragmatic principles but result from the presence of an exhaustification operator in the syntactic representation of the sentence. This operator functions like ‘only’ in two important respects; (i) it may be placed at different scope sites within a sentence; (ii) in all cases it is interpreted relative to alternatives to its argument. To illustrate this point, for (1) the exhaustification operator would be represented as taking only a constituent, ‘x hit some of the targets’ in its scope, leading to alternatives like, ‘x hit all of the targets.’ For sentences where there is apparently a ‘global’ SI, like the items in our Experiment 1, the operator takes scope over the whole sentence. For example, when participants infer that ‘The student is intelligent’ implies she is not brilliant, this would be explained in terms of an operation on the whole sentence, with ‘The student is brilliant’ as alternative. Thus there are two key differences to dual route theories described above. The first is that the grammatical approach posits only a *single* mechanism to account for both local effects of the kind involved in Experiment 3 and ‘global’ effects tested in the inference task, Experiment 1. The second is that alternatives are employed in the derivation of both global and local effects. By contrast, while the ‘dual route’ approach being considered here also allows that an enrichment mechanism can be involved in items in both Experiments 3 and 1,

this enrichment mechanism does not rely on alternatives. In addition, a second mechanism, which does rely on alternatives, only applies in the case of ‘global’ SIs, of the kind studied in Experiment 1.

There is little scope in this paper for a thorough empirical exploration of these two approaches.<sup>7</sup> Here, we make two comments by way of comparison. First, the grammatical account could be integrated into a framework for reasoning with uncertainty since it implies a variety of interpretive possibilities for a sentence depending on whether the operator is inserted and where. Thus it is conceivable that the relation between the results of Experiment 1 and Experiment 3 above could be explained. However, that would require extra assumptions which link rates of insertion of the linguistic operator at the root level of a sentence (as would occur in Experiment 1) and in the scope of negation (as occurs in our Experiment 3).

Second, there is an important point of contrast between this grammatical account of our data and the one outlined in Bergen et al. (2016) and Potts et al. (2016). The latter approach proposes a simple narrowing mechanism to account for local enrichment, while the grammatical theory holds that upper-bound excluding local enrichments of expressions with scalar terms (compared with the many other kinds of local enrichment) are mediated by a syntactically represented exhaustification operator. Thus, the grammatical account would predict an effect of the distinctness of alternatives for local enrichments, comparable to that found for global enrichments. It is possible to investigate this prediction with our data. We can consider whether variation in ratings on our ‘so’ task (Experiment 3) are predicted by factors that are related to distinctness. To do this, we used a multiple regression analysis to test if semantic distance and boundedness significantly predicted participants’ ratings on the ‘so’ task. The results of the regression indicate that the two predictors did not significantly explain the variance [ $R^2 = 0.05$ ,  $F(2,40) = 1.04$ ,  $p = 0.36$ ]. Neither semantic distance [ $\beta = -0.25$ ,  $t(40) = -1.34$ ,  $p = 0.19$ ] nor boundedness [ $\beta = 0.23$ ,  $t(40) = 0.84$ ,  $p = 0.41$ ] significantly predicted the ratings of ‘so’ task. Thus, a preliminary exploration of whether there is the predicted relationship between distinctness of alternatives and local enrichability was unable to find such a relation. This is unexpected if local enrichment relies on alternatives to the same extent as global. As mentioned, the RSA-LU approach assumes a general narrowing option for semantic interpretation as one of two routes to account for scalar enrichment, and this does not rely on alternatives.

To draw out the points of theoretical interest here, let us sum up what we have learnt from the scalar diversity effect. To date, previous studies (replicated here) have shown that factors relating to the distinctness of alternatives can explain some of Scalar Diversity, and this is predicted if SIs are derived by general Gricean reasoning or via a linguistically represented exhaustification operator. However, such factors explain by no means all of the scalar diversity effect. We outlined dual-route approaches above and showed that one version of that approach

successfully explains more of the Scalar Diversity. Unlike the grammatical approach, RSA-LU suggests that mechanisms for deriving local enrichments do not rely on alternatives and thus the second source of potential variation, liability for UBELE, would be independent of factors such as the distinctness of alternatives. An analysis of results from Experiment 3 suggest this may be the case.

To turn to our final point of discussion, we point out that RSA-LU as stated does not shed much light on what factors might lead to the application of this ‘free enrichment’ mechanism used in achieving scalar effects. To put this another way, while the variability in local enrichment of the kind studied in Experiment 3 can partially explain variability in the inference task results, we are left with the question what explains the variability in the application of this second mechanism. For now, we have to leave this as a matter for future research.<sup>8</sup> But, to re-iterate the point of discussion above, we learn from a comparison among theories which can account for local effects that a dual-route approach that does not rely on alternatives is better supported.

## AUTHOR CONTRIBUTIONS

CS carried out the experiments and analyzed the data. CS and RB wrote the first draft of the manuscript. CS, YT, and RB have all contributed to the experimental design and the final version of the manuscript.

## FUNDING

CS was supported by a scholarship under the State Scholarship Fund from China Scholarship Council. The publication of this article was funded by the University College London.

## ACKNOWLEDGMENTS

We thank Anton Benz, Robyn Carston, Chris Cummins, Judith Degen, Jakub Dotlacil, Nicole Gotzner, Uli Sauerland, and Bob van Tiel, as well as the reviewers for valuable comments and suggestions.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02092/full#supplementary-material>

<sup>7</sup> A more detailed empirically based comparison can be made on the basis of Sun and Breheny (2018b,c) but we leave this for future discussion.

<sup>8</sup> In principle there are analytical possibilities as to how UBELE may arise without the use of alternatives. One possibility is that scalar terms (like, ‘warm’) may be upper-bounded in virtue of the presence of a maximality operation as suggested for the analysis of quantificational expressions with numerals (see Kennedy, 2015; Buccola and Spector, 2016). The research on numerals suggests there is a strong bias to understand those expressions via something like maximality (Geurts, 2006; Breheny, 2008). Other scalar expressions to which maximality may apply may simply vary in this bias.

## REFERENCES

- Beltrama, A., and Xiang, M. (2012). "Is good better than excellent? An experimental investigation on scalar implicatures and gradable adjectives," in *Proceedings of the Sinn und Bedeutung 17*, eds E. Chemla, V. Homer, and G. Winterstein (Chicago, IL: University of Chicago), 81–98.
- Benz, A., Ferrer, C. B., and Gotzner, N. (2017). "Scalar diversity and negative strengthening," in *Proceedings of Sinn und Bedeutung 22*, ZAS & University of Potsdam, Potsdam.
- Bergen, L., and Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 1450–1460. doi: 10.1037/a0027850
- Bergen, L., Levy, R., and Goodman, N. D. (2016). Pragmatic reasoning through semantic inference. *Semant. and Pragmat.* 9, 1–83. doi: 10.3765/sp.9.20
- Blakemore, D. (2002). *Relevance and Linguistic Meaning – The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511486456
- Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *J. Semant.* 25, 93–139. doi: 10.1093/jos/ffm016
- Breheny, R., Katsos, N., and Williams, J. (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. doi: 10.1016/j.cognition.2005.07.003
- Buccola, B., and Spector, B. (2016). Modified numerals and maximality. *Linguist. Philos.* 39, 151–199. doi: 10.1007/s10988-016-9187-2
- Carston, R. (1988). "Language and cognition," in *Linguistics: the Cambridge survey*, ed. F. Newmeyer (Cambridge: Cambridge University Press), 38–68. doi: 10.1017/CBO9780511621062.003
- Chemla, E. (2013). *Apparent Hurford Constraint Obviations are based on Scalar Implicatures: An Argument Based on Frequency Counts*. Ms. CNRS, ENS, LSCP. Available at: <http://www.emmanuel.chemla.free.fr/Material/Chemla-HurfordCounts.pdf>, 2013
- Chemla, E., and Spector, B. (2011). Experimental evidence for embedded scalar implicatures *J. Semant.* 28, 359–400. doi: 10.1093/jos/ffq023
- Chemla, E., Cummins, C., and Singh, R. (2017). Training and timing local scalar enrichments under global pragmatic pressures. *J. Semant.* 34, 107–126.
- Chierchia, G., Fox, D., and Spector, B. (2012). "Scalar implicature as a grammatical phenomenon," in *Semantics: An International Handbook of Natural Language Meaning*, Vol. 3, eds P. Portner, C. Maienborn, and K. von Stechow (Berlin: Mouton de Gruyter), 2297–2331.
- Cohen, L. J. (1971). "The logical particles of natural language," in *Pragmatics of Natural Language*, ed. Y. Bar-Hillel (Dordrecht: Reidel), 50–68. doi: 10.1007/978-94-010-1713-8\_3
- Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semant. Pragmat.* 8, 1–55. doi: 10.3765/sp.8.11
- Doran, R., Baker, R. E., McNabb, Y., Larson, M., and Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *Int. Rev. Pragmat.* 1, 211–248. doi: 10.1163/187730909X12538045489854
- Doran, R., Ward, G., Larson, M., McNabb, Y., and Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88, 124–154. doi: 10.1353/lan.2012.0008
- Fox, D. (2007). "Free choice and the theory of scalar implicatures," in *Presupposition and Implicature in Compositional Semantics*, eds U. Sauerland and P. Stateva (Houndmills: Palgrave Macmillan) 71–120. doi: 10.1057/9780230210752\_4
- Gazdar, G. (1979). *Pragmatics: Presupposition, Implicature, and Logical Form* (Cambridge, MA: Academic Press).
- Geurts, B. (2006). "Take 'five': the meaning and use of a number word," in *Non-Definiteness and Plurality*, eds S. Voegelé & L. Tasmowski (Philadelphia, PA: John Benjamins), 311–329. doi: 10.1075/la.95.16geu
- Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511975158
- Grice, H. P. (1975). Logic and conversation. *Syntax Semant.* 3, 41–58.
- Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English*. California, CA: University of California Los Angeles.
- Kennedy, C. (2015). A "de-Fregean" semantics (and neo-Gricean pragmatics) for modified and unmodified numerals. *Semant. Pragmat.* 8, 1–44. doi: 10.3765/sp.8.10
- Noveck, I., and Sperber, D. (2007). "The why and how of experimental pragmatics: The case of 'scalar inferences,'" in *Advances in Pragmatics*, ed. N. Roberts (Basingstoke: Palgrave).
- Potts, C., Lassiter, D., Levy, R., and Frank, M. C. (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *J. Semant.* 33, 755–802. doi: 10.1093/jos/ffv012
- Potts, C., and Levy, R. (2015). "Negotiating lexical uncertainty and speaker expertise with disjunction," in *Proceedings of the 41st Annual Meeting of the Berkeley Linguistics Society* (Berkeley, CA: BLS). doi: 10.20354/B4414110013
- Simons, M., and Warren, T. (2018). A closer look at strengthened readings of scalars. *Q. J. Exp. Psychol.* 71, 272–279. doi: 10.1080/17470218.2017.1314516
- Sun, C., and Breheny, R. (2018b). *Approaching scalar diversity through (RSA with) Lexical Uncertainty*. Available at: <http://www.osf.io/rd6sc>
- Sun, C., and Breheny, R. (2018a). *Rates of scalar inferences beyond "some" – A corpus study*. Available at: <http://www.osf.io/sb2u6>
- Sun, C. and Breheny, R. (2018c). "Shared mechanism underlying unembedded and embedded enrichments: evidence from enrichment priming," in *Proceedings of Sinn und Bedeutung 22(ZAS Papers in Linguistics 61)*, Vol. 2, eds U. Sauerland and S. Solt (Potsdam: ZAS & University of Potsdam), 425–441.
- van Tiel, B., Van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *J. Semant.* 33, 137–175. doi: 10.1093/jos/ffu017
- Wilson, D. (1975). *Presuppositions and Non-Truth-Conditional Semantics*. Cambridge, MA: Academic Press.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Sun, Tian and Breheny. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Development of Quantitative and Temporal Scalar Implicatures in a Felicity Judgment Task

Walter Schaeken<sup>1\*</sup>, Bojoura Schouten<sup>2</sup> and Kristien Dieussaert<sup>3</sup>

<sup>1</sup> Laboratory of Experimental Psychology, KU Leuven, Leuven, Belgium, <sup>2</sup> Health Care, Faculty of Medicine and Life Sciences, Hasselt University, Hasselt, Belgium, <sup>3</sup> Quality Care, UC Leuven-Limburg, Leuven, Belgium

## OPEN ACCESS

### Edited by:

Penka Stateva,  
University of Nova Gorica, Slovenia

### Reviewed by:

Paul Pierre Marty,  
Massachusetts Institute  
of Technology, United States  
Alexandre Cremers,  
University of Amsterdam, Netherlands

### \*Correspondence:

Walter Schaeken  
walter.schaeken@kuleuven.be;  
walter.schaeken@ppw.kuleuven.be

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 11 June 2018

**Accepted:** 21 December 2018

**Published:** 18 February 2019

### Citation:

Schaeken W, Schouten B and  
Dieussaert K (2019) Development  
of Quantitative and Temporal Scalar  
Implicatures in a Felicity Judgment  
Task. *Front. Psychol.* 9:2763.  
doi: 10.3389/fpsyg.2018.02763

Experimental investigations into children's interpretation of scalar terms show that children have difficulties with scalar implicatures in tasks. In contrast with adults, they are for instance not able to derive the pragmatic interpretation that "some" means "not all" (Noveck, 2001; Papafragou and Musolino, 2003). However, there is also substantial experimental evidence that children are not incapable of drawing scalar inferences and that they are aware of the pragmatic potential of scalar expressions. In these kinds of studies, the prime interest is to discover what conditions facilitate implicature production for children. One of the factors that seem to be difficult for children is the generation of the scalar alternative. In a Felicity Judgment Task (FJT) the alternative is given. Participants are presented with a pair of utterances and asked to choose the most felicitous description. In such a task, even 5-year-old children are reported to show a very good performance. Our study wants to build on this tradition, by using a FJT where not only "some-all" choices are given, but also "some-many" and "many-all." In combination with a manipulation of the number of successes/failures in the stories, this enabled us to construct control, critical and ambiguous items. We compared the performance of 59 5-year-old children with that of 34 11-year-old children. The results indicated that performance of both age groups was clearly above chance, replicating previous findings. However, for the 5-year-old children, the critical and ambiguous items were more difficult than the control items and they also performed worse on these two types of items than the 11-year-old children. Interestingly with respect to the issue of scalar diversity, the 11-year-old children were also presented temporal items, which turned out to be more difficult than the quantitative ones.

**Keywords:** pragmatics, experimental pragmatics, scalar implicature, Felicity Judgment Task, informational strength, alternatives

## INTRODUCTION

Consider a brainstorm session for some new research lines, where the head of the research group offers the following feedback: "Some of John's ideas were interesting." The use of "some" seems to lead to the inference that the speaker did not find all of John's ideas interesting. Different theories try to explain this kind of inferences. "Some" seems to invoke "all," which is the more informative. Therefore, "some" is strengthened by the negation of "all." The latter step can be made on the basis of pragmatic reasoning or can be based on grammar.



In Grice's terms (Grice, 1975), the explanation goes as follows. Given the cooperation principle guiding communication ("Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged"), one should try to say no more and no less than is required for the purpose of the exchange (the Quantity-maxim). Therefore, the head of the research group who said that *some* of the ideas were interesting does not think that the alternative and more informative *all*-sentence is true. Moreover, if the addressee assumes that the head of the research group has an opinion about the truth of the *all*-sentence (see Sauerland, 2004 for a definition of the Opinionated Speaker; see also Fox, 2007), the addressee will conclude that the head believes that the *all*-sentence is false and that, therefore, the head thinks that not all of John's ideas were interesting. It is important to note that from a logical point of view one can use "*some*" when "*all*" is the case. Indeed, the lower-bounded semantics of "*some*" is "*at least some and possibly all*" (Horn, 1972). The scalar implicature (SI) corresponds to the upper-bounded meaning ("*but not all*") and can be seen as a pragmatic enrichment of the semantic content of the quantifier. Hence, in a situation where the assertion "*all of John's ideas*" is true, the *some*-sentence is acceptable according to the semantic, lower-bounded interpretation of the scalar term, but unacceptable according to its pragmatic, upper-bounded interpretation. As said before, grammatical accounts (e.g., Chierchia, 2004; Fox, 2007) share some basic aspects with a Gricean account, but are clearly different in their assumption that underinformative sentences are ambiguous between different syntactic structures. In grammatical accounts, a covert syntactic operator is introduced, whose meaning is close to "*only*." Of the possible alternatives, the operator excludes all those that are more informative as the proposition expressed by the sentence without the operator (Geurts, 2010). In our example, appending the operator leads to the proposition that the head of the research group liked some of the ideas and the negation of the proposition that she liked all of them, which can be paraphrased into "*she thinks that only some of the ideas were interesting*."

Experimental research has been devoted to the interpretation of scalars, with a strong focus on "*all-some*," probably because this scale offers a sharply defined and easily testable division between the encoded and the inferred meaning. When adults are presented with problems like the one above ("*some of the ideas were interesting*"), they overwhelmingly chose the pragmatic interpretation, that is, the inference from "*some*" to "*not all*" (e.g., Noveck, 2001; Bott and Noveck, 2004; De Neys and Schaeken, 2007; Marty and Chemla, 2013; Heyman and Schaeken, 2015; van Tiel and Schaeken, 2017). On classical tasks, like the Truth Value Judgment Task (TVJT) where one has to indicate whether an utterance is true or false, young children perform poorer than adults in deriving these SIs. They more often prefer the logical answer; hence they accept underinformative scalar sentences (see e.g., Chierchia et al., 2001; Noveck, 2001; Papafragou and Musolino, 2003; Foppolo et al., 2012; Janssens et al., 2015).

However, these findings do not mean that young children are unable to show more adult-like behavior when interpreting scalar statements. Several factors seem to be able to lift the performance of young children (for an overview and a nice

series of experiments, see also Foppolo et al., 2012). One of the factors is awareness of the goal and training, as demonstrated by Papafragou and Musolino (2003). Before the start of the experiment, the researchers caused an enhanced awareness of the goals of the task and gave a short training to detect infelicitous statements. As a result, children's sensitivity to SI significantly improved, although they still fell short of a fully mature performance. Another factor is the nature of the task. Pouscoulous et al. (2007) did not ask for a truth evaluation, but asked children to perform an action. In order to realize this, they presented the children with five boxes and five tokens. Pouscoulous and her colleagues requested children to adapt the boxes to make them compatible with a statement. For example, the children saw that all five boxes contained a token and were told '*I would like some boxes to contain a token.*' Pouscoulous et al. (2007) reasoned that if the children believed that "*some*" is compatible with "*all*," they should leave the boxes unaltered; otherwise they should remove at least one token. The results showed that the number of derived implicatures in children increased. The nature of the answer is also an important factor. Katsos and Bishop (2011) focused on the fact that underinformative statements are true but suboptimal: in a binary judgment task, one cannot express being aware of the suboptimality. Indeed, one is forced to choose between "*true*" and "*false*." If one is tolerant to this suboptimality and focuses more on the fact that these statements are logically correct, one goes for "*true*." Katsos and Bishop (2011) offered a third response option (corresponding to "*both true and false*") and observed that both adults and young children went overwhelmingly for this middle option, thereby showing sensitivity for informativeness.

The research sketched above shows that the failure observed in classic TVJT-tasks does not reflect a genuine inability to derive SIs. This motivated us to move away from this demanding classic task and to use another task in our experiments, that is, a Felicity Judgment Task (FJT; see e.g., Chierchia et al., 2001). In this task, participants are presented with two alternative descriptions of the same situation and they have to decide which one is the best. One advantage of this task is that the scalar alternatives are explicitly presented, and therefore participants do not have to generate them.

Indeed, one factor that recently received attention is the cognitive availability of the scalar alternatives, that is, the ability to generate the relevant alternative that is going to be used to undergo the SI-process. Consider the task of a pre-schooler who observes a situation where three mice enter a hole. Next the child is asked to evaluate a sentence like "*some of the mice entered the hole*." The pragmatic response "*no, that's not a good sentence*" requires them to generate the stronger alternative ("*all of the mice entered the hole*") and compare the information strength. Interesting in this respect is a study by Barner et al. (2011). Four-year-old children were for instance presented with a situation where Cookie Monster was holding three pieces of fruit (and no other pieces of fruit were available in the context). When they were asked whether Cookie Monster was holding only some of the food, the majority said "*yes*." When asked whether Cookie Monster was holding only the banana and the apple, they overwhelmingly said "*no*." Hence, when the

alternatives were provided contextually, as in the last question, children were able to assign strengthened interpretations to utterances when these included the focus element “only.” For the context-independent scale *some/all*, children were not able to do this. In a sentence-picture verification task, Skordos and Papafragou (2016) manipulated the accessibility of the alternative by varying the order of trials. They compared performance of 5-year-old children in the condition in which the trials with “*some*” were presented before the trials with “*all*” with the mixed condition (in which trials with “*some*” and “*all*” were intermixed in a pseudorandom order). In the latter condition, children derived more SIs, probably due to the fact that alternatives were more accessible. In two follow-up experiments, Skordos and Papafragou (2016) showed the importance of relevance. Children used the explicitly mentioned stronger alternative for SI-generation only when the alternative was relevant. In two experiments, with a modified TVJT, Tieu et al. (2016) showed that as early as 4 years old, children can compute free choice inferences. However, they were not able to compute SIs. As an explanation, they offered the restricted alternatives hypothesis: Children have the ability to compute inferences arising from alternatives whose construction does not require access to the lexicon. Because the alternatives from which free choice inferences arise are contained within the assertion, they can be computed. The alternatives of SIs are typically not contained within the assertions and therefore these implicatures are hard. Tieu et al. (2016) also state explicitly that mentioning alternatives helps children to compute the corresponding inferences.<sup>1</sup>

In the current study we use an adapted version of the FJT to investigate further the role of alternatives. Chierchia et al. (2001) investigated if, on their way to full mastery of scalar terms, children might pass through a stage in which they know already some aspects of them. More specifically, Chierchia et al. (2001) examined situations where the children knew that “*and*” truly applies, and tested if children prefer “*and*” above “*or*” through a FJT. Fifteen 5-year-old children were presented with two alternative descriptions of the same situation and they had to decide which one was the best. Remarkably, with the presence of the relevant alternative representations, the children consistently applied SIs. It has to be emphasized that this task does not require the actual derivation of SIs: Comparing the informativity of the competing utterances and applying the Maxim of Quantity will lead to the appropriate response. Foppolo et al. (2012) presented a rather small set of 17 5-year-old children with a similar task, now employing the terms “*some*” and “*all*.” In line with Chierchia et al. (2001), the children’s performance in this FJT was above 95% correct overall. Hence, these children showed comprehension of the ordering of informational strength. Of course, this does not prove that children can derive SIs easily or independently,

but it shows their sensitivity to the informational strength of the competing utterances and the importance of the cognitive availability of alternatives.

In Experiment 1, with 5-year-old children as participants, we build on this research by introducing – in addition to choices between “*some*” and “*all*” – also choices between “*some*” and “*many*” and between “*many*” and “*all*,” which makes a more fine-grained analysis possible. In Experiment 2, we present the same problems, but to older children, that is, 11-year-old children, to test developmental patterns. Moreover, we added temporal scales (with “*sometimes*,” “*often*,” and “*always*”) to test scalar diversity.

# EXPERIMENT 1: FIVE-YEAR-OLD CHILDREN AND QUANTITATIVE SCALAR IMPLICATURES IN A FELICITY JUDGMENT TASK

As a starting point, Experiment 1 uses the FJT by Foppolo et al. (2012), in which statements with “*some*” and “*all*” were compared as alternative descriptions of pictures in which the statement with “*all*” was the most appropriate. We asked, however, a finer-grained research question: How determining is the generation of alternatives, compared to the evaluation of the information strength itself? In order to have part of the answer to this question, we broadened the FJT of Foppolo et al. (2012). In addition to choices between “*some*” and “*all*,” we also presented choices between “*some*” and “*many*” and “*many*” and “*all*,” and this in situations where “*all*,” “*many*” or “*some*” was the most appropriate according to our intuition. Pezzelle et al. (2018) showed that, for sets with four or more objects, quantifiers primarily represent proportions and not absolute cardinalities. Additionally, even without relying on any quantitative or contextual information, quantifiers lie on an ordered scale, that is, “*none*, *almost none*, *few*, *the smaller part*, *some*, *many*, *most*, *almost all*, *all*.” Consequently, in our study “*some*” should be proportionally less than “*many*.”

Table 1 gives an overview of the different types of items. The three possible pairs constructed with “*some*,” “*many*,” and “*all*” were all confronted with situations with two, five, and six

TABLE 1 | The nine different items in our adapted Felicity Judgment Task.

The presented scalar-pairs	The type of item*		
	Critical	Ambiguous	Control
Some-All	SA6	SA5	SA2
Some-Many	SM5	SM6	SM2
Many-All	MA6	MA2	MA5

\*For each item a situation was presented in which six actions were taken, of which six, five or two were successful. The number refers to the number of successes of the main character. Hence, Some-All 6 successes (SA6); Some-All 5 successes (SA5); Some-All 2 successes (SA2); Some-Many 6 successes (SM6); Some-Many 5 successes (SM5); Some-Many 2 successes (SM2); Many-All 6 successes (MA6); Many-All 5 successes (MA5); Many-All 2 successes (MA2).

<sup>1</sup>Two recent studies with adults also highlighted the effect of activating alternatives. In an eye tracking study, Foppolo and Marelli (2017) obtained new evidence for the incremental derivation of the pragmatic some-but-not-all interpretation of “*some*.” They interpret these findings within the grammatical account of SI (e.g., Chierchia et al., 2012): when scalar alternatives are active, the SIs are factored in locally and incrementally during the online processing of scalar quantifiers. With a structural priming paradigm, Rees and Bott (2018) convincingly demonstrated that adults are sensitive to the salience of alternatives when deriving scalar implicatures.

successes out of six. For instance, there was a boy throwing rings around the trunk of an elephant. He had six attempts and he succeeded in two ( $\approx$  “some”), five ( $\approx$  “many”) or six ( $\approx$  “all”) attempts. This leads to nine combinations. These combinations can be divided in three categories.

The first category consists of three control items (SA2, SM2, MA5), which test the knowledge of the terms, by presenting a pair of assertions, from which one is false and one correct. For instance for item SA2, when there are two successes, the children have to choose between “*some marbles landed in the whole*” and “*all marbles landed in the whole*.” We expect children to perform well on these items, because we expect these items to test the basic lexical/semantic knowledge of the terms used.

The second category consists of the three more or less typical critical items (SA6, SM5, MA6), where an underinformative assertion (“some” or “many”) is paired with a strong true alternative (“many” or “all”). For instance for item SA6, when there are six successes, the participants have to choose between “*some arrows landed in the rose*” and “*all arrows landed in the rose*.” If the difficulty of SIs really lies in the generation of alternatives and not in the evaluation of the informational strength, then these items should be answered well. However, given the absence of a comparison process for the control items and a potentially still fragile evaluation system, performance might be lower for the critical items than for the control items.

Finally, the third category contains three ambiguous situations (SA5, SM6, MA2), where none of the alternatives gives a very appropriate description. In item SA5, an underinformative assertion is paired with an assertion that is too strong: in the case of five successes, the underinformative “some” is paired with the too strong “all.” Consequently, the underinformative “some” is the most appropriate choice. In item SM6, two underinformative assertions are paired: in the case of six successes, the underinformative “some” is paired with the underinformative “many.” Although both assertions are underinformative, one can still make a distinction between them: the difference in informational strength with respect to the six successes ( $\approx$  “all”) is the smallest with “many,” which is therefore the most appropriate choice. In item MA2, two too strong assertions are presented: in the case of two successes, “many” is paired with “all.” Although both assertions are too strong, the difference in informational strength with respect to the correct two successes ( $\approx$  “some”) is the smallest with “many,” which is therefore the most appropriate choice. Hence, these ambiguous items can be solved only if one is able to compare in a more finely grained fashion the informational structure. Given a potentially still fragile evaluation system, performance is expected to be lower than for the control items and maybe even lower than for the critical items, because no clear right answer was presented.

In sum, in the current FJT we wanted to investigate if 5-year-old children can select the most appropriate term when presented with a choice. On the basis of the literature on the importance of alternatives and on the basis of the work of Foppolo et al. (2012), we expected the children to perform well. We broadened the task, by using also the term “many.” We

expected on the basis of this broadening that the difficulty of the task would increase. Moreover, the work on alternatives shows that the mere presence of alternatives is not a wonder solution. Consequently, we expected the control items (SA2, SM2, MA5) to be easier than the critical items (SA6, SM5, MA6) and the ambiguous items (SA5, SM6, MA2).

## Methods

### Participants

We tested 59 5-year-old children (27 boys and 32 girls; mean age = 61 months,  $SD = 3$  months). They were all recruited from two primary schools in Belgium. All were native Dutch speakers, including some bilingual children. This research has been reviewed and approved by the ethical review board SMEC of the University of Leuven. A written informed consent was obtained from the participants’ parents.

### Materials and Procedure

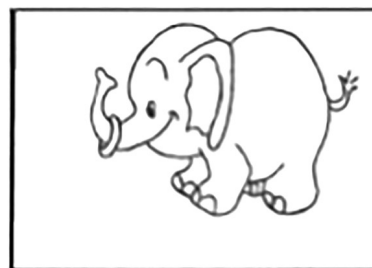
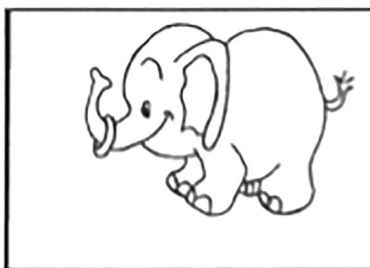
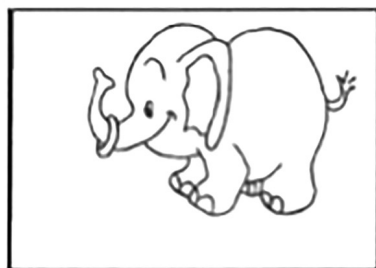
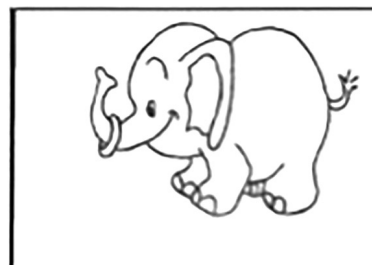
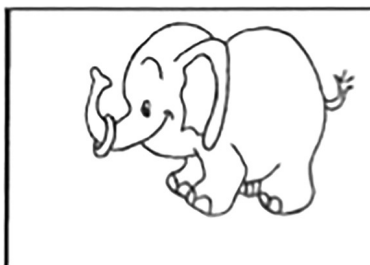
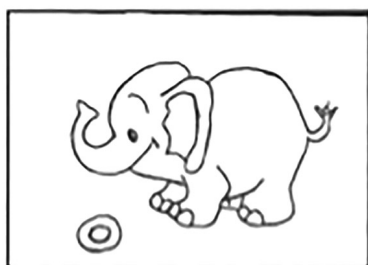
We tested children with a version of the FJT in which we presented two statements, which contained either “some” or “many” or “all” (“*sommige*,” “*vele*,” “*alle*” in Dutch, the language of the experiment; see **Appendix A** for the material) as alternative descriptions. These statements were accompanied by drawings in which two, five, or six successes were achieved. The children had to decide which statement did fit the drawing best. The children received in total nine stories in a random order.

The participants were tested individually in a quiet space. At the beginning of the experiment, they were told that the investigator would tell a few stories, which she would illustrate with drawings. Next, two animals were introduced, Kwaak the frog and Botje the fish. These two plush hugs were presented to the children as good friends of the researcher. They would both make a statement about each of the stories. It was the child’s task to judge each time which puppet said it better (= Felicity Judgment Task). Moreover, we took care to assert that there was not one puppet that was always uttering the best statements. Before the experiment started, two practice items were given to familiarize the children with the procedure (see **Appendix A**).

Each experimental item started with a story that was told and which was illustrated by means of drawings, as illustrated in **Figure 1**. First, the context of the story is told and shown with the contextual drawing. Next it is told how the situation unfolds while six action drawings are shown. For instance, it was told that Victor, a small boy, and Olli, the elephant, are good friends, while a drawing is shown of the two together. Then it is told that they play a game. Victor has to throw six rings around Olli’s trunk. Next, each attempt (success or failure) is described and illustrated with a drawing. For instance, “*The first time Victor fails and the ring is not around Olli’s trunk. Victor tries again and... it works, the ring is around the trunk. The next time also. And again he succeeds. Also the fifth time is the ring around Olli’s trunk. Now Victor throws for the last time and... yes! Once again the ring is sitting around the trunk!*” After this story, both Kwaak and Botje make a statement about the story, and the participant has to indicate which puppet said it better. With the above story, the two statements might be (with between square brackets the English translation):



*“Victor and Olli the elephant are friends. They play a game. Victor has to try to throw as many rings as possible over the trunk of Olli.”*



*“The first time Victor fails and the ring is not around Olli’s trunk. Victor tries again and ... it works, the ring is around the trunk. The next time also. And again he succeeds. Also the fifth time the ring is around Olli’s trunk. Now Victor throws for the last time and ... yes! Once again the ring is around the trunk!”*

**FIGURE 1** | Illustration of one item (with five successes) with the accompanying pictures.

Kwaak: “Victor gooide sommige ringen rond Olli’s slurf.”

[Victor threw some rings around Olli’s trunk]

Botje: “Victor gooide vele ringen rond Olli’s slurf.”

[Victor threw many rings around Olli’s trunk]

## RESULTS

Table 2 presents the percentage of appropriate choices for the nine experimental items and Figure 3 depicts the results graphically (together with part of the data of Experiment 2). There was no difference in performance between different versions presented. Overall, the children’s performance in this

FJT was quite good, with 87% correct overall and with at least 70% correct. In other words, the 5-year-old children were able to choose clearly above chance which element of the scale *< some, many, all >* from a pair is more appropriate in a given context (Binomial probability = 0.001 for the lowest score, i.e., 70%). Moreover, it is not only that the children, as a group, are better than chance. Only one child scored less than chance level, an additional two children answered less than 2/3 of the problems correctly (but were above 1/2) and three children precisely answered 2/3 of the problems correctly. In other words, 90% of the children answered more than 2/3 of the problems correctly. Even if we look at the problem types separately, a similar picture emerges. On the critical and ambiguous problems, four children



**TABLE 2** | The proportion of appropriate answers and the standard deviation in Experiment 1.

The presented scalar-pairs	The type of item		
	Critical	Ambiguous	Control
Some-All	0.95 [0.222] (SA6)	0.70 [0.464] (SA5)	0.97 [0.183] (SA2)
Some-Many	0.73 [0.448] (SM5)	0.93 [0.254] (SM6)	0.88 [0.326] (SM2)
Many-All	0.86 [0.345] (MA6)	0.88 [0.326] (MA2)	0.93 [0.254] (MA5)

scored less than chance level (these were different children for the critical and ambiguous problems), respectively 19 and 20 children answered 2/3 of the problems correctly and respectively 36 and 35 children answered all three problems correctly. For the control items, two children scored less than chance level, nine children answered 2/3 of the problems correctly and 48 children answered all three problems correctly.

Given the binary nature of the dependent variable, we performed a mixed effects logistic regression (Baayen et al., 2008; Jaeger, 2008; Bates et al., 2015). The model fitting procedure was implemented in R using the `glmer()` function from the `lme4` package (Bates et al., 2015). The dependent variable was the appropriateness score (1 for appropriate and 0 for inappropriate). The independent variables were Type (with the levels Control, Critical, and Ambiguous) and Quantifier-Pair (with the levels Some-All, Many-All, Some-Many). All models included random intercepts for participants and following Baayen et al. (2008) we additionally opted for a random interaction between Type and participant identifier. We started with the most complex fixed effects structure, including the two-way interaction between Type and Quantifier-Pair and main effects. We conducted likelihood ratio tests ( $\alpha = 0.05$ ) with the mixed function from the `afex` package to determine the strongest model (Singmann et al., 2018). The model with the interaction was significantly better than the others [ $\chi^2(4) = 28.23$ ,  $p < 0.00001$ ]. For a complete description of the final model, see Table 3. The control items were significantly easier than the critical items (85% vs. 93%;  $Z = 2.34$ ,  $p = 0.0497$ ) and the ambiguous items (84% vs. 93%;  $Z = 2.18$ ,  $p = 0.0292$ ). We analyzed the significant interaction further by pairwise contrasts, using Bonferroni corrected `lsmeans()`. This revealed three significant differences for the interaction between Type and Quantifier-pair. For the SA-pairs, the ambiguous item (SA5) was more difficult than the critical item (SA6; 70% vs. 95%;  $Z = 3.36$ ,  $p = 0.0024$ ) and the control item (SA2; 70% vs. 97%;  $Z = -3.42$ ,  $p = 0.0019$ ). For the SM-pairs, the critical item (SM5) was more difficult than the ambiguous item (SM6; 73% vs. 93%;  $Z = -2.50$ ,  $p = 0.0375$ ).

