

**SPEAKER AND LISTENER EFFECTS ON
THE PROCESSING OF PRAGMATIC MEANING**

by

Sarah Fairchild

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology

Spring 2018

© 2018 Sarah Fairchild
All Rights Reserved

**SPEAKER AND LISTENER EFFECTS ON
THE PROCESSING OF PRAGMATIC MEANING**

by

Sarah Fairchild

Approved:

Robert Simons, Ph.D.
Chair of the Department of Psychological and Brain Sciences

Approved:

George H. Watson, Ph.D.
Dean of the College of Arts and Sciences

Approved:

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Anna Papafragou, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Paul Quinn, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Peter Mende-Siedlecki, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Satoshi Tomioka, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

I am incredibly grateful to my advisor Anna Papafragou for her guidance and support throughout my graduate career. She has been an ideal role model for me, and I truly admire her as a deep thinker and writer. I can't thank her enough for helping me to develop the crazy ideas I brought her, teaching me how to communicate my research through her constructive feedback, and for always encouraging me to keep going.

I thank my committee members Paul Quinn, Peter Mende-Siedlecki, and Satoshi Tomioka for generously sharing their time and providing helpful insight on this project. Many thanks also go to the rest of the Cognitive faculty in the Department of Psychological and Brain Sciences, who have helped me grow as a scientist throughout my time at UD. In particular, I thank Helene Intraub for providing valuable research and career guidance.

Thank you to the members of the Language and Cognition Lab for being supportive colleagues, and for always providing thoughtful and creative feedback. I also thank the undergraduates in the lab who I have the pleasure of working with (including those who contributed to this dissertation – Hannah Schwartz, Kate Gordon, Ally Waller, Harry Rowe, and Jenna Miller). I was continually impressed by your motivation and enthusiasm, and learned so much about both research and teaching from you all.

I extend very special thanks to my undergraduate advisor, Janet van Hell. She introduced me to the world of language science, sparked my passion for research and

travel, and invested an incredible amount of time and resources into my training. She taught me so much about research, writing, mentoring, and life (especially the value of having one outside of the lab), and I would not be where I am today without her continued support.

Finally, thank you to my friends and family who have been more than understanding of my sometimes hermit-like lifestyle during the past four years – I look forward to catching up with everyone soon. :) And, of course, I have two fellow graduate students and friends at UD that deserve more thanks than I can fit in this acknowledgements section. Lee – thank you for always being ready for a coffee break, for running countless failed experiments with me, and for giving me many great game, TV show, and podcast recommendations to momentarily distract me from work. Steve – thank you for everything. Thank you for loving and supporting me in every aspect of work and life, for listening to many iterations of every presentation I’ve ever given, for encouraging me to start playing music again, and for going on adventures with me around our neighborhood and around the world.

This dissertation was made possible by the National Science Foundation (#1632849), a Graduate Fellowship from the Office of Graduate and Professional Education, and a Dissertation Award from the Department of Psychological and Brain Sciences.

TABLE OF CONTENTS

LIST OF TABLES.....	x
LIST OF FIGURES	xii
ABSTRACT.....	xiv

Chapter

1	INTRODUCTION	1
1.1	Speaker Effects on the Processing of Pragmatic Meaning	3
1.2	Listener Effects on the Processing of Pragmatic Meaning	6
1.3	Overview.....	9
2	SPEAKER EFFECTS ON PRAGMATIC MEANING:.....	11
3	SCALAR IMPLICATURE IN NATIVE AND NON-NATIVE SPEAKERS	11
2.1	Experiment 1	19
2.1.1	Participants.....	19
2.1.2	Materials and Procedure	20
2.1.2.1	Sentence Rating Task.....	20
2.1.2.2	Autism-Quotient Questionnaire	23
2.1.2.3	Chinese Cultural Attitudes Questionnaire	23
2.1.3	Results.....	24
2.1.3.1	Overall Analysis.....	24
2.1.3.2	Responder Bias Analysis	26
2.1.3.3	Individual Differences Analyses.....	28
2.1.4	Discussion	29
2.2	Experiment 2.....	31
2.2.1	Participants.....	32

2.2.2	Materials and Procedure	32
2.2.3	Results.....	34
2.2.3.1	Overall Analysis.....	34
2.2.3.2	Responder Bias Analysis	36
2.2.4	Discussion.....	37
2.3	Experiment 3	38
2.3.1	Participants.....	40
2.3.2	Materials and Procedure	40
2.3.3	Results.....	43
2.3.3.1	Overall Analysis.....	43
2.3.3.2	Responder Bias Analysis	45
2.3.4	Discussion.....	46
2.4	General Discussion	47
2.4.1	Theories of Non-Native Language Processing	47
2.4.2	The Pragmatics of Accent.....	49
2.4.3	Extensions and Future Directions	52
4	SPEAKER EFFECTS ON PRAGMATIC MEANING:.....	55
5	JUSTIFYING UNDER-INFORMATIVENESS IN NATIVE VS. NON-NATIVE SPEAKERS.....	55
3.1	Experiment 4.....	58
3.1.1	Participants.....	58
3.1.2	Materials and Procedure	58
3.1.3	Results.....	59
3.1.4	Discussion.....	61
3.2	Experiment 5.....	62
3.2.1	Participants.....	62
3.2.2	Materials and Procedure	62
3.2.3	Results.....	62
3.2.4	Discussion.....	63
3.3	Experiment 6.....	63

3.3.1	Participants.....	64
3.3.2	Materials and Procedure	64
3.3.3	Results.....	66
3.3.4	Discussion.....	67
3.4	Experiment 7.....	67
3.4.1	Participants.....	67
3.4.2	Materials and Procedure	67
3.4.3	Results.....	69
3.4.4	Discussion.....	71
3.5	General Discussion	71
6	LISTENER EFFECTS ON PRAGMATIC MEANING:.....	75
7	INDIVIDUAL DIFFERENCES IN SCALAR IMPLICATURE AND OTHER PRAGMATIC DOMAINS.....	75
4.1	Executive Function and Scalar Implicature	76
4.2	Theory of Mind and Scalar Implicature.....	77
4.3	The Present Study	79
4.4	Experiment 8.....	81
4.4.1	Participants.....	82
4.4.2	Materials and Procedure	82
4.4.2.1	Dual Scalar Implicature Task.....	83
4.4.2.2	Auditory Digit Span Task.....	84
4.4.2.3	Simple Scalar Implicature Task.....	85
4.4.2.4	Mind in the Eyes and Strange Stories Task	86
4.4.3	Results.....	87
4.4.3.1	Dual Scalar Implicature Task.....	87
4.4.3.2	Auditory Digit Span, Mind in the Eyes, and Strange Stories Tasks	89
4.4.3.3	Simple Scalar Implicature Task	89
4.4.4	Discussion	93
4.5	Experiment 9.....	96
4.5.1	Participants.....	98
4.5.2	Materials and Procedure	99

4.5.2.1	Metaphor Task	99
4.5.2.2	Indirect Request Task	99
4.5.2.3	Simple Scalar Implicature Task	101
4.5.2.4	Auditory Digit Span Task	101
4.5.2.5	Mind in the Eyes and Strange Stories Tasks.....	102
4.5.3	Results.....	102
4.5.3.1	Metaphor Task	102
4.5.3.2	Indirect Request Task	104
4.5.3.3	Simple Scalar Implicature Task.....	106
4.5.3.4	Correlations Among Tasks	108
4.5.4	Discussion.....	110
4.6	General Discussion	112
4.6.1	Executive Function and Pragmatic Judgments	114
4.6.2	Theory of Mind and Pragmatic Judgments.....	116
4.6.3	Pragmatic Judgments Within and Across Tasks.....	118
4.6.4	Future Directions	119
8	SUMMARY & CONCLUSIONS.....	121
5.1	Speaker Effects on Pragmatic Processing.....	122
5.2	Listener Effects on Pragmatic Processing.....	124
5.3	Broader Implications.....	125
	REFERENCES	129
Appendix		
A	STIMULI FOR EXPERIMENTS 1 AND 2	145
B	STIMULI FOR EXPERIMENT 3	148
C	STIMULI FOR THE DUAL SCALAR IMPLICATURE TASK IN EXPERIMENT 8	156
D	STIMULI FOR THE SIMPLE SCALAR IMPLICATURE TASK IN EXPERIMENTS 8 AND 9	158
E	STIMULI FOR THE METAPHOR TASK IN EXPERIMENT 9.....	159
F	STIMULI FOR THE INDIRECT REQUEST TASK IN EXPERIMENT 9.....	160
G	IRB APPROVAL FOR HUMAN SUBJECTS RESEARCH.....	163
H	IRB APPROVAL FOR HUMAN SUBJECTS RESEARCH.....	165