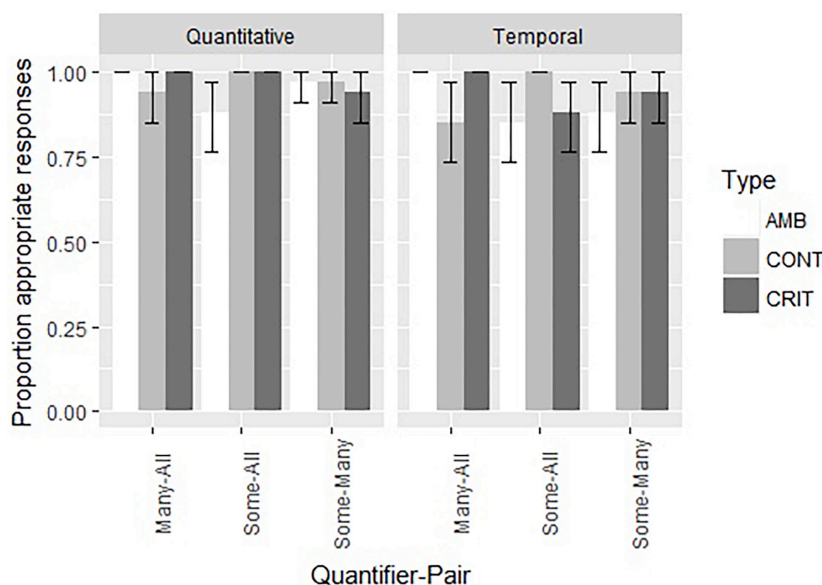
## Discussion Experiment 1

We tested if adding an extra term, that is, “many,” would lead to similar results as the Foppolo et al. (2012) study. Despite this extra term, the children’s performance in our Felicity Judgment Task was still convincingly above chance level, with all pairs

answered appropriately above 70% and with an overall score of 87%. Although these results are in general less good than the ones of Foppolo et al. (2012), where an overall rate of 95% was observed, our results still show that children are able to choose which element of the scale  $< \text{some, many, all} >$  is more appropriate in a given context. In other words, when they are offered with an alternative, they can more or less easily decide which one fits the situation best.

Nevertheless, some interesting differences were observed. As predicted, the critical items were more difficult than the control items. For the latter, the lexical/semantic knowledge does not leave room for doubt about which is the most appropriate answer. For the critical items, the informational strength of the two alternatives has to be compared, in order to provide the correct answer. The necessity of the comparison process for the critical items seems to have caused the lower performance on the critical items. Compared to the control items, performance was also lower for the ambiguous items, which can only be solved by a more sophisticated comparison process: neither of the alternatives is perfect, so a fine-grained comparison is needed. Interestingly, we did not observe a significant difference between the critical and the ambiguous items. In other words, the informational strength evaluation process was sophisticated enough to handle both kinds of items.

The two most difficult items were the ambiguous SA5 and the critical SM5 item. In the ambiguous SA5 item, an underinformative assertion is paired with an assertion that is too strong. Therefore, this item is somewhat different from the two other ambiguous items, where the alternatives are either both too strong or both too weak in terms of informational strength. For the latter two items, one only has to take the distance from the “correct” answer to make the decision. This strategy does not work for the SA5 item, because it leads to the inappropriate (and false) “all” choice. “All” is indeed in terms of distance closer to five than “some.” In other words, it makes sense that this item is more difficult than other items: rather sophisticated inferencing is needed to produce the appropriate answer. Another reason why this item might be more difficult is that “some” not only leads to the implicature “not all,” but also to the implicature “not many,” which then blocks the children. However, the derivation of the “not many” implicature by children is unlikely given what we know of their ability to derive SIs. If they would do it anyway, there is a good chance they would see the violation of the implicature as less problematic than the falsity of “all.” Why the critical SM5 is also more difficult than other items is less clear. A difference between SM5 and the two other critical items, that is SA6 and MA6, is that the latter two are connected to the endpoint, that is, to the strongest case (six successes). SM5 however is linked to five successes, which is at the top of the scale, but is not an endpoint. This might cause some extra insecurity and therefore explains the lower appropriateness-scores for this item. Support for this hypothesis comes from the work of Van Tiel et al. (2016). They observed for adults large differences between rates of scalar inferences on different scales (between 4 and 100%). One important factor causing these differences was the openness/closeness of the scales. Closed scales (like e.g.,  $< \text{some, all} >$ , where “all” is the end point) lead to more scalar inferences



**FIGURE 2 |** The proportion of appropriate choices and the standard error for both the nine quantitative and the nine temporal items in Experiment 2.

than open scales (like e.g., *<cool, cold>*, where “cold” is not an end point). Unlike *<some, all>*, *<some, many>* is an open scale and maybe therefore more difficult.

## EXPERIMENT 2: ELEVEN-YEAR-OLD CHILDREN AND QUANTITATIVE AND TEMPORAL SCALAR IMPLICATURES IN A FELICITY JUDGMENT TASK

Although performance was already high, for some items there was clearly room for improvement. In Experiment 2, we investigated whether 11-year-old children would perform better than the 5-year-old children. With respect to the more traditional TVJT, there is a clear developmental trend observed in the literature (see e.g., Pouscoulous et al., 2007). Therefore, we also expected a better performance by the 11-year-old children on our FJT.

Additionally, we wanted to gather some extra data with respect to the issue of scalar diversity. Until recently, the uniformity of SIs had not been questioned. Doran et al. (2009) tested this assumption by looking not only to the scale *<some, all>* but also to scales like *<possibly, definitely>*, *<beginner, intermediate, advanced>* and *<warm, hot>*. They observed in adults a significant variability between the rates of pragmatic answers that these scalar terms elicit. Likewise, a survey of ten experiments by Geurts (2010, pp. 98–99) showed that, for disjunction sentences (containing “or”), the mean rate of SIs was much lower than for the sentences containing “some”: 35% against 56.5%. Van Tiel et al. (2016) build further on the work by Doran et al. (2009). Apart from the effect of closed versus open scales, they observed that giving the adjectives a richer context leads to more scalar inferences. Also, word class and

semantic distance had a significant effect on the rate of pragmatic responses, while there was no effect of focus, word frequency, or strength of association between stronger and weaker terms. In other words, different types of scales are not all the same and we cannot use one type as the prototypical type. The *<some, all>* scale triggers unusually high levels of pragmatic answers. It is worth noting that recently Benz et al. (2018) provided some support for a modified version of the uniformity hypothesis on the basis of their work on negative strengthening.

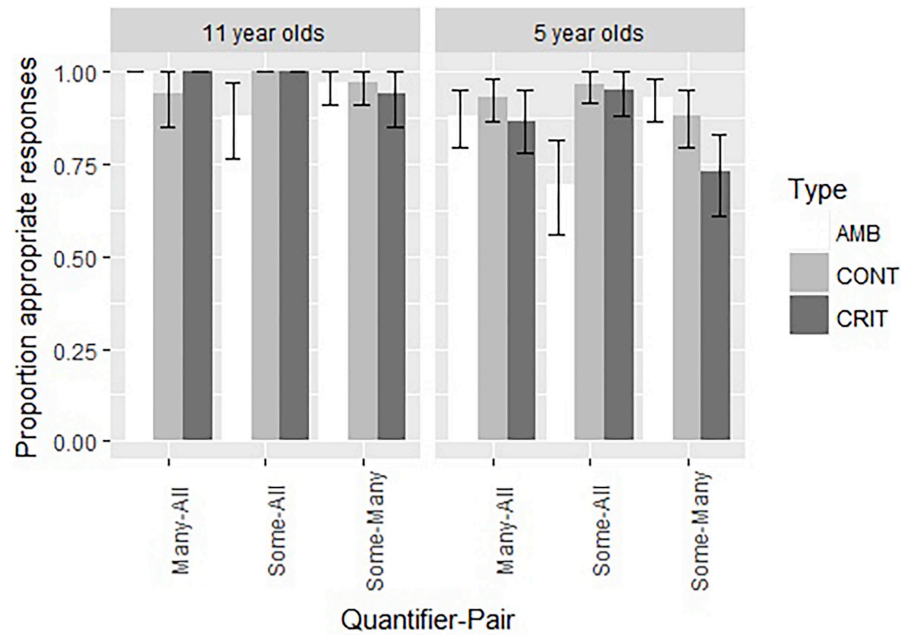
To the best of our knowledge, there is no research on scalar diversity with the FJT. Chierchia et al. (2001) already showed good performance with the scale *<or, and>*, Foppolo et al. (2012) with *<some, all>*, but the two scales were not compared. In the current experiment, we directly compared performance on the quantitative scale *<some, many, all>* with the temporal scale<sup>2</sup> *<sometimes, often, always>*. We opted for these two scales for two reasons. First, they allowed us to use the same materials and procedure. Second, we wanted a scale which was not too difficult for children and Van Tiel et al. (2016) observed for these two scales in adults a high performance. Given the high accuracy of the 5-years-old children in Experiment 1 on the quantitative SIs, we expected not too many difficulties with the temporal SIs.

## Methods

### Participants

We tested 34 11-year-olds (15 boys and 19 girls; mean age = 11 years; 4 months, *SD* = 5 months). They were all recruited from two primary schools in Belgium. All were native

<sup>2</sup>In order to avoid confusion, we explicitly mention that we use the term “temporal scalar implicature” differently from other authors, e.g., Altschuler and Schwarzschild (2012) and Thomas (2012), who use it to describe the situation where one infers from “John was in his office” that “John isn’t in his office now/anymore.”



**FIGURE 3 |** A comparison of the proportion of appropriate choices and the standard error for the nine quantitative items in Experiment 1 (5-year-old children) and Experiment 2 (11-year-old children).

**TABLE 3 |** A complete description of the final model for Experiment 1: Type \*Quantifier-Pair + (1| Participant) + (1| Type: Participant).

**Estimators of the relative quality of the statistical model: statistical model:**

AIC	BIC	logLik	deviance	df.resid
391.2	438.2	−184.6	369.2	520

**Scaled residuals:**

Min	1Q	Median	3Q	Max
−4.8245	0.1746	0.2574	0.3483	1.0312

**Random effects:**

Groups	Name	Variance	Standard Deviation
Type: Participant	(Intercept)	0.3536	0.5946
Participant	(Intercept)	0.4528	0.6729

Number of obs: 531, Type: Participant, 177; Participant, 59

**Fixed effects\*:**

	Estimate	Standard Error	z-value	Pr(>  z )
(Intercept)	2.42135	0.27328	8.860	<2e-16*
Many-All	0.23252	0.25152	0.924	0.3552
Some-Many	0.04477	0.22243	0.201	0.8405
Critical Items	0.57824	0.26523	2.180	0.0292*
Ambiguous Items	−0.23347	0.23457	−0.995	0.3196
Many-All: Critical	0.48963	0.39929	1.226	0.2201
Some-Many Ambiguous guous	0.07821	0.34245	−0.228	0.8194
Many-All: Critical	0.84582	0.35910	2.355	0.0185*
Some-Many Ambiguous	−0.10110	0.29951	−0.338	0.7357

\*The fixed effects are Quantifier-Pair (with the levels Some-All, Many-All, and Some-Many) and Type (with the levels Control, Critical, and Ambiguous).

Dutch speakers, including a few bilingual children. A written and informed consent was obtained from the participants' parents.

## Materials and Procedure

The same materials and procedure were used as in Experiment 1. The only difference was that the participants had to solve both the Quantitative Scale (QS, as in Experiment 1, with “all,” “many,” and “some”) and the Temporal Scale (TS, with “always,” “often,” “sometimes”). For the exploration of the temporal implicatures the statements that were presented after the context story were rephrased. Consider the example we used in Experiment 1. The same drawings (see **Figure 1**) were used. First, the same context story was given (Victor, a small boy, and Olli, the elephant, are good friends). Next the ring-throwing game was introduced, with the same sentences and drawings, “*The first time Victor fails and the ring is not around Olli's trunk. Victor tries again and... it works, the ring is around the trunk. The next time also. And again he succeeds. Also the fifth time is the ring around Olli's trunk. Now Victor throws for the last time and....yes! Once again the ring is sitting around the trunk!*” After the story, the two puppets made a statement about the story:

Kwaak: “Victor heeft soms de ring rond Olli's slurf geworpen.”

[Victor has *sometimes* the ring around Ollie's trunk thrown]

Botje: “Victor heeft altijd de ring rond Olli's slurf geworpen.”

[Victor has *always* the ring around Olli's trunk thrown]

Eighteen participants started with the Quantitative Scale and received afterward the Temporal Scale, while 16 participants started with the Temporal scale. To make the comparison easier, we will use the label SA for both the *some-all* (preceded by Q\_) and the *sometimes-always* pairs (preceded by T\_), SM for both the *some-many* (preceded by Q\_) and the *sometimes-often* pairs (preceded by T\_), and MA for both the *many-all* (preceded by Q\_) and the *often-always* pairs (preceded by T\_).

## Results Experiment 2

We observed no difference in measurements between the two blocks (starting with the quantitative items vs. starting with the temporal items). Therefore, we collapsed the data over the two orders. Likewise, there was, as in Experiment 1, no difference in performance between the different versions presented. **Table 4** presents the percentage of appropriate choices for the nine experimental items for the two scales and **Figure 2** depicts the results graphically.

Overall, performance on the Quantitative items in this FJT was very good, with 97% correct overall and with at least 88% correct (Binomial probability = 0.001). For the Temporal items, a similar pattern was observed: performance was very good, with 93% correct overall and with at least 85% correct (Binomial probability = 0.001). All children answered more than 2/3 of the quantitative items correctly; 33 children answered more than 2/3 of the temporal items correctly, two children answered precisely 2/3 of the temporal items correctly. If we look at the problem types separately, a similar picture emerges. On the quantitative and temporal critical items, respectively two and six children answered 2/3 of the problems correctly, and respectively 33 and 29 children answered all three problems correctly. On the

**TABLE 4 |** The proportion of appropriate answers and the standard deviation in Experiment 2.

The presented scalar-pairs	The type of item		
	Critical	Ambiguous	Control
Q_Some-All*	100 [0.000]	88 [0.327]	100 [0.000]
Q_Some-Many	94 [0.239]	97 [0.172]	97 [0.172]
Q_Many-All	100 [0.000]	100 [0.000]	94 [0.239]
T_Some-All	88 [0.327]	85 [0.360]	100 [0.000]
T_Some-Many	94 [0.239]	88 [0.327]	94 [0.239]
T_Many-All	100 [0.000]	100 [0.000]	85 [0.360]

\*Q\_ indicates the quantitative items, T\_ indicates the temporal items.

quantitative and temporal ambiguous items, respectively five and seven children answered 2/3 of the problems correctly, and respectively 30 and 28 children answered all three problems correctly. On the quantitative and temporal control items, three children answered 2/3 of the problems correctly, three children answered only 1 of the temporal items correctly, and respectively 32 and 29 children answered all three problems correctly.

As for Experiment 1, we performed a mixed effects logistic regression, with the model fitting procedure `glmer()` function from the `lme4` package (Bates et al., 2015). The dependent variable was the appropriateness score (1 for appropriate and 0 for inappropriate). The independent variables were Type (with the levels Control, Critical, and Ambiguous), Quantifier-Pair (with the levels Some-All, Many-All, and Some-Many), and Diversity (Quantitative and Temporal). All models included random intercepts for participants. This model and other more complex models failed to converge, possibly due to ceiling effects. Non-parametric analyses confirm a lack of differences between the different conditions and the very high performance (see **Appendix B**). However, of the simple models, the one with Diversity as a factor was the best [model fitting verified through the Akaike information criterion (AIC) and the BIC]. The estimation of the fixed effects of Diversity was consistent, even if we had a more complex random structure, for instance by including a random interaction between participants and Type. However, we opted for the model without problems with convergence and degenerated random effects, which is the one with only a random intercept for participants. This model indicates that the temporal items were more difficult than the quantitative ones (93% vs. 97%;  $Z = 2.20$ ,  $p = 0.0278$ ), a difference also confirmed through non-parametric analyses. For a complete description of this simple final model, see **Table 5**.

## Comparison Results Experiments 1 and 2

We also compared the performance on the quantitative items between the younger age group of Experiment 1 and the older group of Experiment 2. These data are presented graphically in **Figure 3**. As before, we performed a mixed effects logistic regression, with the model fitting procedure `glmer()` function from the `lme4` package (Bates et al., 2015). The dependent variable was the appropriateness score (1 for appropriate and 0 for inappropriate). The independent variables were Type (with



**TABLE 5 |** A complete description of the simple model for Experiment 2: Diversity + (1| Participant).

Estimators of the relative quality of the statistical model: statistical model:					
AIC	BIC	logLik	deviance	df.resid	
240.9	254.2	−117.5	234.9	609	
Scaled residuals:					
Min		1Q	Median	3Q	Max
−4.5978		0.1051	0.1638	0.2175	0.6007
Random effects:					
Groups		Name	Variance	Standard. deviation	
Participant		(Intercept)	1.457	1.207	
Number of obs: 612, Participant: 34					
Fixed effects°:					
			Estimate	Standard Error	z value
					Pr( >   z   )
(Intercept)			4.0209	0.4940	8.139
Diversity			−0.8880	0.4036	−2.200

<sup>°</sup>The fixed effect is Diversity (with the levels Quantitative and Temporal).

the levels Control, Critical, and Ambiguous), Quantifier-Pair (with the levels Some-All, Many-All, and Some-Many), and Age (5-year-olds vs. 11-years-old). As for Experiment 1, all models included random intercepts for participants and a random interaction between Type and participant identifier. We started with the most complex fixed effects structure, including the three-way interaction, two-way interactions and main effects. We conducted likelihood ratio tests ( $\alpha = 0.05$ ) with the mixed function from the afex package to determine the strongest model. The best model contained Age ( $\chi^2(1) = 23.32$ ,  $p < 0.00001$ ) and the interaction between Quantifier-Pair and Type ( $\chi^2(4) = 25.23$ ,  $p < .00001$ ). For a complete description of the final model, see **Table 6**. The 11-year-olds performed significantly better than the 5-year-olds (97% vs. 85%;  $Z = 4.39$ ,  $p < 0.00001$ ). We analyzed the significant interaction further by pairwise contrasts, using Bonferroni corrected lsmeans (). This revealed three significant differences for the interaction between Type and Quantifier-pair. For the SA-pairs, the ambiguous item (SA5) was more difficult than the critical item (SA6; 79% vs. 98%;  $Z = 3.36$ ,  $p = 0.0024$ ) and the control item (SA2; 79% vs. 97%;  $Z = -3.42$ ,  $p = 0.0019$ ). For the SM-pairs, the critical item (SM5) was more difficult than the ambiguous item (SM6; 85% vs. 95%;  $Z = -2.50$ ,  $p = 0.0375$ ).

## GENERAL DISCUSSION

In two experiments, we tested the ability of both 5-year-old and 11-year-old children to select the most appropriate item in a FJT. The set-up of our experiments was inspired by Foppolo et al. (2012), but we broadened the typical <some, all> scale to a <some, many, all> scale. Two aspects seem immediately relevant.

First, both age groups performed well above chance level. When asked to choose which of the two alternatives is the best description, the children were good in making the right decision. In other words, with the alternatives explicitly presented, even young children are able to pick the pragmatically most appropriate option. Second, despite performing at a high level, the 5-year-old children were less able to choose the appropriate answer compared to the 11-year-old children. Interestingly, this difference was not observed on the control items, but only on the critical and on the ambiguous items. Likewise, for the group of 5-year-old children separately, the critical and the ambiguous items were more difficult than the control items.

These findings are important for the literature about the role of alternatives. Our data confirm the claim that the explicit presence of alternatives eases pragmatic reasoning for young children (see e.g., Barner et al., 2011). Young children seem to be able to pick the most appropriate answer, which is the only correct one in the case of the control items and the one with the most information strength in the case of the critical and most ambiguous items. However, our data also point to the importance of the comparison process. The critical and the ambiguous items were more difficult than the control items. So, the mere presence of the most appropriate alternative is not enough to elicit performance at ceiling level. For the critical and the ambiguous items, the information strength of the two alternatives has to be compared and this seems to have increased the difficulty level. We have to emphasize that even for these items the performance was clearly above chance level: children can reliably solve these problems. However, performance was lower on these items than on the control items, which can be interpreted as a sign of the processing load of the comparison process or of the intrinsic difficulty of the comparison itself. This interpretation is in line with the constraint-based approach of

SIs (Degen and Tanenhaus, 2015, 2016), which claims that the probabilistic support for the implicature in context determines the probability of a SI and the speed at which it is derived (see e.g., Breheny et al., 2006 for earlier results in this direction). Greater contextual support leads to a higher probability for the implicature and a faster derivation. The explicit use of a third alternative in the experiments (not only “some” and “all,” but also “many”) could have complicated the process. It is indeed conceivable that, in contrast to the Foppolo et al. (2012) study, the children in the current study spontaneously assumed that a bigger set of alternatives was available for the speaker, which in turn affected the difficulty of the inferences drawn. Degen and Tanenhaus (2016) for instance showed that the availability of lexical alternatives outside the <all-some> scale, that is, number alternatives, increased the difficulty of interpreting “some.” In our experiment, the introduction of “many” might have played a similar role. It is possible that this was especially the case for the youngest children. Moreover, the observed difficulty with the critical and the ambiguous items is in agreement with the idea that the contextual support for their appropriate choices is less strong than for the appropriate choices for the control items.

Given that especially two items (i.e., SA5 and SM5) were more difficult for the youngest children, we believe that it's not

so much the general processing load of the comparison process itself which caused the effect, but the intrinsic difficulty of some comparisons. For both items, it can be argued that the most appropriate choice received less contextual support compared to the other items. In hindsight, it is therefore not surprising that the ambiguous SA5 and the critical SM5 turned out to be the hardest ones. The ambiguous SA5 item is the only ambiguous item where an underinformative assertion is paired with a too strong assertion. For this item, the child had to realize that the shorter distance between “all” (i.e., six successes) and five successes, compared to the distance between “some” (i.e., at least one success) and the five successes, has to be neglected, given the fact that “all” is too strong in this case. For the other two ambiguous items, the alternatives were either both too strong or both too weak in terms of informational strength, which enabled the children to focus only on the distance from the “correct” answer for their decision. In the discussion of Experiment 1, we mentioned another potential explanation for the difficulty of SA5. “Some” might not only elicit a “not all” implicature, but also a “not many” implicature, which consequently might have blocked the children. However, this clearly is a rather sophisticated inferencing, which you would not expect from the youngest children, but maybe from the older ones. Given the

**TABLE 6 |** A complete description of the final model for Experiment 1: Age + Type \*Quantifier-Pair + (1| Participant) + (1| Type: Participant).

Estimators of the relative quality of the statistical model:					
AIC	BIC	logLik	deviance	df.resid	
472.9	529.7	−224.5	448.9	825	
Scaled residuals:					
Min	1Q	Median	3Q	Max	
−7.7700	0.1206	0.2103	0.3084	0.9694	
Random effects:					
Groups	Name	Variance	Standard Deviation.		
Type: Participant	(Intercept)	0.00533	0.07303		
Participant	(Intercept)	0.529134	0.72742		
Number of obs: 837, Type: Participant, 177; Participant, 59					
Fixed effects*:					
	Estimate	Standard Error	z value	Pr(>  z )	
(Intercept)	2.19039	0.42243	5.185	2.16e-07*	
Age 11	1.61636	0.37028	4.365	1.27e-05*	
Some-All	1.09034	0.70911	1.538	0.1241	
Some-Many	−1.04144	0.47935	−2.173	0.0298*	
Ambiguous	0.15599	0.55943	0.279	0.7804	
Control	0.33225	0.58069	0.572	0.5672	
Some-All: Ambiguous	−2.58529	0.86495	−2.989	0.0028*	
Some-Many Ambiguous guous	1.42184	0.78602	1.809	0.0705	
Some-All: Control	0.09507	1.10093	0.086	0.9312	
Some-Many: Control	0.70954	0.75133	0.944	0.3450	

\*The fixed effects are Age (with the levels Age\_5 and Age\_11), Quantifier-Pair (with the levels Many-All, Some-All, and Some-Many), and Type (with the Critical, Ambiguous, and Control).

fact that the 11-year-old did not struggle so much with this item, we believe that this explanation is unlikely, although it cannot be completely ruled out on the basis of our study. Future research should look further into this issue. A related factor is the potential effect of order of presentation, which might definitely be of importance for the ambiguous and critical items. As written in the results sections, in our experiments there was no order effect. However, we only presented nine different items and we did therefore not present similar items after each other. Suppose participants receive a few ambiguous items of the SA5-type. This item forces them to accept “some” with five successes (or “all” with five successes). Multiple presentations of this item might consequently have an influence on the subsequent items with “some.” Using reaction times as an extra dependent variable is clearly advisable here. The critical SM5 item is also special, because the endpoint, that is, the strongest case (six successes), is not part of the comparison process. Van Tiel et al. (2016) already observed for adults that scales with an endpoint lead to more scalar inferences than scales without.

Experiment 2 showed that with development, children are able to deal with these more difficult items. For the 11-year-old children, there was no difference between the control, critical, and ambiguous quantitative items, and also pairwise comparisons between the nine different items revealed no significant differences. In other words, at that age, when presented with two alternatives, irrespective of the difficulty of the comparison process, the 11-year-old children are able to pick the most appropriate quantitative description. This is maybe not very surprising because at age eleven children seem to be able to perform a large range of pragmatic inferences (but not all, see e.g., Janssens et al., 2015 on conventional implicatures). For instance, the age of ten is critical for metaphor (Lecce et al., 2018), idiom (Kempler et al., 1999), and irony understanding (Glenwright and Pexman, 2010). It will be interesting to see in future research how children younger than five behave on the current task: Which items will be the most difficult for them and from which age is performance above chance level? We know that the classic TVJT with SIs is often too difficult for 3-year old children (e.g., Hurewitz et al., 2006; Janssens et al., 2014), but with contextually grounded, *ad hoc* implicatures children by age three and a half, and perhaps even slightly earlier, can cope with it (see e.g., Stiller et al., 2015). Similarly, Tieu et al. (2016) showed that 4-year-old children could compute free choice inferences but not SIs. Given the high performance on our task, we can expect already above chance performance for the 3.5 year old children. Future research could also investigate how performance is with other numbers. Here we opted for a maximum of six potential successes, given the young age of our participants in Experiment 1. Not only will it be interesting to see how children cope with situations with a higher number of potential successes, this manipulation would also give the opportunity to play a bit more with the set-sizes attached to “some” and “many.” Additionally, such a manipulation would provide evidence about which conditions trigger which quantifiers easier, because it is perfectly conceivable that some set-sizes are better fits for “some” or “many” than others (see also Degen and Tanenhaus, 2015). There is some work on this with adults (see e.g., Newstead and Coventry, 2000;

Coventry et al., 2005, 2010; Van Tiel, 2014; Pezzelle et al., 2018), but to the best of our knowledge not with children. Especially relevant for our results might be the observation of Pezzelle et al. (2018) that both low- and high-magnitude quantifiers are ordered along a scale, but that the high-magnitude quantifiers are extremely close to each other, which indicates that their representations overlap. This kind of overlap or “confusion” might be bigger for young children, and might explain some of the difficulties that they experience with “many.” Finally, manipulating the range of number of items also opens an extra link with the work of Degen and Tanenhaus (2015, 2016), which showed that “some” competes with numbers in the subitizing range, which caused a slower processing.

The results of Experiment 2 teach us that, although we explicitly opted for very similar scales that elicited high numbers of scalar responses from adults (Van Tiel et al., 2016), the temporal items were somewhat more difficult than the quantitative ones for the 11-year-old children. We want to emphasize, however, that the difference is small and only present in a simple model of the data and needs replication in subsequent research. A reason for the observed difference might be found in the stories that we used to introduce each pair of utterances. In these stories, we mentioned a success or a failure, one after the other, and so on, until six events were described. Although this can be seen as a temporal framework, no explicit temporal information was given. The mere mentioning of the different attempts one after the other might have therefore advantaged the quantitative implicatures. If that is the case, we can expect a bigger difference between the two scales for the younger children. This is also of interest for future research. Nowadays there seems to be great concern for the diversity of scalar expressions, with Van Tiel et al. (2016) as a great example (see also e.g., Doran et al., 2009; Geurts, 2010). However, from a developmental point of view, clearly much more research is necessary. Also interesting in this respect is the observation that the most difficult critical item in Experiment 1 was one where the endpoint of the scale was not involved in the comparison process. Van Tiel et al. (2016) already argued that scales with and without an endpoint differ from each other.

A last consideration from our data concerns our use of the extra term “many.” This is not the first demonstration that a small change in a simple experiment investigating SIs can lead to important differences in behavioral patterns. The introduction by Katsos and Bishop (2011) of a middle option in the classic binary TVJT (‘I do agree’ vs. ‘I disagree’ became ‘I totally agree,’ ‘I agree a bit,’ and ‘I totally disagree’) proved to be crucial in developmental studies. In the binary task children accept underinformative sentences while adults reject them. When a middle option is present, both adults and children clearly prefer this middle option. Hence, it seems that in the binary task children are not insensitive to underinformativeness, but they do not show it, whereas in the ternary task sensitivity to informativeness is demonstrated through the possibility of showing tolerance to violations of informativeness, by choosing the middle value for underinformative statements. Wampers et al. (2017) and Schaeken et al. (2018) evidenced that, with such a ternary task, respectively patients with psychosis and children with autism

spectrum disorder produce less pragmatic responses, while such a difference was not observed with the classic binary task. In other words, a more nuanced task revealed a previously not visible effect, casting new light on the range of pragmatic difficulties in atypical populations. Similarly, in the current study, the introduction of some extra pairs revealed a subtle but important shortcoming in the 5-year-old children, which was absent in the older children and which was not visible in a more simple experiment.

In sum, the current research elucidated the underlying processes connected with scalar alternatives. In a Felicity Judgment Task, where the alternative is given, both the 5- and 11-year-old children performed above chance on all items. However, for the 5-year-old children, the critical and ambiguous items were more difficult than the control items and they also performed worse on these two items than the 11-year-old children. Interestingly with respect to the issue of scalar diversity, the 11-year-old children were also presented temporal items, which turned out to be more difficult than the quantitative ones.

## ETHICS STATEMENT

This research has been reviewed and approved by the ethical review board SMEC (Sociaal-Maatschappelijke Ethische

Commissie; Social and Societal Ethics Committee) of the University of Leuven. Informed consent was obtained from the participants' parents in accordance with the Declaration of Helsinki.

## AUTHOR CONTRIBUTIONS

All authors contributed to this article, both substantially and formally. WS and KD designed the study and interpreted the data. BS prepared the experiment, constructed the stimuli, performed the experiments, and statistical analysis under supervision of WS and KD. BS wrote the first draft of the method and result section. WS wrote the introduction and the general discussion, and revised the methods and results sections. All authors approved the final version of the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02763/full#supplementary-material>

## REFERENCES

- Altschuler, D., and Schwarzschild, R. (2012). "Moment of change, cessation implicatures and simultaneous readings," in *Proceedings of Sinn und Bedeutung*, eds E., Chemla, V., Homer, and G., Winterstein, Paris, 45–62.
- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005
- Barner, D., Brooks, N., and Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition* 188, 87–96. doi: 10.1016/j.cognition.2010.10.010
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). lme4: Linear Mixed-Effects Models Using Eigen and S4 (Version 1.1-7).
- Benz, A., Bombi, C., and Gotzner, N. (2018). "Scalar diversity and negative strengthening," in *Proceedings of Sinn und Bedeutung* 22, Vol. 1, eds U. Sauerland and S. Solt (Berlin: ZAS), 191–204.
- Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006
- Breheny, R., Katsos, N., and Williams, J. (2006). Are generalized scalar implicatures generated by default? an online investigation into the role of context in generating pragmatic inferences. *Cognition* 100, 434–463. doi: 10.1016/j.cognition.2005.07.003
- Chierchia, G. (2004). "Scalar implicatures, polarity phenomena and the syntax/pragmatics interface," in *Structures and Beyond*, ed. A. Belletti (Oxford: Oxford University Press), 39–103.
- Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., and Meroni, L. (2001). "The acquisition of disjunction: evidence for a grammatical view of scalar implicatures," in *Proceedings of the 25th Annual Boston University Conference on Language Development*, eds A. H. J. Do, L. Dominguez, and A. Johansen (Somerville, MA: Cascadilla Press), 157–168.
- Chierchia, G., Fox, D., and Spector, B. (2012). "The grammatical view of scalar implicatures and the relationship between semantics and pragmatics," in *Semantics: An International Handbook of Natural Language Meaning*, Vol. 3, eds C. Maienborn, K. von Steussner, and P. Portner (Berlin: Mouton de Gruyter), 2297–2332.
- Coventry, K. R., Cangelosi, A., Newstead, S., Bacon, A., and Rajapakse, R. (2005). "Grounding natural language quantifiers in visual attention," in *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, eds B. G. Bara, L. Barsalou, and M. Bucciarelli (Mahwah, NJ: Lawrence Erlbaum Associates).
- Coventry, K. R., Cangelosi, A., Newstead, S. E., and Bugmann, D. (2010). Talking about quantities in space: vague quantifiers, context and similarity. *Lang. Cogn.* 2, 221–241. doi: 10.1515/langcog.2010.009
- De Neys, W., and Schaecken, W. (2007). When people are more logical under cognitive load. *Exp. Psychol.* 54, 128–133. doi: 10.1027/1618-3169.54.2.128
- Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171
- Degen, J., and Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: a visual world eye-tracking study. *Cogn. Sci.* 40, 172–201. doi: 10.1111/cogs.12227
- Doran, R., Baker, R. M., McNabb, Y., Larson, M., and Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *Int. Rev. Pragmat.* 1, 211–248. doi: 10.1163/187730909X12538045489854
- Foppolo, F., Guasti, M. T., and Chierchia, G. (2012). Scalar implicatures in child language: give children a chance. *Lang. Learn. Dev.* 8, 365–394. doi: 10.1080/15475441.2011.626386
- Foppolo, F., and Marelli, M. (2017). No delay for some inferences. *J. Semant.* 34, 659–681. doi: 10.1093/jos/ffx013
- Fox, D. (2007). "Free choice and the theory of scalar implicatures," in *Presupposition and Implicature in Compositional Semantics*, eds U. Sauerland and P. Stateva (Basingstoke: Palgrave Macmillan).
- Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511975158
- Glenwright, M., and Pexman, P. M. (2010). Development of children's ability to distinguish sarcasm and verbal irony. *J. Child Lang.* 37, 429–451. doi: 10.1017/S0305000909009520
- Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics, Speech Acts*, Vol. 3, eds P. Cole and J. L. Morgan (New York, NY: Academic Press), 41–58.
- Heyman, T., and Schaecken, W. (2015). Some differences in some: examining variability in the interpretation of scalars using latent class analysis. *Psychol. Belg.* 55, 1–18. doi: 10.5334/pb.bc
- Horn, L. (1972). *On the Semantic Properties of Logical Operators in English*. Ph.D. dissertation, University of California, Los Angeles, Los Angeles, CA.