LIST OF TABLES

Table 2.1: Speaker bios for Experiment 1.	21
Table 2.2: Parameter estimates for a mixed-effects regression model predicting Sentence Ratings from Speaker and Sentence Type in Experiment 1.	26
Table 2.3: Speaker bios for Experiment 2.	33
Table 2.4: Parameter estimates for a mixed-effects regression model predicting Sentence Ratings from Speaker and Sentence Type in Experiment 2.	36
Table 2.5: Sample stimuli for Experiment 3.	41
Table 2.6: Parameter estimates for a mixed-effects regression model predicting Sentence Ratings from Speaker and Sentence Type in Experiment 3.	45
Table 3.1: Breakdown of justifications given in Experiment 4 by Speaker and Type.	60
Table 3.2: Results of the binary logistic regression model for Experiment 4.	61
Table 3.3: Results of the binary logistic regression model for Experiment 5.	63
Table 3.4: Results of the binary logistic regression model for Experiment 6.	66
Table 3.5: Results of the binary logistic regression model for Experiment 7.	71
Table 4.1: Examples of stimuli used in Experiment 8.	84
Table 4.2: Linear mixed-effects regression for the Dual Scalar Implicature Task results of Experiment 8, with Cognitive Load, Sentence Type, and their interaction included as fixed effects and crossed random intercepts for Participant and Item.	89
Table 4.3: Scores on all EF and ToM tasks in Experiment 8.	89

Table 4.4: Multiple linear regression predicting Pragmatic Bias (the number of “Bad” ratings of Under-Informative sentences in the simple Scalar Implicature task) in Experiment 8 from composite ToM and EF scores.	93
Table 4.5: Examples of stimuli used in the indirect request comprehension task in Experiment 2 (borrowed from Van Ackeren et al., 2012).	101
Table 4.6: Scores on all EF and ToM tasks in Experiment 9.	102
Table 4.7: Multiple linear regression analyses predicting scores on metaphor, indirect request and scalar implicature tasks in Experiment 9 from EF and ToM scores.	104

LIST OF FIGURES

Figure 2.1: Mean Sentence Ratings by Speaker for all Sentence Types in Experiment 1. Error bars indicate +/-1 S.E.M. Asterisks denote significance as follows: * $p < .05$, ** $p < .01$, *** $p < .001$	25
Figure 2.2: NNS Effect by Under-Informative rating in the Native Speaker condition in Experiment 1. A NNS Effect of 0 indicates no difference in ratings between speaker conditions, whereas a positive NNS Effect indicates higher ratings of under-informativeness for Non-Native Speakers as compared to Native Speakers.....	28
Figure 2.3: Mean Sentence Ratings by Speaker for all Sentence Types in Experiment 2. Error bars indicate +/-1 S.E.M. Asterisks denote significance as follows: * $p < .05$, ** $p < .01$, *** $p < .001$	35
Figure 2.4: NNS Effect by Under-Informative rating in the Native Speaker condition in Experiment 2. A NNS Effect of 0 indicates no difference in ratings between speaker conditions, whereas a positive NNS Effect indicates greater lenience towards under-informativeness from Accented Non-Native Speakers as compared to Native Speakers.....	37
Figure 2.5: Mean Sentence Ratings by Speaker for all Passage Types in Experiment 3. Error bars indicate +/-1 S.E.M. Asterisks denote significance as follows: * $p < .05$, ** $p < .01$, *** $p < .001$	44
Figure 2.6: NNS Effect by Some/All rating in the Native Speaker condition in Experiment 3. A NNS Effect of 0 indicates no difference in ratings between speaker conditions, whereas a positive NNS Effect indicates greater tolerance of under-informativeness from Accented Non-Native Speakers as compared to Native Speakers.....	46
Figure 3.1: Proportion of Inability Justifications in Experiments 4 (left), 5 (center) and 6 (right). Error bars represent +/- 1 S.E.M.....	61
Figure 3.2: Novel object used in Experiment 6 (“Zeg”) and its three functions.	65
Figure 3.3: Speaker descriptions used in Experiment 6.....	65

Figure 3.4: Mean Helpfulness Ratings (left) and the proportion of Same Teacher Choices (right) in Experiment 7. Error bars represent 95% confidence intervals.....	70
Figure 4.1: Examples of Control (left) and Load (right) patterns used in Experiment 8.	84
Figure 4.2: Example trial from the Mind in the Eyes Task (correct answer highlighted).....	87
Figure 4.3: Mean Sentence Ratings by Sentence Type and Cognitive Load condition in the Dual Scalar Implicature Task in Experiment 8. Error bars represent +1 S.E.M.....	88
Figure 4.4: Results of the Simple Scalar Implicature Task in Experiment 8. Error bars represent +1 S.E.M.....	93
Figure 4.5: Results of the Metaphor Task in Experiment 9. Error bars represent +1 S.E.M.	103
Figure 4.6: Results of the Indirect Request Task in Experiment 9. Error bars represent +1 S.E.M.	106
Figure 4.7: Results of the Simple Scalar Implicature Task in Experiment 9. Error bars represent +1 S.E.M.....	107
Figure 4.8: Correlations among pragmatic tasks in Experiment 9 (Top Left: Indirect Request and Metaphor, Top Right: Scalar Implicature and Metaphor, Bottom: Scalar Implicature and Indirect Request).	109

ABSTRACT

Successful communication requires a listener to reason about a speaker's intended meaning – the pragmatic level of meaning – in addition to the semantic meaning of the utterance. Pragmatic competence is an important communicative skill, yet it is also one which varies widely across contexts and individuals. What a listener interprets a speaker's utterance to mean is informed by knowledge of that speaker's abilities and preferences, as well as the surrounding context. Furthermore, it is well-documented that even adult listeners vary in the ease with which they make pragmatic inferences. While prior research has investigated the effects of linguistic context on pragmatic inference, it is not yet fully understood how variation in speaker and listener identities affect pragmatic processing. Investigating the effects of such variation is important to our understanding of human communication, which is not complete until we can account for the ways in which diverse speakers and listeners may intend and interpret meaning differently. This issue is particularly relevant given the growing number of opportunities to converse with individuals of different linguistic and cultural backgrounds. Accordingly, this dissertation explores variation in pragmatic inference from both speaker and listener perspectives. In Chapters 2 and 3, we investigate how pragmatic inference and later behavior is influenced by a speaker's linguistic abilities by manipulating (non-)native speaker identity. In Chapter 4, we investigate the relationship between an individual listener's pragmatic competence and their executive function and theory of mind abilities. The findings of this dissertation enrich our understanding of communication by demonstrating systematic differences

in pragmatic inference that are dependent on characteristics of both the speaker and the listener.

Chapter 1

INTRODUCTION

Much of our everyday communication requires “reading between the lines,” or making inferences about what our conversational partner intended but did not actually say. For example, at Thanksgiving when your aunt says “Can you pass the potatoes?” she is not asking about your physical ability – rather, she is making a request. In conversation, we seem to make these inferences with ease, enriching literal meanings with pragmatic meanings driven by our expectations about how communication works. In this case, you expect that your aunt will ask a question relevant to the current context and so you are easily able to infer that she is making an indirect request, and hand her the potatoes. Literal statements are enriched with pragmatic meanings in many other situations such as metaphor, irony, and even humor.

These inferences are clearly central to communication, yet there is considerable variation in when and by whom pragmatic inferences are made. Speaker knowledge is one contextual factor that has been shown to impact pragmatic inferences: the way we arrive at a literal vs. a pragmatic meaning of an utterance differs depending on who we are talking to and what that person knows about the situation at hand. Additionally, listeners differ in how well they are able to make pragmatic inferences. For example, children (e.g., Papafragou & Musolino, 2003) and adults with Autism (e.g., Noveck, Guelminger, Georgieff, & Labruyere, 2007) make fewer pragmatic inferences, and – importantly for the present work – even neurotypical adults vary in how likely they are to enrich a literal meaning with a

pragmatic one (e.g., Bott & Noveck, 2004). Specifying the nature of this variation bears directly on the practical issue of how intended meanings are transmitted in light of the inherent diversity of communicators. However, solid empirical evidence elucidating the cause of this variation is lacking. Additionally, little is known about the extent to which pragmatic abilities in one domain – indirect requests, for example – generalize to other pragmatic domains such as metaphor comprehension. Accordingly, this dissertation aims to study variation in pragmatic processing from multiple angles: Chapters 2 and 3 focus on speaker effects on the processing of pragmatic meaning while Chapter 4 addresses listener effects on pragmatic inference.