- Hurewitz, F., Papafragou, A., Gleitman, L., and Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *J. Lang. Learn. Dev.* 2, 77–96. doi: 10.1207/s15473341l1d0202-1
- Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007
- Janssens, L., Drooghmans, S., and Schaecken, W. (2015). But: do age and working memory influence conventional implicature processing? *J. Child Lang.* 42, 695–708. doi: 10.1017/S0305000914000312
- Janssens, L., Fabry, I., and Schaecken, W. (2014). 'Some' effects of age, task, task content and working memory on scalar implicature processing. *Psychol. Belg.* 54, 374–388. doi: 10.5334/pb.ax
- Katsos, N., and Bishop, D. V. M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015
- Kempler, D., VanLancker, D., Marchman, V., and Bates, E. (1999). Idiom comprehension in children and adults with unilateral brain damage. *Dev. Neuropsychol.* 15, 327–349. doi: 10.1080/87565649909540753
- Lecce, S., Ronchi, L., Sette, P. D., Bischetti, L., and Bambini, V. (2018). Interpreting physical and mental metaphors: is theory of mind associated with pragmatics in middle childhood? *J. Child Lang.* 1–15. doi: 10.1017/S030500091800048X [Epub ahead of print].
- Marty, P., and Chemla, E. (2013). Scalar implicatures: working memory and a comparison with 'only'. *Front. Psychol.* 4:403. doi: 10.3389/fpsyg.2013.00403
- Newstead, S. E., and Coventry, K. R. (2000). The role of context and functionality in the interpretation of quantifiers. *Eur. J. Cogn. Psychol.* 12, 243–259. doi: 10.1080/095414400382145
- Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1
- Papafragou, A., and Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition* 86, 253–282. doi: 10.1016/S0010-0277(02)00179-8
- Pezzelle, S., Bernardi, R., and Piazza, M. (2018). Probing the mental representation of quantifiers. *Cognition* 181, 117–126. doi: 10.1016/j.cognition.2018.08.009
- Pouscoulous, N., Noveck, I. A., Politzer, G., and Bastide, A. (2007). A developmental investigation of processing costs and implicature production. *Lang. Acquis.* 14, 347–375. doi: 10.1080/10489220701600457
- Rees, A., and Bott, L. (2018). The role of alternative salience in the derivation of scalar implicatures. *Cognition* 176, 1–14. doi: 10.1016/j.cognition.2018.02.024
- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguist. Philos.* 27, 367–391. doi: 10.1023/B:LING.0000023378.71748.db
- Schaecken, W., Van Haeren, M., and Bambini, V. (2018). The understanding of scalar implicatures in children with autism spectrum disorder: dichotomized responses to violations of informativeness. *Front. Psychol.* 9:1266. doi: 10.3389/fpsyg.2018.01266
- Singmann, H., Bolker, B., Westfall, J., and Aust, F. (2018). *Afex: Analysis of Factorial Experiments (Version 0.21-2)*. Available at: <https://github.com/singmann/afex>.
- Skordos, D., and Papafragou, A. (2016). Children's derivation of scalar implicatures: alternatives and relevance. *Cognition* 153, 6–18. doi: 10.1016/j.cognition.2016.04.006
- Stiller, A. J., Goodman, N. D., and Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Lang. Learn. Dev.* 11, 176–190. doi: 10.1080/15475441.2014.927328
- Thomas, G. (2012). *Temporal Implicatures*. Doctoral dissertation, MIT, Cambridge.
- Tieu, L., Romoli, J., Zhou, P., and Crain, S. (2016). Children's knowledge of free choice inferences and scalar implicatures. *J. Semant.* 33, 269–298. doi: 10.1093/jos/ffv001
- Van Tiel, B. (2014). *Quantity Matters: Implicatures, Typicality and Truth*. Ph.D. thesis, Radboud University, Nijmegen.
- van Tiel, B., and Schaecken, W. (2017). Processing conversational implicatures: alternatives and counterfactual reasoning. *Cogn. Sci.* 41, 1119–1154. doi: 10.1111/cogs.12362
- Van Tiel, B., van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *J. Semant.* 33, 137–175.
- Wampers, M., Schrauwen, S., De Hert, M., Gielen, L., and Schaecken, W. (2017). Patients with psychosis struggle with scalar implicatures. *Schizophr. Res.* 195, 97–102. doi: 10.1016/j.schres.2017.08.053

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Schaecken, Schouten and Dieussaert. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Cross-Linguistic Variation in the Meaning of Quantifiers: Implications for Pragmatic Enrichment

Penka Stateva<sup>1,2\*</sup>, Arthur Stepanov<sup>1†</sup>, Viviane Déprez<sup>3</sup>, Ludivine Emma Dupuy<sup>3</sup> and Anne Colette Reboul<sup>3</sup>

<sup>1</sup> Center for Cognitive Science of Language, University of Nova Gorica, Nova Gorica, Slovenia, <sup>2</sup> 2018–2019 EURIAS Fellow at the Collegium – Lyon, Institute for Advanced Studies, University of Lyon, Lyon, France, <sup>3</sup> UMR5304, Institut des Sciences Cognitives Marc Jeannerod, Bron, France

## OPEN ACCESS

### Edited by:

Andrea Moro,  
Istituto Universitario di Studi Superiori  
di Pavia (IUSS), Italy

### Reviewed by:

Giosuè Baggio,  
Norwegian University of Science  
and Technology, Norway  
Valentina Bambini,  
Istituto Universitario di Studi Superiori  
di Pavia (IUSS), Italy

### \*Correspondence:

Penka Stateva  
penka.stateva@ung.si

<sup>†</sup> Joint first authors

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 26 November 2018

**Accepted:** 10 April 2019

**Published:** 15 May 2019

### Citation:

Stateva P, Stepanov A, Déprez V,  
Dupuy LE and Reboul AC (2019)  
Cross-Linguistic Variation  
in the Meaning of Quantifiers:  
Implications for Pragmatic  
Enrichment. *Front. Psychol.* 10:957.  
doi: 10.3389/fpsyg.2019.00957

One of the most studied scales in the literature on scalar implicatures is the quantifier scale. While the truth of *some* is entailed by the truth of *all*, *some* is felicitous only when *all* is false. This opens the possibility that *some* would be felicitous if, e.g., almost all of the objects in the restriction of the quantifier have the property ascribed by the nuclear scope. This prediction from the standard theory of quantifier interpretation clashes with native speakers' intuitions. In Experiment 1 we report a questionnaire study on the perception of quantifier meanings in English, French, Slovenian, and German which points to a cross-linguistic variation with respect to the perception of numerical bounds of the existential quantifier. In Experiment 2, using a picture choice task, we further examine whether the numerical bound differences correlate with differences in pragmatic interpretations of the quantifier *some* in English and *quelques* in French and interpret the results as supporting our hypothesis that *some* and its cross-linguistic counterparts are subjected to different processes of pragmatic enrichment.

**Keywords:** quantifier, numerical bound, scalar implicature, R/I-implicature, M-implicature

## INTRODUCTION

In a broad sense, natural language quantification includes expressions of explicit quantities or numerical proportions (e.g., 50%), as well as a set of expressions that do not directly refer to numbers but express quantities or proportions as more or less vague estimations thereof. Such are the quantificational determiners *some*, *few*, *many*, *half*, *most* (*at least/at most/as many as*) *n* (for a natural number *n*), *all*, among others. The standard approach in formal semantics that goes back to Barwise and Cooper's (1981) seminal work, treats these determiners as relations between sets of individuals. In this framework, for instance, the determiner *some*, as in *Some balloons are red*, relates the set of balloons and the set of relevant red objects in a way which requires that the intersection of the two sets is not empty for the sentence to be True in a given situation. Similar semantic definitions are offered for the whole class of other determiners. They are all defined as relations between two sets of individuals. Some examples are given in (1):

- (1) a.  $\llbracket \text{some} \rrbracket = \{ \langle A, B \rangle : A \cap B \neq \emptyset \}$
- b.  $\llbracket \text{every} \rrbracket = \{ \langle A, B \rangle : A \cap B = \emptyset \}$
- c.  $\llbracket \text{every} \rrbracket = \{ \langle A, B \rangle : A \subseteq B \}$
- d.  $\llbracket \text{most} \rrbracket = \{ \langle A, B \rangle : |A \cap B| > \frac{1}{2} |A| \}$
- e.  $\llbracket \text{many} \rrbracket = \{ \langle A, B \rangle : \frac{|A \cap B|}{|A|} > n_c, \text{ for some number } n \text{ in a context } c \}$

In addition, pragmatic theories which come in some varieties (cf. the classical theory of Grice, 1989; the neo-Gricean theory of Horn, 1984, 2004; Levinson, 2000, the grammatical theory of Chierchia et al., 2012; Chierchia, 2013, the Relevance theory of Sperber and Wilson, 1995) specify a further component (through a different mechanism for each theory) in the meaning of the quantificational expressions that enriches the proposition of which it is part with some pragmatic inference. The most typical example involves enrichment through scalar implicatures. In Horn's terminology, these implicatures result from (i) the fact that quantifiers are part of a set that forms an entailment scale (see de Carvalho et al., 2016 for evidence of the psychological reality of scales) and as such are always under consideration as possible alternatives and (ii) speakers' adherence to a pragmatic principle that requires maximal informativeness (Quantity Maxim of Grice/Q-Principle of neo-Griceans) or to the requirements of the exhaustivity operator in the grammatical theory of implicatures. As an illustration we can consider again the example with *some*. The literal meaning of *Some balloons are red* is complemented by a pragmatic inference that *Not all balloons are red* so that the resulting meaning is *Some but not all balloons are red*. The scalar implicature is derived by negating the scalar alternative, *All balloons are red*, to the sentence containing *some* because it is stronger or more informative since it asymmetrically entails the original sentence, but was not chosen by the cooperative speaker. A similar meaning enrichment process applies to all items on the closed quantificational scale which do not occupy its end-points.

However, even if we assume that literal meanings of quantifiers are often strengthened by scalar implicatures, speakers who evaluate the truth of sentences like *Some balloons are red* are expected to always judge as well acceptable the sentence in all contexts in which the size of the set of red balloons relates to the size of the whole set of balloons by a proportion which could be expressed by any number between 0 and 1. That means that situations in which red balloons are 1 or 99% of all balloons are predicted to be just as good as situations in which red balloons are 20% of all balloons in terms of verifying that sentence. This prediction is not always borne out by speakers' reported intuitions concerning respective contexts. Moreover, according to the standard theories, no cross-linguistic variation is expected in the evaluation of translational equivalents. In other words, quantifiers like *some* or *most* are expected to cover exactly the same range of proportions in different languages.

The goal of this paper is to subject to scrutiny these predictions of the standard semantic-pragmatic treatment of quantifiers. To this end, we report the results of two experiments. Experiment 1 is a cross-language questionnaire study spanning the Germanic, Romance, and Slavic language groups. Two main findings of this experiment are the following: (i) meaning strengthening through scalar implicatures is not sufficient to account for the observed numerical bounds of quantifiers, and (ii) at least the English quantifier *some* is not conceptualized in the same manner cross-linguistically and should not, therefore, receive the same analysis as its counterparts in other languages. In Experiment 2, using a picture-choice test, we further experimentally explore the implications of these results for the theory of scalar implicatures. Specifically, we observe a different

pattern of comprehension of sentences containing the English *some* and its French counterpart *quelques*. We interpret the difference as supporting our claim that the meanings of *some* and its crosslinguistic variants result from applying different mechanisms of pragmatic enrichment.

## QUANTIFIERS AND NUMERICAL BOUNDS

### The Psychometric Approach

Quantifier processing has also been in the focus of cognitive psychology. Previous experimental research on the "psychometric" dimension of quantifiers established that the meanings of quantifiers lie on some sort of scale, and suggested that a mapping should hold between a quantifier and its respective range of numerical values (Moxey and Sanford, 2000). Furthermore, the respective numerical range-referring representations of quantifier meanings have been formulated as membership functions used in fuzzy logic, whereby different values pertaining to the quantifier are graded, e.g., between 0, meaning no fit, to 1, implying a perfect fit (Wallsten et al., 1986). For instance, the probability quantifier *likely* might be given a value of 0 for  $p = 0.2$ , one of 0.1 for  $p = 0.3$ , and 1.0 for  $p = 0.8$ . Membership functions encode information about the form of the mapping from an expression to amounts (e.g., variance, skew, kurtosis) as well as central tendency information. These membership functions were found to be stable for a given individual and suggested to be a good substitution for an internalized scale (Wallsten et al., 1993).

However, it was soon recognized that the "psychometric" approach in this form faces serious difficulties, in that that direct assignment of the empirically established range to the respective quantificational expression is very difficult or impossible to implement. Membership functions were found to depend greatly on a number of potentially confounding factors. One such factor is contrast effects that arise because of the within-subject experimental design, whereby subjects are asked to provide values for different quantifiers in a single trial (e.g., Daamen and de Bie, 1991). Another factor has to do with the set size from which proportions are drawn: e.g., low-quantity determiners such as *few* were found to denote a greater proportion when they described small set sizes, compared to larger ones (Newstead et al., 1987). Yet another problem arises from the conflict with base-rate expectations concerning the event described by the quantifier-bearing sentence. For instance, the values assigned to *many* in *Many people enjoyed the party* is higher than in *many doctors are female*, because the former (people enjoying parties), but not the latter, event has a higher base-rate expectation (Moxey and Sanford, 1993). One also faces a serious methodological problem when trying to marry the "psychometric" approach in its present form to the currently standard truth-conditional formal semantics, which interprets sentence meanings in terms of binary truth values 0 and 1. This binary system is in conflict with the rationale behind the membership function allowing an intermediate degree of fit. Irrespective of these shortcomings, it is important to

note, however, that the psychometric approach was based on the valid observation that quantifier meanings predicted by the standard semantic-pragmatic approach are not strictly validated by speakers' intuitions. There is no controversy as to the numerical bounds and set-theoretic meaning of the universal quantifier *every/all* and of the negative one *no* but the rest of the quantifiers apparently need to be reanalyzed.

## The Typicality Approach

The interpretation of quantifiers has recently been reconsidered within a framework based on typicality measures (van Tiel, 2014; van Tiel and Geurts, 2014). This line of research relies on a distinction between typicality and category membership (cf. Fuhrmann, 1991, a.o.). The typicality theory of quantifier interpretation is related to a general mechanism of ascribing typicality differences among members of the same category. One example discussed in van Tiel (2014) regards an experimental study reported in Rosch (1975) where results point to a stable ordering of members of the category BIRD with the robin being evaluated as the most typical in comparison to the rest of the birds denoted by relevant hyponyms of *bird*. In a similar vein, the typicality approach to quantifier interpretation assumes that quantified statements are assigned functions from situations to typicality values. As the authors argue, typicality values can be related to probability values but only if the cardinality of the total set is known. This makes the typicality-based proposal more advantageous than similar proposals of interpreting quantified statements as functions from situations to probability values (cf. Yildirim et al., 2013) since speakers need not necessarily have knowledge about the relevant set cardinality in all situations in which quantifiers are used.

van Tiel and Geurts (2014) investigate typicality judgments associated with the quantifiers *all*, *every*, *few*, *many*, *more than half*, *most*, *some*, *none* not *all*, *not many* in a large-scale study involving 340 English-speaking participants. They construct visual contexts with 10 black or white circles. The number of black and white circles in each context was manipulated to represent all 11 different possibilities. Using a 7-point Likert scale, participants evaluated the fit between respective quantified sentences and each context. This task was intended to provide typicality judgments. These were contrasted to truth-value judgments which were elicited by using the same material and a task to provide a binary judgment (True/False). The results were interpreted to indicate that typicality judgments were influenced by two factors: set-theoretic definitions and distance from prototype. A necessary condition for a prototype is to be a situation in which the quantified sentence is true according to the respective set-theoretic definition. But, they were also found to depend on competing quantifiers, i.e., a prototypical situation related to a quantifier *q* must be maximally distinct from a prototypical situation related to any competing quantifier *q'*.

Here we focus on three important consequences of the typicality-based analysis of quantifiers. First, the proposal does not make a clear prediction about the interaction between typicality inferences and pragmatic inferences resulting from quantifier alternative competition, i.e., scalar implicatures in

non-embedded contexts (see also Cummins, 2014). Second, the proposal leaves no obvious space for cross-linguistic variation. Inasmuch as quantifier numerical bounds are related to prototypes, these are expected to have general cognitive foundations. And finally, if all of the quantifiers in the reported studies involve the same mechanism of association with prototypical values, prototypes should be relatively stable and clearly distinguished even for quantifiers with partially overlapping set-theoretic definitions. This last expectation was not borne out in some cases in the study reported in van Tiel and Geurts. In addition, the claim that prototypes depend on competing quantifiers might need a more detailed formulation given that the study does not distinguish between cases with linguistically provided alternatives and cases with implicitly available alternatives. The last consideration is validated by an experimental study on the processing of two Slovenian counterparts of the determiner *many*, namely *precej* and *veliko* (see Stateva and Stepanov, 2017) and by reported experimental work on processing implicatures within a paradigm that provides alternatives explicitly (cf. Felicity Judgment Task in Foppolo et al., 2012, a.o.).

## Quantifiers as Representations of Proportions: Pezzelle et al. (2018)

The discussion above aimed at motivating the cross-linguistic perspective in studying the perception of quantifiers since potential differences might pinpoint the nature of mechanisms affecting perception. Another important perspective is suggested in Pezzelle et al. (2018), namely the role of proportions as opposed to numerosity in quantifier perception. The study features two experiments, one investigating visually grounded representations and the second one, abstract representations of similarity/difference between quantifiers. Both experiments examine the perception of Italian quantifiers and encompass a list of nine quantifiers including the positive end-point of the proportional scale corresponding to *tutti* (*all*) and the negative end-point corresponding to *nessuno* (*none*). The grounded task used visual stimuli representing a set of objects, part of which were animals in all items. In each trial, the participants were supposed to pick one out of the set of nine quantifiers which best expressed the approximate representation of animals within the whole set of objects. The second experiment asked for metalinguistic judgments about closeness within pairs of quantifiers on a scale from 1 to 7. Both experiments revealed that mental representations of quantifiers represent (non-fixed) proportions rather than cardinalities. The data showed that quantifiers represent an ordered but non-linear scale. Interestingly, the upper part of the scale corresponding to high magnitudes, i.e., *all*, *almost all*, *most*, and *many* involved more overlaps (lower degree of differentiation) in comparison to the lower which was interpreted to indicate a stronger numerical factor in low-magnitude quantifiers. Consequently, the latter type of quantifiers are better differentiated in mental representations.

Using a different protocol we also aim to investigate the mental representations corresponding to quantifiers in four Indo-European languages and compare the results especially to



those in Pezzelle et al. (2018). Our main task, however, is to identify the mechanisms behind the different processing patterns.

## The Present Study

We examine the interpretation of quantifiers in two experiments whose aim is to shed further light on a number of relevant questions given the discussion so far. In particular, we aim to identify the main pragmatic factors that influence the processing of quantifiers cross-linguistically. Toward this goal, we address the following questions:

- Is it possible to identify the numerical ranges assigned to different quantifiers and their translational equivalents in other languages? Are numerical ranges encoded in meanings or are they epiphenomenal?
- Are cross-linguistically related quantifiers processed identically? Can we maintain a universal theory of quantifiers on the basis of similarities in the respective numerical values?
- Which pragmatic processes are relevant for the interpretation of quantifiers?
- How are quantifiers with overlapping lexical meanings distinguished?

The main predictions of the present study are rather straightforward. If the classical theory of Barwise and Cooper (1981) and others is on the right track, then, with respect to the quantifier *some*, we should not expect to encounter any specific numerical limitations in the range of evaluated proportions, in English as well as in other tested languages. As pointed out in the Section “Introduction,” given the definition in (1a), situations in which quantified objects constitute between 1 and 99% are predicted to be more or less appropriate for the use of this quantifier. This is not the case for the use of *most* where the definition (1d) restricts the use to the numerical proportions over 50%: therefore, its use in proportions less than 50% should be unacceptable. With respect to quantifier *half*, we obviously expect a peak in acceptability around 50%, while lower and higher proportions should not be acceptable. With respect to *few*, following the standard theory, we view *few* as a negative counterpart of *many* [cf. (1e)] and therefore expect, its upper bound to be well below 50%. In line with neo-Gricean reasoning, we assign *few* to the negative scale <none, hardly any, few> and predict that its lower bound is affected by a scalar implicature negating the two stronger alternatives in the ordered set. Finally, following Penka (2006) which defines *almost* as a member of a Horn-set on a par with *most*, we expect a numerical range for *almost* above that for *most* and excluding the top of the proportional scale.

The predictions concerning the scalar implicature component of the quantifier’s meaning are important in one additional aspect. As both neo-Gricean and Relevance theories predict, meaning strengthening through scalar implicatures should be sufficient to account for the numerical ranges of the quantifier *some* and its crosslinguistic counterparts, that is, the numerical range of *some* must not overlap with numerical range of

other quantifiers like *few*, *half*, *most*, or *almost all* if pragmatic enrichment applies.

We were also interested in testing the prediction made by the typicality approach that, inasmuch as quantifier numerical bounds are related to prototypes, the latter are expected to have general cognitive foundations and therefore, no cross-linguistic variation is expected in the meaning of the respective quantifiers, including their numerical ranges.

## EXPERIMENT 1

Experiment 1 addresses a similar question to the one of van Tiel and Geurts (2014), namely whether speakers make reference to particular numerical values in their use of different quantifiers. The experimental design is therefore similar to theirs but it, nevertheless, bears some important differences. The main one is that this is a cross-linguistic study involving four languages belonging to different language groups within the Indo-European family: Germanic, Romance, and Slavic. We thus have a possibility to compare how close or different respective lexical counterparts are. The second difference is that we use verbal contexts making reference to a relatively big cardinality of the respective total sets to avoid interference of possible world knowledge.

## Design and Materials

We investigate the cross-linguistic distribution of quantificational determiners by running a series of similarly designed experiments in four languages: English, French, German, and Slovenian. The quantifiers used in the questionnaires per language are listed in Table 1.<sup>1</sup>

Several clarifications concerning the choice of the target items are in order. First, the reader might wonder why *almost* and its translational equivalents were included in the experimental paradigm given that the classical theory of quantifiers does not normally extend to this determiner. Our decision was partly influenced by a proposal in Penka (2006) based on the argument that *almost* is part of the entailment (Horn-) scale along with determiners like *all* and *most*. If this is the case, then it must belong to that natural class. In addition, we wanted to find out if *almost* acts as an alternative to *most* in forcing it to be restricted to a lower interval than the one predicted by its set-theoretic meaning. Yet another reason for

<sup>1</sup>The German data were collected by Stateva and Gergel for a study published as Gergel and Stateva (2014) focusing on the differences between the German determiners *allermeisten* (“most”) and *fast alle* (“almost all”). Except for *fast alle*, the data collected on the German quantifiers as in Table 1 were not discussed in that study and were largely treated there as filler conditions.

**TABLE 1** | Experiment 1: Quantifiers per language used in target sentences.

English	<i>few</i>	<i>some</i>	<i>half</i>	<i>most</i>	<i>almost (all)</i>
French	<i>un peu</i>	<i>quelques</i>	<i>la moitié</i>	<i>la plupart</i>	<i>presque (tous)</i>
German	<i>wenige</i>	<i>einige</i>	<i>halbe</i>	<i>meisten</i>	<i>fast (alle)</i>
Slovenian		<i>nekaj</i>	<i>polovica</i>	<i>večina</i>	<i>skoraj (vse)</i>

6% completed

**133 men sought a life partner.**  
**97 of these men utilized an online dating site. [Q337]**

Please evaluate how well each of the following sentences describes the situation above:

	not well	1	2	3	4	5	very well
Few men utilized an online dating site.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
Some men utilized an online dating site.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
Half of the men utilized an online dating site.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
Almost all men utilized an online dating site.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
Most men utilized an online dating site.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

Next

**FIGURE 1 |** Experiment 1: A sample stimulus screen (the English portion).

including this item was that we are not familiar with many experimental studies about the numerical bounds of *almost* (cf. Pezzelle et al., 2018).

Second, as **Table 1** makes evident, the general cross-language comparison involved five tested quantifiers per language, with one exception: we did not include the Slovenian counterpart of *few*, *malo*, which can be considered a limitation in our design. The appropriate slot in the Slovenian part of the questionnaire was used to test the quantifier *precej* that roughly corresponds to English *many*, instead. All *precej*-sentences were treated as fillers for the purposes of the present study. This quantifier, however, was in the focus of a similarly designed experiment in Stateva and Stepanov (2017); see this work for details.

Third, we did not include in our testing the universal quantifier and the existential quantifier *all* and *no*, because of their extremely narrow-ranged associated proportions, namely 100% in one case and 0% in the other, and therefore trivial (or close to trivial) associated intuitions. We did, however, include the quantifier *half* which is also associated with a fairly trivial proportional range (around 50%) but, because of the more complex actual numerical proportions that we used in this experiment, speakers do not necessarily have a direct access to the result of the respective calculation; as a result, a limited amount of vagueness can also be expected.

Fifty items were prepared as experimental materials. Each item contained a two-sentence context. The first sentence established an event and made reference to the cardinality of a set of individuals. The second sentence referred to one of its subsets. The numbers used in all first sentences of the contexts ranged from 100 to 200. The ratio between first and second number in contexts was manipulated in order for the proportion scale to

be covered from 1 to 99% with an increment of 2 within the 50 contexts.<sup>2</sup> Each context was accompanied by five sentences describing it by using a different quantifier. Furthermore, each sentence was accompanied by a 1–5 Likert scale with annotated end-points *not well* (1) and *very well* (5) in the English, French and Slovenian versions of the questionnaire. German participants had the labels *not well* and *very well* correspond to the scale from 5 to 1, respectively, following a similar convention in the German educational system.<sup>3</sup> An example of the stimulus materials (from the English portion of the experiment) is given in **Figure 1**.

## Participants

One hundred eight (24 males) adult self-reported native speakers of English ( $N = 28$ ), French ( $N = 30$ ), German ( $N = 25$ ), and Slovenian ( $N = 25$ ) were recruited for this experiment, and gave an informed consent to participate in the study. The distribution of participants by age groups is shown in **Table 2** (participants gave categorical responses regarding their age range in this experiment; the null hypothesis of similar age distribution across different language groups was not confirmed by Fisher's exact test using Monte Carlo-simulated  $p$ -value computation:  $p < 0.001$ ).<sup>4</sup> Approximate age means for each language group were calculated by summing over mid-point

<sup>2</sup>We avoided the use of round numbers in the contexts for two reasons. First, we wanted the participants not to be tempted to provide judgments with the help of explicit ratio calculations. Second, since it has been shown that round numbers invite more approximate interpretations (Krifka, 2007b), we believed that by avoiding round numbers, we prompt as precise interpretations as possible, especially in view of the task to evaluate vague quantifiers.

<sup>3</sup>Scores were reassigned as follows for the analysis: 1→5, 2→4, 4→2 and 5→1.

<sup>4</sup>All simulations and modeling in this study were performed in the R environment 3.5.1 (R Core Team, 2018).

of each group range multiplied by frequency of its occurrence and dividing the resulting product by the total number of participants. The English and German participants were undergraduate students at Rutgers University (United States) and the University of Graz (Austria), respectively, and they received course credit. The Slovenian and French participants were students and employees at the University of Nova Gorica (Slovenia) and The University of Lyon (France), respectively. They participated voluntarily and received no compensation. All participants had normal or corrected to normal vision and they were naïve as to the purpose of the study and the research question.

## Procedure

The participants were instructed to read each context carefully and then evaluate, following their first intuition, how well each of the accompanying sentences described the respective context, by clicking on the respective number on the corresponding 5-point scale. All participants received all 50 items in this task. The experiment was administered via the web-based software SoSciSurvey.<sup>5</sup> The contexts as well as the five target sentences in each context were presented in a pseudo-randomized order for each participant. The participants were allowed to take a break, if necessary, after completing the evaluation of a whole context. Note that this is a task related not only to (semantic) knowledge of quantifier meaning but also a task on pragmatic knowledge of quantifier use. As such, its design involves reasoning, similarly to other tasks targeting pragmatic knowledge.<sup>6</sup> There were no time limits on finishing the task or evaluation of a particular context. Response times of evaluating all five sentences on each screen were also recorded, mostly for informational purposes (we postpone exploration of the detailed time course in evaluating sentences with specific quantifiers in this type of task for future research; see, e.g., Bott and Noveck, 2004; Hackl, 2009, for relevant discussion).

## Results

Average times spent by the participants on a single screen, broken down by language, are shown in **Table 3** (average times less than 7 s and greater than 300 s were trimmed). A one-way ANOVA showed an effect of language on evaluation time ( $df = 3$ ,  $F = 4.95$ ,  $p = 0.003$ ). *Post hoc* pairwise comparison tests (Tukey-type simulations) showed that French was mostly the culprit, with an average evaluation taking about 9 s longer than in German and about 11 s longer than in English. Speakers of the other three languages did not significantly differ in their evaluation time ( $df = 2$ ,  $F = 1.82$ ,  $p = 0.16$ ).

For the score analysis, we assumed the mid-scale judgment of 3 points as a threshold for a positive judgment on appropriateness

**TABLE 2 |** Experiment 1: Mean times (standard error) spent by the participants on a single screen, broken down by language.

Language	Mean (SE)
English	25.28 (2.32)
French	36.03 (2.35)
German	26.73 (1.65)
Slovenian	30.90 (2.28)

**TABLE 3 |** Experiment 1: Participation by age group and language.

Language/age group	18–20	21–24	25–30	31–35	36–55	Approximate mean
English	9	18	1	0	0	21.6
French	1	17	3	2	7	28.8
German	4	11	6	4	0	24.7
Slovenian	1	3	18	3	0	27.2

of the respective contexts and excluded datapoints below this threshold. The rationale for not using the set of datapoints collected over the entire set of conditions comes from the perspective seeing quantifiers as markers of numerical proportions. To illustrate the point informally, consider the determiner *half*. It is clear that when an expression such as “half of the dots are red” is evaluated against a finite set of red dots within a particular range, it is only within a very narrow subrange of conditions that this expression will receive high scores, whereas in the vast majority of other cases, it will receive low scores (this was, in fact, the case in our study). Taking the entire set of data points into consideration in this case would lead to the misleading conclusion that speakers generally dislike this determiner, whereas in fact the scores simply reflect the natural situation that the use of this determiner is licensed within a very narrow numerical range. Similar considerations apply in the case of the determiner *all*, as well as for all cardinal quantificational determiners. By analogy, we believe this holds also in the case of the other quantifiers, even though the particular numerical range for this determiner may be hard to establish *a priori* because of their vague character. Thus it would not be appropriate to compare the alleged differences in the use of quantifiers across numerical ranges in which their use is not licensed in principle. In contrast, dividing the Likert acceptability scale in half provides at least a rough estimation of the acceptability boundary. Doing so thus extends the usual tradition of collecting speakers’ evaluations in terms of binary judgments, but also adds the functionality for estimation of the size of the observed differences across different conditions.<sup>7</sup>

<sup>7</sup>As the experiment also probes into pragmatic knowledge about quantifier meaning, we need to acknowledge the paramount importance of the methodology of judgment collection. Appealing to a pragmatic tolerance principle that potentially obscures existing sensitivity in binary tasks, Katsos and Bishop (2011) argue against the elicitation of a binary judgment in developmental pragmatics (but see also Noveck, 2018). The non-binary judgment approach fits our research question better. Given that we expect some degree of similarity in the perception of different quantifiers, we believe that considering an interval of acceptable points, rather than a point of acceptability is the more informative choice.

<sup>5</sup><https://www.sosicurvey.de/>

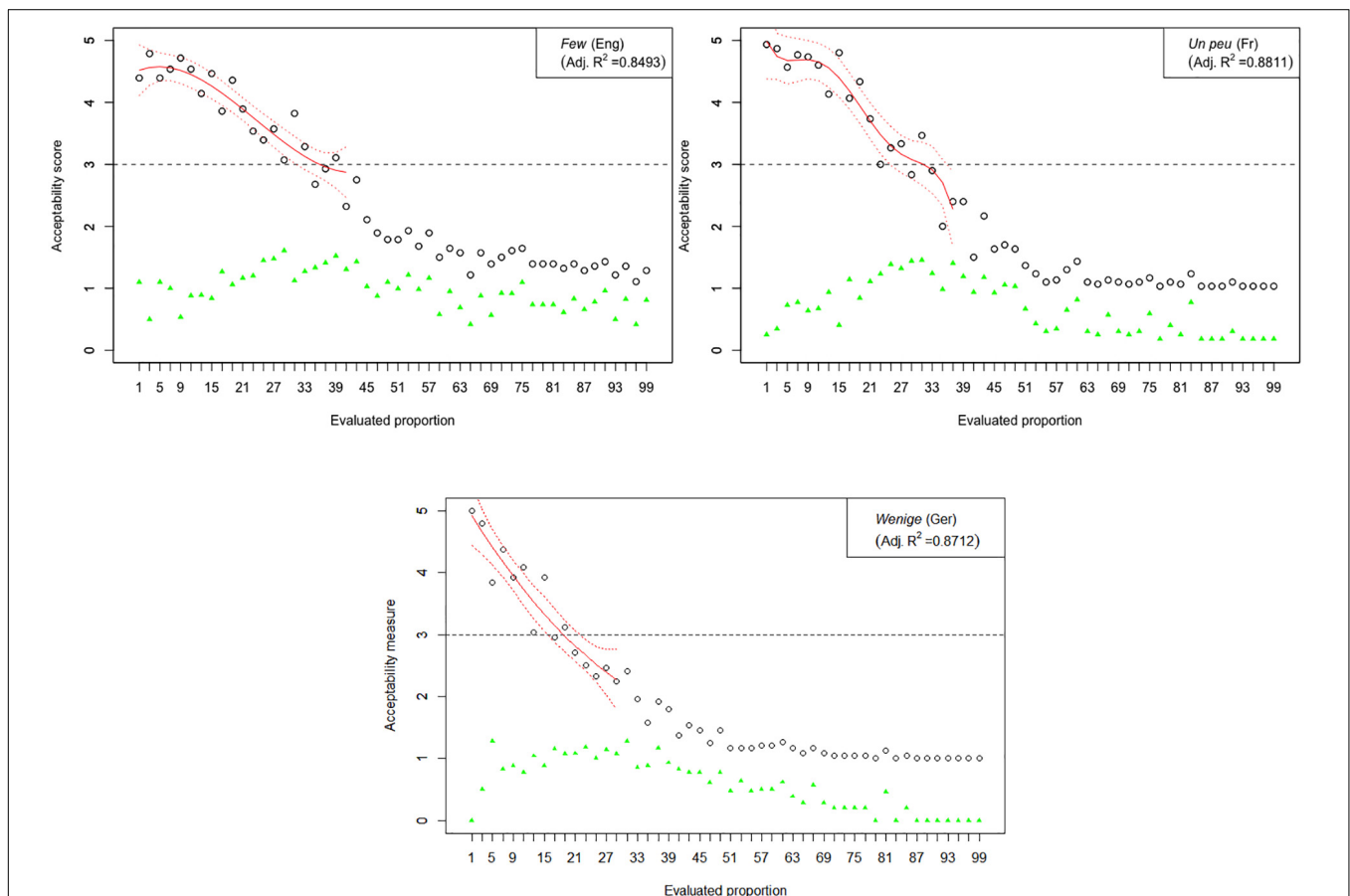
<sup>6</sup>Standard protocols in experimental pragmatics like the Truth Value Task or Felicity Judgment Task involve evaluation of alternatives on the basis of pragmatic reasoning restricted by pragmatic principles like Quantity (Grice, 1989), Maximize Presupposition (Heim, 1991), etc. This methodology allows for testing implicated meaning which is also relevant to quantifier comprehension.

The results of Experiment 1 are graphically represented in **Figures 2–6**. The graphs in the figures summarize acceptability scores in the upper half of the Likert scale per language and per respective quantifier together with respective polynomial fit curves and confidence intervals. Regression models were used to fit the data using polynomial functions. As can be seen from the figures, different quantifiers were judged acceptable in different ranges of proportions. In particular, the numerical proportions characterized by respective cross-linguistic counterparts of *few* appear to be restricted well below 50%, with the score peaking in the first quarter (<25%) of the proportional range. On the other hand, *most* and *almost all* are predictably evaluated higher with proportions of 50% and above. The scores on *most* tend to a plateau in the upper part of the numerical range (>50%), whereas the acceptability on *almost all* increases more steeply toward the last quarter (>75%). The determiner *half* received most of the acceptable scores midrange, peaking around 50% and sharply dropping before and after that.