The specific case of pragmatic inference at the center of this dissertation is a case of conversational implicature – an inference made by the listener based on an expectation about what the speaker intended to communicate. According to Grice (1975), listeners expect that their interlocutors aim to produce utterances that are true (Maxim of Quality), informative (Maxim of Quantity), relevant (Maxim of Relevance), and clear (Maxim of Manner). Because people strongly expect speakers to follow these maxims, they will often pragmatically enrich the literal semantic meaning of an utterance that appears to be in violation of the maxims, making an inference about what the speaker intended. For example, a sentence such as “Some giraffes have long necks” appears to violate the Maxim of Quantity: it is under-informative, because the speaker used the weaker term in a logical scale (‘some’) when s/he could have used a stronger, more informative scalar term (‘all’). In many contexts, this utterance will lead the hearer to infer that not all giraffes have long necks (an inference known as *scalar implicature*; see Grice, 1975; Sperber & Wilson, 1986; Horn, 1972; Horn, 1984; Hirschberg, 1985; Carston, 1995; Levinson, 2000). The theoretical framework

of scalar implicature assumes that the hearer goes through several steps to arrive at the pragmatic meaning: first, the literal meaning is processed, then the other alternatives on the quantifier scale are accessed (e.g., *many*, *most*, *all*), then the hearer takes the so-called “epistemic step” and reasons about the intended meaning of the speaker in order to rule out the stronger alternatives (Sauerland, 2004; Hochstein, Bale, Fox, & Barner, 2014). Judgments of these under-informative statements with scalar quantifiers have been used extensively in the literature to assess pragmatic ability, making these sentences an excellent test case for the present set of studies. Additionally, scalar implicature is known to be heavily context-dependent, with characteristics of both the speaker and the hearer affecting such context. In the remainder of this introduction, we sketch the theoretical background to speaker and listener effects on the processing of pragmatic meaning and relate such effects to the case of scalar implicature.

1.1 Speaker Effects on the Processing of Pragmatic Meaning

Because scalar implicature is thought to require consulting another person’s mental state (their intentions, specifically), a number of studies have investigated whether the knowledge state of the speaker affects scalar inferences. For example, in a study measuring reading times, Bergen and Grodner (2012) showed that participants made stronger “not all” implicatures when the speaker was highly knowledgeable of the topic at hand, for example: “I *meticulously* compiled the investment reports. Some of the real estate investments lost money.” In this case, the speaker was very likely to have meant that the stronger alternative “all” was false, since they were thorough in their work. Conversely, participants generated weaker implicatures in cases when the speaker was less knowledgeable, for example: “I *skimmed* the investment reports. Some of the real estate investments lost money.” Here, the speaker may not have

known whether all of the real estate investments lost money, and thus may not have intended the implicature. Similar results demonstrating comprehenders' sensitivity to speaker knowledge have been found in a different paradigm that manipulated the visual access of the speaker (Breheny, Ferguson & Katsos, 2013).

Thus, speaker knowledge of the context at hand affects processing of pragmatic meaning. This type of speaker knowledge can be considered a *situational* property of the speaker - the speakers in Bergen and Grodner (2012) and Breheny et al. (2013) were less knowledgeable on a portion of trials, and fully knowledgeable on others. This can be contrasted with *stable* properties of the speaker like gender, intelligence, or language background – characteristics that remain fairly stable over time and are less likely to bear directly on the content of utterances. For example, the same message would be communicated if a woman said “Some giraffes have long necks” as compared to if a man uttered the same statement. However, these stable speaker properties have the potential to impact pragmatic processing: if a person known to have poor world knowledge says “Some giraffes have long necks,” it is unclear whether he or she intended the “not all” implicature, or whether they simply do not know that all giraffes have long necks. Prior research has not focused on stable speaker properties, therefore it is as of yet unknown whether listeners take into account these types of speaker characteristics when considering the intended meaning of an utterance. It may be that situational knowledge and stable, general knowledge are integrated similarly by listeners during pragmatic processing. Alternatively, a speaker's knowledge of the present situation could influence pragmatic inferences to a greater extent than a speaker's general knowledge as situational knowledge may be more critical to the immediate success of the conversation.

One way that stable speaker knowledge (linguistic knowledge, specifically) can be manipulated is by comparing interpretation of infelicities such as “Some giraffes have long necks” uttered by native and non-native speakers of English. While a native listener might infer that a native speaker made a statement that carries a false implicature (“Not all giraffes have long necks”), they might be more tolerant of such statements from a non-native speaker with poorer linguistic knowledge who may not have intended the implicature. Manipulating non-native speaker status is an excellent test case for stable speaker properties as much research in recent years has demonstrated that individuals are sensitive to the language background of a speaker, and integrate that information into other types of linguistic processing. For example, listeners are less sensitive to grammatical errors produced by non-native speakers. In an ERP study, Hanulikova, Van Alphen, Van Goch, and Weber (2012) found that participants showed the typical P600 response to syntactic violations (grammatical gender errors) and N400 to semantic violations (e.g., the Dutch translation of “It was very cold last night, so I put a thick **evening* on my bed”) when these errors were produced by a native speaker. When listening to non-native speech, semantic errors still elicited an N400 but syntactic violations from foreign-accented non-native speakers failed to elicit a P600, a finding that has been confirmed by more recent work (Grey & Van Hell, 2017). There is also clear practical motivation for manipulating non-native speaker status: as of the most recent census, there were about 51 million Americans (~16% of the population) who spoke a language other than English at home (U.S. Census Bureau, 2011), and thus it is important to understand how non-native speakers are understood by native listeners as these interactions are becoming increasingly frequent in our society. Manipulating non-native speaker status in the

study of scalar implicature would not only reveal whether or not comprehenders are sensitive to stable properties of the speaker in pragmatic judgments, it will also extend previous work on the comprehension of non-native speech to the domain of pragmatics.

1.2 Listener Effects on the Processing of Pragmatic Meaning

Individuals vary in the extent to which they successfully make pragmatic inferences. In response to the sentence “Some giraffes have long necks,” some listeners infer the pragmatic meaning (not all have long necks) while others respond to the literal meaning (some have long necks, in fact they all do). Papafragou and Musolino (2003) found that adults rejected such statements 92.5% of the time, Noveck (2001) reported that adults rejected such sentences 59% of the time, and in a study by Guasti et al. (2005) adults rejected under-informative statements only 50% of the time. In observing this variation more closely, some studies have found that groups of adults have consistent response patterns within a single task, with Logical Responders accepting under-informative sentences and Pragmatic Responders rejecting them (e.g., Noveck & Posada, 2003; Bott & Noveck, 2004); other studies, however, have included an Inconsistent Responders category (Politzer-Ahles, Fiorentino, Jiang, & Zhou, 2013; Heyman & Schaeken, 2015). Assuming that individuals can be neither perfectly logical nor perfectly pragmatic, and that pragmatic responding should best be treated as a continuum (cf. Degen & Tanenhaus, 2015), an important question remains as to *why* such individual variation in pragmatic judgments emerges. A reasonable hypothesis is that stable participant characteristics (alongside task characteristics) can

shift individuals' responses to scalar statements from more to less pragmatic. We focus on two such individual characteristics, Executive Function and Theory of Mind.¹

Differences in Executive Function (EF) – the collection of control processes such as working memory, inhibition, and task-switching – might be responsible for variability in pragmatic judgments, given the complex, multi-step process of computing an implicature that we have outlined above (e.g., Grice, 1975; Horn, 1972, 1984; Sauerland, 2004; cf. also Carston, 1995; Sperber & Wilson, 1986, for a different but still richly inferential model). Eye-tracking and reading time studies indicate that processing scalar implicatures often requires additional time and is therefore cognitively costly compared to processing the literal, semantic content of the utterance (e.g., Noveck & Posada, 2003; Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006; Huang & Snedeker, 2009; Bott, Bailey, & Grodner, 2012; but see Grodner, Klein, Carbary, & Tanenhaus, 2010; Degen & Tanenhaus, 2015; and Huang & Snedeker, 2018 for contexts that decrease the processing demands of scalar implicatures). Individuals with more available cognitive resources – especially working memory - may be more likely to recruit these resources and complete the steps necessary to calculate the implicature.

Alternatively, differences in Theory of Mind (ToM) – the ability to reason about another person's mental state – may better explain variation in scalar implicature. Recall that classic models of pragmatics assume that scalar implicature computation involves the listener's assessment of the speaker's epistemic state (Grice,

¹ Other participant characteristics such as personality traits have also been proposed as sources of variation for scalar implicature computation (Heyman & Schaeken, 2015; Antoniou et al., 2016). However, the evidence for such proposals is weak at best (e.g., Antoniou et al., 2016), hence we do not discuss them further here.