The results of Experiment 1 revealed that, despite the relatively small sample sizes in this experiment, speakers of all four languages follow consistently similar patterns of evaluating the quantifiers with the meaning of *few*, *half*, *most*, and *almost*

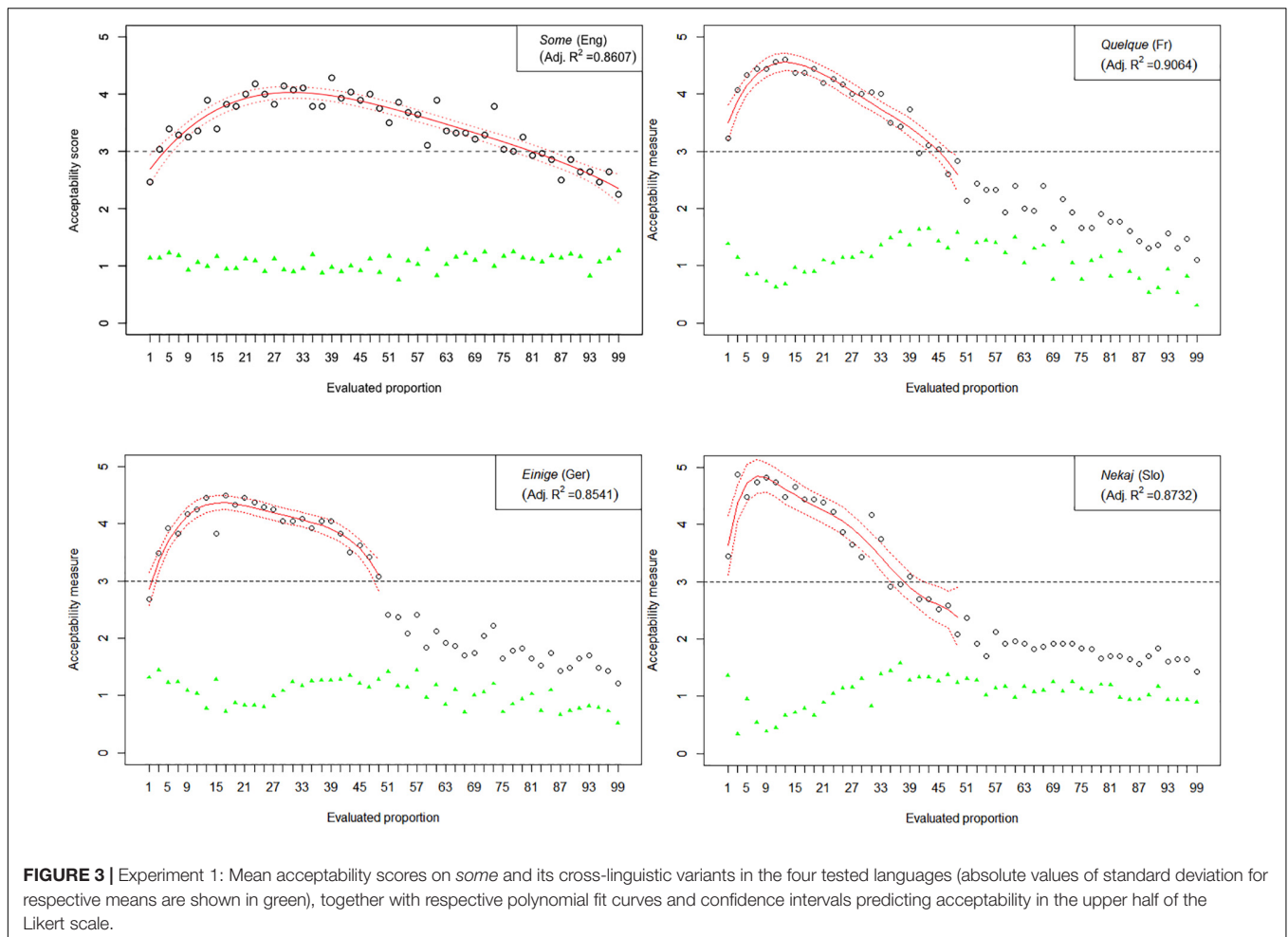
*all*. An important exception in this picture concerns the English quantifier *some* (**Figure 3**). The numerical proportions whose characterization by non-English counterparts of *some* was acceptable ranged from 3% to slightly less than 50% of the total number of items in the three languages under consideration, namely, German, French, and Slovenian. In contrast, English speakers found proportions in the range between 3 and about 80% of the total number of items at issue, as acceptable to be characterized by *some*. In other words, the range of proportions that can be characterized by the meaning of *some* is 60% larger in English than in the other three tested languages.

Another notable anomaly pertaining to the English quantifier *some* that we observed in contrast with its cross-linguistic counterparts concerns the correlation of mean score values with respective pooled variance. Mean scores and respective measures of variance such as standard deviations were previously shown to be inherently correlated in studies using Likert scales, whereby standard deviations tend to be smaller if mean values are closer to the extreme points of the Likert scale and increase toward the middle. This trend, when observed over the entire evaluation, can be described with a quadratic regression model and graphically represented by a parabola with a peak around



**FIGURE 2 |** Experiment 1: Mean acceptability scores on *few* and its cross-linguistic variants in the three tested languages (absolute values of standard deviation for respective means are shown in green), together with respective polynomial fit curves and confidence intervals (in red) predicting acceptability in the upper half of the Likert scale.





mid-range (Lipovetsky, 2017). In our study, all except one of the tested quantifiers demonstrated a reliable quadratic trend, peaking around the mid-scale (3) and declining on both ends (1 and 5).<sup>8</sup> The only exception was indeed English *some*, where no discernible trend could be identified, suggesting that means and standard deviations are not correlated here (adjusted  $R^2 < 0.27$ ). This state of affairs is depicted in Figure 7, in which the results from the English *some* are contrasted with its cross-linguistic counterparts; the latter also serve as representative examples of the polynomial trend observed in the evaluations of the other tested quantifiers.

## Discussion

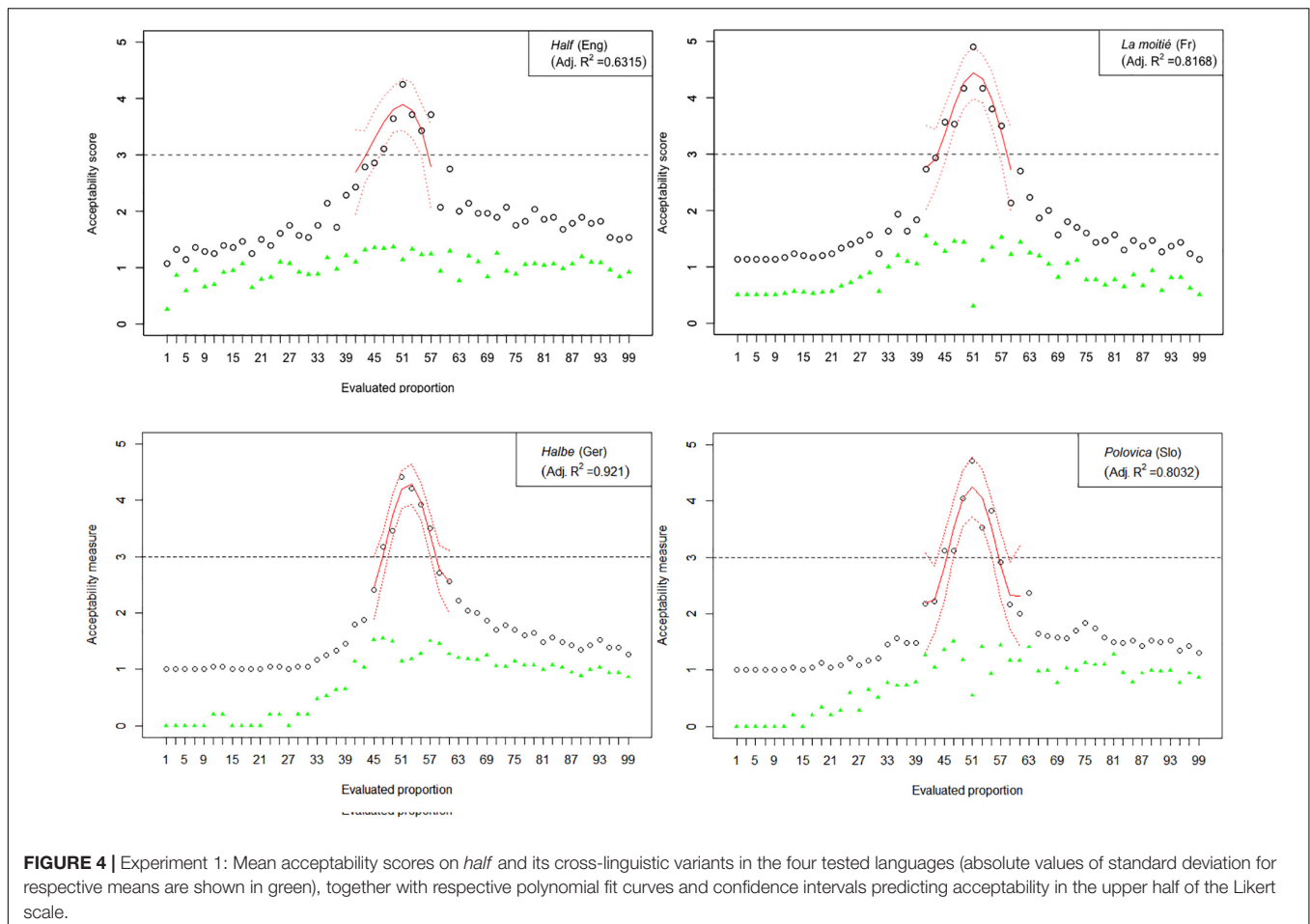
The reported study demonstrated a lot of unanimous decisions of quantifier evaluation across languages. In most cases it looks like participants intuitively follow a similar mechanism of rough estimation of the proportions and match the outcome against a given quantifier. We can hypothesize that if numerical bounds are related to these determiners are stable, these bounds have

a universal character (as much as such a generalization is warranted with observations from a limited language sample of four languages).

However, we have also discovered a divergence of behavioral responses with respect to the determiner meaning *some*. The divergence is twofold: (i) the range of proportions that can be characterized by the meaning of *some* is 60% larger in English than in the other three tested languages; and (ii) the mean scores received from evaluating sentences with English *some* did not correlate with respective standard deviations, in contrast with its respective cross-linguistic counterparts. We take these results to indicate that English speakers are likely to treat this determiner differently compared to speakers of the other tested languages. In particular, the first observation suggests that one way in which this can be different is that English speakers do not associate *some* with the same numerical range as speakers of the other languages.

As concerns the second difference, we tentatively suggest that a reduced standard deviation at the extreme ends of a Likert scale indicates a greater speakers' confidence or certainty in their judgments, while intermediate scorings are more volatile. Thus a standard deviation can be seen as a measure at the continuum between less and more confident judgments. Both

<sup>8</sup>We stress, in this regard, that all observed correlations were *post hoc* and as such could not have affected our choice of the 3-point acceptability threshold itself (this thresholding methodology was also used in a similar experiment reported in Stateva and Stepanov, 2017).



speakers' confidence and volatility should be seen as qualities at the population, rather than individual, level. At the individual level, the ability to give a "confident" judgment in either direction is a function of a well-defined task. One will not be able to produce a confident judgment if the task conditions are in some sense vague, or allow for more than one "correct" answer, to the speaker. We will argue that this is precisely the case with English *some*, whereby our context conditions allowed for interpreting *some* in more than one way, differently from the way speakers of the other languages interpret it. To anticipate the forthcoming discussion, this alternative way of interpretation is associated with scalar implicatures, possible but not necessary for the speakers of English. In contrast, we will argue that the presence or absence of scalar implicatures enriching the meaning of counterparts of *some* in the other three languages does not affect perceived numerical bounds due to the an additional mechanism of pragmatic strengthening. So far, however, we believe that the point of divergence related to *some* could serve as a basis for a more general evaluation of the nature of quantifiers and a focus on the properties of *some* and its counterparts in the other languages will ultimately shed light on the four questions which motivated this study.

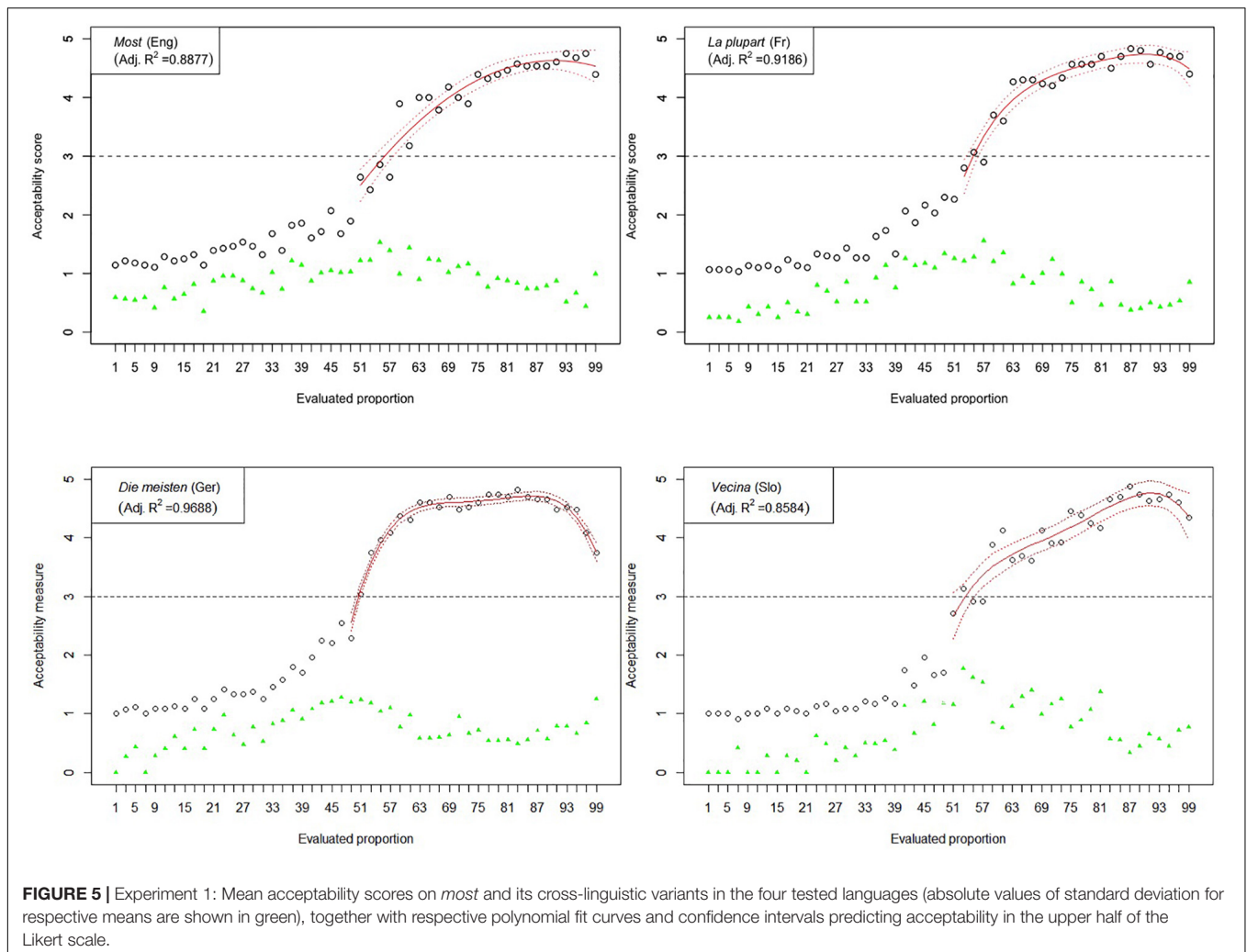
In order to get a clearer understanding of this peculiar difference between *some*, on the one hand and the other

existential quantifiers in French, Slovenian, and German, on the other we will look for other patterns of divergence. Recall that the standard semantic-pragmatic theory views the existential quantifier as a trigger of the quantity related implicature. Below we report the results of a second experiment which juxtaposes English and French, the latter as a representative of the group of languages that showed a similar pattern of processing their existential quantifier in Experiment 1.

## EXPERIMENT 2

### Implications for the Derivation of Scalar Implicatures

As we saw, there seems to be a difference between English *some* and its counterpart in French, Slovenian, and German. While in the other languages the quantifier is best used for an interval between *a few* and *half*, English *some* is best used for an interval between *a few* and *almost all*. This opens a lot of interesting questions, which have to do with whether this should be seen as a refutation of Grice's Modified Ockham Razor (in as much as the lexical meaning of the quantifier does not correspond to the logical entailment from *all* to *some*) or as a matter



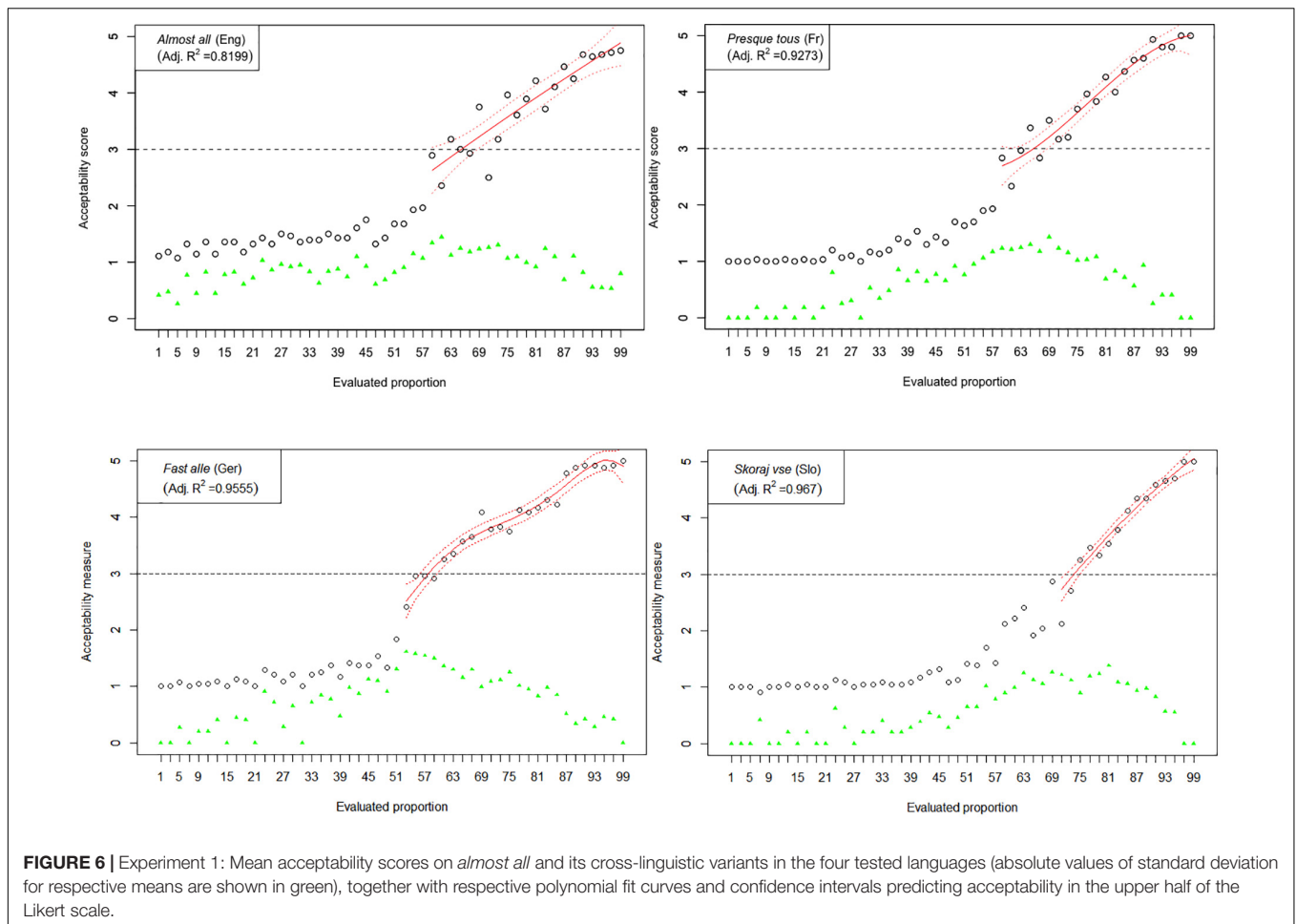
**FIGURE 5 |** Experiment 1: Mean acceptability scores on *most* and its cross-linguistic variants in the four tested languages (absolute values of standard deviation for respective means are shown in green), together with respective polynomial fit curves and confidence intervals predicting acceptability in the upper half of the Likert scale.

of typicality (see van Tiel, 2014), though the latter possibility would raise the further question of why English would pattern differently from other European languages, including German. But the main question we want to raise here is whether this difference between English *some* and its counterparts impacts the derivation of scalar implicatures. We address this question by comparing English *some* and French *quelques* in a simple picture choice test.

## Experimental Design

We choose a picture choice test paradigm in preference to the more frequently used sentence evaluation task paradigm (see, e.g., Noveck, 2001; Bott and Noveck, 2004) for a number of reasons. Notably, in a sentence evaluation task, the relevant condition is the one where *some* is under-informative, as it is the only one that allows one to differentiate between the pragmatic and the semantic interpretations. However, there are quite a few problems with that task, the first being that the rate of pragmatic answers (which ranges between 40 and 60%) is not clearly different from chance, given that participants have to choose between two answers (putting chance at 50%). This

suggests that the infelicity of the experimental condition leads participants to random answers. Another problem is that it is not clear that the task allows a reliable distinction between pragmatic answers (negative) and semantic answers (positive) (see Guasti et al., 2005; Mazzaggio et al., unpublished). Thus, a picture choice task, which offers a reliable distinction between the pragmatic and the semantic answers and avoids the difficulty linked to infelicity seemed by far a better choice. In essence, participants are presented with a sentence with a quantified NP (in the object position) and are asked to choose which among two pictures best corresponds to the sentence. In the *some* condition, one picture illustrates the pragmatic interpretation and the other illustrates the semantic interpretation. We tested French and English native speakers, as we will now describe. To avoid the confound raised by the entailment from *all* to *some*, participants were allowed a single answer. In addition to the *some* experimental condition, we also had an *only some* experimental condition. As Marty and Chemla (2013) have noted, the pragmatic interpretation of *some* has the same content as *only some*, the difference between the two being only the fact that the pragmatic interpretation is implicit. They



**FIGURE 6 |** Experiment 1: Mean acceptability scores on *almost all* and its cross-linguistic variants in the four tested languages (absolute values of standard deviation for respective means are shown in green), together with respective polynomial fit curves and confidence intervals predicting acceptability in the upper half of the Likert scale.

used thus a comparison between *some* and *only some* and we followed their example.

## Experimental Material

The experiment was composed of three main conditions, exemplified in **Figures 8–10**:

- one control condition, using *all* (four items);
- two test conditions:
  - *only some* (eight items);
  - *some* (eight items);
- Four filler conditions with four items each:
  - *half*;
  - *exactly one*;
  - *exactly two*;
  - *exactly three*.

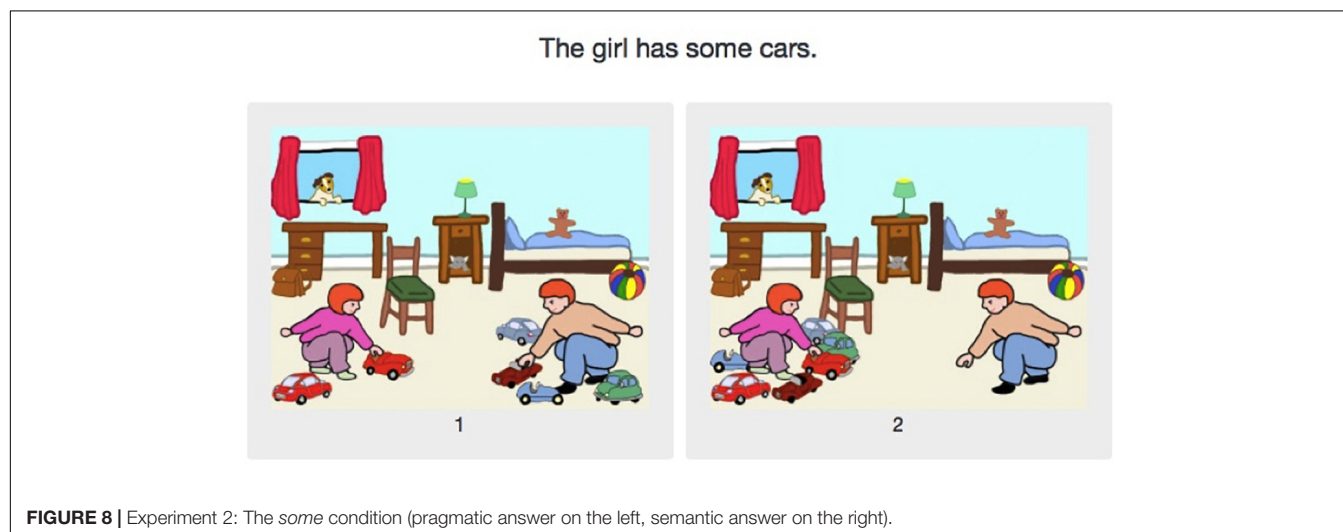
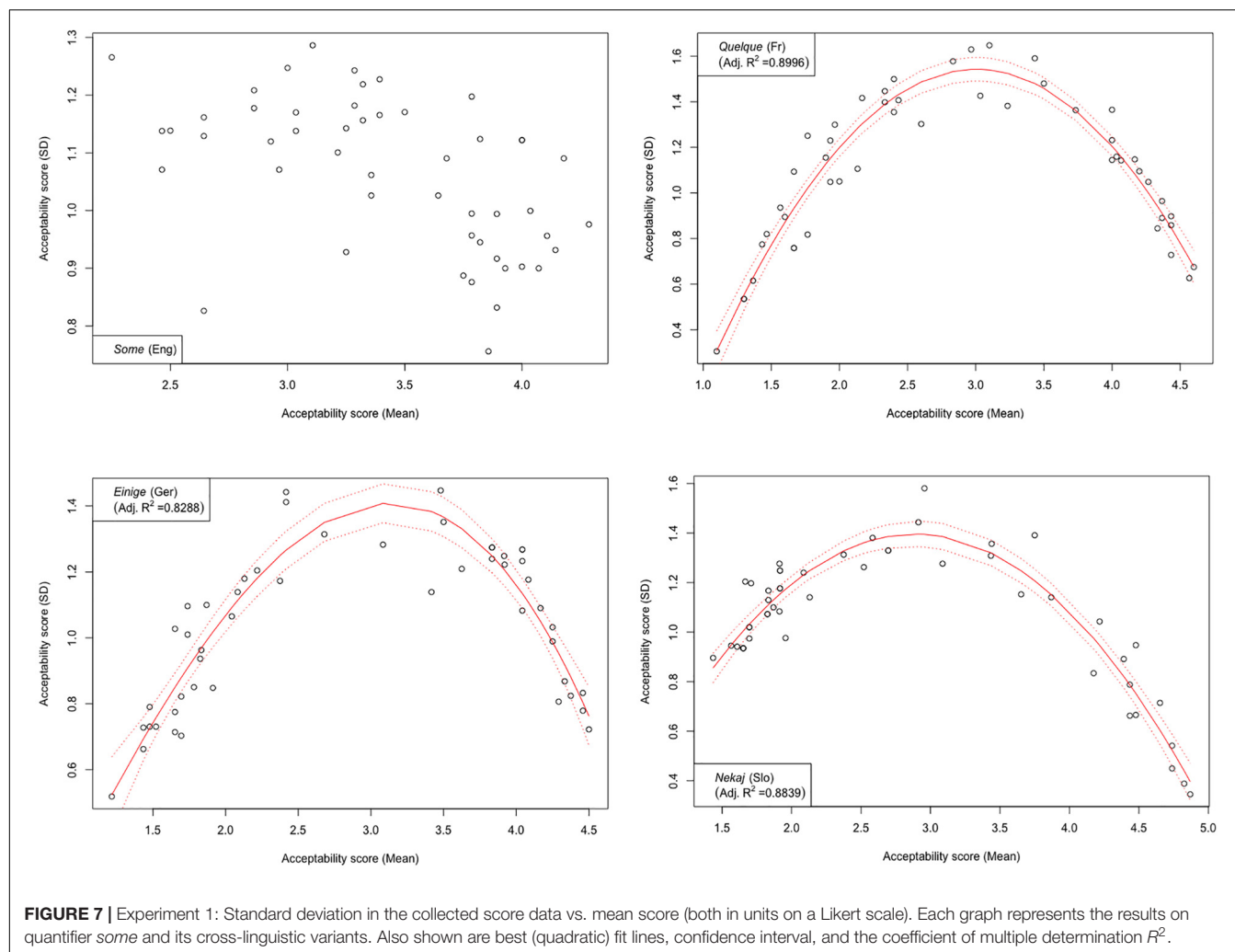
In the *some* condition, one image corresponds to the pragmatic interpretation and the other to the semantic interpretation, as **Figure 8** exemplifies. In all other conditions, including fillers, one image verified the sentence, while the other falsified it. All images presented two characters and six objects, which were either in the possession of a single character or shared

among both characters. In the two test conditions (*some* and *only some*), one picture showed the character named in the test sentence with all the objects, while the other showed them with only two of the six objects. **Figure 9** exemplifies an example of the other experimental condition, *only some*. As for the control condition, the evaluated sentences contained *all*, as illustrated in **Figure 10**. The “correct” choice was presented either on the left or on the right in a counterbalanced way. The filler conditions, while not necessarily standardly used in such experiments, also use quantifiers, albeit non-monotonous ones (for *exactly n*), and *half*. As exactly the same fillers were used for the French and for the English groups, it is very unlikely that the choice of fillers had any influence on the wide discrepancy between the French and the English results. Response times were not recorded in Experiment 2, as the question we were interested in was whether the difference between English *some* and French *quelques* evidenced in Experiment 1 would influence the rate of pragmatic answers in this simple picture choice task. It is not clear why response times as such would be directly relevant to that question.

## Participants

Twenty nine French participants were students at the University of Lyon, aged between 18 and 30 (mean age = 21.9; 17





females). They were all native speakers of French. In addition, 34 English participants were recruited through the Prolific platform. They were all students, aged between 18 and 30

(mean age = 23.1; 18 females). There was no significant difference in age across both tested groups (two-tailed  $t$ -test:  $t = 1.49$ ,  $p = 0.14$ ).

The boy in green has only some balls.



**FIGURE 9** | Experiment 2: The *only some* condition (the image verifying the sentence is on the left, the image falsifying it on the right).

The girl has all the dolls.



**FIGURE 10** | Experiment 2: The *all* condition (the image verifying the sentence is on the left, the image falsifying the sentence in on the left).

## Procedure

The experiment was presented online on the Qualtrics platform.<sup>9</sup> It began with a short introduction, where participants indicated sex, age, student status and confirmed that they were native speakers of French or, respectively, English. They were given instructions as well as an example of the task. They then proceeded to the experiment itself. The whole process lasted 10–15 min at the most.

## Results and Discussion

### Data Treatment and Exclusion

Exclusion was based on more than five items failed in either the control or the filler conditions. No participants were excluded.

### Response Analysis

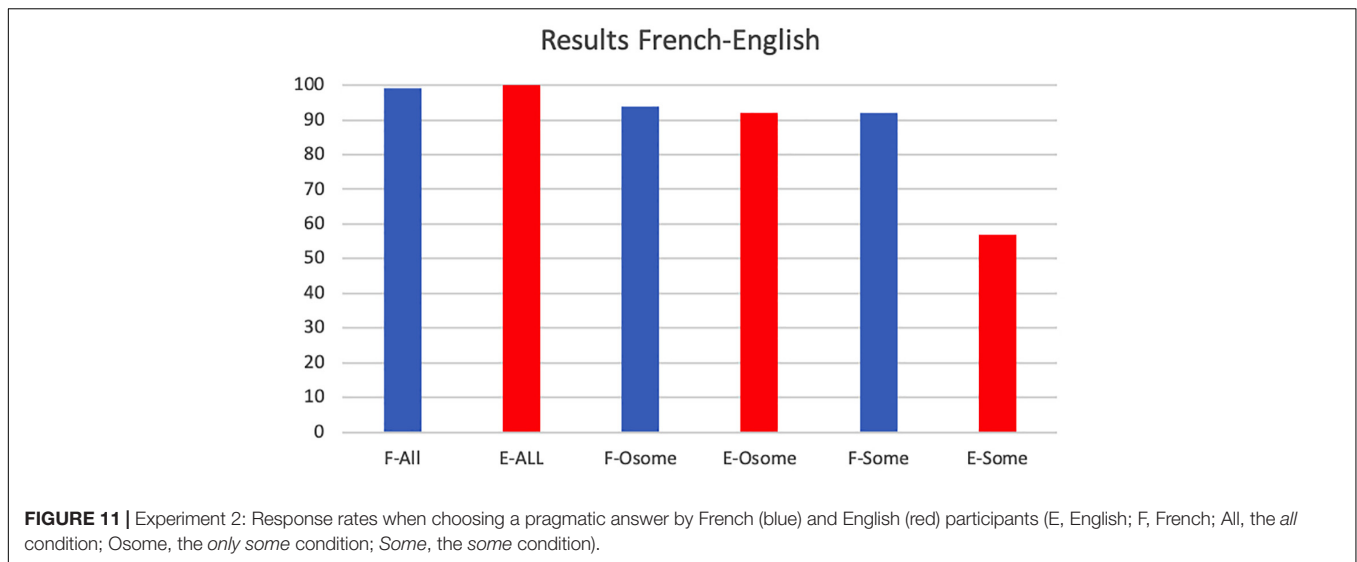
The rates of response in choosing a pragmatic answer are summarized in **Figure 11**. Comparing the response rates of choosing a pragmatic interpretation in English and French, we

find that French and English participants behave similarly in the *all* control condition choosing pragmatic answers in virtually all cases (French: 99.13%, English: 100%;  $\chi^2$  test:  $\chi^2(1) = 0.002$ ,  $p = 0.96$ , no significant difference at the 0.05 level) and in the *only some* test condition (French: 94.39%, English: 92.64%;  $\chi^2(1) = 0.021$ ,  $p = 0.88$ ). However, they behaved very differently in the *some* test condition, whereby French participants chose the pragmatic interpretation at a higher rate than did the English participants (French: 92.24%, English: 57.30%;  $\chi^2(1) = 11.901$ ,  $p = 0.0005$ ) and at a rate similar to that of their interpretation of *only some*. Correspondingly, French participants did not differ in their response rate on *some* and *only some* conditions [ $\chi^2(1) = 0.029$ ,  $p = 0.86$ ], while English participants showed a significantly greater preference for the targeted answer on the *only some* condition compared to the *some* condition [ $\chi^2(1) = 13.9040$ ,  $p = 0.0003$ ].

### Discussion

We tested French and English participants in the simple image-choice task for three main conditions: an *all* control condition,

<sup>9</sup><http://www.qualtrics.com>



a target *only some* condition, and a target *some* condition. This last condition was intended to establish whether French and English participants draw the scalar implicature at the same rate despite the difference in the interval inside which, respectively, *quelques* and *some* are best used in the two languages. It appears that they do not.

## GENERAL DISCUSSION

There are two general patterns that emerge from the cross-linguistic studies we report. The first one is that French, Slovenian and German counterparts of *few*, *half*, *most*, and *almost* are assigned very similar numerical bounds. The second one is that *some* and its variants like *quelques* are different in more than one way, namely: (i) with respect to numerical bounds, and (ii) with respect to their potential to trigger a scalar implicature.

In what follows we attempt to account for these facts by arguing that the set of quantifiers viewed as a natural class by the standard semantic-pragmatic theory is, in fact, diverse and the set-theoretic semantics is not appropriate for all of its members. As a consequence, the mechanism of pragmatic enrichment that these items trigger is of a different nature. Finally, we will argue that it is possible that languages do not assign the same kind of semantic definition to determiners that might, from a cross-linguistic perspective, look like translational equivalents.