1975; Horn, 1972, 1984; Sauerland, 2004; Carston, 1995; Sperber & Wilson, 1986). Recent psycholinguistic evidence supports this assumption (Bergen & Grodner, 2012; Breheny, Ferguson, & Katsos., 2013; see also Hochstein, Bale, Fox, & Barner, 2014 and Papafragou, Friedberg, & Cohen, in press, for developmental evidence). For instance, adults are more likely to draw scalar inferences when the speaker is knowledgeable as compared to when the speaker has partial knowledge about the topic (and therefore may not know whether a stronger statement is true; Bergen & Grodner, 2012). Given that the ability to take someone else's perspective varies across individuals (e.g., Apperly, 2012), it is likely that individuals with poorer ToM abilities are less likely to calculate scalar implicatures and thus might surface as logically-biased responders. However, this hypothesis has not been tested directly.

Empirical investigations of individual differences in scalar implicature have been limited in scope and have yielded mixed results. Of the two studies that took multiple measures of participants' cognitive abilities, one did not test ToM (Heyman & Schaeken, 2015) and the other was a smaller-scale study ($N = 63$) that could not include ToM in the analyses due to this sample size limitation (Antoniou, Cummins, & Katsos, 2016). Although both studies analyzed participants' EF (working memory specifically), their results are inconsistent with one another: Heyman and Schaeken (2015) reported no association with EF, while Antoniou et al. (2016) found EF to be predictive of pragmatic responding on a scalar implicature task.

Furthermore, many prior investigations of scalar implicature have used this type of inference as a marker of general pragmatic competence, yet it is unknown how an individual's ability to compute a scalar implicature relates to their other pragmatic abilities. Assuming that understanding an indirect request, for example, also requires

processing the literal meaning of the utterance then reasoning about the speaker's mental state to arrive at the intended meaning, indirect request judgments should be related to scalar implicature judgments, but there have been no empirical investigations to confirm this hypothesis.

1.3 Overview

The body of this dissertation consists of three independent sets of experiments that investigate the contributions of speaker (Chapters 2 and 3) and listener (Chapter 4) identity to the processing of pragmatic meaning. Each of these chapters was originally developed as a journal publication that is either submitted (Fairchild & Papafragou, Resubmitted; Fairchild & Papafragou, Submitted) or in preparation (Fairchild & Papafragou, In Prep), and their structure has remained somewhat independent as we collected them into a single work here. This can explain some repetition in the introductory material for each chapter. In Chapter 2, we present a series of sentence rating experiments in which we manipulate the (non-native) identity of the speaker uttering under-informative sentences (e.g., "Some dogs are mammals.") This chapter demonstrates an effect of speaker identity on scalar implicature and is also highly relevant to theories of non-native speech processing which are discussed therein. In Chapter 3, we manipulate the non-native speaker identity of under-informative speakers in order to examine how under-informativeness is processed and justified in more ecologically valid learning and social interaction contexts. In Chapter 4, we examine listener effects on pragmatic judgments by investigating the unique contributions of ToM and EF to scalar implicature computation, and expanding the results to cover metaphor and indirect request comprehension. This chapter aims to disentangle the roles of ToM and EF in scalar implicature calculation but also to

ascertain the relationships among various pragmatic domains and the cognitive abilities they presuppose. Finally, in Chapter 5 we summarize and synthesize our major findings.

Chapter 2

SPEAKER EFFECTS ON PRAGMATIC MEANING:

SCALAR IMPLICATURE IN NATIVE AND NON-NATIVE SPEAKERS

There are over 51 million Americans who speak a language other than English at home (U.S. Census Bureau, 2011). Worldwide, it is estimated that about half of the population is bilingual (Grosjean, 2010). Psycholinguistic research on bilingualism has flourished in recent decades, with investigations focusing on the mechanisms allowing bilinguals to switch effortlessly between their two languages (e.g., Poplack, 1980; Clyne, 1987; Milroy & Muysken, 1995; Myers-Scotton & Jake, 2000; Moreno, Federmeier, & Kutas, 2002), the organization of the bilingual mental lexicon (e.g., De Bot & Schreuder, 1993; Kroll & Stewart, 1994; Green, 1998; Wei, 2001; Pavlenko, 2009), and the potential cognitive advantages of bilingualism (e.g., Bialystok, Craik, Klein, & Viswanathan, 2004; Costa, Hernández, Costa-Faidella, & Sebastián-Gallés, 2009; Prior & MacWhinney, 2010; Hilchey & Klein, 2011; Sebastián-Gallés et al., 2012; Paap & Greenberg, 2013). However, the prevalence of bilingualism has consequences that extend beyond the bilingual individuals themselves. Monolinguals frequently interact with individuals who speak more than one language, many of whom are non-native speakers of the language. Because second language phonology is notoriously difficult for adults to acquire (Flege, Yeni-Komshian, & Liu, 1999; Piske et al., 2001; Golestani & Zatorre, 2009), many of these non-native speakers speak with a foreign accent. A recent line of research exploring how speech from non-native

speakers is processed by native listeners suggests that a foreign accent impacts communication in a number of ways.

Most obviously, foreign-accented speech poses a challenge for intelligibility. Non-native speakers may not be able to properly produce the phonemic inventory of their second language, causing listeners to have difficulty in comprehension. Indeed, participants are slower to process sentences uttered by non-native speakers, and rate such foreign-accented sentences as less comprehensible (Munro & Derwing, 1995). Unsurprisingly, research also demonstrates that word identification is impaired by non-native speech in both adult (Bent & Bradlow, 2003; Clarke & Garrett, 2004) and infant (van Heugten & Johnson, 2014) native listeners, although listeners are generally able to quickly adapt to a foreign accent (Clarke & Garrett, 2004; Baese-Berk, Bradlow, & Wright, 2013; van Heugten & Johnson, 2014).

Foreign-accented speech has broader effects on language comprehension: a recent study using Event-Related Potentials (ERPs) found that neural signatures differed in response to native and non-native (Turkish-accented) speech errors in Dutch (Hanulikova, Van Alphen, Van Goch, & Weber, 2012). Participants showed the typical P600 response to syntactic violations (grammatical gender errors) and N400 to semantic violations (e.g., the Dutch translation of “It was very cold last night, so I put a thick **evening* on my bed”) when these errors were produced by a native speaker. When listening to non-native speech, semantic errors still elicited an N400 but syntactic violations from foreign-accented non-native speakers failed to elicit a P600 – a finding that was replicated for Chinese-accented English. The explanation that the authors adopt for these results is that listeners expect that non-native speakers will produce syntactic violations because of their lower second language proficiency, but

do not expect non-native speakers to produce semantically bizarre utterances. Additional work using ERPs to investigate the effects of a foreign accent on semantic processing has found differences in the N400 component to native and non-native semantic errors, although results conflict as to whether the component was attenuated or amplified in non-native speech (Goslin, Duffy, & Floccia, 2012; Romero-Rivas, Martin, & Costa, 2015). Relatedly, very recent work shows that listeners tend to respond to semantically implausible sentences (e.g., “The mother gave the candle the daughter”) as though they were their plausible counterparts (“The mother gave the candle to the daughter”) when the sentences are uttered by non-native speakers (Gibson et al., 2017).

Non-native speaker status also has consequences about how the content of an utterance is evaluated offline. For example, non-native speakers are deemed less credible than their native speaker counterparts: general knowledge statements like “Ants don’t sleep” that are true but not widely known are judged as less likely to be true when they are spoken by a non-native speaker with a thick foreign accent compared to when they are spoken by a native speaker (Lev-Ari & Keysar, 2010). Moreover, native listeners judge non-native speakers’ narrative stories as more vague, are less likely to detect changes to a non-native speech stream in a change detection paradigm, and have poorer memory for sentences spoken by non-native speakers (Lev-Ari & Keysar, 2012). Even very young children hold negative biases towards non-native speakers – for example, they are less willing to befriend a non-native speaker and are more likely to trust a novel label offered by a native speaker as compared to one offered by a non-native speaker (Kinzler, Dupoux, & Spelke, 2007)

There are several possible explanations as to why these differences arise between non-native, foreign-accented speech and native speech (cf. Lev-Ari, 2015). One type of theory – which can be called the Intelligibility-Based account – is that a foreign accent is an additional processing demand that alters language comprehension because it is highly variable and perceptually distinct from the listener’s own accent (e.g., Davis et al., 2005; Floccia, Goslin, Girard, & Konopczynski, 2006). A prediction that this account makes is that foreign-accented speech should be processed similarly to regional-accented speech and noisy speech. Intelligibility factors can explain the attenuated P600 (and intact N400, an earlier component that is arguably more automatic in nature than the P600 which reflects a reanalysis process; Kutas & Federmeier, 2011) to non-native speech observed by Hanulikova et al. (2012) as a result of cognitive overload, with few resources available for reanalysis of syntactic errors. As for offline effects, the Intelligibility-Based account would argue that non-native speakers are rated as more vague (Lev-Ari & Keysar, 2012) because they are actually more difficult to understand. Likewise, an Intelligibility-Based account could argue that participants in Gibson et al.’s (2017) study may have responded to semantically implausible sentences as though they were semantically plausible because of the costs imposed by processing a foreign accent.

Alternatively, what can be called Expectation-Based accounts argue that listeners have different expectations about the speech of non-native speakers from the outset; specifically, the expectation is that non-native speech is highly variable and that grammatical (and possibly semantic) errors will occur more often than in native speech (Niedzielski, 1999; Lev-Ari, 2015). These expectations cause individuals to rely more on top-down extra-linguistic information such as visual context and

background knowledge of the situation. Expectation-Based accounts would argue that the results of Hanulikova et al. (2012) and Gibson et al. (2017) stem from the expectation that non-native speakers make more grammatical errors, leading to processing differences in the earliest moments of speech comprehension. Similarly, an Expectation-Based account would argue that listeners judge non-native speakers' narratives as more vague (Lev-Ari & Keysar, 2012) because of expectations held about the quality of non-native speech that affect speech processing.

Currently, researchers have yet to manipulate expectations about (native vs. non-native) speaker identity directly while keeping intelligibility constant. This type of manipulation would provide the strongest support for the role of expectations about speaker identity on the way language is processed. The difficulty in accomplishing this is that non-native speech, as has been noted above, is more challenging to understand compared to native speech. Thus in practice, the role of expectations about the speech of non-native speakers is difficult to isolate from the role of the processing cost incurred by a foreign accent. Here we resolve this difficulty by using a sentence rating task to compare how readers react to *written* sentences that they believe were uttered by a non-native vs. a native speaker. The advantage of using written materials is that processing demands arising from the sentences themselves are equivalent across the native and non-native speaker manipulations (across participants, the same sentences are attributed to different types of speaker). Thus any asymmetries in how sentences are processed across speaker conditions can be unambiguously attributed to expectations about speaker identity. The present paradigm therefore allows us to isolate and probe the role of expectations on non-native speech processing in the absence of intelligibility factors.

Unlike previous work that has focused on syntactic or semantic processing of native vs. non-native speech (e.g., Hanulikova et al., 2012; Gibson et al., 2017), in the present study we focus specifically on how comprehenders interpret the pragmatic meaning of utterances produced by native vs. non-native speakers. Pragmatic aspects of meaning go beyond the semantic, literal meaning of a sentence and include contextual inferences that hearers compute as part of what the speaker intended to convey. Pragmatic aspects of meaning are driven by expectations about how rational communication works. Following Grice (1975), one can assume that interlocutors are mutually invested in a cooperative activity. According to Grice, listeners expect that their interlocutors aim to produce utterances that are true (Maxim of Quality), informative (Maxim of Quantity), relevant (Maxim of Relevance), and clear (Maxim of Manner). Because people strongly expect speakers to follow these maxims, they will often pragmatically enrich the literal semantic meaning of an utterance that appears to be in violation of the maxims, making an inference about what the speaker intended. For example, a sentence such as “Some giraffes have long necks” appears to violate the Maxim of Quantity: it is under-informative, because the speaker used the weaker term in a logical scale (‘some’) when s/he could have used a stronger, more informative scalar term (‘all’). In many contexts, this utterance will lead the hearer to infer that not all giraffes have long necks (an inference known as *scalar implicature*; see Grice, 1975; Sperber & Wilson, 1986; Horn, 1972; Horn, 1984; Hirschberg, 1985; Carston, 1995; Levinson, 2000).

In the literature, judgments about under-informative sentences have been used as a test of whether a logical or pragmatic interpretation of the sentence has been reached: one might accept a sentence such as “Some giraffes have long necks” since

the sentence is semantically/logically true; alternatively, one might reject the sentence since it pragmatically gives rise to a scalar implicature that is itself false (“Not all giraffes have long necks”; Noveck, 2001). In general, judgment tasks that have used a 3-point or 5-point Likert scale have shown that adults (and even 5-year-old children) judge under-informative statements as more acceptable than completely false statements but not as good as completely true (and informative) statements (e.g., Katsos & Bishop, 2011; Davies, Andres-Roqueta, & Norbury, 2016). However, the degree to which comprehenders adopt logical or pragmatic interpretations varies with task demands and individual preferences (e.g., Bott & Noveck, 2004; Noveck, 2001; Guasti et al., 2005; Ozturk & Papafragou, 2016; Noveck & Posada, 2003; Feeney, Scafton, Duckworth, & Handley, 2004; Hunt et al., 2011; Tavano & Kaiser, 2010). It is unclear what individual characteristics contribute to this variability, but several options have been proposed, including social-communicative ability (Nieuwland, Ditman, & Kuperberg, 2010), executive function (De Neys & Schaeken, 2007), and participants’ uncertainty about the Question under Discussion (Degen & Tanenhaus, 2015).

In this chapter, we use a (non-binary) pragmatic judgment task to assess how expectations about non-native speakers affect comprehenders’ interpretation of utterances produced by native and non-native speakers. Across three experiments, we present adult native speakers of English with written under-informative sentences and attribute these sentences to either native or non-native speakers of English. Participants then rate the sentences on the basis of how much sense they make. If altering beliefs (and corresponding expectations) about the speaker can change sentence interpretation, then judgements should change depending on speaker status.

One possibility is that participants might judge under-informative (but true) sentences more negatively when uttered by non-native compared to native speakers for reasons related to biases against non-native speakers (cf. Lev-Ari & Keysar, 2010; Kinzler et al., 2007). Alternatively, under-informative statements might be given higher ratings when believed to have been produced by a non-native compared to a native speaker of English. Since non-native speakers are expected to be less accurate in their lexical (and other linguistic) choices, they may be seen as more likely to (unintentionally) produce under-informative utterances. Sins of information omission may thus be more likely to be forgiven in non-native speakers. This line of reasoning is in accordance with previous findings showing that listeners penalize grammatical violations less for non-native than for native speakers (Hanulikova et al., 2012; Gibson et al., 2017),

Because of the well-established variability in how people judge under-informative statements, we further investigate whether sensitivity to speaker identity in pragmatic judgments varies across the continuum of responding preferences (i.e., more logical vs. more pragmatic responders). One might expect that speaker sensitivity is higher in individuals who consistently respond to the pragmatically-enriched meaning of an utterance compared to those who tend to respond only to the literal meaning of an utterance within a task. This is because comprehenders who tend to adopt a pragmatic final interpretation recognize that the choice of one scalar term (e.g., ‘some’) over another (e.g., ‘all’) has pragmatic implications, and have reasoned about the alternatives that the speaker could have used but did not, as well as the reasons that the speaker must have had for using a less-than-optimal alternative (see Horn, 1972; 1984; Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Gualmini, Crain, Meroni, Chierchia, & Guasti, 2001; Barner, Brooks, & Bale, 2011; Ozturk &

Papafragou, 2015; Skordos & Papafragou, 2016). In the case of logical responders, alternatives to the present utterance may never have been considered (Bott & Noveck, 2004) or pragmatically-enriched meanings may have been considered but later rejected in favor of a literal interpretation. Thus, more pragmatically-inclined responders may be more sensitive to properties of the speaker's identity and how these properties affect the choice of a linguistic stimulus and its intended meaning compared to people who tend to adopt a logical/semantic interpretation.

2.1 Experiment 1

In Experiment 1, we administered a Sentence Ratings task to compare how under-informative statements (among other types of statements) are processed when attributed to native vs. non-native speakers. We then investigated whether such speaker sensitivity varies across individuals. We also measured participants' general social-communicative ability and cultural attitudes towards non-native speakers, and related these measures to participants' ratings of pragmatic infelicities from different kinds of speakers.

2.1.1 Participants

One hundred and fourteen native speakers of English aged 18-38 years ($M = 28.14$, $SD = 4.16$) living in the United States, 50 of whom were female, were recruited from Amazon's Mechanical Turk to participate in the experiment. Participants were compensated at a rate of \$0.10 per minute for a total of \$1.50.