We start with the point that not all quantifiers quantify over sets of individuals. There have been numerous proposals in the semantic literature that argue against the standard set-theoretic analysis and in favor of a degree-based analysis for some items. Classical cases involve *most*, *many*, *much*, *few*, *half* (cf. Rett, 2008, 2015; Hackl, 2009; Solt, 2009, 2015, etc.). However, if we assume that these particular quantifiers have a different semantic nature, we might face a challenge in restricting the application of those pragmatic principles which rely on the availability of semantic alternatives that constitute a natural class. Crucially, this

affects the derivation of quantity-based implicatures. Of course, all determiners that we have considered so far, including the ones we did not test like *every*, *all*, *no*, etc. have the same brevity, or roughly, morphological complexity and thus satisfy the basic criterion for serving as a source of a scale of alternatives (cf. Levinson, 1983, 2000). However, a stricter requirement on their semantic make-up can leave some of these determiners outside of the set of possible alternatives to *all/every*, for example. But even if this is so, the results from our Experiment 1, as well as the results from the other psychometric and typicality-based studies indicate that the meanings of degree-based determiners are pragmatically enriched because they differ from the respective truth-conditional meanings. So if quantity implicatures are not always available for pragmatic enrichment in the domain of quantifiers, how can one account for pragmatic strengthening in all degree-based quantifiers?

A possible answer comes from a proposal in Stateva and Stepanov (2017). That proposal extends Krifka's (2007a) analysis of negated antonyms (like *happy*, *not happy*, *unhappy*, *not unhappy*) to the domain of the Slovenian degree quantifiers *precej* and *veliko*, both of which are counterparts of the English *many*. The gist of the proposal is that *precej* and *veliko* are semantically equivalent but their meanings are differentiated as a result of pragmatic enrichment through an M-implicature and an R/I-implicature, respectively. R/I-implicatures are associated with a stereotypical interpretation while M-implicatures are related to non-stereotypical interpretations (Horn, 1984; Levinson, 2000). A prerequisite for the Krifka-type analysis is a state of affairs in which there is at least one pair of antonyms so that together they exhaust a relevant degree scale as contradictories. Since degree predicates are vague, the cut-off point is related to epistemic uncertainty for the speaker (Williamson, 1994). In the availability of synonyms in the positive or the negative extension of the scale, as is the case with the two Slovenian positive amount words *precej* and *veliko* that are antonyms to the negative *malo* "few," a stereotypical interpretation, i.e., an interpretation which is related to a segment of the positive scale which is at a

safe distance from the potential cut-off point is assigned to one of the synonyms as an R/I-implicature. The stereotypical interpretation is then always closer to the endpoint of the scale than the non-stereotypical interpretation which results from the application of an M-implicature. If we generalize on the basis of the Slovenian case involving the two quantifiers *precej* and *veliko*, we will have a potential mechanism of pragmatic enrichment of other degree quantifiers which are part of a paradigm that contains at least one antonym and at least one synonym to them.

Very importantly, the above suggestion does not exclude quantity implicatures in the degree domain in general. Under a strict version of restricting scalar alternatives, we expect that, for example, *most* and *all* should not be members of a Horn-set given that one involves quantification over degrees and the other, quantification over individuals but the two degree-based quantifiers *most* and *almost all* would. This would explain why the upper part of the degree scale is not accessible for *most* (although the truth-conditional meaning of *most* is compatible with it). Arguably, *most* triggers a scalar implicature that negates the *almost-all* alternative.

We now have the ingredients for a proposal that explains the facts from the reported experiments. We would like to suggest that the existential quantifiers that we tested are of different semantic nature and because of that they are subject to different processes of pragmatic strengthening. To English *some* we attribute the standard semantic meaning as relating two sets of individuals. The results from both Experiment 1 and Experiment 2 suggest that *some* is pragmatically enriched with a scalar implicature. This hypothesis is confirmed (i) by the larger acceptability interval on the proportion scale for *some* in comparison to the rest of the tested existential quantifiers where *some* covers also very high ratios, as predicted, and (ii) by the lower rate of scalar implicature derivation associated with *some* in comparison to *quelques* which is also expected given the optional character of scalar implicatures. As for the counterparts of *some* in French, Slovenian, and German, we would like to suggest that they are degree-based quantifiers, lexically synonymous to the lexical items corresponding to *few* in each of the languages and antonyms of the lexical items corresponding to *many*. In this analysis, the French *quelques*, Slovenian *nekaj* and German *einige* are associated with the lower part of the degree scale while the respective counterparts of *many* in each language are associated with intervals above the cut-off point. All three languages have a lexical version of *few* which competes with *quelques*, *einige*, or *nekaj* for the stereotypical or non-stereotypical interpretation. Our results from Experiment 1 suggest that *quelques*, *einige*, and *nekaj* are pragmatically enriched with the non-stereotypical implicature and are thus at a greater distance from the scale and-point in comparison to the stereotypically interpreted counterpart of *few*. Some overlap within synonym pairs in each of the languages is always expected due to epistemic uncertainty because of the vague character of quantifiers that do not denote end points. In much the same vein in which speakers are uncertain about the cut-off point on a relevant scale between two antonyms and simultaneously have a whole set of potential cut-off points under consideration

speakers entertain a set of cut-off points within the scale part associated with the pair of synonyms. As a result of epistemic uncertainty, there are overlaps in all zones coinciding with potential points of delineation.

This explanation gets further support from the results of Experiment 2. Recall that in the Picture-Choice task, French speakers were at ceiling with the choice of the pragmatic meaning while English speakers had a significantly lower rate of choosing the pragmatic answer in comparison to the French speakers. These facts are consistent with the hypothesis that the pragmatically enriched English target sentence results from a scalar implicature negating the *all*-alternative which is only optional (Chierchia, 2013). When the implicature is forced by the explicit use of *only*, speakers responded in accord with expectations and performed at ceiling, too. The relevant pragmatic alternative for the French speakers in the target condition is, in fact, not based on *all* but rather on *few* and so the non-targeted answer did not interfere in this case.

The proposal makes a prediction for the relation between *many* and *some* and their respective counterparts. These items are a pair of antonyms in French/Slovenian/German-type languages. This entails some overlap region on the degree scale corresponding to the zone where different cut-off points are under consideration because of epistemic uncertainty but the overlap cannot be too large. As for English, *some* and *many* can partially overlap to a greater extent. We have indirect confirmation of this prediction from Experiment 1: as we saw previously, unlike its counterparts, English *some* is acceptable in contexts with very high proportions bordering the region reserved for the upper part of *most* and *almost all*. This interval can be reasonably expected to contain the interval allotted to *many*.

This is the stronger version of the proposal we want to push forward. A weaker version of it would not exclude scalar implicatures based on entailment relations even among the members of the class of quantifiers that are triggers of R/I-implicatures or M-implicatures in French/Slovenian/German, i.e., among the counterparts of *some*, *few*, and *many*. To give substance to this possibility we can refer to Chemla (2007) and Buccola et al. (2018) which suggest that scales of alternatives are based on concepts rather than lexical elements. If this is so, a Horn-set of alternatives can well be formed by quantifiers that do not denote functions of the same semantic type. Under this weaker proposal, however, the availability of M-implicature for the French, Slovenian and German counterparts of *some* in contrast to the R/I-implicature triggered by the counterparts of *few* would trivialize the effect of quantity induced implicatures which, in this case, would not be necessary to explain the facts about the existential quantifier cross-linguistic differences we observe in Experiments 1 and 2.

The results from Experiment 1 are in line with the findings of Pezzelle et al. (2018). They demonstrate that quantifiers are perceived as part of an ordered scale which involves overlaps (i.e., similarities) of different dimensions. We argued that quantifier differentiation depends on more than one mechanism of pragmatic enrichment on a par with the semantic makeup of quantifiers as potential alternatives.



As we stated above, our research question about quantifier meanings in language use is focused on potential cross-linguistic variation. The data we collected suggests that such differences exist and they have a systematic character as confirmed by both experiments reported here. Let us, however, consider briefly some other studies bearing on cross-linguistic variation among quantifiers. Katsos et al. (2016) reports a study on quantifier acquisition by 5-year-old children in 31 languages, among which three of the languages discussed here: English, French, and German. That study includes a task on the counterparts of *some* and *most* which makes the comparison between both studies possible. However, the German existential quantifier tested in Katsos et al. (2016) is *ein Paar* and since it is different from *einige* used in our study, we will limit our attention to English and French only. In contrast to our study, Katsos et al. (2016) does not report any relevant cross-linguistic variation. Whether this result contradicts the results we report, however, can only be appreciated if we scrutinize the research questions and the tested hypotheses of that study. Katsos et al. (2016) investigate whether the order of acquisition of quantifiers is similar across languages given a number of factors related to the formal properties of different determiners. More specifically, these are monotonicity (upward vs. downward), totality [related to scale endpoint (e.g., *no*, *all*) or non-end-point (*some*, *most*)] (morphological) complexity and finally, truth versus felicitousness, i.e., whether pragmatic meaning is acquired after semantic meaning.<sup>10</sup> In effect, the comparisons track the acquisition order among quantifiers within each language but not the order of acquisition of translational equivalents among languages which would have been indicative of potential cross-linguistic differences among translational equivalents. In particular, results reported from testing 17 English speaking children and 15 French speaking children show that in both languages, accuracy of *some/quelques* is higher in comparison to *most/la plupart*, respectively, as predicted by the hypothesis that *most/la plupart*, being the superlative form of *many*, is morphologically more complex than *some/quelques*. However, this finding is orthogonal to our study because there is no a priori reason to assume that a quantifier based on degrees, as we argue, is more complex, and therefore, more difficult to acquire than a quantifier that relates sets of individuals. Consequently, the comparable accuracy of English and French participants on the conditions related to the acquisition of *some* in English and *quelques* in French, respectively, cannot be interpreted as a counterargument against the proposal we are advancing. What is more, the data obtained by Katsos et al. (2016) is not inconsistent with the data obtained within the study we report.

Before we conclude this discussion, we would like to mention two facts that could serve as independent evidence for our proposal. The first one is based on an observation about the morphological makeup of the plural morphology paradigm in Bulgarian. Bulgarian features two plural nominal agreement patterns in the masculine paradigm. The default case is a plural ending that agrees with the plural morpheme of any adjectival

modifier within the nominal phrase. The second one, known as the “count form,” is non-agreeing, and is selected if the noun is preceded by a numeral (cf. Stoyanov, 1980; Stateva and Stepanov, 2016, etc.). Both plural patterns are exemplified in (2a) in (2b), respectively:

- (2) a. Červen-i (dârven-i) prozorec-i  
red-pl wooden-pl window-pl  
“red (wooden) windows”
- b. Pet (dârven-i) prozorec-a  
five wooden-pl window-count  
“five (wooden) windows”

Interestingly, the count form is also used when the noun contains the existential quantifier *njakolko* “some” but not when it contains the universal one *vsichki* “all,” as shown in (3):

- (3) a. Njakolko (dârven-i) prozorec-a/\*prozorec-i  
some wooden-pl window-count/window-pl  
“some wooden windows”
- b. Vsichki (dârven-i) prozorec-i/\*prozorec-a  
all wooden-pl window-pl/window-count  
“all wooden windows”

The parallel between numerals and *njakolko* indicates that they belong to the same natural class to the exclusion of *vsichki*. The possibility of having a numeral-like existential quantifier in one language suggests a similar possibility for other languages even in the absence of morphological makeup indicative of the specific semantic nature of the quantifier.

Second, our proposal can account for the observation that numerical bounds of vague quantifiers depend on the cardinality of the total set. If pragmatic enrichment of degree-based quantifiers that come in pairs of antonyms and synonyms depends on delineation between lower and upper scale parts, as well as on interval assignment to stereotypical and non-stereotypical, we can expect that partitioning in a closed scale of this kind will involve a lot of overlaps. This is so because each interval to which a quantifier is related in this case ends up being too small to be distinguished from the neighboring ones, especially in view of epistemic uncertainty. It follows then that different numerical bounds are associated with the same quantifier in small and larger sets where competing alternative quantifiers are assigned to greater scale intervals.

## CONCLUSION

We conclude by going back to the questions we posed in the Section “The Present Study.” We started with the question of whether it is possible to identify the different quantifiers’ numerical bounds and whether these are encoded in quantifier meanings or are epiphenomenal. The answer that follows from our discussion is that numerical ranges are epiphenomenal: they result from pragmatic strengthening and no additional meaning component needs to be postulated in order to account for the difference between lexical meanings and actual judgments in tasks.

<sup>10</sup>Upward monotone quantifiers license inferences to supersets while downward monotone quantifiers license inferences to subsets.

We believe that the cross-linguistic perspective that we added to this study sheds light on the question of whether quantifier meanings can be given the status of a semantic universal (Determiner Universal, Barwise and Cooper, 1981). If our interpretation is correct, the existential quantifier is a source of considerable cross-linguistic variation. An anonymous reviewer raises a question related to it. It pertains to the source of this difference from the point of view of language change. While it is not that difficult to assume that a Slavic language like Slovenian, or a Romance language like French might differ in some fundamental aspect from English, which belongs to the Germanic family, it is much less obvious why English and German would not pattern together. Assuming that there has been a common source for the existential quantifier, it is important to look for an answer to the question about the trigger of the semantic shift and the trajectory leading to these two different patterns. While we acknowledge the importance of this question, it falls beyond the focus of our current study and therefore we leave it for future research.

We identified two types of pragmatic enrichment processes that are operative in the domain of quantifiers: quantity-based enrichment through scalar implicatures, and stereotypical and non-stereotypical meaning enrichment through R/I-implicatures and M-implicatures. We argued that if we assume a cognitive-based definition of pragmatic alternative, both processes are operative but in some cases the effect of quantity induced implicatures is trivialized.

Finally, we come to the question of overlapping meanings. We argued that meaning overlap is language dependent and is less likely to be expected in cases that involve pragmatic strengthening through R/I- and M-implicatures.

## ETHICS STATEMENT

The experiments in this study were carried out in accordance with the Declaration of Helsinki and the existing European

and international regulations concerning ethics in research. All participants gave an informed consent prior to the beginning of testing. In addition, the English portion of Experiment 1 was approved by the IRB at Rutgers, the State University of New Jersey. Experiment 2 was carried out in accordance with the recommendations of the Comité de Protection des Personnes Sud Est II, who gave it its agreement (IRB number: 11263).

## AUTHOR CONTRIBUTIONS

PS and AS designed the Experiment 1 and collected the Slovenian data. PS participated in collecting the German data. LD and VD collected the French and English data, respectively, for Experiment 1. AR designed the Experiment 2 and collected the French and English data for it. PS, AS, and AR analyzed all of the data and wrote the manuscript.

## FUNDING

PS acknowledges the support of the EURIAS Fellowship Programme and the European Commission (Marie-Sklódowska-Curie Actions – COFUND Programme – FP7). AS acknowledges the financial support from the Slovenian Research Agency (research core funding No. P6-0382).

## ACKNOWLEDGMENTS

For insightful comments and suggestions, the authors gratefully acknowledge Aurelién Belot, John Hamman, two reviewers, and the audience of the Workshop “Logical Words” which took place at the University of Geneva in January 2019 where a portion of this work was presented.

## REFERENCES

- Barwise, J., and Cooper, R. (1981). Generalized quantifiers in natural language. *Linguist. Philos.* 4, 159–219.
- Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006
- Buccola, B., Kriz, M., and Chemla, E. (2018). *Conceptual Alternatives: Competition in Language and Beyond*. Available at: <https://ling.auf.net/lingbuzz/003208/current.pdf> (accessed June 20, 2018).
- Chemla, E. (2007). French both: a gap in the theory of antipresupposition. *Snippets* 15, 4–5.
- Chierchia, G. (2013). *Logic in Grammar: Polarity, Free Choice, and Intervention*. Oxford: Oxford University Press.
- Chierchia, G., Fox, D., and Spector, B. (2012). “Scalar implicature as a grammatical phenomenon,” in *An International Handbook of Natural Language Meaning Semantics*, eds P. Portner, C. Maienborn, and K. von Stechow (Berlin: Mouton de Gruyter), 2297–2331.
- Cummins, C. (2014). Typicality made familiar. *Semant. Pragmat.* 7, 1–15.
- Daamen, D. D. L., and de Bie, S. E. (1991). “Serial context effects in survey interviews,” in *Context Effects in Social and Psychological Research*, eds N. Schwarz and S. Sudman (New York, NY: Springer-Verlag).
- de Carvalho, A., Reboul, A., Van der Henst, J.-B., Cheylus, A., and Nazir, T. (2016). Scalar implicatures: the psychological reality of scales. *Front. Psychol.* 7:1500. doi: 10.3389/fpsyg.2016.01500
- Foppolo, F., Guasti, M. T., and Chierchia, G. (2012). Scalar implicatures in child language: give children a chance. *Lang. Learn. Dev.* 8, 365–394. doi: 10.1080/15475441.2011.626386
- Fuhrmann, G. (1991). Note on the integration of prototype theory and fuzzy-set theory. *Synthese* 86, 1–27. doi: 10.1007/bf00485412
- Gergel, R., and Stateva, P. (2014). “A compositional analysis of almost : diachronic and experimental comparative evidence,” in *Pre-Proceedings of the International Conference Linguistic Evidence 2014* (Tübingen: Eberhard Karls Universität), 150–156.
- Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., and Meroni, L. (2005). Why children sometimes but not always compute scalar implicatures. *Lang. Cogn. Process.* 20, 667–676.
- Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Nat. Lang. Semant.* 17, 63–98. doi: 10.1007/s11050-008-9039-x
- Heim, I. (1991). “Artikel und Definitheit,” in *Semantik: Ein Internationales Handbuch der Zeitgenössischen Forschung*, eds A. von Stechow, and D. Wunderlich (Berlin: De Gruyter), 487–535.

- Horn, L. (1984). "Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature," in *Meaning, Form, and Use in Context: Linguistic Applications*, ed. D. Schiffrin (Washington DC: Georgetown University Press), 11–89.
- Horn, L. R. (2004). "Implicature," in *The Handbook of Pragmatics*, eds L. R. Horn and G. Ward (Oxford: Blackwell Publishing).
- Katsos, N., and Bishop, D. V. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015
- Katsos, N., Cummins, C., Ezeizabarrena, M. J., Gavarró, A., Kraljević, J. K., Hrzica, G., et al. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proc. Natl. Acad. Sci. U.S.A.* 113, 9244–9249. doi: 10.1073/pnas.1601341113
- Krifka, M. (2007a). "Negated antonyms: Creating and filling the gap," in *Presupposition and Implicature in Compositional Semantics*, eds U. Sauerland and P. Stateva (Basingstoke: Palgrave Macmillan), 163–177. doi: 10.1057/9780230210752\_6
- Krifka, M. (2007b). "Approximate interpretations of number words: A case for strategic communication," in *Cognitive Foundations of Interpretation*, eds G. Bouma, I. Krämer, and J. Zwarts (Amsterdam: Royal Netherlands Academy of Arts and Sciences), 111–126.
- Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Levinson, S. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Language, Speech, and Communication. Cambridge, MA: The MIT Press.
- Lipovetsky, S. (2017). Factor analysis by limited scales: which factors to analyze? *J. Mod. Appl. Stat. Methods* 16, 233–245. doi: 10.22237/jmasm/1493597520
- Marty, P. P., and Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Front. Psychol.* 4:403. doi: 10.3389/fpsyg.2013.00403
- Moxey, L., and Sanford, A. J. (1993). Prior expectation and the interpretation of natural language quantifiers. *Eur. J. Cogn. Psychol.* 5, 73–91. doi: 10.1080/09541449308406515
- Moxey, L., and Sanford, A. J. (2000). Communicating quantities: a review of psycholinguistic evidence of how expressions determine perspectives. *Appl. Cogn. Psychol.* 14, 237–255. doi: 10.1002/(sici)1099-0720(200005/06)14%3A3%3C237%3A%3Aaid-acp641%3E3.0.co%3B2-r
- Newstead, S. E., Pollard, P., and Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Appl. Ergon.* 18, 178–182. doi: 10.1016/0003-6870(87)90001-9
- Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicatures. *Cognition* 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1
- Noveck, I. (2018). *Experimental Pragmatics: The Making of Cognitive Science*. Cambridge: Cambridge University Press.
- Penka, D. (2006). "Almost there: The meaning of almost," in *Proceedings of Sinn und Bedeutung 10th Annual Meeting of the Gesellschaft für Semantik* (Berlin: Zentrum für allgemeine Sprachwissenschaft).
- Pezzelle, S., Bernardi, R., and Piazza, M. (2018). Probing the mental representation of quantifiers. *Cognition* 181, 117–126. doi: 10.1016/j.cognition.2018.08.009
- R Core Team (2018). *R: A Language and Environment For Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rett, J. (2008). *Degree Modification in Natural Language*. Ph.D. thesis, Rutgers University, New Brunswick.
- Rett, J. (2015). *The Semantics of Evaluativity*. New York, NY: Oxford University Press.
- Rosch, E. (1975). Cognitive representations of semantic categories. *J. Exp. Psychol.* 104, 192–233. doi: 10.1037//0096-3445.104.3.192
- Solt, S. (2009). *The Semantics of Adjectives of Quantity*. Ph.D. thesis, City University of New York: New York, NY.
- Solt, S. (2015). Q-adjectives and the semantics of quantity. *J. Semant.* 32, 221–274.
- Sperber, D., and Wilson, D. (1995). *Relevance: Communication and Cognition*. Oxford: Basil Blackwell.
- Stateva, P., and Stepanov, A. (2016). Agreement errors and structural distance: a corpus study of Bulgarian. *Zeitschrift für Slawistik* 61, 448–462.
- Stateva, P., and Stepanov, A. (2017). Two "many"-words in slovenian: experimental evidence for pragmatic strengthening. *Acta Linguist. Acad. Int. J. Linguist.* 64, 435–473. doi: 10.1556/2062.2017.64.3.7
- Stoyanov, S. (1980). *Gramatika na Balgarskiya Knizhoven Ezik. Fonetika i Morfologia [Grammar of the Bulgarian Literary Language. Phonetics and Morphology]*. Sofia: Nauka i izkustvo.
- van Tiel, B. (2014). Embedded scalars and typicality. *J. Semant.* 31, 147–177. doi: 10.1093/jos/fft002
- van Tiel, B., and Geurts, B. (2014). "Truth and typicality in the interpretation of quantifiers," in *Proceedings of Sinn und Bedeutung 18*, eds U. Etxeberria, A. Fălăuş, A. Irurtzun, and B. Leferman (Vitoria-Gasteiz: University of Basque Country), 433–450.
- Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., and Forsyth, B. (1986). Measuring the vague meanings of probability terms. *J. Exp. Psychol. Gen.* 115, 348–365. doi: 10.1037//0096-3445.115.4.348
- Wallsten, T. S., Budescu, D. V., and Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgements. *Manag. Sci.* 39, 176–190. doi: 10.1287/mnsc.39.2.176
- Williamson, T. (1994). *Vagueness*. London: Routledge.
- Yildirim, I., Degen, J., Tanenhaus, M. K., and Jaeger, T. F. (2013). "Linguistic variability and adaptation in quantifier meanings," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, eds M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Austin, TX: Cognitive Science Society), 3835–3840.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Stateva, Stepanov, Déprez, Dupuy and Reboul. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Some Pieces Are Missing: Implicature Production in Children

Sarah F. V. Eiteljoerge<sup>1,2,3</sup>, Nausicaa Pouscoulous<sup>3\*</sup> and Elena V. M. Lieven<sup>4</sup>

<sup>1</sup> Psychology of Language, University of Göttingen, Göttingen, Germany, <sup>2</sup> Leibniz ScienceCampus Primate Cognition, Göttingen, Germany, <sup>3</sup> Psychology and Language Sciences, University College London, London, United Kingdom, <sup>4</sup> ESRC International Centre for Language and Communicative Development, School of Health Sciences, University of Manchester, Manchester, United Kingdom

## OPEN ACCESS

### Edited by:

Anne Colette Reboul,  
Claude Bernard University Lyon 1,  
France

### Reviewed by:

Alex de Carvalho,  
University of Pennsylvania,  
United States  
Joanna Blochowiak,  
Université Catholique de Louvain,  
Belgium

### \*Correspondence:

Nausicaa Pouscoulous  
n.pouscoulous@ucl.ac.uk

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 18 May 2018

**Accepted:** 19 September 2018

**Published:** 24 October 2018

### Citation:

Eiteljoerge SFV, Pouscoulous N and  
Lieven EVM (2018) Some Pieces Are  
Missing: Implicature Production in  
Children. *Front. Psychol.* 9:1928.  
doi: 10.3389/fpsyg.2018.01928

Until at least 4 years of age, children, unlike adults, interpret *some* as compatible with *all*. The inability to draw the pragmatic inference leading to interpret *some* as *not all*, could be taken to indicate a delay in pragmatic abilities, despite evidence of other early pragmatic skills. However, little is known about how the production of these implicature develops. We conducted a corpus study on early production and perception of the scalar term *some* in British English. Children's utterances containing *some* were extracted from the dense corpora of five children aged 2;00 to 5;01 ( $N = 5,276$ ), and analysed alongside a portion of their caregivers' utterances with *some* ( $N = 9,030$ ). These were coded into structural and contextual categories allowing for judgments on the probability of a scalar implicature being intended. The findings indicate that children begin producing and interpreting implicatures in a pragmatic way during their third year of life, shortly after they first produce *some*. Their production of *some* implicatures is low but matches their parents' input in frequency. Interestingly, the mothers' production of implicatures also increases as a function of the children's age. The data suggest that as soon as they acquire *some*, children are fully competent in its production and mirror adult production. The contrast between the very early implicature production we find and the relatively late implicature comprehension established in the literature calls for an explanation; possibly in terms of the processing cost of implicature derivation. Additionally, *some* is multifaceted, and thus, implicatures are infrequent, and structurally and contextually constrained in both populations.

**Keywords:** scalar implicatures, production, *some*, corpora, pragmatic development, language acquisition

## 1. INTRODUCTION

A lot of information conveyed in conversation is not communicated explicitly, but implicitly; it is left for the audience to infer. For instance, if a student says she "read some of the papers assigned," the listener may infer that she has not read all of them even though this was not been stated. Deriving the implicit interpretation of an utterance seems challenging for young children (Noveck, 2001; Papafragou and Musolino, 2003). Most work on how children come to grips with implicit meaning was carried out on scalar terms such as *some*. These expressions are part of a semantic informativeness scale (e.g., *some*, *most*, *all*) and the use of a weaker term on the scale (*some* or *most*) will often be taken to imply the negation of the stronger term (*all*) giving rise to a scalar implicature.



In experimental contexts, children, unlike adults, interpret *some* as compatible with *all*, and are not found to be adult-like until seven (Noveck, 2001; Papafragou and Musolino, 2003; Guasti et al., 2005; Huang and Snedeker, 2009b). While the age at which children draw scalar implicatures has been pushed down in some paradigms, they are still not found to interpret *some* in a pragmatic way until at least 4 years of age (Pouscoulous et al., 2007; Katsos and Bishop, 2011).

One of the keys to the enigma of scalar implicature development has to be production. Indeed, little is known about how the most popular scalar term, *some*, is produced by children. In the hope to shed light on implicature competence in early childhood we conducted a corpus study looking at the production of the quantifier *some* by five British English children aged two to five and their caregivers.

Most experimental work on children's understanding of implicit meaning has focused on children's interpretation of scalar implicatures. These occur when a speaker chooses to use a weaker expression (e.g., *some*) where she could just as easily have used a stronger one (i.e., *all*) and the hearer thereby understands that she has reasons not to use the stronger one—either because she did not have sufficient information or because she knew that it was inappropriate to use the stronger expression.

According to Grice's (1957; 1989) widely accepted model, implicatures—including scalar implicatures—are propositions that the speaker intends to communicate even though she does not express them explicitly. Hearers can infer the intended implicature by assuming that the speaker is cooperative and that she tries, as much as possible, to follow the conversational maxims of quantity, relevance, truth, and manner. In the case of scalar expressions such as *some* the hearer assumes that the speaker abides by the first sub-maxim of quantity ("Make your contribution as informative as is required"), at least so long as she can honour the second sub-maxim of quality, as well ("Do not say that for which you lack adequate evidence"). Therefore, in the example above, the hearer can infer that the speaker intends to convey the upper-bounded reading of *some* (not *all*) either because she does not know if the student read all the papers, or because she knows that the student did not read all of them. Depending on the context, scalar terms may therefore have two different interpretations, either a lower-bounded reading where *some* is compatible with *all* or an upper-bounded one, which excludes *all*. It is important to bear in mind that in real conversational uses a context might neither clearly prompt nor exclude a *some*-related implicature; in such contexts the relevance of the stronger alternative (*all*) may be uncertain and hearers' intuitions might diverge on whether a scalar implicature was intended by the speaker.

Scalar implicatures are particularly interesting for two reasons. First, they have stirred up a lot of theoretical controversy in recent years (for a review, see Geurts, 2010). It is hotly debated whether these implicatures are an output of grammar (Chierchia et al., 2012) or of fully-fledged pragmatic inferences. Amongst the defenders of the latter position, some view them as regular implicatures ("particularised" implicatures, in Gricean terms), which are derived only when prompted by the context (Noveck and Sperber, 2007; Geurts, 2010), while others argue

they are "generalised" implicatures—i.e., they arise unless the context blocks them (Horn, 1989) or even by default (Levinson, 2000). Second, scalar implicatures often arise from the use of specific terms such as *some*, which makes them much easier to use in experimental settings. And, indeed, fueled by the theoretical debates, scalars have given rise to an important body of adult empirical work (for a review, see Breheny, forthcoming). The assumption behind much work on pragmatic development is that the findings on scalar implicatures can be generalised to other types of implicit meaning. Most studies on scalar expressions focus on the quantifiers *some*. In practice, this means our knowledge on children and implicatures is largely based on their understanding of *some* (for other implicatures, see Noveck et al., 2009; Schulze et al., 2013; Wilson, 2017).

Noveck (2001) conducted the first systematic experiments on children's treatment of scalar expressions. He asked 8- to 10-year-olds to assess sentences of the form "Some giraffes have long necks," which are logically true, but pragmatically underinformative, since "all giraffes have long necks." Most children accepted the pragmatically underinformative utterances as true (at rates of 89%), while adults tended to reject them as false (41% accepted these as true). Unlike adults, children accept (rather than reject) utterances expressed with relatively weak terms when a stronger one is called for, and thus appear to be more literal than adults. These results were supported at the time by classic studies that inadvertently included scalar expressions (Paris, 1973; Smith, 1980; Braine and Romain, 1981). Since then, several studies further demonstrated the phenomenon using a range of experimental methods (Papafragou and Musolino, 2003; Feeney et al., 2004; Guasti et al., 2005; Huang and Snedeker, 2009b). The effect seems to hold cross-linguistically with quantifiers (Katsos et al., 2016) and can be generalised to other scalar expressions; it has been found with 5-year-olds with *or* (not *and*) (Chierchia, 2004), *might* (not *must*) (Noveck, 2001), *start* (not *finish*) as well as numerals (Papafragou and Musolino, 2003). In all these experiments, the great majority of children accepted the weaker term as compatible with a stronger one, whereas adults would either consider them to be incompatible or at the very least be equivocal. Taken together, these findings might suggest that young children are unable to derive pragmatic inferences prompted by scalar expressions (for reviews on developmental findings on scalars, see Siegal and Surian, 2004; Pouscoulous and Noveck, 2009; Katsos, 2014; Papafragou and Skordos, 2016).

Children's performance on these implicature comprehension tasks is not due to semantic shortcomings. Indeed, children acquire *some* and *all* at around age 2 in both comprehension (roughly 16 months) and production (at roughly 26 months, Fenson et al., 1994). Furthermore, control conditions on most of the experiments described above indicate that children have a good semantic grasp of the two quantifiers (although, for a more nuanced picture, see Barner et al., 2009; Horowitz et al., 2017). Yet, other factors may influence children's performance on linguistic tasks—and in particular their understanding of pragmatic phenomena. Most studies mentioned above involve some type of sentence verification task. Children have to judge the truth or, at least, the adequacy of an utterance, a task which

taps into their metalinguistic abilities. These develop through childhood, and children have been shown to understand a pragmatic phenomenon at an earlier age when assessed on non-metalinguistic tasks (such as act-out tasks or picture selection tasks) than when their comprehension of the same phenomenon is established based on tasks involving metalinguistic skills (see, e.g., Bernicot et al., 2007). In some paradigms, children have been shown to derive scalar implicatures, suggesting their poor performance is not due to semantic or pragmatic inability. Indeed, 5-year-olds' performance improves when they are trained to detect pragmatic infelicities (Papafragou and Musolino, 2003; Guasti et al., 2005). Importantly, it also does when the implicature outcome is made more salient and relevant in context (Papafragou and Musolino, 2003; Guasti et al., 2005; Foppolo et al., 2012; Skordos and Papafragou, 2016). Even 4-year-olds have been shown to derive scalar implicatures in two paradigms. In one of them, the child's understanding was assessed using a ternary scale rather than a binary choice; children could reward the speaker's utterance with a small, medium, or large strawberry rather than decide they were right or wrong (Katsos and Bishop, 2011). In the other, a simplified act-out paradigm was designed aiming to reduce task cognitive load and the effort involved in deriving the scalar implicature (Pouscoulous et al., 2007). Thus, children have been found to compute scalar implicatures linked to *some* from 4 years onwards but not younger (Pouscoulous et al. 2007; Katsos and Bishop 2011; see Stiller et al., 2014, for comprehension of non-lexicalised scalar implicatures in 3-year-olds).

There is therefore still a gap between the moment children produce and understand *some* and the point where they have been shown to derive its upper-bounded reading in an experimental context. Four main accounts of this phenomenon have been put forward. According to Katsos and Bishop (2011), young children understand the scalar implicature linked to *some*, but they are pragmatically more tolerant than adults. This leads them to accept utterances with *some* in contexts where *all* would be more appropriate even though they perceive the term as under-informative. Skordos and Papafragou (2016) on the other hand, emphasise the importance of conversational relevance in accessing the stronger alternative (*all*), and thus deriving the scalar implicature. Specifically, they maintain that children's ability to consider the stronger alternative depends fundamentally on how relevant this alternative is in context. When the lexical alternative is explicitly present or when it is simply contextually relevant, children consider it and infer the scalar implicature. A third strand has argued that the processing cost of implicatures is too high for young children; while they have the ability to understand scalar implicatures, they often lack the resources to make a relatively effortful inference (Reinhart, 2004; Pouscoulous et al., 2007). Indeed, evidence suggests that even for adults, scalar implicatures can be cognitively taxing (Noveck and Posada, 2003; Bott and Noveck, 2004; Breheny et al., 2006; De Neys and Schaeken, 2007). Finally, lexicalist accounts claim that while young children know the meaning of quantifiers such as *some* and *all*, they have not yet acquired the overarching informativeness scale. This prevents them from comparing *some* to *all*, and thus, from deriving the scalar

implicature (Barner et al., 2010, 2011; Hochstein et al., 2014). It is worth noting that these accounts are not necessarily mutually exclusive. The first three, in particular, are sometimes presented by their supporters as potentially complementary (Katsos, 2014; Papafragou and Skordos, 2016). The debate to establish the best account of children's early difficulties with scalar implicatures is still very much raging. Yet despite our knowledge of implicature acquisition being largely based on children's understanding of *some*, we know very little about its production by children—and only slightly more for adults.

A single study has looked at scalar implicature production in children. Katsos and Smith (2010) investigated how 7-year-olds fare with scalar implicatures from a speaker's as well as a hearer's perspective. In addition to a usual binary truth value judgment task, children were asked to provide descriptions themselves. While the 7-year-olds' performance on the sentence verification task resembles what was found in other studies, they produced informative sentences at very high rate. These findings could be taken to point toward a speaker/comprehender asymmetry—where children find production easier than comprehension—as is sometimes alluded to for other pragmatic phenomena (e.g., informativeness, Davies and Katsos 2010, and presuppositions, Berger and Höhle, 2012). Importantly, the authors do not attribute this apparent comprehension-production asymmetry to a lack of pragmatic competence, but to a different metalinguistic attitude in children when they have to judge utterances.