2.1.2 Materials and Procedure

2.1.2.1 Sentence Rating Task

Eighty sentences were created for the Sentence Ratings task (see Appendix A), half beginning with *some* and half beginning with *all*. Sentences were based on general knowledge and were evenly distributed across four Sentence Types: True but Under-Informative sentences with *some* (henceforth Under-Informative; “Some people have noses with two nostrils”), True and Felicitous sentences with *some* (henceforth, True (Some); “Some people have dogs as pets in the house”), True and Felicitous sentences with *all* (henceforth, True (All); “All snow is cold and can melt into water”), and False sentences with *all* (henceforth, False; “All women are doctors who went to medical school”). The critical trials consisted of the Under-Informative sentences that were literally true but pragmatically odd (in the example above, all people have noses with two nostrils), and the other three Sentence Types were treated as control sentences. The four sentence types did not differ from one another in sentence length as measured in words or syllables (all p 's > .1).

Speaker bios were created to accompany the sentences. Each bio either gave a short description of Emma, a native English speaker with a strong Boston accent (Native Speaker condition), or Yuqi, a native speaker of Mandarin Chinese with a strong Chinese accent (Non-Native Speaker condition). Thus, in both cases the speaker had an accent, and the only difference between the two was the non-native speaker status. There were two versions of each Speaker condition, in which the speaker's hobbies and major varied. This was done so as not to present two nearly identical bios to the same participant. Thus there were four total bios, presented in Table 2.1.

Table 2.1: Speaker bios for Experiment 1.

Native Speaker	Non-Native Speaker
<p>Emma is a college student at the University of Delaware, majoring in History/Sociology. She is doing well in her classes and plans to be a high school teacher after graduation. Emma moved with her family to Delaware from Boston, and her classmates often tease her about her strong Boston accent. She laughs it off, because she knows they are just having fun. In her spare time, Emma likes to hike/run and play the piano/guitar.</p>	<p>Yuqi is a college student at the University of Delaware, majoring in History/Sociology. She is doing well in her classes and plans to be a high school teacher after graduation. Yuqi moved with her family to Delaware from China, and her classmates often tease her about her strong Chinese accent. She laughs it off, because she knows they are just having fun. In her spare time, Yuqi likes to hike/run and play the piano/guitar.</p>

Although the speaker bios for the Native and Non-Native Speaker conditions were nearly identical, we wanted to ensure that participants did not assume that one of the two speakers (or one of the two versions of the speakers) was more knowledgeable, particularly in terms of the topics in the critical Under-Informative sentences. Thus, we recruited an additional 60 participants from Mechanical Turk living in the United States. Participants read one of the four speaker bios and were then presented with each of the topics in the Under-Informative sentences (20 total). For example, “Dogs” would be the topic for the sentence “Some people have dogs as pets in the house.” For each topic, participants rated on a scale from 0 to 100 how much they felt the person in the description knew about the topic. Mean ratings ($M = 61.42$, $SD = 9.38$) did not differ across the four speaker bios, nor did ratings for any one topic (all p 's $> .1$). Thus, any potential differences between speakers in the Sentence Ratings task is unlikely to be attributed to perceptions of the speaker's general world knowledge.

The Sentence Ratings task consisted of two blocks: a Native Speaker block and a Non-Native Speaker block (counterbalanced across participants). Sentences within each block were evenly distributed across the four Sentence Types (10 of each), and were presented in a random order. Thus, both Speaker (Native, Non-Native) and Sentence Type (Under-Informative, True (Some), True (All), False) were treated as within-subjects factors. At the start of each block, one of the four speaker bios appeared on the screen. Participants were instructed to read carefully in order to answer the comprehension questions that followed, and were given as much time as they needed to read the paragraph before moving on. The speaker bio was followed by three multiple-choice questions about the speaker, presented in a random order (“Where is Emma/Yuqi from?”, “What is Emma/Yuqi majoring in?”, “What does Emma/Yuqi like to do in her spare time?”). Performance on these comprehension questions was quite high (88%), indicating that participants had fully read and understood the speaker bios. All participants correctly answered at least one of the two comprehension questions for each Speaker. Participants were then instructed that they would be reading 40 sentences that were originally uttered by the person they had just read about, and that their job was to rate how “Good” each sentence was on a five-point scale where 1 is “Very bad” and 5 is “Very good.” Participants were instructed that a good sentence is one that makes perfect sense, and a bad sentence is one that makes no sense at all. Additionally, participants were told that because a given utterance can make more or less sense, they should make use of the intermediate values on the scale for sentences that were neither very good nor very bad.

On each trial, a sentence appeared in the center of the screen with the ratings scale below. The speaker bio was always present at the top of the screen, in a muted

gray color. Participants could move the marker on the scale to indicate their desired rating. The marker snapped into one of five possible positions as it was moved (i.e., movement was not continuous). The five locations were not marked on the scale, but participants were instructed beforehand that there were five possible choices. As participants made their response, a face attached to the scale changed its expression (a frown for low ratings, a smile for high ratings, with three intermediate faces). Participants could take as long as they needed to make a response.

2.1.2.2 Autism-Quotient Questionnaire

Following the Sentence Ratings task, participants completed the Communicative Subscale of the Autism-Quotient Questionnaire (AQ-COMM; Baron-Cohen et al., 2001). The questionnaire consists of 10 statements designed to probe social communication skills (e.g., “I am often the last to understand the point of a joke,” “I know how to tell if someone listening to me is getting bored”). For each statement, participants indicated how true it was of themselves. The standard scoring method was used, calculating a total score out of 10 of the number of autistic traits the person possessed.

2.1.2.3 Chinese Cultural Attitudes Questionnaire

Finally, participants completed a Chinese Cultural Attitudes questionnaire, adapted from the American Attitudes Toward Chinese Americans & Asian Americans survey conducted by the Committee of 100. The questionnaire assesses how strongly individuals believe in cultural stereotypes of Chinese-Americans, both positive and negative. Participants were asked to rate how strongly they agreed with fourteen statements about Chinese-Americans (e.g., “Chinese-Americans are overly aggressive

in the workplace,” “Chinese-Americans have strong family values”). A total score was calculated for each participant based on the average agreement with cultural stereotypes.

2.1.3 Results

2.1.3.1 Overall Analysis

Linear mixed-effects regressions were performed on Sentence Rating data for Experiment 1 and all subsequent experiments in this chapter using the *nlme* package (Pinheiro et al., 2017) for the R Project for Statistical Computing v3.2.2 (R Core Team, 2015). This method of analysis has several benefits, particularly for repeated-measures data like ours. For instance, variability across both participants and items can be accounted for in the same model, rather than needing to conduct separate by-participants and by-items analyses. Speaker (Native Speaker, Non-Native Speaker), Sentence Type (Under-Informative, True (Some), True (All), False), and the interaction between the two were included in the model as fixed effects, with crossed random intercepts for Participants and Items. Mean Sentence Ratings are presented in Figure 2.1. Sentence Ratings differed significantly across Speakers, $\chi^2(1) = 12.577, p < .001$, and Sentence Types, $\chi^2(3) = 3751.703, p < .001$. Planned contrasts (presented in Table 2.2) indicate that Sentence Ratings were higher in the Non-Native Speaker ($M = 3.41, SD = 1.44$) condition as compared to the Native Speaker ($M = 3.32, SD = 1.46$) condition, $p < .001$. Additionally, Under-Informative ($M = 2.93, SD = 1.40$) sentences were rated higher than False ($M = 2.24, SD = 1.30$) sentences, $p < .001$, but worse than True (Some) ($M = 4.20, SD = 0.97$) sentences, $p < .001$. True (All) ($M = 4.09, SD = 1.05$) sentences were rated higher than True (Some) sentences, $p < .001$.

These main effects were qualified by a significant interaction between Speaker and Sentence Type, $\chi^2(3) = 10.715, p = .013$. Post-hoc tests (Bonferroni-corrected for multiple comparisons) indicated that ratings of Under-Informative sentences were higher in the Non-Native Speaker ($M = 3.02, SD = 1.39$) condition than in the Native Speaker ($M = 2.85, SD = 1.41$) condition ($p = .004$). In other words, participants were more accepting of under-informativeness when it was attributed to a Non-Native speaker. Ratings of True (Some), True (All) and False sentences did not differ by Speaker (all p 's $> .05$), indicating that the effect was selective to under-informative sentences (and did not extend to true or completely false statements).