The ideal way to investigate the production of *some* is to study corpora of real use in addition to experimental methods. Three corpus studies have looked at adult production of *some*. The first is a small scale study in Huang and Snedeker (2009b), where they extracted 50 random instances of *some* from the British National Corpus and analysed them depending on whether they referred to a subset or not. More convincingly, Degen (2015) extracted 1748 occurrences of *some*-NPs from a telephone dialogue corpus. She excluded 359 *some*-NPs headed by singular count nouns and 26 cases where the NP consisted only of *some*. The remaining 1363 *some* instances were used in a web-based study. Participants recruited on Amazons Mechanical Turk were asked to judge the probability of an implicature being intended by assessing the similarity on a 7-point-Likert scale between the original *some* utterance and an "implicature paraphrase" resulting from inserting *but not all* after *some*—e.g., "I like to read some of the philosophy stuff" and "I like to read some, but not all, of the philosophy stuff." Sun (2017) uses a very similar procedure to get implicature plausibility rates for several triggers extracted from twitter, including 200 instances of *some*. These studies were designed to test what Degen calls the "Frequency Assumption"; an implicit assumption found in much of the theoretical and empirical literature on scalars that lexicalised scalar terms, such as *some*, will more often than not give rise to implicatures. The findings show that the upper-bound reading of *some* is found in naturally occurring speech, but is not prevalent; a conclusion with important (negative) consequences for theories relying on a dominant upper-bound interpretation of scalar terms, such as the defaultism of Levinson (2000) or syntax-based approaches (Chierchia, 2006; Chierchia et al., 2012). These results also have implications for children's acquisition.

Indeed, a low implicature rate in adult speech might account, in part, for their difficulties with the lower-bound interpretation of *some*.

At this juncture of our understanding of scalar implicature and its development, a study of naturalistic child and parent production seems essential. Such data are very difficult to get in experimental settings, particularly for children, and a child corpus analysis seems a more convincing way forward. Yet, while focusing on a corpus reflecting children's natural spontaneous speech, as well as their environment, comes with a host of advantages, it brings its own issues, too. How are we to assess the speaker's intention to produce an implicature? Degen (2015) solves this impasse by postulating that in communication, hearer's recognition of speaker's intention is, overall, a fair approximation of the speaker's intention: the audience's intuitions about implicatures correspond by and large to the speaker's intention to produce them. Unfortunately, when looking at younger children's production we cannot rely on implicature plausibility ratings from untrained Mechanical Turk participants. But, we can code for the plausibility of an implicature being intended by the use of *some*, based on the context of utterance and tests such as whether it refers to a subset (Huang and Snedeker, 2009b) or the *not all* paraphrase (Degen, 2015).

In the following, we therefore present a corpus study on young children's production of *some*, adding a missing piece to the current literature and our understanding of early pragmatic abilities. Children's utterances containing *some* were extracted from dense corpora of five children aged 2;00 to 5;01 ( $N = 5,276$ ), and analysed alongside an equivalent portion of their mothers' utterances with *some*. These were coded into structural and contextual categories allowing for judgments on the probability of a scalar implicature being intended (coding scheme partly based on Degen, 2015).

## 2. DATA AND METHODS

### 2.1. The Corpus

We looked at the production of *some* in dense corpora of five British English speaking children aged 2;00 to 5;01. Three sets (Thomas, Fraser and Eleanor) are part of the CHILDES database (MacWhinney, 2000; Lieven et al., 2009), while two (Gina and Helen) were accessed with the kind permission of the Child Study Centre, University of Manchester (De Ruiter et al., 2017). All families were from the Greater Manchester area in the United Kingdom. For each child, the corpus included dense recordings of 5 hours per week for the first 6 weeks following each of their birthdays, as well as 5 hours within one week during each of the subsequent months of the year. The interactions between children and their parents (mostly their mothers, a father appears once) took place at home usually during play, reading, or snack time. The children were recorded from 2;00 to 3;01 years for Eleanor and Fraser, from 2;00 to 4;11 years for Thomas, and from 3;00 to 4;07 years for Gina and Helen.

### 2.2. Coding

Children's utterances containing *some* were extracted with three lines of context before and after each *some* occurrence ( $N = 5,276$ ). For each child, data were organised into age windows of 3 months allowing for an analysis of individual developmental trajectories. To examine inputs in the early years, we extracted the mothers' first sentences with *some* in a number equivalent to their child's production ( $N = 5,430$ ). To further investigate input development, we extracted another 300 *some* utterances produced by each of the mothers after their child's birthdays ( $N = 3,556$ ; Total number of utterances coded for mothers = 9,030). For one mother, the recording stopped after 256 utterances after the child's last birthday, meaning that 300 utterances could not be reached. All 14,306 utterances were categorised following structural and contextual categories allowing for judgments on the probability of a scalar implicature.

All utterances were first coded following a structural grid, according to the type of syntactic structure the word *some* appeared in. Eleven structural categories were established: Seven were marked as *Included* and four as *Excluded*. Utterances falling under the *Included* categories were subsequently coded according to the contextual coding scheme while utterances falling within the *Excluded* categories could not be coded further due to missing or incomplete information (e.g., errors, ambiguities). In a second phase, the *Included* cases were coded according to their likelihood of carrying an implicature from *some* to *not all*. Four contextual categories were devised to reflect judgment on the probability of an implicature being intended: *Implicature Impossible*, *Implicature Implausible*, *Implicature Possible*, and *Implicature Plausible*. The coding scheme was adapted in part from Degen (2015), and was used equally for children and adult uses of *some*. The data and coding of the corpus reported in this paper are accessible to readers on the Open Science Framework database at [osf.io/g6psr](https://osf.io/g6psr).

#### 2.2.1. Structural Categories

All the extracted *some* utterances were coded as belonging to one of the mutually exclusive, structural categories outlined in Table 1. There are seven *Included* categories.

1. In the *Mass* category, *some* precedes a mass noun including object mass nouns (e.g., coffee and furniture).
2. The *Count as mass* category includes count nouns that appear in a mass noun-like structure (e.g., *Want some banana*).
3. The *Adjective* category includes *some* utterances headed by an adjectival noun (often colours, e.g., *Need some blue*).
4. Similarly, in the *Plural noun* category the phrase is headed by a plural noun (e.g., *some people*).
5. The category *Singular NP* covers utterances with a singular count noun. Although the structure is similar to the *Count as mass* category they differ in the quantity of the referent; in the *Singular NP* cases it only refers to one single entity and not to a mass (cf. "Some guy predicted the end of the world today," Degen, 2015, p. 5, Ex. 12).
6. The *Plural NP* category includes cases where *some* is followed by a count noun in its plural form (e.g., *I need some blocks*).

**TABLE 1 |** Structural categories, their definition, and examples.

	Category	Structure	Example
Included	Mass	mass NP object mass NP	Mummy want <i>some</i> tea. (E., 2;00) Get <i>some</i> fruit from there. (E., 2;11)
	Count as mass	sg count NP for quantity	I like <i>some</i> banana. (E., 2;00)
	Adjective	adjectival NP	I need <i>some</i> yellow. (E., 2;00)
	Plural noun	pl NP for pl quantity	<i>Some</i> people love Peppa Pig. (H., 3;00)
	Singular NP	sg count NP	<i>Some</i> little boy kissed a chair. (H., 4;01)
	Plural NP	pl count NP	I want <i>some</i> dinosaurs. (E., 2;01)
	Of XP	partitive preposition	Mum keeps <i>some</i> of these balls. (E., 3;01)
Excluded	Solitary <i>some</i>	no spelled-out NP	Po like <i>some</i> . (E., 2;00)
	More	might mean <i>more</i>	I need <i>some more</i> . (E., 2;00)
	Structure unclear	pl NP for sg quantity conjunctive NPs	Need <i>some</i> scissors. (E., 2;00) I've got <i>some</i> fish and chips cook. (E., 2;08)
	Transcription	<i>some</i> replaced	I've got <i>some</i> ja triangle. (E., 2;00)
	unclear	incomplete phrase transcription failure	Let's play <i>some</i> +... [+ IN] (E., 2;03) Mummy, let's go <i>some</i> paint xxx. (E., 2;00)
		unclear utterance	I can do <i>some</i> [=? the] shopping. (E., 2;05)

sg, singular; pl, plural; NP, Noun Phrase (fully compatible with DP analysis).

7. Finally, the *Of XP* category covers prepositional phrases (e.g., I need *some* of these toys).

There are four *Excluded* categories. Utterances falling in one of these categories were not analysed further.

- In the *Solitary some* category, *some* is not followed by a noun.
- The *More* category, includes utterances with *some more*. These seem to mean *more* in the context of language acquisition as children are often asking for *some more* of food for example. Although it could be argued that *more* is used here as a modification, an implicature is implausible in most such cases.
- In the category *Structure unclear*, two different types of uses are pooled.
  - Plural nouns such as scissors and pants were excluded, because it could not be established whether the noun refers to a single quantity or to a mass.
  - Some* introducing *conjunctive* phrases were also excluded due to the structural ambiguity. Indeed, it could not be established whether *some* should be linked to the first conjunct or the whole conjunctive phrase.
- The category *Transcription unclear* also includes several cases.
  - When the sentence includes the word *some*, but is continued with a replacement, the word *some* is not used to quantify anymore (e.g., I want *some*, a bread).
  - Incomplete phrases were excluded when the referent for *some* was missing (e.g., I want *some* +IN). When the referent for *some* was uttered in the next line of the transcription, the utterance was included since the referent of the *some* phrase was readily available (e.g., "I want *some* +IN. *some* grapes").

- (c) Partly unintelligible sentences (transcribed with xxx) were also excluded.

- (d) When the transcription left a doubt about *some* being uttered, the utterance was excluded as well.

All occurrences of *some* were also independently coded according to additional, non-mutually exclusive, structural categories which impact discourse accessibility and therefore the likelihood of an upper-bounded reading of *some*. In doing so, we followed the approach of Degen (2015), and collected data which could inform how structural linguistic elements may influence implicature probability. These categories also provide further dimensions on which to compare child and adult production. For example, it has been argued that the subject position tends to support implicature interpretation (Degen, 2015). Breheny et al. (2006) suggested that a scalar implicature is more likely when in focus as focus highlights relevant content. This would then underline the contrast between *some* and *all*. The same holds for phrases that are topicalised, as the topic position is often associated with focus which can support contrasting *some* with *not all* (e.g., Some of the grapes the girls ate). Third, we coded whether the phrase was modified. On the one hand, modification can increase the salience of a novel mention in a discourse (Degen, 2015). On the other hand, modification can also counteract implicature plausibility when a set (e.g., of blocks) is then already subsetted (e.g., blue) which reduces the salience of *some* (e.g., I need *some* blue blocks). Forth, we coded whether *Of XP* phrases were headed by a pronoun or demonstrative. As Degen (2015) notes, pronoun and demonstrative phrases with *some* are ungrammatical when used without the partitive (Example 39 on p. 22: "And *some* \*(of) them fizzled out," Degen, 2015). Nonetheless, in her study, sentences with and without pronouns or demonstratives receive similarly high ratings.



**TABLE 2 |** Contextual categories indicating implicature plausibility.

Category	Description	Examples
Implicature impossible	No available set	I did <i>some</i> trumps. (E., 2;00) Blowing <i>some</i> bubbles. (T., 3;01)
Implicature implausible	Set possible but not referred to	Squirrel wants <i>some</i> nuts. (E., 2;00)
Implicature possible	Maybe referring to subset of set	Po's got <i>some</i> biscuits in his house. (E., 2;00)
Implicature plausible	Referring to subset of present set	I lost <i>some</i> pieces. (F., 3;00)

### 2.2.2. Contextual Categories

Utterances falling in one of the *Included* categories (see **Table 1**) were then assigned to one of four, mutually exclusive, contextual categories, which reflect their likelihood to carry an implicature based on structure and the extracted context ( $\pm 3$  utterances): *Implicature Impossible*, *Implicature Implausible*, *Implicature Possible*, and *Implicature Plausible* (see **Table 2**).

1. For utterances categorised as *Implicature Impossible*, no quantifiable set could be identified of which *some* could have been a subset. With no clear set in the discourse, the speaker cannot intend to refer to a subpart through a scalar implicature ("I need *some* help"). For instance, this category includes cases of spontaneously occurring natural phenomena (like trumps or clouds).
2. In utterances categorised as *Implicature Implausible*, a quantifiable set could be found, but the speaker was unlikely to be referring to it in this context. For instance, it would be possible in some contexts to use the sentence "We need to buy some batteries" to refer to a subset of batteries. Yet, in the corpus, the context suggested a more general meaning of getting batteries.
3. In the occurrences categorised as *Implicature Possible*, a quantifiable set could be identified and it was possible that the speaker was using the quantifier *some* to refer to a subset via an implicature, for instance in "I ate some biscuits". Yet, the available context does not provide sufficient elements to disambiguate between the two readings roughly paraphrased as "I ate biscuits" and "I ate some, but not all biscuits".
4. Finally, *Implicature Plausible* utterances involved a clearly identifiable set, which was relevant to the conversational exchange and to a subset the speaker seemed to be referring to. Thus, the speaker seemed to have used *some* intending the hearer to derive the scalar implicature and understand *not all*. For instance, when in the context of playing with jigsaw puzzles, a child utters "The puzzle is missing some pieces." Even in such cases, there can be no guarantee that the speaker intended to convey an implicature, rather we establish that the utterance is highly compatible with an implicature interpretation.

As mentioned in the introduction, we had to assess the likelihood of the speaker intending to convey an implicature

based on the hearer's understanding of this intention—more specifically, we have to rely on the coder's pragmatic inferences. Therefore, to avoid false positives and inflating the proportion of intended implicatures, the less implicature-compatible category was chosen when in doubt about the most appropriate contextual category for an utterance.

To correctly categorise all phrases, certain tests were applied. As seen above, to establish implicature plausibility, Degen (2015) used similarity ratings with paraphrases where *some* was replaced by *some but not all* (e.g., "I ate some biscuits" and "I ate some, but not all, biscuits"). We used the paraphrase test as a guideline: high similarity would correspond to a categorisation as *Implicature Possible* or even *Implicature Plausible* when the context strongly supported an implicature reading.

However, note that *all* is not necessarily the upper bound in all discourses as it can also be interpreted differently in certain pragmatic contexts. For example, when *all* is used to exaggerate, it can actually mean *some* or *most* (e.g., "She ate *all* the biscuits!" when meaning that this person did not leave many biscuits for the rest of the group. See also section 4).

Another paraphrase test we used as a guideline was the omission of the quantifier. When *some* can be left out [as in "I need (some) help" or "We need to buy (some) batteries"], it seems to be used as an indefinite marker and the occurrence would be categorised as *Implicature Impossible* or *Implicature Implausible*. To decide between these two categories, the content was taken into account. When no set could be defined (as in "help"), then the *Implicature Impossible* category was chosen. When a set could be identified, but was either non-quantificational or not the topic of discussion (e.g., an existing set of batteries in the store, but not relevant to the dialog) the utterance would fall into the *Implicature Implausible* category.

Context remained crucial to judgments about categorisation. Take a child saying "I want to eat some grapes," for instance. It is possible that there is a set of grapes in the kitchen. In most cases, it would be unlikely that the child is referring to that set. The implicature would thus be deemed *Implausible*. On the other hand, if the mother just uttered "See, there are some grapes on the table, the rest is in the kitchen," now the context establishes clear, relevant subsets and the implicature of the mother's utterance seems *Plausible*. The same holds if the child said "I want to eat some grapes. The others are for you," thereby actively differentiating between subsets.

A second coder independently coded 1,730 out of the 14,306 utterances of the overall corpus data; roughly 20% of *Included* and 9.5% of *Excluded* utterances split proportionally across children and adults, which sums up to roughly 12% of the whole corpus. Interrater reliability for all utterances was at 85% indicating very high agreement overall (contextual categories: 81% and Cohens Kappa of 0.7; structural categories: 89% and Cohens Kappa of 0.87. Cohens Kappa was calculated using confusion matrices with the package caret in R; Kuhn, 2013, for the use of Cohen's Kappa to assess interrater reliability, see Landis and Koch, 1977; Viera and Garrett, 2005; Cameron-Faulkner et al., 2007; Spooren and Degand, 2010).

**TABLE 3 |** Results for the structural categories of the mothers' data.

	Category	N	%
Included	Mass	1,614	28.38
	Count as mass noun	480	8.44
	Adjective	23	0.40
	Plural noun	110	1.93
	Singular NP	63	1.11
	Plural NP	1,605	28.22
	Of X	277	4.87
Excluded	Solitary some	456	8.02
	More	689	12.12
	Structure unclear	156	2.74
	Transcription unclear	214	3.76

Percentages are in proportion to all 5,687 utterances. N of Included utterances: 4,172; N of Excluded utterances: 1,515.

**TABLE 4 |** Results for the contextual categories of the mothers' data.

Category	N	%
Implicature impossible	710	17.02
Implicature implausible	2,774	66.49
Implicature possible	420	10.07
Implicature plausible	268	6.42

Percentages are in proportion to all 4,172 Included utterances.

### 3. RESULTS

#### 3.1. Mothers' Usage

Categorisation of the 5,687 utterances coded for the mothers can be seen in the **Tables 3** and **4**. Note that the number of appearances deviates from the extracted utterances as *some* could appear more than once in a sentence.

Regarding the structural categories, the categories *Mass* and *Plural NP* dominated. Adjectival phrases were rare. Exclusion was highest for the *More* category.

The *some* phrase appeared rarely in subject position ( $N = 63$ , 1.15%), and was almost never topicalised ( $N = 3$ , 0.07%), and therefore mostly realised in object position. A small part of utterances was modified pre- or post-phrasal ( $N = 450$ , 10.79%). Around a quarter of all *Of XP* utterances were headed by a pronoun or a demonstrative ( $N = 75$ , 27.08%).

As in Degen (2015), structural properties seemed to relate to implicature plausibility as can be seen in **Figure 1**: While *some* in subject position supported an implicature reading (*Implicature Plausible* ratings), modifications were mostly found in the *Implicature Implausible* category.

In the contextual categories, *Implicature Plausible* utterances represented a small proportion of the *Included* set (6.42%), while most utterances were categorised as *Implicature Implausible* (66.49%).

Looking at the relation between structural and contextual categories we find more *Implicature Plausible* ratings in certain

**TABLE 5 |** Contextual categorisation of the individual *Included* structural categories of the mothers' data.

Category	Total N	Impossible		Implausible		Possible		Plausible	
		N	%	N	%	N	%	N	%
Mass	1,614	362	22.43	1,121	69.46	114	7.06	17	1.05
Count as mass noun	480	143	29.79	333	69.38	0	0	4	0.83
Adjective	23	0	0	23	100	0	0	0	0
Plural Noun	110	0	0	96	87.27	6	5.46	8	7.28
Singular NP	63	0	0	63	100	0	0	0	0
Plural NP	1605	205	12.77	1,138	70.9	218	13.58	44	2.74
Of XP	277	0	0	0	0	82	9.62	195	70.4

**TABLE 6 |** Results for the structural categories of the children's data.

	Category	N	%
Included	Mass	1,080	20.34
	Count as mass noun	279	5.25
	Some adjective	45	0.85
	Plural noun	75	1.41
	Singular NP	140	2.64
	Plural NP	1,078	20.3
Excluded	Of X	186	3.50
	Solitary some	754	14.20
	More	754	14.20
	Structure unclear	100	1.88
	Transcription unclear	819	15.42

Percentages are in proportion to all 5,310 utterances. N of Included utterances: 2,883; N of Excluded utterances: 2,427.

structural categories and close to none in others (see **Table 5**). For example, there were no *Implicature Plausible* cases amongst *Singular NP*. Cases of *Plural NP*, however, could belong to any of the four contextual categories. Furthermore, *Of XP* utterances were prone to be categorised as *Implicature Plausible* ( $N = 195$ , 70.4%). Thus, the partitive structure seems to support implicature interpretation. On the other hand, structures suggesting a singular quantity are difficult to combine with a partitive reading and are unlikely to give rise to an implicature reading. A structure such as the *Plural NP* category is more flexible; it allows for more variation in implicature readings, and its interpretation is therefore highly dependent on context.

#### 3.2. Children's Usage

**Table 6** provides the structural categorisation and **Table 7** the contextual categorisation for all 5,310 *some* utterances of the children. Again, the number of appearances deviates from the extracted utterances as *some* could appear more than once in a sentence.

Note that children, as their mothers, used *some* in several different structural forms and that, again as their mothers, there is a predominance of *Mass* and *Plural NP* usage. Adjectival phrases were rare for children, too. Exclusion was highest for the *Transcription unclear* category. Overall, more utterances had to

**TABLE 7 |** Results for the contextual categories of the children's data, indicating plausibility of implicatures for included utterances.

Category	N	%
Implicature impossible	282	9.78
Implicature implausible	2,040	70.76
Implicature possible	322	11.17
Implicature plausible	239	8.29

Percentages are in proportion to all 2,883 included utterances.

be excluded than in the mothers' data suggesting that the data of the children were noisier, as would be expected considering their age.

As for their mothers, the *some* phrase appeared rarely in subject position ( $N = 86$ , 2.98%), and was never topicalised, and therefore mostly realised in object position. A small part of utterances was modified pre- or post-phrasal ( $N = 213$ , 7.39%). More than half of all *Of XP* utterances were headed by a pronoun or a demonstrative ( $N = 113$ , 60.75%).

As in Degen (2015) and our adult data, structural properties seemed to relate to implicature plausibility as can be seen in **Figure 1**: While *some* in subject position supported an implicature reading (*Implicature Plausible* ratings), modifications were mostly found in the *Implicature Implausible* category.

Interestingly, in the children's contextual categorisation, implicature production can clearly be observed. A total of 19.46% of the *Included* cases were categorised as *Implicature Possible* or *Implicature Plausible*, despite the fact that the *Implicature Implausible* was still the most largely represented.

Here again, implicature plausibility diverged depending on the structural category as can be seen in **Table 8**. For example, *Singular NP* provided no *Implicature Plausible* cases, indicating that its structure is a cue against implicature plausibility as suggested by Degen (2015, p. 5). The *Plural NP* category however, provided utterances belonging to all four contextual categories. Thus, such a structure allows for more variation in implicature readings; whether it gives rise to an implicature interpretation or not is therefore highly dependent on context. As observed in the mothers' production, the *Of XP* category was prone to carry implicatures ( $N = 158$ , 84.95%). Therefore, the partitive structure supported implicature readings also in the children's data.

We also looked at children's individual production of *some* over time within the corpus to establish when different types of uses, as well as implicature production, first appear (see **Table 9**). The resulting developmental picture shows that children begin using *some* in its many forms during their third year of life. Importantly, this includes implicature production. Indeed, as can be seen in **Table 9**, the first *Implicature Plausible* instances of *some* produced by the three 2-year-olds appear 3 to 9 months after their first use of *some* in the corpus.

Altogether, the findings indicate children's competence regarding different types of *some* including pragmatic production. To see whether their behaviour mirrors the input provided by their mothers, we next turn to the comparison of these results with child-directed speech.

**TABLE 8 |** Contextual categorisation of the individual *Included* structural categories of the children's data.

Category	Total N	Impossible		Implausible		Possible		Plausible	
		N	%	N	%	N	%	N	%
Mass	1,080	135	12.5	829	76.76	103	9.54	13	1.2
Count as mass noun	279	54	19.36	225	80.65	0	0	0	0
Adjective	45	0	0	45	100	0	0	0	0
Plural Noun	75	0	0	30	40	17	22.67	28	37.34
Singular NP	140	0	0	140	100	0	0	0	0
Plural NP	1,078	93	8.63	769	71.34	176	16.33	40	3.71
Of XP	186	0	0	0	0	28	15.05	158	84.95

**TABLE 9 |** Overall data of the individual children.

Child	Recording	Total	Incl	Excl	1st some	1st Category	1st implicature
Eleanor	2:00 - 3:01	937	497	440	2:00:03	Mass	2:04:02
Fraser	2:00 - 3:01	627	359	268	2:00:28	Mass	2:03:06
Thomas	2:00 - 4:11	1770	906	864	2:00:13	Mass	2:09:11
Gina	3:00 - 4:07	971	504	467	3:00:01	Plural NP	3:00:04
Helen	3:00 - 5:01	1005	617	388	3:00:02	Plural NP	3:00:10

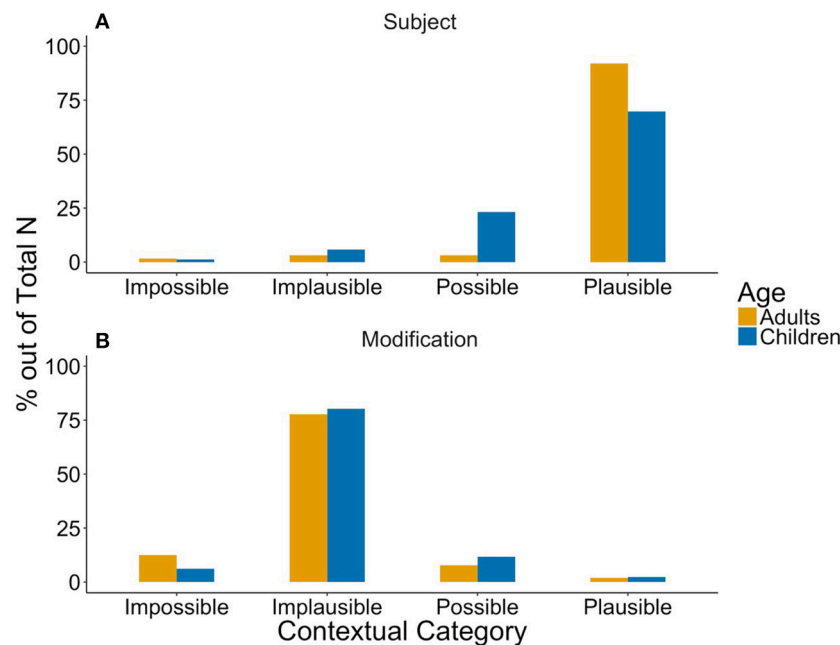
Incl, N of utterances in included categories; Excl, N of utterances in excluded categories.

### 3.3. Comparison of the Children and Their Mothers

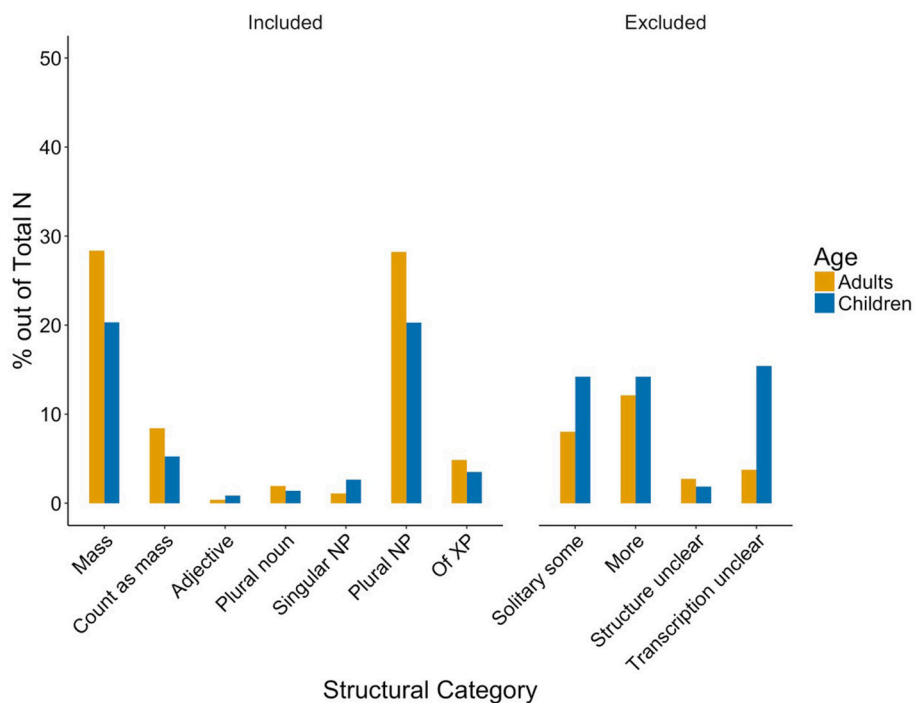
Children's production and mothers' child-directed speech did not differ significantly from each other in either structural or contextual categories (Mann Whitney *U*,  $ps > 0.1$ , Kilgariff, 2001), indicating similar usage patterns across groups (see **Figures 2 and 3**). Thus, implicature production was similarly low. Even when pooling *Implicature Possible* and *Implicature Plausible* utterances, only 16.49% of adults' and 19.46% children's uses of *some* in the *Included* categories potentially carry an implicature (cf. Degen, 2015, for similarly low rates).

Interestingly, mothers' usage of *some* changed as a function of the child's age. To analyse how the mothers' implicature production changes, we further coded roughly 300 utterances of the mother after each birthday of her child. To model the data, we fitted a generalized linear mixed model using *lme4s* *lmer* function (Bates et al., 2015) with Gaussian error structure and identity link function in R (R Core Team, 2016). Contextual Category and the child's age, and their interaction were included as fixed effects of interest. We also included Child as a random factor to allow for random slopes across participants. The number of utterances in each category at each age was transformed to percentages to standardize the dependent measure across mothers and time points. A reduced model was fit that did not include Contextual Category. A comparison between the reduced model and the full model then allows for conclusions about differential effects in the different contextual categories across the ages. Results can be seen in **Table 10 and Figure 4**.

Comparing the full with the reduced model revealed that Contextual Category significantly improved the model fit ( $\chi^2$



**FIGURE 1 |** Structural influences on the implicature plausibility of *some* in **(A)** subject position (Adult  $N = 63$ , Child  $N = 86$ ), and **(B)** *some* being modified (Adult  $N = 450$ , Child  $N = 214$ ) in caregivers' (yellow) and children's (blue) production.

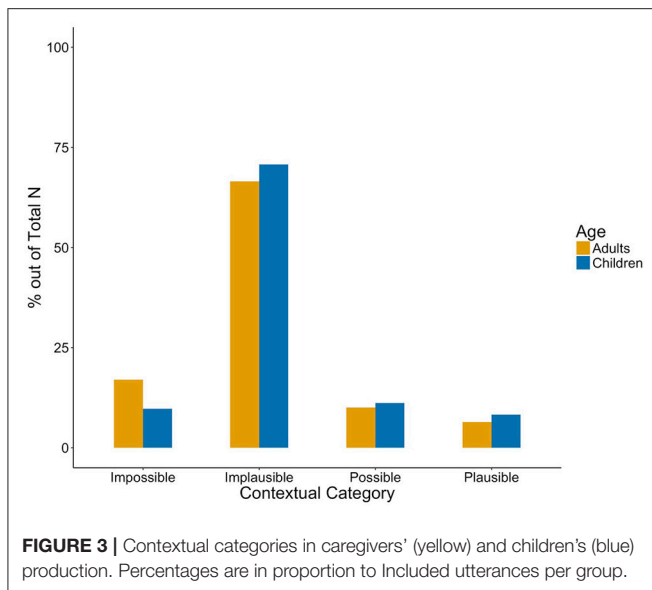


**FIGURE 2 |** Structural categories in caregivers' (yellow) and children's (blue) production. Percentages are in proportion to all utterances per group.

$= 150.47$ ,  $df = 6$ ,  $p < 0.001$ ). Using drop1, the model revealed a significant interaction of Contextual Category\*Age ( $\chi^2 = 39.22$ ,  $df = 3$ ,  $p < 0.001$ ), suggesting differences between

contextual categories at different ages. To analyse these effects further, we split the data according to the different contextual categories. In the model examining the data from the contextual





category Impossible alone (*Impossible* split model), there was no significant effect of age. For the *Implausible* split model, the effect of age was significant ( $\chi^2 = 6.09$ ,  $df = 1$ ,  $p = 0.014$ ). For the *Possible* split model, the effect of age only tended toward significance ( $\chi^2 = 3.4$ ,  $df = 1$ ,  $p = 0.065$ ). For the *Plausible* split model, the effect of age was significant ( $\chi^2 = 29.67$ ,  $df = 1$ ,  $p < 0.001$ ). Thus, with each birthday of the child, the mother's number of *Implicature Plausible* instances increased, and the number of *Implicature Implausible* ones decreased. However, neither *Implicature Impossible* nor *Implicature Possible* utterances changed significantly in number across the ages.

### 3.4. Further Observations

Before we turn to the possible implications of these results, we would like to present a few additional qualitative observations. These are potentially very interesting and would deserve a systematic investigation that goes beyond the scope of the current study. First, we highlight some cases where children contrast directly the quantifier *some* with other relevant quantifiers. Second, we discuss how modification and *some* in subject position might interact with each other. Finally, we present a few cases where children's utterances were erroneous.

In order to assess how competent young children are with scalar implicatures linked to *some* it is worth looking at whether they spontaneously contrast *some* with other quantifiers on the same semantic "scale." We found some cases in the corpus where children contrast *some* directly either with *all* or with other quantifiers. Below are four such examples from Thomas and Fraser between 2;02 and 3;08. Further examples can be found in the **Supplementary Material**.

- (1) Contrasting *some* with *all* (Fraser, 3;00)  
 \*FAT: Put all these pieces away.  
 CHI: You don't put all of them away.  
 FAT: Why?  
 CHI: Just do [/] just do (.) *some* at the time.

CHI: Not all of them.  
 FAT: Not all of them?  
 CHI: No.

- (2) Contrasting *some* with *all* (Mum of Fraser, 2;02)  
 \*MOT: you dropped *some* pennies.  
 CHI: *all* my pennies.
- (3) Contrasting *some* with *lots* (Thomas, 3;08)  
 \*CHI: just put some things back in the box.  
 INV: do you want to put them back in the box?  
 CHI: yes.  
 CHI: *some of them* but not *lots of them*.  
 INV: okay.  
 INV: are you keeping [/] are you keeping everything tidy?  
 INV: yeah?
- (4) Contrasting *some* with *some* (Fraser, 3;00)  
 \*FAT: Yeah.  
 CHI: But some girls don't.  
 FAT: No.  
 CHI: But some girls do.  
 FAT: \*chuckles\* Some boys don't like milk either.  
 CHI: But, but...  
 FAT: Makes them poorly.

Moreover, we observe a structural hierarchy of implicature plausibility. Indeed, in both adults and children *some* in subject position supported implicature readings, while modifications hindered implicature readings. While these general lines are very clear, the combination of different factors results in a more complex picture. The combination of modification and the *some* phrase in subject position reduces the implicature likelihood (e.g., "Some blue blocks are missing" in the context of many blocks. See also Example 6). In contrast, the combination of modification and the *some of* partitive phrase in subject position makes implicature readings more likely again (e.g., "Some of the blue blocks are missing" in the context of many blocks. See also Example 7). *Some of* highlights the partitive interpretation (84.95% in children), and thus, this structure might serve as a cue to implicature interpretation that outweighs modification; even though the phrase *some of* is not sufficient for an implicature interpretation in and of itself (See Example 8, and see also Degen and Tanenhaus, 2011). Of course, these observations are based on few utterances and more detailed exploration is needed.

- (5) Modification (Helen, 4;00)  
 \*MOT: that's to put that plant in, isn't it?  
 CHI: oh yeah.  
 CHI: but the plant comes out.  
 CHI: I've got *some* even better funny ones.
- (6) Modification in subject position hindering implicature (Helen, 4;08)  
 \*CHI: and *some new people* are coming.  
 MOT: are they?  
 CHI: yeah some new school children that go to

**TABLE 10 |** Generalized Linear Mixed Model testing the relative change in the frequency of utterances of the mothers across the childrens ages in the contextual categories *Impossible*, *Implausible*, *Possible*, and *Plausible*. res = lmer(Utterances ~ Category\*Age + (1 + Age | Child); data = d2; REML = F; control = contr).

		Estimates	SE	Lower CL	Upper CL	$\chi^2$	p
Full model <sup>(1)</sup>	(Intercept)	0.09	0.05	-0.01	0.18	(3)	(3)
	Cat: Implausible	0.83	0.07	0.69	0.96	(3)	(3)
	Cat: Possible	-0.03	0.07	-0.16	0.12	(3)	(3)
	Cat: Plausible	-0.15	0.07	-0.28	-0.01	(3)	(3)
	Age	0.03	0.01	-0.00	0.06	(3)	(3)
	Cat: Implausible:Age	-0.12	0.02	-0.16	-0.08	(3)	(3)
	Cat: Possible:Age	-0.01	0.02	-0.05	0.03	(3)	(3)
	Cat: Plausible:Age	0.02	0.02	-0.02	0.07	(3)	(3)
Impossible <sup>(2)</sup>	(Intercept)	0.17	0.04	0.09	0.25	(3)	(3)
	Age	-0.00	0.02	-0.04	0.03	0.02	0.89
Implausible <sup>(2)</sup>	(Intercept)	0.87	0.07	0.73	1.00	(3)	(3)
	Age	-0.08	0.02	-0.12	-0.03	6.09	0.01
Possible <sup>(2)</sup>	(Intercept)	0.05	0.03	-0.02	0.13	(3)	(3)
	Age	0.02	0.01	-0.00	0.04	3.40	0.07
Plausible <sup>(2)</sup>	(Intercept)	-0.06	0.02	-0.09	-0.03	(3)	(3)
	Age	0.05	0.00	0.04	0.06	29.67	<0.001

<sup>(1)</sup> df = 3.<sup>(2)</sup> df = 1.<sup>(3)</sup> Not shown because of having a very limited interpretation as this value is only in relation to the reference level.