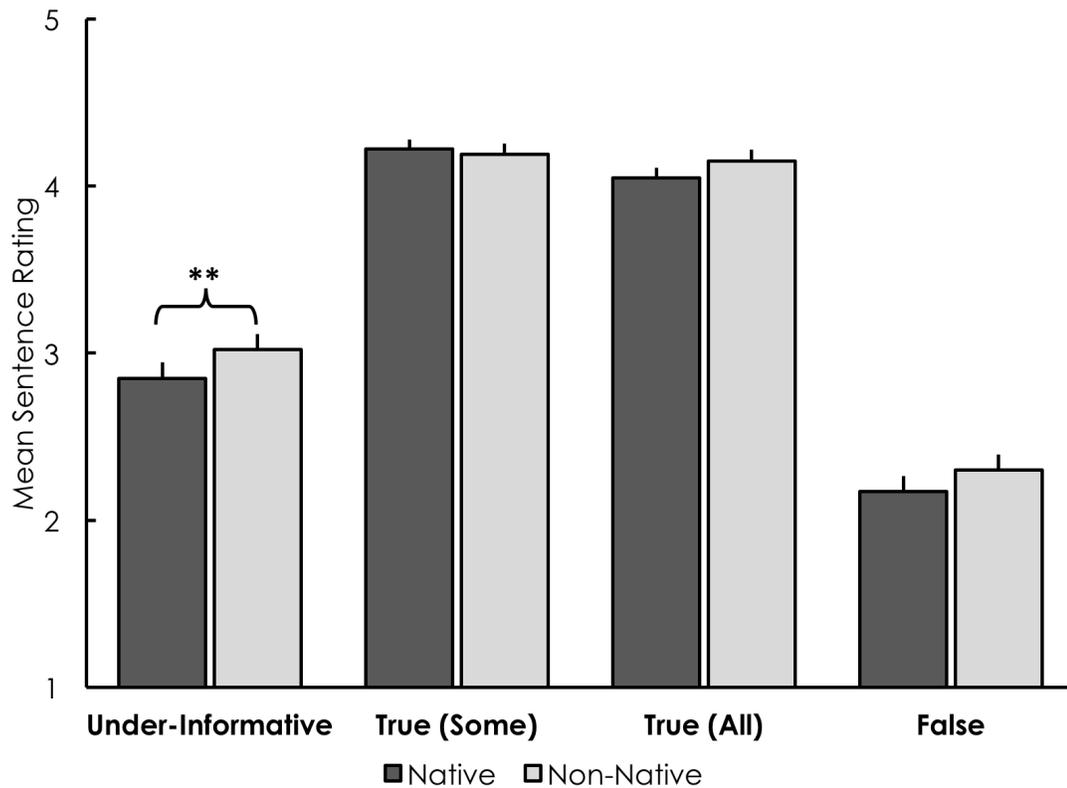


Figure 2.1: Mean Sentence Ratings by Speaker for all Sentence Types in Experiment 1. Error bars indicate +/-1 S.E.M. Asterisks denote significance as follows: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 2.2: Parameter estimates for a mixed-effects regression model predicting Sentence Ratings from Speaker and Sentence Type in Experiment 1.

Effect	β	S.E.	t	p
Intercept	3.370	0.037	92.256	< .001
Speaker (Native vs. Non-Native)	0.043	0.012	3.545	< .001
Sentence Type (Under-Inf. vs. False)	1.137	0.021	54.540	< .001
Sentence Type (Under-Inf. vs. True (Some))	-1.568	0.024	-65.162	< .001
Sentence Type (True (Some) vs. True (All))	-0.731	0.021	-35.076	< .001

2.1.3.2 Responder Bias Analysis

To further investigate the source of the forgiveness of under-informativeness in non-native speakers and its variation across individuals, a Non-Native Speaker Effect (hereafter NNS Effect) score was calculated for each participant by subtracting the mean rating for Under-Informative sentences in the Native Speaker condition from the mean rating for Under-Informative sentences in the Non-Native Speaker condition. Thus, individuals with positive scores were more lenient towards non-native speakers as compared to native speakers, while individuals with negative scores tended to penalize under-informativeness from non-native speakers more than from native speakers. To determine whether non-native speaker sensitivity varied across pragmatically- and logically-biased individuals, or whether the NNS Effect was stable across participants, we conducted a linear regression predicting the NNS Effect from the mean Under-Informative rating in the Native Speaker condition. This predictor was chosen because the Native Speaker condition reflects how a participant would judge under-informative utterances without other influences (and most closely corresponds to logical vs. pragmatic responders in the literature). As mentioned already, we expected that participants with a greater bias towards responding pragmatically (i.e., giving a low rating to under-informative sentences) in the Native

Speaker condition may be more likely to take into account the speaker's identity (including the ability to handle linguistic alternatives) and be more accepting when a non-native speaker produces an under-informative utterance.

In our data, there was great variability in Under-Informative sentence ratings in the Native Speaker condition, with mean ratings ranging the entire span of the scale (1 – 5; $SD = 1.01$). Overall, as we had anticipated, sensitivity to under-informativeness in the Native Speaker condition significantly predicted the NNS Effect, $F(1, 112) = 13.75, p < .001, R^2 = .11$. As can be seen in Figure 2.2, the more a participant adopted a pragmatic interpretation of the under-informative utterances and judged them as not making sense when a native speaker uttered them, the more likely the participant was to give the benefit of the doubt to non-native speakers for such cases, $\beta = -0.184, SE = 0.050, t = -3.708, p < .001$.

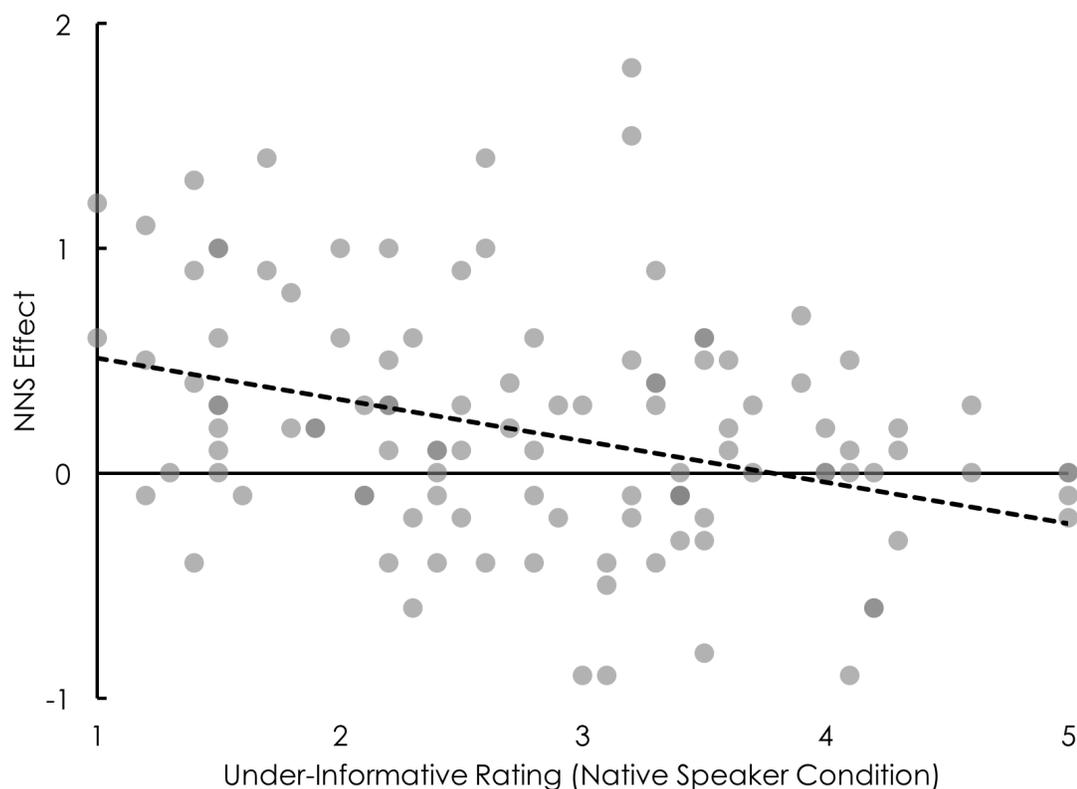


Figure 2.2: NNS Effect by Under-Informative rating in the Native Speaker condition in Experiment 1. A NNS Effect of 0 indicates no difference in ratings between speaker conditions, whereas a positive NNS Effect indicates higher ratings of under-informativeness for Non-Native Speakers as compared to Native Speakers.

2.1.3.3 Individual Differences Analyses

To further investigate potential individual differences in performance on the Sentence Ratings task, Kendall's tau correlation analyses (chosen to account for the positively-skewed distribution of AQ-COMM scores) were performed with the AQ-COMM score, the Chinese Cultural Attitude score, and Under-Informative Sentence Ratings for each Speaker as variables. AQ-COMM scores were marginally correlated with Under-Informative sentence ratings in the Native Speaker condition, $\tau_b(112) =$

.130, $p = .058$, and not significantly correlated with such ratings in the Non-Native Speaker condition, $\tau_b(112) = .070, p > .1$. Chinese Cultural Attitude scores were not significantly correlated with Under-Informative sentence ratings for either the Native Speaker condition, $\tau_b(112) = -.036, p > .1$, or the Non-Native Speaker condition, $\tau_b(112) = .004, p > .1$. Thus, Under-Informative sentence ratings are not significantly associated with social-communicative ability as measured by the AQ-COMM or an individual's cultural attitudes towards Chinese-American bilingual speakers.