Wwww\_Mwww [% school].

MOT: right.

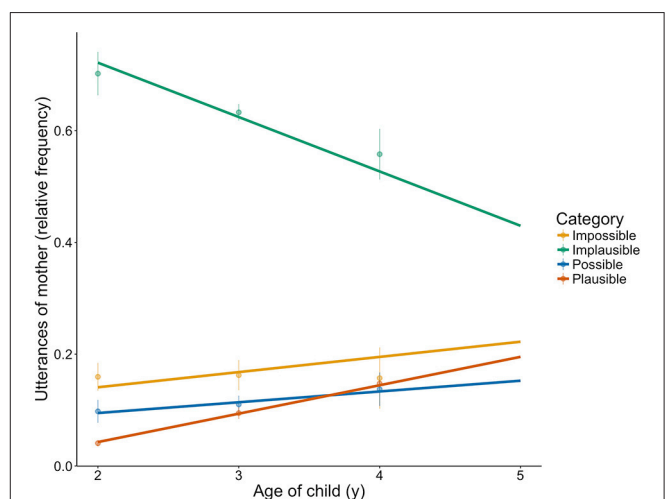
- (7) *Some of* in subject position overriding modification (Mother of Thomas, 2;03)  
 \*MOT: *some of* the little bubble bath tab eh [/] bubble tabs that we've bought haven't been very good, but this one is special. Teletubby double bubble it's called.

- (8) Uncertain *some of* case (Gina, 4;01)  
 \*CHI: I wanna touch *some of* this.  
 CHI: I wanna touch someone with this.  
 CHI: I wanna touch some of this.  
 MOT: no it's bacon.

A final observation concerns the type of errors children produced. For all children, the category *Singular NP* seemed to be used erroneously: they used *some* as a determiner with count nouns (e.g., *some garden*). This resembles a mass noun construction, but would usually be expressed with a simple determiner such as *a*, as we can see in Example 9. This might indicate an overgeneralisation of the frequent *count as mass noun* pattern.

- (9) Erroneous *Singular NP* utterance (Eleanor, 2;04)  
 \*CHI: I've got *some garden*.  
 \*MOT: you've got a garden?  
 CHI: yeah.  
 MOT: I like gardens.

Another type of mistake was the production of multiple quantifiers in a row, such as in Example 10.

**FIGURE 4 |** Relative change in the frequency of utterances of the mothers across the childrens age span in the contextual categories *Impossible*, *Implausible*, *Possible*, and *Plausible*. The lines reflect the fitted model of the GLMM including Contextual Category and Age, as well as their interaction. res = lmer(Utterances ~ Category\*Age + (1 + Age | Child); data = d2; REML = F; control = contr).

- (10) Several quantifiers (Thomas, 3;01)  
 \*CHI: I want that's lots *some* few things here.  
 MOT: oh alright.  
 MOT: you want to look at those books up there?

These cases are mainly present around age 3, when children seem to have acquired the basics of the adult system (Lieven and Behrens, 2012). This pattern of error is particularly

interesting and could enlighten our understanding of the development of language structure. In particular, a closer look to these cases could have an impact on syntax-based approaches to scalars (e.g., Chierchia, 2004, 2006), which we discuss briefly below.

## 4. DISCUSSION

In this study we investigated young children's implicature production by looking at the production of *some* in five young children and their mothers. Overall, 14,306 utterances containing *some* were extracted from dense corpora of five British English children aged 2;00 to 5;01 ( $N = 5,276$ ) and alongside that, an equivalent portion for their parents was analysed ( $N = 9,030$ ). All instances of *some* were categorised according to mutually exclusive structural and contextual categories. Structural categories were based on syntactic form while contextual categories considered the contextual environment of the utterance and allowed for judgments on the probability of a scalar implicature being intended.

Analysis of the parents' production revealed that few uses of *some* could be meant to carry an implicature. Our highest implicature plausibility category (*Implicature Plausible*) represents 6.42% of the adult data (8.29% of the children data). A generous approximation of potential intended implicatures pooling together the *Implicature Possible* and the *Implicature Plausible* categories gathers 16.5% of the adult *some* cases (19.5% of the children's). Importantly, the adult results also imply that children are rarely confronted with upper-bound *some*.

Interestingly, the parents' implicature production increased as a function of the children's age. We note an increase of the *Implicature Plausible* cases and a decrease of the *Implicature Implausible* instances over the years. This might be due in part to the large number of *Implicature Implausible* utterances related to food (i.e., "Want some banana"), while *Implicature Plausible* cases highlight a contrast. The change, then, might be brought about by conversations evolving from a focus on more basic desires, such as nutrition, to more complex arguments about variations in the world ("Some girls have brown hair"). While this aspect of our findings would need to be investigated further in future research, the changes in parents' production suggest an evolving learning environment for the child.

The low frequency of implicatures in child-directed speech corroborates the findings of Degen (2015) and Sun (2017) in other adult corpora. Unfortunately, because of differences in methodology our data are not directly comparable to theirs. Degen and Sun both relied on on-line participants ratings on a seven point Likert scale to assess the likelihood of an implicature being intended, while we assessed implicature plausibility according to coding on a four categories scheme performed by one or two coders. The proportion of combined *Implicature Possible* and *Implicature Plausible* cases we find (16.5%, for two out of four categories) is lower than that of the ratings higher than midpoint in either studies (44.7% for Degen and 64% for Sun). Note that Sun's is already higher than Degen's and that the short study by Huang and Snedeker

(2009b) reports that a relatively high 42% *some* occurrences "unambiguously referred to a subset" (Huang and Snedeker, 2009b, p. 410). It is unclear that looking at midpoint ratings is the best way to compare these different data sets. For instance, Degen finds that only 14.7% (Degen, 2015, p. 12) of her data corresponds to the highest ratings while, under what she considers to be the best analysis of the components, 28% are generated by an upper-bound interpretation (Degen, 2015, p. 16). Yet, even from this angle, our data seem to foster less upper-bound *some* instances than these other studies.

The discrepancy in the various findings might stem from two sources: differences in the nature of the corpora, on the one hand, and differences in the way implicature plausibility was established on the other. First, corpora varied greatly in kind and in size: we coded 4,172 *included some* instances taken from child-directed speech in every day activities, while Huang and Snedeker (2009b) looked at 50 occurrences of *some* from the British National Corpus, Degen (2015) analysed 1,748 from telephone dialogues and 200 cases taken from tweets were rated in Sun (2017). This diversity might influence *some* distributions. For instance, Sun (2017, p. 80) notes that a higher percentage of partitive *some* in her corpus might, in part, explain why she finds higher implicature plausibility ratings than Degen (2015). It is also possible that parents addressing a young child intend less upper-bound readings of *some*. Such an interpretation fits well with our finding that parent *Implicature Plausible* instances increase as their children grow. We found many utterances of the "want some grapes" type in child-parent interactions; probably substantially more than we would in adult conversation. Yet, without further evidence, this conclusion is premature since several other parameters might explain a somewhat lower frequency of upper-bound *some* cases in our data.

Second, diverging findings might come down to differences in data collection (rating vs. coding), implicature assessment tests (existence of a subset vs. *not all* paraphrases) or exclusion criteria for irrelevant cases. For example, Sun (2017) filtered out occurrences falling under the scope of negation, in questions or conditional antecedents, and Degen (2015) took out singular *some* cases, while we did neither. The crucial parameter in explaining the difference between our results and those of Degen and Sun is probably how implicature plausibility was coded for. Indeed, untrained Mechanical-Turkers are likely to be more lenient in their assessment than linguistically trained coders instructed to be conservative when granting implicature plausibility (to prevent overestimating implicature production in toddlers). Importantly, discussion about differences in findings and methods of assessments should not distract us from the striking convergence of all available adult corpus studies on a low proportion of upper-bound interpretation for *some*.

The relatively low frequency of adult implicature production found in all four corpora clearly speaks against what Degen (2015) coined the Frequency Assumption. No matter how one looks at the data it is impossible to claim that the predominant reading of *some* is prone to implicature. This important, and now

robust, finding is difficult to reconcile with theories assuming that *some* commonly induces implicatures, such as syntactic accounts (e.g., Chierchia et al., 2012) or Horn's (1984, 1989) Generalised Conversational Implicature thesis and Levinson's default theory (Levinson, 2000), which maintains that *some* will give rise to a scalar implicature by default, unless the context blocks the inference. Additionally, as Degen (2015) argues, the low frequency of *some*-related implicatures in corpus research also has consequences for the so-called Literal-First-Hypothesis (Huang and Snedeker, 2009a). According to this thesis, the interpretation of upper-bound *some* follows a two-stage processing model where it always appears with a delay, after the lower-bound reading has been computed. This hypothesis is not directly contradicted by the low frequency of upper-bound readings in corpora, but it makes it more difficult to test. Indeed, while several researchers have shown that deriving a scalar implicature linked to *some* comes at a cognitive cost and is processed slower (Breheny et al., 2006; De Neys and Schaeken, 2007; Huang and Snedeker, 2009a, 2011; Degen and Tanenhaus, 2011; Bergen and Grodner, 2012; Bott et al., 2012), this could be due to the low frequency of the reading rather than to a two-stage processing. After all, as Degen (2015) points out, frequency is a well-established factor in psycholinguistics and there is no reason to assume it would not influence pragmatic aspects, too.

Interestingly, structural elements influence implicature plausibility both in the production by parents and children. Here, too, our data corroborates the work of Degen (2015). For instance, *some* in the subject position increases the likelihood of an implicature (as in Degen, 2015, p. 28). In contrast, modification reduces implicature likelihood; although, this finding is not as pronounced in Degen's analysis (Degen, 2015, p. 29). Additionally, some structural categories seemed related to implicature plausibility. *Singular Some* cases, for instance, did not include any *Implicature Plausible* cases and indeed they were part of the excluded categories in Degen (2015). On the other hand, the majority of *Some Of* cases did support an implicature reading; as was found both by Degen (2015, p. 23) and Sun (2017, p. 80). While partitive *some* does not always promote an implicature, it often does and more often so than non-partitive *some* (see also Degen and Tanenhaus, 2011).

It is worth noting that we found a high proportion of *some* uses in constructions typical of English (as opposed to many other Indo-European languages) and where *some* cannot necessarily be linked to implicature production. Specifically, the determiner *some* is frequent in English (e.g., "I need some batteries"). While it might be meant to carry an implicature when a set of batteries is present, it can also be a simple determiner phrase when no set is referenced (28.22% of all utterances in adults and 20.41% of all utterances in children, see also Bagassi et al., 2009; Degen, 2015, for a thorough discussion). This reading is widespread in English, but would be conveyed without recourse to the quantifier *some* in other languages (see **Supplementary Material** for examples and their translations). Such instances were categorised as *Implicature Implausible* and might induce a lower rate of implicature plausibility than in other

languages. In her work, Degen (2015) concludes that implicatures are highly dependent on syntactic, semantic, and pragmatic influences from the context and appear to be probabilistic in nature—i.e., rather than being an all-or-nothing phenomenon it makes sense to ask to what degree they arise (see also Degen and Tanenhaus, 2015). Our results support her argument: in the present study, implicature-compatible utterances in both child-directed speech and children's production are low in frequency, but seem dependent on syntactic and contextual information.

The most surprising aspect of the data, of course, is that children produce *Implicature Plausible* instances of *some* very early on and at rates matching those of their parents. The children's production of *some* mirrors that of their parents' in all aspects. Although, this has also been found for other structural phenomena in language (e.g., Kidd et al., 2007), the degree of resemblance between adult and child production both in the structural and in the contextual categories is remarkable (see **Figures 2** and **3**). The overall pattern of the findings suggests children master the use of *some* early on with a distribution of *some* mimicking child-directed speech. This is what one would expect considering work on frequency matching between parents' and children's speech (Ambridge et al., 2015). It seems natural that the children use *some* highly frequently in non-implicature, more low scope formulaic utterances such as "I want some banana," since parents use these constructions very frequently. The real surprise, then, is that children produce scalar implicatures, which are regarded as a complex pragmatic inference, so early. Although parents' production suggests children are rarely confronted with instances of *some* meant to carry implicatures, utterances favouring a lower-bound interpretation nonetheless appear in their third year of life (or were present as soon as the recording started), shortly after their first production of *some* (Eleanor 2;04;02, Fraser 2;03;06, Thomas 2;09;11, Gina 3;00;04, Helen 3;00;10; see **Table 2**). As for their parents, *some* is produced in many different syntactic structures; implicatures appear to be rare and dependent on linguistic structure and context. Nevertheless, almost as soon as they acquire *some*, we see the children producing it competently, including upper-bound uses.

How can we account for such an early production of implicatures? There is ample evidence that children calculate intentions in communicative contexts even preverbally (e.g., Tomasello, 2008). Indeed, much work, in language acquisition also suggests that they could not learn to speak without impressive pragmatic abilities (e.g., Bloom, 2000; Tomasello, 2003; Clark, 2016). Once they have figured out the semantics of *some*, children might therefore be able to work out how to produce the implicature. An additional element is necessary, of course, the understanding that *some* might be on a semantic scale with other quantifiers (*all*, *many*, *most*), or at least that its meaning can contrast with theirs. Examples (1)–(3) above indicate they do so early on. Yet, such an interpretation of early scalar implicature production and, more generally, our findings contrast with work showing that *some*-related implicatures are understood relatively late in childhood, and thus, call



for an explanation. On the one hand, our production results corroborate the study by Katsos and Smith (2010) suggesting that implicature production arises early. On the other hand, the earliest children have been found to understand *some*-related scalar implicatures is 4 (Pouscoulous et al., 2007; Katsos and Bishop, 2011), while our findings suggest that they can produce *some* with an upper-bound reading from the age of two. The gap between these two sets of evidence must be bridged.

An account along lexicalist lines (e.g., Barner et al., 2011) might find it difficult to contend with such early implicature production. If toddlers have not associated *some* with its lexical scale (*many, most, all*), this should affect their ability to produce, as well as comprehend, implicatures. Importantly, examples where children's use of *some* is directly contrasted with another member of the semantic scale (*all* or other, see Examples 1 - 4 and **Supplementary Material**), reinforce a picture where children master the contrast set of *some* from a very early age—as young as 2;03 for some of them. These cases indicate that the *Implicature Plausible* instances found in child production are not merely an artifact of our way of categorising *some*-utterances, but truly reflect the ability of very young children to intend scalar implicatures linked to *some*. They also speak further against a lexicalist account of scalar implicature acquisition. Therefore, an approach on the development of scalars integrating several contextual factors might be more appropriate to reconcile the experimental comprehension findings with our production data.

Several elements may explain children's behaviour in comprehension experiments such as their pragmatic tolerance (Katsos and Bishop, 2011), the relevance of the implicature in context (Papafragou and Musolino, 2003; Guasti et al., 2005; Skordos and Papafragou, 2016), and children's limited processing resources when faced with an infrequent, relatively effortful inference (Reinhart, 2004; Pouscoulous et al., 2007). Indeed, pragmatic tolerance constrains experimental measures of implicature comprehension, since children might be inclined to judge a sentence as correct despite pragmatic infelicity. But, of course, pragmatic tolerance would have no impact on production. Similarly, while implicature comprehension might be affected by how relevant the scalar implicature is in context, relevance does not influence production: if a speaker intends to produce an implicature, then it is *a priori* relevant to them. These factors combined with children's limited exposure to *some*-related implicatures may be sufficient to account for the discrepancy between production and comprehension. In this view, children are capable of producing and inferring *some*-related implicatures from their third year of life and any difficulty in understanding them in experimental settings is to be attributed to factors outside their semantic and pragmatic competence.

This type of account also resonates with experimental findings suggesting a much earlier comprehension of linguistic pragmatic phenomena than previously thought. Indeed, while preschoolers find most pragmatic inferences challenging on traditional metalinguistic tasks such as explaining or judging the truth value of an utterance, a few recent studies indicate that they fare much better with paradigms using act-out or

picture selection tasks: 3-year-olds understand other pragmatic phenomena (e.g., Berger and Höhle 2012 on presupposition; Falkum et al. (2017) on metonymy; Pearson (1990) on metaphor), but also other implicatures (Schulze et al., 2013, on relevance implicatures) and even other types of scalar implicatures (Stiller et al., 2014, on *ad hoc* scalar implicatures).

In the past decade a lot of work has been devoted to children's comprehension of *some*. In fact, our knowledge of implicature acquisition is largely based on their understanding of this one expression. A systematic corpus analysis of how toddlers hear and produce it should therefore be essential to any informed argument in the debate. The findings indicate that children begin producing and interpreting implicatures in a pragmatic way during their third year of life, very soon after they first produce *some*. Thus, almost as soon as they acquire *some*, children produce it competently and mirror adult behaviour. Their production of *some* implicatures is low but matches their parent's input in frequency. In both children and adults *some* appears to be multifaceted and implicatures are infrequent, and both structurally contextually constrained. Our findings add to a growing body of evidence showing that the upper-bound reading of *some* is much less frequent in adult speech than some scholars would have had us believe. Our study is also the first to go against the popular belief in some psychology and linguistics circles that children do not produce implicatures, much less so lexicalized scalar implicatures, at an early age. Yet, it does by no means answer all the questions. The method we used has its flaws in that it relies on coder judgment; it has its strengths, too, in the nature and size of the corpus we used. The similarity between other adult findings and ours, and the striking resemblance between our adult and children results give us reasonable confidence in the soundness of our paradigm. In any case, this work should be expanded by experimental research looking at children's production of *some* and other implicatures. An important question which still requires a more fine-tuned answer – both empirically and theoretically – is how children can appear to fare so poorly with implicatures in experimental paradigms if the basic mechanisms are in place so early.

## DATA AVAILABILITY STATEMENT

A file including data and coding for this study can be found on the Open Science Framework at [osf.io/g6psr](https://osf.io/g6psr).

## AUTHOR CONTRIBUTIONS

NP, EL, and SE designed the study. EL provided the dataset. SE coded and analysed the data; SE, NP, and EL wrote the manuscript. All authors discussed the results and commented on the manuscript.

## FUNDING

This research was supported by a Travel Grant to SE by the Leibniz ScienceCampus Primate Cognition, Germany. Elena Lieven's research is funded by the Economic and Social

Research Council (ESRC) ES/L008955/1 for the ESRC LuCiD Centre.

Caterina Paolazzi and Giulio Dulcinati for their help with coding.

## ACKNOWLEDGMENTS

We are grateful to the Child Study Centre at the University of Manchester who kindly granted us access to the Gina and Helen datasets. We also wish to thank

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01928/full#supplementary-material>

## REFERENCES

- Ambridge, B., Kidd, E., Rowland, C. F., and Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *J. Child Lang.* 42, 239–273. doi: 10.1017/S030500091400049x
- Bagassi, M., D'Addario, M., Macchi, L., and Sala, V. (2009). Children's acceptance of underinformative sentences: the case of *some* as a determiner. *Think. Reason.* 15, 211–235. doi: 10.1080/13546780902864306
- Barner, D., Brooks, N., and Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition* 118, 84–93. doi: 10.1016/j.cognition.2010.10.010
- Barner, D., Brooks, N., and Bale, A. C. (2010). Quantity implicature and access to scalar alternatives in language acquisition. *Semant. Linguist. Theor.* 20, 525–543. doi: 10.3765/salt.v20i0.2571
- Barner, D., Chow, K., and Yang, S.-J. (2009). Finding one's meaning: a test of the relation between quantifiers and integers in language development. *Cogn. Psychol.* 58, 195–219. doi: 10.1016/j.cogpsych.2008.07.001
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bergen, L., and Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 1450. doi: 10.1037/a0027850
- Berger, F., and Höhle, B. (2012). Restrictions on addition: children's interpretation of the focus particles *auch* and *nur* only in German. *J. Child Lang.* 39, 383–410. doi: 10.1017/S0305000911000122
- Bernicot, J., Laval, V., and Chaminaud, S. (2007). Nonliteral language forms in children: in what order are they acquired in pragmatics and metapragmatics? *J. Pragmatics* 39, 2115–2132. doi: 10.1016/j.pragma.2007.05.009
- Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.
- Bott, L., Bailey, T. M., and Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *J. Mem. Lang.* 66, 123–142. doi: 10.1016/j.jml.2011.09.005
- Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006
- Braine, M. D., and Romain, B. (1981). Development of comprehension of *or*: evidence for a sequence of competencies. *J. Exp. Child Psychol.* 31, 46–70.
- Breheny, R., Katsos, N., and Williams, J. (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. doi: 10.1016/j.cognition.2005.07.003
- Breheny, R. E. (forthcoming). "Scalar implicatures in a gricean cognitive system," in *Handbook of Experimental Pragmatics*, eds N. Katsos and C. Cummins (Oxford University Press).
- Cameron-Faulkner, T., Lieven, E., and Theakston, A. (2007). What part of *no* do children not understand? A usage-based account of multiword negation. *J. Child Lang.* 34, 251. doi: 10.1017/S0305000906007884
- Chierchia, G. (2004). "Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface," in *Structures and Beyond*, Vol. 3, ed A. Belletti (Oxford, UK: Oxford University Press), 39–103.
- Chierchia, G. (2006). Broaden your views: implicatures of domain widening and the logic of language. *Linguist. Inquiry* 37, 535–590. doi: 10.1162/ling.2006.37.4.535
- Chierchia, G., Fox, D., and Spector, B. (2012). "The grammatical view of scalar implicatures and the relationship between semantics and pragmatics," in *An International Handbook of Natural Language Meaning*, Vol. 3, eds P. Portner, C. Maienborn and K. von Stechow (Berlin: Mouton de Gruyter), 2297–2332.
- Clark, E. V. (2016). *First Language Acquisition*, 3rd Edn. Cambridge, UK: Cambridge University Press.
- Davies, C., and Katsos, N. (2010). Over-informative children: production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua* 120, 1956–1972. doi: 10.1016/j.lingua.2010.02.005
- De Neys, W., and Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Exp. Psychol.* 54, 128–133. doi: 10.1027/1618-3169.54.2.128
- De Ruiter, L., Theakston, A., Brandt, S., and Lieven, E. (2017). "The relationship between parental input and children's spontaneous use of adverbial clauses containing *after*, *before*, *because*," in *Poster Presented at the 14th International Congress for the Study of Child Language (IASCL)*, July 17–21 (Lyon).
- Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semant. Pragmatics* 8, 1–55. doi: 10.3765/sp.8.11
- Degen, J., and Tanenhaus, M. K. (2011). "Making inferences: the case of scalar implicature processing," in *33rd Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society, 3299–3304.
- Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171
- Falkum, I. L., Recasens, M., and Clark, E. V. (2017). The moustache sits down first?: on the acquisition of metonymy. *J. Child Lang.* 44, 87–119. doi: 10.1017/S0305000915000720
- Feeney, A., Scafton, S., Duckworth, A., and Handley, S. J. (2004). The story of some: everyday pragmatic inference by children and adults. *Can. J. Exp. Psychol.* 58, 121. doi: 10.1037/h0085792
- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., et al. (1994). Variability in early communicative development. *Monogr. Soc. Res. Child Dev.* 59, 1–173; discussion 174–185. doi: 10.2307/1166093
- Foppolo, F., Guasti, M. T., and Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Lang. Learn. Dev.* 8, 365–394. doi: 10.1080/15475441.2011.626386
- Geurts, B. (2010). *Quantity Implicatures*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511975158
- Grice, H. P. (1957). Meaning. *Philos. Rev.* 66, 377–388.
- Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., and Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Lang. Cogn. Process.* 20, 667–696. doi: 10.1080/01690960444000250
- Hochstein, L., Bale, A., Fox, D., and Barner, D. (2014). Ignorance and inference: do problems with gricean epistemic reasoning explain children's difficulty with scalar implicature? *J. Semant.* 33, 1–29. doi: 10.1093/jos/ffu015
- Horn, L. (1984). "Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature," in *Meaning, Form, and Use in Context: Linguistic Applications*, ed D. Schiffrin (Washington, DC: Georgetown University Press), 11–42.
- Horn, L. (1989). *A Natural History of Negation*. Chicago, IL: Chicago University Press.

- Horowitz, A. C., Schneider, R. M., and Frank, M. C. (2017). The trouble with quantifiers: exploring children's deficits in scalar implicature. *Child Dev.* doi: 10.1111/cdev.13014. [Epub ahead of print].
- Huang, Y. T., and Snedeker, J. (2009a). Online interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cogn. Psychol.* 58, 376–415. doi: 10.1016/j.cogpsych.2008.09
- Huang, Y. T., and Snedeker, J. (2009b). Semantic meaning and pragmatic interpretation in 5-year-olds: evidence from real-time spoken language comprehension. *Dev. Psychol.* 45, 1723–1739. doi: 10.1037/a0016704
- Huang, Y. T., and Snedeker, J. (2011). Logic and conversation revisited: evidence for a division between semantic and pragmatic content in real-time language comprehension. *Lang. Cogn. Process.* 26, 1161–1172. doi: 10.1080/01690965.2010.508641
- Katsos, N. (2014). "Scalar implicature," in *Pragmatic Development in First Language Acquisition*, Vol. 10, ed D. Matthews (Amsterdam; Philadelphia, PA: John Benjamins Publishing Company), 183–197.
- Katsos, N., and Bishop, D. V. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015
- Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kuvac Kuvac Kraljević, J., Hrzica, G., et al. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proc. Natl. Acad. Sci. U.S.A.* 113, 9244–9249. doi: 10.1073/pnas.1601341113
- Katsos, N., and Smith, N. (2010). "Pragmatic tolerance and speaker comprehender asymmetries," in *34th Annual Boston Conference on Language Development*, eds K. Franich, K. M. Iserman, and L. L. Keil (Somerville, MA: Cascadilla Press), 221–232.
- Kidd, E., Brandt, S., Lieven, E. V., and Tomasello, M. (2007). Object relatives made easy: a cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Lang. Cogn. Process.* 22, 860–897. doi: 10.1080/01690960601155284
- Kilgarriff, A. (2001). Comparing corpora. *Int. J. Corpus Linguist.* 6, 97–133. doi: 10.1075/ijcl.6.1.05kil
- Kuhn, M. (2013). *caret: Classification and Regression Training*. R package version 6.0-73. Available online at: <https://CRAN.R-project.org/package=caret>
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159.
- Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.
- Lieven, E. V., and Behrens, H. (2012). "Dense sampling," in *Research Methods in Child Language: A Practical Guide*, ed E. Hoff (Oxford: Wiley Online Library), 226–239. doi: 10.1002/9781444344035.ch15
- Lieven, E. V., Salomo, D., and Tomasello, M. (2009). Two-year-old children's production of multiword utterances: a usage-based analysis. *Cogn. Linguist.* 20, 481–507. doi: 10.1515/COGL.2009.022
- MacWhinney, B. (2000). *The CHILDES Project: The Database*, Vol. 2. New York, NY: Psychology Press.
- Noveck, I., Chevallier, C., Chevaux, F., Musolino, J., and Bott, L. (2009). "Children's enrichments of conjunctive sentences in context," in *Utterance Interpretation and Cognitive Models*, eds P. De Brabanter and M. Kissine (Bingley, UK: Emerald Group), 211–234.
- Noveck, I., and Sperber, D. (2007). "The why and how of experimental pragmatics: the case of 'scalar inferences,'" in *Pragmatics*, ed N. Burton-Roberts (London, UK: Palgrave Macmillan), 184–212.
- Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1
- Noveck, I. A., and Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain Lang.* 85, 203–210. doi: 10.1016/S0093-934X(03)00053-1
- Papafragou, A., and Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition* 86, 253–282. doi: 10.1016/S0010-0277(02)00179-8
- Papafragou, A., and Skordos, D. (2016). *Scalar Implicature*. Oxford: Oxford University Press.
- Paris, S. G. (1973). Comprehension of language connectives and propositional logical relationships. *J. Exp. Child Psychol.* 16, 278–291.
- Pearson, B. Z. (1990). The comprehension of metaphor by preschool children. *J. Child Lang.* 17, 185–203.
- Pouscoulous, N., and Noveck, I. A. (2009). "Going beyond semantics: the development of pragmatic enrichment," in *Language Acquisition*, ed S. Foster-Cohen (London, UK: Palgrave Macmillan), 196–215.
- Pouscoulous, N., Noveck, I. A., Politzer, G., and Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Lang. Acquis.* 14, 347–375. doi: 10.1080/10489220701600457
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Reinhart, T. (2004). The processing cost of reference set computation: acquisition of stress shift and focus. *Lang. Acquis.* 12, 109–155. doi: 10.1207/s15327817la12021
- Schulze, C., Grassmann, S., and Tomasello, M. (2013). 3-year-old children make relevance inferences in indirect verbal communication. *Child Dev.* 84, 2079–2093. doi: 10.1111/cdev.12093
- Siegal, M. and Surian, L. (2004). Conceptual development and conversational understanding. *Trends Cogn. Sci.* 8, 534–538. doi: 10.1016/j.tics.2004.10.007
- Skordos, D., and Papafragou, A. (2016). Children's derivation of scalar implicatures: alternatives and relevance. *Cognition* 153, 6–18. doi: 10.1016/j.cognition.2016.04.006
- Smith, C. L. (1980). Quantifiers and question answering in young children. *J. Exp. Child Psychol.* 30, 191–205.
- Spooren, W. and Degand, L. (2010). Coding coherence relations: reliability and validity. *Corpus Linguist. Linguist. Theor.* 6, 241–266. doi: 10.1515/clt.2010.009
- Stiller, A. J., Goodman, N. D., and Frank, M. C. (2014). Ad-hoc implicature in preschool children. *Lang. Learn. Dev.* 11, 176–190. doi: 10.1080/15475441.2014.927328
- Sun, C. (2017). *Scalar Implicature: Gricean Reasoning and Local Enrichment*. PhD thesis, University College London.
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Approach to Child Language Acquisition*. Cambridge, MA: Harvard University Press.
- Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT press.
- Viera, A. J., and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam. Med.*, 37, 360–363.
- Wilson, E. A. (2017). *Children's Development of Quantity, Relevance and Manner Implicature Understanding and the Role of the Speakers Epistemic State*. Ph. D. thesis, University of Cambridge.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Eiteljoerge, Pouscoulous and Lieven. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Competition and Symmetry in an Artificial Word Learning Task

Brian Buccola<sup>1\*†</sup>, Isabelle Dautriche<sup>2†</sup> and Emmanuel Chemla<sup>1</sup>

<sup>1</sup> Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Études Cognitives, École Normale Supérieure, PSL Research University, Paris, France, <sup>2</sup> Centre for Language Evolution, University of Edinburgh, Edinburgh, United Kingdom

## OPEN ACCESS

### Edited by:

Penka Stateva,  
University of Nova Gorica, Slovenia

### Reviewed by:

Chiara Gambi,  
Cardiff University, United Kingdom  
Jacques Moeschler,  
Université de Genève, Switzerland

### \*Correspondence:

Brian Buccola  
brian.buccola@gmail.com

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

Received: 01 May 2018

Accepted: 22 October 2018

Published: 13 November 2018

### Citation:

Buccola B, Dautriche I and Chemla E  
(2018) Competition and Symmetry in  
an Artificial Word Learning Task.  
Front. Psychol. 9:2176.  
doi: 10.3389/fpsyg.2018.02176

Natural language involves competition. The sentences we choose to utter activate alternative sentences (those we chose not to utter), which hearers typically infer to be false. Hence, as a first approximation, the more alternatives a sentence activates, the more inferences it will trigger. But a closer look at the theory of competition shows that this is not quite true and that under specific circumstances, so-called *symmetric* alternatives cancel each other out. We present an artificial word learning experiment in which participants learn words that may enter into competition with one another. The results show that a mechanism of competition takes place, and that the subtle prediction that alternatives trigger inferences, and may stop triggering them after a point due to symmetry, is borne out. This study provides a minimal testing paradigm to reveal competition and some of its subtle characteristics in human languages and beyond.

**Keywords:** competition, symmetry, alternatives, psycholinguistics, semantics, pragmatics

## 1. COMPETITION IN LANGUAGE

### 1.1. First Examples and Description

In using language to communicate, the words and phrases that a speaker decides to use nearly always acquire an interpretation that goes beyond their strict, literal meaning. For example, if Alice utters to Bob the sentence in (1), then Bob might infer, among other things, that the animal Alice saw was not a cat or dog (or, at least, that Alice does not believe it was), but rather some more unusual animal like a raccoon, even though cats and dogs obviously count as animals, too.

(1) I saw an animal on my neighbor's porch this morning.

Similarly, if Alice utters to Bob the sentence in (2), then Bob will likely infer that Alice did not see both a dog *and* a cat on her neighbor's porch, even though, strictly speaking, seeing both animals counts as an instance of seeing one or the other.

(2) I saw a cat or a dog on my neighbor's porch this morning.

The process by which hearers draw these inferences has been the subject of much research and debate in semantic and pragmatic theory. However, starting with the pioneering work of Grice (1975), there is a consensus that, at its root, the process involves the hearer reasoning not just about what the speaker said, but also what the speaker could have said but chose *not* to say. That is, the things we say, as well as alternative things we could have said but chose not to, together affect the overall meanings of our utterances. In the case of (1), for example, if Alice had in fact seen a dog (and Alice knows she saw a dog), then it would be more appropriate for Alice to say so, even if (1) is true. Thus, if Alice chooses to utter (1) rather than the minimally different (3), in which *dog*



replaces *animal*, then it is reasonable to infer that she did so because the animal she saw is *not* a dog (or any other option of the sort worth mentioning).

(3) I saw a dog on my neighbor's porch this morning.

In a parallel way, if Alice had seen both a cat and a dog (and Alice knows she saw both), then it would be more appropriate for Alice to say so, even if (2) is true. Thus, if Alice chooses to utter (2) rather than the minimally different (4), in which *and* replaces *or*, then it is reasonable to infer that she did so because she did *not* see both a cat *and* a dog.

(4) I saw a cat and a dog on my neighbor's porch this morning.

Grice (1975) coined the term *implicature* (and the associated verb *implicate*) to refer to the act of implying one thing by saying another. Thus, for instance, a speaker who utters (2) tends, we say, to implicate that (4) is false.

In sum, then, as speakers, the various things we can say when communicating a message “compete” with one another, so that what we choose to say and what we choose *not* to say together affect the final message we transmit.

## 1.2. Toward a Theory of Competition

As a first approximation toward a theory of competition in language, we might say that the use of an expression  $\varphi$  licenses the inference that  $\varphi$  was, in some sense, “better” or “more appropriate” than every alternative of  $\varphi$  that could have been used instead. We refer to this as the Competition Principle. (Our formulation in (5) is more general than the sorts of formulations found in the literature — e.g., Davis, 2014 and the references therein — and our reason for this is so that we may apply it to situations beyond traditional communicative settings.)

(5) **Competition Principle.** The use of  $\varphi$  implies that each alternative  $\psi$  of  $\varphi$  is less appropriate than  $\varphi$ .

This principle presupposes several notions that need to be spelled out: the notion of *use*, the notion of *appropriateness*, and the notion of an *alternative*.

In the context of the examples of above, and indeed in most of the relevant literature, to use an expression simply means to utter it, broadly speaking (i.e., to vocalize it, to sign it, to write it, and so on). In the context of our experimental task, this notion will take on a slightly broader meaning, which we will discuss later on.

The notion of appropriateness encompasses several possible things, because alternatives may be inappropriate (or less appropriate) for different reasons. For example, an alternative  $\psi$  of  $\varphi$  may be inappropriate simply because  $\psi$  is false (while  $\varphi$  is true), or  $\psi$  may be inappropriate because, although true,  $\psi$  is less informative, or specific, than  $\varphi$ . (This aspect of the Competition Principle is traditionally grounded in Grice's maxims of Quality and Quantity, respectively. We collapse them here for the sake of simplicity.)

Finally, the notion of alternative raises the question of what exactly “counts” as an alternative of  $\varphi$ . This is an important question that has received quite a bit of attention in the literature, the consensus being that alternatives need to be constrained in

one way or another (for specific proposals, see, e.g., the *Horn scales* of Horn, 1972, and the theory of structurally defined alternatives of Katzir, 2007). We will not have much to add to this debate. For concreteness, we will adopt the simplistic view that the alternatives of  $\varphi$  are obtained by (recursively) replacing lexical elements in  $\varphi$  with other lexical elements from the given language. (For our experimental task, the choice of theory is immaterial, roughly because it will involve single-word expressions anyway.)

Putting everything together, we can say that, because of the Competition Principle, an utterance of (1) licenses the inference that the alternative in (3) is false (hence, that Alice saw an animal, but not a dog), and an utterance of (2) licenses the inference that the alternative in (4) is false (hence, that Alice saw a cat or a dog, but not both).

## 1.3. Symmetry

We have seen that alternatives create inferences. From the discussion so far, one may think that the more alternatives a sentence has, the more inferences one will draw from the use of that sentence. But this is not always so, because alternatives may cancel each other out, when a certain logical relation, known as *symmetry*, obtains between them (relative to the uttered sentence).

Abstractly first, symmetry arises when a sentence  $\varphi$  has two alternatives,  $\psi_1$  and  $\psi_2$ , such that  $\psi_1$  and  $\psi_2$  can each be individually negated without contradicting  $\varphi$ , but their combined negation contradicts  $\varphi$ . In symbols,  $\varphi \wedge \neg\psi_1 \wedge \neg\psi_2$  is a contradiction, while  $\varphi \wedge \neg\psi_1$  and  $\varphi \wedge \neg\psi_2$  are not. In such cases, we say that  $\psi_1$  and  $\psi_2$  are symmetric alternatives (relative to  $\varphi$ ), and that they create symmetry, because they cannot both be negated in a way that is compatible with  $\varphi$  — negating one forces the other to be true (Fox, 2007).

Concretely now, let  $\varphi$  be (2), and suppose that its two alternatives are (6a) ( $= \psi_1$ ) and (6b) ( $= \psi_2$ ) below. Then it is not possible for both (6a) to be false (Alice did not see a cat) and (6b) to be false (Alice did not see a dog), while at the same time the original sentence is true (Alice saw one or the other). So, disjunction ( $\varphi = \psi_1 \vee \psi_2$ ) is a concrete case where two alternatives ( $\psi_1$  and  $\psi_2$ ) cannot both be negated, hence are symmetric.

(6) a. I saw a cat on my neighbor's porch this morning.  
b. I saw a dog on my neighbor's porch this morning.

In cases of symmetry, one might expect that in some contexts,  $\varphi$  could imply  $\neg\psi_1$  (rather than  $\neg\psi_2$ ), while in other contexts,  $\varphi$  could imply  $\neg\psi_2$  (rather than  $\neg\psi_1$ ). In actual fact, however, we observe that context cannot “break” symmetry (Fox and Katzir, 2011). Instead, hearers draw speaker uncertainty inferences regarding symmetric alternatives.

For example, (2), in addition to conveying that Alice did not see both a cat and a dog, also conveys that Alice is uncertain which of the two animals (a cat or a dog) she actually saw. How does the Competition Principle help us to understand this uncertainty inference? If Alice utters (2), and if (6a) and (6b) are alternatives of (2) (Sauerland, 2004), then the Competition Principle us that each of them was less appropriate than (2).

However, by “less appropriate,” we cannot mean false, because it cannot be that (2) is true while (6a) and (6b) are both false (again, that would be a contradiction). So, it must mean something else. One natural possibility is that (2) is appropriate because Alice is certain that it is true, whereas each of (6a) and (6b) is less appropriate in virtue of Alice *not* being certain that it is true. If so, then this amounts to the observed uncertainty inference regarding the two symmetric alternatives (6a) and (6b).<sup>1</sup>

In short, more alternatives does not always equal more inferences. Sometimes, more alternatives introduces symmetry, which cancels out inferences that otherwise may have obtained (or converts them from plain negated inferences to uncertainty inferences).

## 1.4. Symmetry as a Diagnosis of Competition

In actual language use, symmetry does not seem to appear or disappear from context to context, but instead is rather stable across contexts. Abstractly, a more informative alternative  $\psi$  of  $\varphi$  either always has a symmetric partner (hence, the use of  $\varphi$  yields speaker uncertainty about  $\psi$ ), or never does (hence, the use of  $\varphi$  yields the inference that  $\psi$  is false, provided the speaker is competent about  $\psi$ , and  $\psi$  is relevant). For example, when it comes to disjunction, as in (2), the conjunctive alternative, (4), never has a symmetric partner — this would be something like (7) below — so as a result, (2) invariably triggers the inference that (4) is false, rather than speaker uncertainty about (4) and (7).<sup>2</sup> Conversely, a disjunction like (2) always has its individual disjuncts, (6a) and (6b), as alternatives, hence always exhibits symmetry, so as a result, (2) invariably triggers speaker uncertainty about (6a) and (6b), rather than the inference that one (or the other) of them is false.

- (7) I saw a cat or a dog but not both on my neighbor's porch this morning.

A consequence of all this is that it can be relatively tricky to observe competition directly. If  $\varphi$  typically implies  $\neg\psi$ , then maybe this is simply because  $\varphi$  literally entails that  $\psi$  is false, or because  $\psi$  is extremely unlikely to begin with (given  $\varphi$ ). For example, for (1), one might argue the inference that Alice did not see a dog is simply a contextual one (it's less likely for her to have seen a dog than, say, a raccoon — a weak argument, admittedly). Conversely, for (2), one might argue that *or* is inherently exclusive, i.e., that  $\varphi$  *or*  $\psi$  literally means “ $\varphi$  or  $\psi$  but not both”.

In a similar fashion, if  $\varphi$  typically implies speaker uncertainty about  $\psi_1$  and  $\psi_2$ , then maybe this is simply because  $\varphi$  literally entails such uncertainty. For example, perhaps the literal meaning of *or* encodes something about the knowledge state

of the speaker who uses it, so that it actually entails speaker uncertainty about the individual disjuncts.

In short, because competition is difficult to observe directly, one may wonder whether there is any competition going on in these cases to begin with. Of course, linguists have developed intricate diagnostics to argue that these *are* examples of competition, e.g., embedding them in downward-entailing (roughly, negative) contexts and observing that the relevant inferences disappear. For example, *I did not see a cat or a dog on my neighbor's porch this morning* does not trigger any speaker uncertainty inferences, nor does it convey the denial of speaker uncertainty about the individual disjuncts (if *or* literally encoded speaker uncertainty, then this sentence could mean “it is not the case that I saw a cat or a dog but I don't know which,” which would be true in a scenario where Alice saw a cat or a dog and Alice knew which — an impossible reading of the sentence).

Nevertheless, our goal here is to explore whether there is a way to observe the Competition Principle more directly. We propose to do so using symmetry as the diagnosis for the presence of competition, by manipulating the presence or absence of symmetry across experimental contexts (something that does not readily happen in everyday linguistic contexts). Specifically, we report on an artificial word learning experiment which had the following goal: to see whether we could create competition between two nonce words — a word  $w$  that applies to more than one kind of object, and a more specific/informative word  $w_1$  that applies to a strict subset of what  $w$  applies to — and observe its effect, and then to remove that effect by introducing a third word,  $w_2$ , such that  $w_1$  and  $w_2$  are symmetric relative to  $w$ .

Our artificial word learning experiment involved tasks in which communicative cooperativeness (hence, traditional Gricean maxims) seemed to play little or no role (there was no speaker-hearer, for instance). Capitalizing on this aspect, a secondary goal of ours was to see whether a general, i.e., not specifically conversational, notion of competition — something like our Competition Principle in (5) — could be detected, especially since it is often assumed in the Gricean literature that Gricean principles are grounded in more general principles of rationality.<sup>3</sup> Up to now, this idea has never been tested. Our results suggest a positive answer: the Competition Principle *does* play a role in non-conversational tasks like the ones we used.

## 2. EXPERIMENT

The Competition Principle seems to be at the heart of pragmatic enrichment during communication in natural language, but it can often be difficult to assess exactly what is in competition, what role symmetry plays, etc. We present an experimental study that investigates whether we may observe the Competition Principle somewhat more directly over the course of acquisition of nonce words, by manipulating the presence or absence of alternatives and symmetry across experimental contexts.

<sup>1</sup>Sauerland (2004) (building on Gazdar, 1979; Soames, 1982; Horn, 1989) provides a more formal implementation of this reasoning process.

<sup>2</sup>The question of why (4) but not (7) is an alternative of (2), and how the theory of alternatives should explain this fact, is an instance of the so-called *symmetry problem* (Fox, 2007; Katzir, 2007; Fox and Katzir, 2011), which does not concern us here.

<sup>3</sup>The idea that conversation is a cooperative enterprise grounded in rational behavior originates with Grice himself (Grice, 1975) (for discussion, see also Levinson, 1983). Recent game-theoretic approaches to this idea include Franke (2011) and Bergen et al. (2016).

## 2.1. Task Summary and Hypothesis

The goal of the task was to learn three new words —  $w$ ,  $w_1$ , and  $w_2$  — where  $w$  applied to (at least) two kinds of objects (e.g., both triangles and circles), while  $w_1$  applied to just one of the two kinds (e.g., triangles), and  $w_2$  applied to just the other of the two kinds (e.g., circles) (see **Figure 1**).

To learn the meaning of words, participants observed a series of displays containing one of the words to be learned and a collection of objects with different properties (see **Figure 2**). They then picked an object from the collection and received feedback.

We tested participants' understanding of  $w$  when presented with both  $w_1$ -type-objects and  $w_2$ -type-objects at different learning stages: after they learned  $w$  only, after they learned  $w$  and  $w_1$  but not  $w_2$ , and after they learned all three words  $w$ ,  $w_1$ , and  $w_2$  (see **Figure 3**). The idea then was to gradually introduce alternatives: first a unique alternative, which may trigger inferences through the Competition Principle, and then yet another alternative that may create symmetry, and could therefore remove the inferential effect of competition. More specifically, our hypothesis was the following: after learning  $w$ , but before learning  $w_1$  or  $w_2$ , participants should choose indiscriminately between the two kinds of objects (or perhaps with some measurable bias); after learning  $w$  and  $w_1$ , participants should choose  $w_2$ -type-objects more so than before, due to competition between  $w$  and  $w_1$ ; and finally after learning  $w$ ,  $w_1$ , and  $w_2$ , participants should go back to choosing indiscriminately, due to symmetry between  $w_1$  and  $w_2$ .

## 2.2. Method

All data and scripts for their analysis are available at <https://semanticsarchive.net/Archive/DJmNjYxY/>.

### 2.2.1. Ethics Statement

This study was carried out in accordance with the recommendations of the Comité d'Éthique de la Recherche en Santé (2013/46). The protocol was approved by the Comité

d'Éthique de la Recherche en Santé (2013/46). In accordance with the Declaration of Helsinki, prior to participating in this online study, all participants were presented with the informed consent document and instructions stating that by clicking "I accept" they indicated their consent to participate in the study.

### 2.2.2. Participants

Fifty-three adults were recruited through Amazon's Mechanical Turk (25 females;  $M = 38$  years; all native speakers of English) and compensated \$1.80 for their participation. Participants were randomly assigned to one of two groups (see Design below): the Competition group ( $N = 26$ ) and the No-competition group ( $N = 27$ ). One additional participant was excluded in the Competition group for failing to pass the learning criteria.

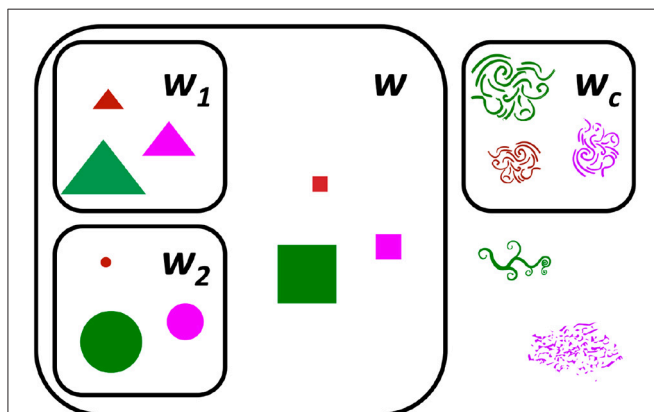
### 2.2.3. Procedure

Participants were tested online. They were instructed that their task was to learn new words by associating them with objects displayed on the screen. In the instructions, participants were given a screenshot of a trial involving a word (not used during the test) and a set of objects. No information about the number of to-be-learned words was given. For each trial, a word was displayed, first alone for 500 ms to attract participants' attention to the word, then together with a collection of 3 objects, aligned horizontally, below the word (see **Figure 2**). Participants were asked to click on the object they believed to be associated with the word. The experiment consisted of several learning and testing phases (see Design below).

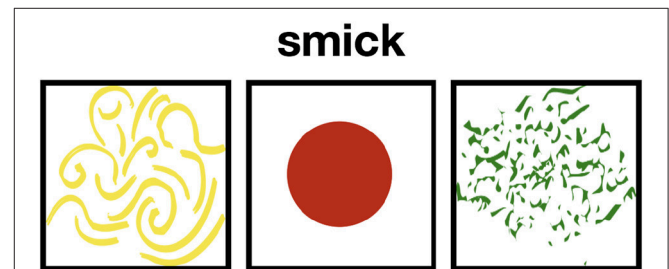
During the learning phases, participants received feedback on their response after each trial. The feedback was displayed in a horizontal bar positioned at the top of the screen. The bar turned green and displayed the prompt "Correct!" for correct responses, and turned red and displayed the prompt "Incorrect" for incorrect responses. Correct responses had 2 s of feedback before the next trial, while incorrect responses had 6 s of feedback to increase attention to the task.

During the testing phases, participants did not receive any feedback: once they responded, the experiment continued with the next trial. Each testing phase was preceded by a warning to participants ("You will not receive feedback for the next couple of events.") displayed for 4 s in the same top horizontal bar used for the feedback.

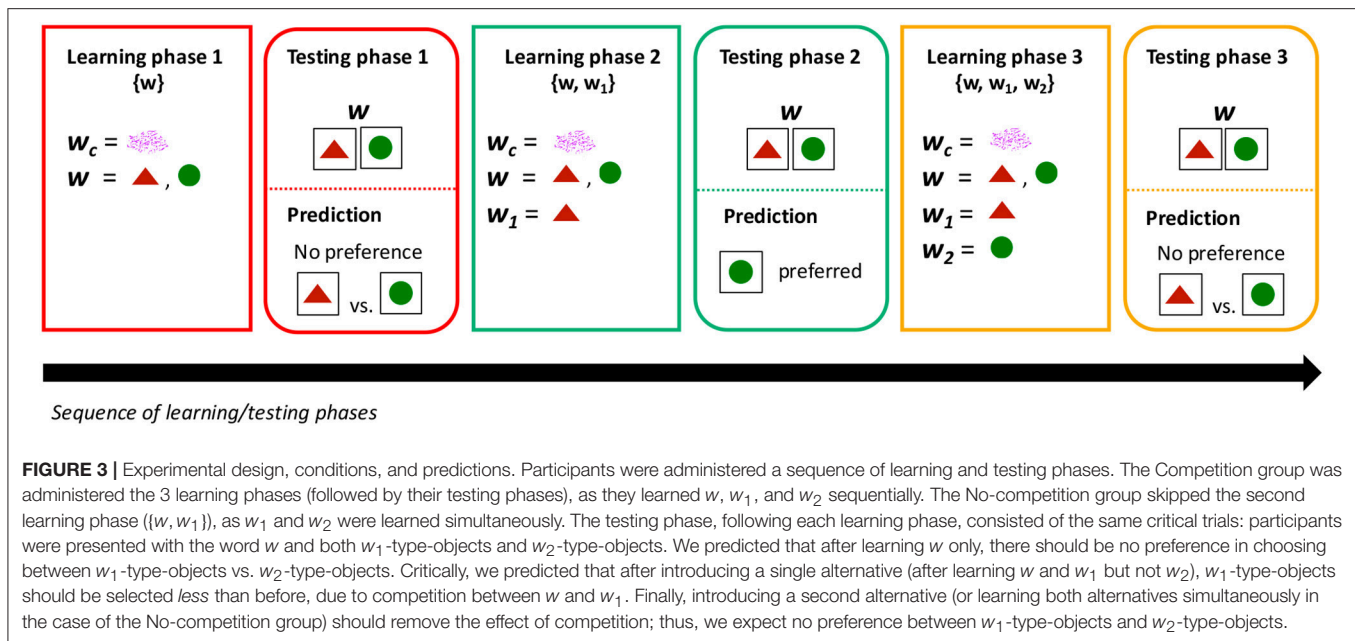
Participants' answers as well as their response times were recorded on each trial. At the end of the experiment, there was



**FIGURE 1** | Participants learned four novel words: three critical words ( $w$ ,  $w_1$ , and  $w_2$ ) and one control word ( $w_c$ ).  $w$  applied to (at least) two kinds of objects (e.g., both triangles and circles), while  $w_1$  applied to just one of the two kinds (e.g., triangles), and  $w_2$  applied to just the other of the two kinds (e.g., circles).



**FIGURE 2** | Example of a trial.



a final questionnaire asking participants about their age, native language, and country.

#### 2.2.4. Stimuli

The space of objects included 3 geometric shapes (circles, triangles, and squares) and 3 organic shapes (clouds of dots, clouds of curly lines, and spiraling branches). For variability, objects also varied across two irrelevant dimensions: colors (red, yellow, blue, green, and pink) and size (small, medium, and big), leading to 15 possible configurations per object.

We chose 4 novel words from a list of pseudowords obeying the rules of English phonotactics (*blicket*, *dax*, *diti*, *smick*, *tupa*, *fep*, *bosa*, *moop*, *zud*, *vash*, and *gaddle*).

#### 2.2.5. Design

Each participant learned four words over the course of the experiment:  $w$ ,  $w_1$ ,  $w_2$ , and  $w_c$ .  $w$  applied to the 3 geometrical shapes (i.e., circles, triangles, and squares), whereas  $w_1$  applied to just one (e.g., triangles), and  $w_2$  to another one (e.g., circles) (see **Figure 1**).  $w_c$  was a control word that applied to one of the 3 organic shapes to encourage participants to pay attention to the words and not click systematically on any of the geometrical shapes present in the display. The target objects associated with  $w_1$ ,  $w_2$ , and  $w_c$  were randomized across participants.

The experiment was divided into several learning phases, each followed by a testing phase. We used a between-subject design in which some subjects received three learning and testing phases (the Competition group), and others two (the No-competition group). In the former case, participants first learned  $w$  and  $w_c$ , then  $w_1$ , and finally  $w_2$ ; in the latter case, participants first learned  $w$  and  $w_c$ , then  $w_1$  and  $w_2$  simultaneously (see **Figure 3** for a graphical illustration of the time course of the experiment).

#### 2.2.6. Learning Phases

All trials featured a single target object with two randomly chosen distractors such that there was only a single correct response. Trials were presented in blocks to control for the amount of learning received for each word. Details describing the exact number of trials per word per block in each learning phase can be found in the **Supplemental Material**. Participants were exposed to a minimum of 3 blocks. The learning phase ended when participants responded correctly for all trials in a block. If they answered more than 250 trials without reaching the learning criteria, the experiment continued normally but we discarded their responses ( $N = 1$ ).

#### 2.2.7. Testing Phases

The testing phases always consisted of 4 critical trials interspaced with the same type of trials seen during the previous learning phase (3 trials per word learned until that point; see the **Supplemental Material** for a precise description). In the critical trials, participants were presented with  $w$ , together with a collection of objects that contained both a  $w_1$ -type-object (e.g., a triangle) and a  $w_2$ -type-object (e.g., a circle). These critical trials were placed at the beginning of the testing phases, and interspaced by one other trial.

#### 2.2.8. Conditions and Predictions

There were three conditions that depended on the training a participant received. In the  $\{w\}$  condition (no alternative), participants had learned  $w$  but not  $w_1$  or  $w_2$ ; in the  $\{w, w_1\}$  condition (one alternative), participants had learned both  $w$  and  $w_1$  but not  $w_2$ ; and in the  $\{w, w_1, w_2\}$  condition (two alternatives, symmetric), participants had learned  $w$ ,  $w_1$ , and  $w_2$ .

The testing phase, with the same critical trials, was administered after each of these different learning phases, allowing us to test the effect of symmetry in participants'



lexicon on their responses on the critical trials. We measured the proportion of  $w_1$ -type-objects vs.  $w_2$ -type-objects that participants picked when presented with the word  $w$  and both kinds of objects. The critical trials and the predictions associated with each condition are illustrated in **Figure 3**. Our predictions were the following: in the  $\{w\}$  condition (after learning  $w$ , but before learning  $w_1$  or  $w_2$ ), participants should choose indiscriminately between the two kinds (or perhaps with some measurable bias); in the  $\{w, w_1\}$  condition (after learning  $w$  and  $w_1$ , but before learning  $w_2$ ), participants should choose  $w_2$ -type-objects more so than before, due to competition between  $w$  and  $w_1$ ; and in the  $\{w, w_1, w_2\}$  condition (after learning  $w$ ,  $w_1$ , and  $w_2$ ), participants should go back to the same response rate observed in the  $\{w\}$  condition, due to symmetry between  $w_1$  and  $w_2$ . Critically, in the No-competition group, who are not learning  $w_1$  and  $w_2$  sequentially but simultaneously (and thus do not receive the  $\{w, w_1\}$  condition), there should be no difference in their response rate between the  $\{w\}$  and the  $\{w, w_1, w_2\}$  conditions, since both  $w_1$  and  $w_2$  immediately compete with  $w$ , and the effects of competition are thus canceled out due to symmetry.

### 2.2.9. Data Analysis

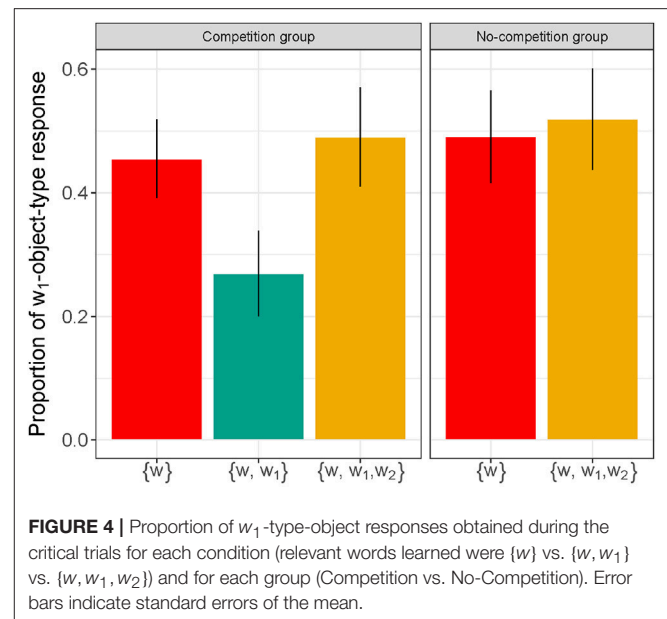
The data analysis was conducted using the `lme4` package (Bates et al., 2015) of R. In a mixed logit regression (Jaeger, 2008), we modeled the selection of  $w_1$ -type-objects (coded as 0 or 1) compared to  $w_2$ -type-objects during the critical test trials. All responses on critical trials were included in the analysis since participants always picked either  $w_1$ - or  $w_2$ -type-objects, and never a distractor object. The model included two categorical predictors with their interaction: Group (Competition vs. No-Competition) and Condition ( $\{w\}$  vs.  $\{w, w_1\}$  vs.  $\{w, w_1, w_2\}$ ) as well as a random intercept and random slopes for Condition for participants. The resulting R syntax for the model was: `w1-type-objects-selection ~ Condition * Group + (1 + Condition | Participant)`.

## 2.3. Results

**Figure 4** reports the average proportion of  $w_1$ -type-object responses during the critical trials by condition ( $\{w\}$  vs.  $\{w, w_1\}$  vs.  $\{w, w_1, w_2\}$ ) and group (Competition vs. No-Competition).

The outputs of the models are in the **Supplemental Material**, with the full script available at [https://semanticsarchive.net/Archive/DJmNjYxY/Competition\\_analysis.R](https://semanticsarchive.net/Archive/DJmNjYxY/Competition_analysis.R). Over the conditions present in the two groups ( $\{w\}$  and  $\{w, w_1, w_2\}$ ), there was no main effect of Group ( $\chi^2 = 0.27$ ;  $p = 0.87$ ), nor a significant interaction between Condition and Group ( $\chi^2 = 0.01$ ;  $p = 0.91$ ), illustrating that both groups responded in the same way in the  $\{w\}$  condition ( $M_{comp} = 0.46$ ;  $SE_{comp} = 0.06$  vs.  $M_{no-comp} = 0.49$ ;  $SE_{no-comp} = 0.07$ ) and in the  $\{w, w_1, w_2\}$  condition ( $M_{comp} = 0.49$ ;  $SE_{comp} = 0.08$  vs.  $M_{no-comp} = 0.52$ ;  $SE_{no-comp} = 0.08$ ).

Critically, there was a main effect of Condition ( $\chi^2 = 13.61$ ;  $p < 0.01$ ). Participants' responses were sensitive to the presence or the absence of symmetry in their lexicon: participants in the Competition group selected less  $w_1$ -type-objects in the  $\{w, w_1\}$  ( $M = 0.27$ ;  $SE = 0.07$ ) condition than in the surrounding  $\{w\}$



**FIGURE 4** | Proportion of  $w_1$ -type-object responses obtained during the critical trials for each condition (relevant words learned were  $\{w\}$  vs.  $\{w, w_1\}$  vs.  $\{w, w_1, w_2\}$ ) and for each group (Competition vs. No-Competition). Error bars indicate standard errors of the mean.

( $\beta = 1.72$ ;  $z = 2.66$ ;  $p < 0.01$ ) and  $\{w, w_1, w_2\}$  ( $\beta = 1.82$ ;  $z = 2.86$ ;  $p < 0.01$ ) conditions. In other words, learning  $w_1$  created visible effects of competition, and further learning its symmetric alternative,  $w_2$ , removed these effects.

## 2.4. Discussion

Participants were sensitive to the presence/absence of alternatives and symmetry in their lexicon: when asked to pick an object corresponding to a word  $w$ , participants preferred to pick a  $w$ -compatible object for which there was no alternative word that also applied, i.e., to pick a  $w_2$ -type-object (for which there was no alternative word yet) rather than a  $w_1$ -type-object (for which there was an alternative word,  $w_1$ ). This competition effect between the referents of the word  $w$ —those that were  $w_1$ -compatible and those that were not—was removed when participants learned another alternative word,  $w_2$ , that applied to just the other kind of objects labeled by  $w$ , due to symmetry between  $w_1$  and  $w_2$ .

Our task involved nonce words that have translation equivalents in the English lexicon (e.g., *shape*, *triangle*, *circle*). Can our result be explained by participants' existing lexicon? We believe it is unlikely. If participants used their existing lexicon in the task, then we would expect no competition in the  $\{w, w_1\}$  condition, as the English lexicon would still be symmetric in this case. Therefore, the presence of a competition effect, and its subsequent removal after introducing  $w_2$ , suggest that participants use only their newly acquired lexicon in the task.

Another possible alternative explanation for the effect is that participants answered strategically with the goal of balancing out their  $w_1$ -type (e.g., triangle) and  $w_2$ -type (e.g., circle) responses. As a result, when they had a choice between triangle and circle, if they had responded triangles often enough, they may have decided to pick the circle. When learning an alternative word during a learning phase, participants were given opportunities

to respond with the shape corresponding to that word, and so in the following testing phase, they may have thus seized opportunities to give the other options. This explanation predicts that the symmetry effect should be mitigated by this behavior, since the third learning phase does not completely erase the imbalance between the two alternatives (triangles have been selected more often than circles across all learning phases). Yet this is not what we observe. Also, although it is phrased differently, this description may actually be just another version of the Competition Principle: the reason why participants want to balance their triangle and circle responses, all things being equal, may very well be because of a competition effect (selecting triangles repetitively when prompted with the alternative word,  $w_1$ , would encourage participants to pick circles over triangles when prompted with a compatible word,  $w$ ). All in all, however, after debriefing a few people who did the tasks in our lab, it seems that the direct competition explanation is a better match for explaining our participants' behavior.

Finally, an anonymous colleague (p.c.) notes that in testing phase 3, perhaps participants construe  $w$  as referring to the third,  $w_3$ -type-object (via competition between  $w$  and both  $w_1$  and  $w_2$ ), and are at chance only because the  $w_3$ -type-object is not an available option on the critical test trials, not because symmetry is at play, as we claim. To spell this idea out a bit more explicitly, once  $w_1$  and  $w_2$  are both learned, then in the critical trial, if competition were at play, then participants would construe  $w$  as " $w$  but not  $w_1$  and not  $w_2$ ," i.e., as  $w_3$ ; but since  $w_3$  is not an available option, the overall result is a kind of "contextual contradiction." As such, competition leads to a crash, and so competition evidently must not be at play (is "turned off"), and so participants choose randomly between  $w_1$  and  $w_2$ , just like in phase 1. If this is correct, then one could still present this situation as a case of symmetry blocking inferences:  $w_1$  and  $w_2$  are symmetric relative to  $w$  and the *context* of the trial (which excludes  $w_3$  as an option), and that is why participants do not invariably go for just one or the other. Put differently,  $w_1$  and  $w_2$  are still symmetric relative to  $w$  in the *context of the trial*, in the sense that " $w$  and not  $w_1$  and not  $w_2$ " is a contextual contradiction given the absence of any  $w_3$ -type-object. (In other cases of symmetry, " $w$  and not  $w_1$  and not  $w_2$ " would be a plain contradiction, as discussed in §1.3 for the case of disjunction,  $\varphi = \psi_1 \vee \psi_2$ .) So, here, contextual symmetry blocks inferences, just as in other cases of symmetry.

In sum, our results suggest that the Competition Principle may be observed directly during an artificial word learning task as a function of the absence or presence of symmetry at different learning stages of an artificial lexicon.

### 3. GENERAL DISCUSSION AND CONCLUSION

Our results suggest that competition (with and without symmetry) arises spontaneously in artificial word learning tasks, even though the experimental context is not a traditional conversational exchange in any obvious sense. This in turn means that participants appear to apply something like the Competition

Principle during the task. Specifically, they presumably apply a kind of reasoning like the following:

- $\{w\}$  condition: No competition. Choose freely between the  $w_1$ -type-object and the  $w_2$ -type-object.
- $\{w, w_1\}$  condition: The trial uses  $w$ , but it could have used  $w_1$  instead. Therefore,  $w_1$  might have been less appropriate. Thus, the  $w_1$ -type-object might be less appropriate than the  $w_2$ -type-object. Choose the  $w_2$ -type-object.
- $\{w, w_1, w_2\}$  condition: The trial uses  $w$ , but it could have used  $w_1$  or  $w_2$ . Therefore,  $w_1$  and  $w_2$  might have each been less appropriate. But it would not follow that the  $w_1$ -type-object is less appropriate than the  $w_2$ -type-object, or vice versa. Thus, neither is more or less appropriate than the other. Choose freely between them.

### 3.1. The Minimal Ingredients for Competition

It is worth stressing that our experimental task involves the absolute minimal ingredients required for observing competition in all of its intricacy, including the role played by symmetry (there are just three words:  $w$ ,  $w_1$ , and  $w_2$ ). That these ingredients turn out to also be sufficient is remarkable, particularly in an experimental context that bears little resemblance to everyday conversational contexts (there is no speaker-hearer, for example). Our results therefore suggest that, when even the minimal ingredients for competition are present, humans instinctively and spontaneously employ something like the Competition Principle.

### 3.2. Beyond Human Reasoning

Non-human animals, such as monkeys, dogs, and birds, are capable of learning words, and they are also capable of applying strategic reasoning in various tasks. It has even been suggested that some monkeys apply a kind of Competition Principle in their natural alarm call system (Schlenker et al., 2014, 2016). A natural question is whether we can directly detect the Competition Principle at play in non-human animal behavior. Our experimental design is sufficiently simple that it should be straightforward to examine this question, something we hope to do in future work.

## AUTHOR CONTRIBUTIONS

All authors contributed to most aspects of the project, including designing the experiment. BB wrote the first draft of the manuscript. ID programmed the experiment and performed the statistical analysis. EC initiated the project.

## ACKNOWLEDGMENTS

We would like to thank Mélissa Berthet, Guillaume Dezechache, Emmanuel Dupoux, Joël Fagot, Andreas Haida, Nathan Klinedinst, Takashi Morita, Philippe Schlenker, Benjamin Spector, and Shane Steinert-Threlkeld for discussion (all errors remain ours). The research leading to these results has received funding from the European Research Council under the

European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 313610, from the ESRC under the Future Research Leaders scheme (ES/N017404/1), and was supported by ANR-10-IDEX-0001-02 PSL\* and ANR-10-LABX-0087 IEC.

## REFERENCES

- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effect models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bergen, L., Levy, R., and Goodman, N. D. (2016). Pragmatic reasoning through semantic inference. *Semant. Pragmat.* 9:20. doi: 10.3765/sp.9.20
- Davis, W. (2014). "Implicature," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Metaphysics Research Lab, Stanford University). Available online at: <https://plato.stanford.edu/archives/fall2014/entries/implicature/>
- Fox, D. (2007). "chapter 4: Free choice and the theory of scalar implicatures," in *Presupposition and Implicature in Compositional Semantics*, Palgrave Studies in Pragmatics, Language and Cognition Series, eds U. Sauerland, and P. Stateva (New York, NY: Palgrave Macmillan), 71–120.
- Fox, D., and Katzir, R. (2011). On the characterization of alternatives. *Natl. Lang. Semant.* 19, 87–107. doi: 10.1007/s11050-010-9065-3
- Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semant. Pragmat.* 4, 1–82. doi: 10.3765/sp.4.1
- Gazdar, G. (1979). *Pragmatics*. New York, NY: Academic Press.
- Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics*, Vol. 3, eds P. Cole and J. L. Morgan (New York, NY: Academic Press), 41–58.
- Horn, L. R. (1972). *On the Semantics of Logical Operators in English*. New Haven, CT: Yale University.
- Horn, L. R. (1989). *A Natural History of Negation*. Chicago, IL: University of Chicago Press.
- Jaeger, T. F. (2008). Categorical data analysis. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007
- Katzir, R. (2007). Structurally-defined alternatives. *Ling. Philos.* 30, 669–690. doi: 10.1007/s10988-008-9029-y
- Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Ling. Philos.* 27, 367–391. doi: 10.1023/B:LING.0000023378.71748.db
- Schlenker, P., Chemla, E., Arnold, K., Lemasson, A., Ouattara, K., Keenan, S., et al. (2014). Monkey semantics. *Ling. Philos.* 37, 439–501. doi: 10.1007/s10988-014-9155-7
- Schlenker, P., Chemla, E., and Zuberbühler, K. (2016). What do monkey calls mean? *Trends Cogn. Sci.* 20, 894–904. doi: 10.1016/j.tics.2016.10.004
- Soames, S. (1982). How presuppositions are inherited. *Ling. Inquiry* 13, 483–545.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02176/full#supplementary-material>

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Buccola, Dautriche and Chemla. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Advantages of publishing in Frontiers



## OPEN ACCESS

Articles are free to read  
for greatest visibility  
and readership



## FAST PUBLICATION

Around 90 days  
from submission  
to decision



## HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,  
and constructive  
peer-review



## TRANSPARENT PEER-REVIEW

Editors and reviewers  
acknowledged by name  
on published articles

## Frontiers

Avenue du Tribunal-Fédéral 34  
1005 Lausanne | Switzerland

**Visit us:** [www.frontiersin.org](http://www.frontiersin.org)

**Contact us:** [info@frontiersin.org](mailto:info@frontiersin.org) | +41 21 510 17 00



## REPRODUCIBILITY OF RESEARCH

Support open data  
and methods to enhance  
research reproducibility



## DIGITAL PUBLISHING

Articles designed  
for optimal readership  
across devices



## FOLLOW US

@frontiersin



## IMPACT METRICS

Advanced article metrics  
track visibility across  
digital media



## EXTENSIVE PROMOTION

Marketing  
and promotion  
of impactful research



## LOOP RESEARCH NETWORK

Our network  
increases your  
article's readership