2.1.4 Discussion

Three main findings arise from the present data. First, under-informative sentences were rated as making more sense than patently false sentences, but less sense than true sentences. Comprehenders thus understood that the under-informative statements were literally true, but were also sensitive to the fact that such statements were sub-optimal ways of conveying information. This finding constitutes a conceptual replication of nuanced pragmatic judgment patterns that have been previously observed for adults - and even children - in a laboratory setting (Katsos & Bishop, 2011).

Second, and critically, ratings of under-informative sentences increased when comprehenders believed these sentences to have come from a non-native compared to a native speaker of English. This effect of speaker identity applied selectively to under-informative statements and did not extend to falsehoods (or simply true sentences), i.e., individuals did not overall give the benefit of the doubt to anything a non-native speaker said.

Third, participants who tended to consistently adopt a pragmatic interpretation of under-informative statements when uttered by a native speaker (and thus gave low

ratings) were more forgiving towards non-native speakers for such sentences but those who tended to adopt the literal meaning for under-informative sentences (and thus gave high ratings) were less so. Assuming that calculating pragmatic inferences from the use of *some* requires reasoning about the communicative intentions of another person, including their access to linguistic alternatives such as *all*, it is reasonable to conclude that those individuals who consistently calculated the pragmatic meaning – unlike less pragmatically-inclined participants - were also sensitive to properties of the speaker (presumably reasoning, for instance, that a non-native speaker might not have been able to access or handle an alternative, pragmatically more felicitous way of phrasing their message).

The fact that comprehenders altered the way they processed under-informative statements simply as a result of information about speaker identity is in line with Expectation-Based accounts of non-native speech processing (predictions from Intelligibility-Based accounts are inert since the sentences were presented in the written modality and there was no processing load difference between the Native and Non-Native Speaker condition). We hypothesize that the observed pragmatic lenience towards non-native speakers is related to comprehenders' beliefs about these speakers' linguistic competence. Suggestive, though not conclusive, evidence for this hypothesis comes from the fact that individuals who are likely to judge that under-informative sentences from native speakers make little sense (presumably because there are alternative, more felicitous means of constructing the sentence) also show the highest pragmatic lenience towards non-native speakers (presumably because these speakers lack the ability to handle such linguistic alternatives).

Several aspects of our findings argue against major alternative explanations of the speaker identity effect. For instance, since native and non-native speakers had been judged as equally knowledgeable of the subject matter in the under-informative sentences, it is unlikely that the effect of speaker identity could be attributed to differences in native vs. non-native speakers' general world knowledge (cf. also the lack of difference in tolerance for false statements attributed to native vs. non-native speakers). Furthermore, since there was no correlation between participants' social-communicative score or attitude towards Chinese individuals and their level of tolerance for pragmatic anomalies, neither general communicative skills nor cultural stereotypes appear to be likely sources of the pattern observed in our data. In the next experiment, we seek to strengthen and clarify the evidence linking forgiveness of non-native speakers' under-informativeness to those speakers' presumed L2 skills.

2.2 Experiment 2

In Experiment 2, we sought to replicate and extend evidence for the conclusion that comprehenders forgive under-informativeness to a greater extent from non-native as compared to native speakers. We followed the same procedure as in Experiment 1, but manipulated the degree to which the non-native speaker had an accent in English. In auditory studies, the strength of a non-native speaker's foreign accent is often interpreted as a marker of their second language proficiency (e.g., Kang, Rubin, & Pickering, 2010). If the results of Experiment 1 were due to expectations about the lower second language proficiency level of the non-native speaker, an accent-free speaker might be treated more closely to a native speaker compared to a non-native speaker with a heavy accent. An alternative possibility is that forgiveness of under-informativeness in non-native speakers emerges as a result of a general belief that non-

native speakers have imperfect linguistic competence; if so, the pattern of results in our earlier experiment should extend to any kind of non-native speaker.

2.2.1 Participants

One hundred and eighty native speakers of English aged 20-35 ($M = 29.33$, $SD = 3.91$) living in the United States, 75 of whom were female, were recruited from Amazon's Mechanical Turk to participate in the experiment. Participants were compensated \$2.00 for their time.

2.2.2 Materials and Procedure

The materials were based on those in Experiment 1 with some minor alterations. First and foremost, we introduced an additional non-native speaker, Peiyao, who was also from China but had “no Chinese accent whatsoever.” Thus, we had three within-subjects Speaker conditions: Native Speaker, Accent-Free Non-Native Speaker, and Accented Non-Native Speaker. For all three, we shortened the descriptions by removing the information about their performance in school and future career. There were three versions of each speaker bio, to add variation to the task, and as in Experiment 1 the majors and hobbies of the speaker were altered to create these different versions. All speaker bios for Experiment 2 are presented in Table 2.3.

Table 2.3: Speaker bios for Experiment 2.

Native Speaker	Accent-Free Non-Native Speaker	Accented Non-Native Speaker
<p>Emma is a college student at the University of Delaware, majoring in History/Sociology/Mathematics. Emma moved with her family to Delaware from Boston, and her classmates often tease her about her strong Boston accent. In her spare time, Emma likes to hike/run/swim and play the piano/guitar/violin.</p>	<p>Peiyao is a college student at the University of Delaware, majoring in History/Sociology/Mathematics. Peiyao moved with her family to Delaware from China, and her classmates often tease her about the fact that she has no Chinese accent whatsoever. In her spare time, Peiyao likes to hike/run/swim and play the piano/guitar/violin.</p>	<p>Yuqi is a college student at the University of Delaware, majoring in History/Sociology/Mathematics. Yuqi moved with her family to Delaware from China, and her classmates often tease her about her strong Chinese accent. In her spare time, Yuqi likes to hike/run/swim and play the piano/guitar/violin.</p>

An additional 40 sentences were created for the purposes of Experiment 2 (see Appendix A), 10 for each Sentence Type (Under-Informative, True (Some), True (All), False). These sentences followed the same constraints as in the previous experiment. Even with the addition of these sentences the four Sentence Types did not differ by length in words or syllables (all p 's > .1).

We administered a Sentence Ratings task that was nearly identical to Experiment 1 except for the addition of a third block, for the Accent-Free Non-Native Speaker condition. As in the previous experiment, participants read a description of a speaker at the beginning of each block and answered comprehension questions about the speaker (performance was very high, 90%). Participants then judged 40 sentences “originally spoken by that person.” They were asked to rate how “Good” each sentence was on a five-point scale where 1 is “Very bad” (makes no sense at all) and 5

is “Very good” (makes perfect sense). The order of the speaker was counterbalanced across participants, and sentences were fully rotated through each speaker condition.

2.2.3 Results

2.2.3.1 Overall Analysis

A linear mixed-effects regression with crossed random intercepts for Participants and Items and Speaker (Native Speaker, Accent-Free Non-Native Speaker, Accented Non-Native Speaker) and Sentence Type (Under-Informative, True (Some), True (All), False) included as fixed effects was performed on participants’ Sentence Ratings in Experiment 2 (see Figure 2.3). Sentence Ratings varied significantly by Speaker, $\chi^2(1) = 13.782, p = .001$, and by Sentence Type, $\chi^2(3) = 12015.314, p < .001$. Planned contrasts (Table 2.4) indicated that ratings were higher in the Accented Non-Native Speaker ($M = 3.38, SD = 1.25$) condition as compared to the Native Speaker ($M = 3.31, SD = 1.30$) condition ($p = .008$), but ratings in the Accent-Free Non-Native Speaker ($M = 3.36, SD = 1.30$) condition did not differ significantly from the Native Speaker condition ($p > .1$). Under-Informative ($M = 2.76, SD = 1.11$) sentences were rated higher than False ($M = 2.06, SD = .94$) sentences, but lower than True (Some) ($M = 4.34, SD = .56$) sentences (both p ’s $< .001$). True (All) ($M = 4.25, SD = .57$) sentences were rated lower than True (Some) sentences ($p < .001$).

These main effects were qualified by a significant interaction between Speaker and Sentence Type, $\chi^2(3) = 22.187, p = .001$. Post-hoc tests revealed that Under-Informative sentences were rated significantly higher in the Accented Non-Native Speaker ($M = 2.86, SD = 1.49$) condition as compared to both the Native Speaker ($M =$

2.66, $SD = 1.49$) condition ($p < .001$) and the Accent-Free Non-Native Speaker ($M = 2.73$, $SD = 1.50$) condition ($p = .020$). Ratings did not differ significantly between the Native Speaker and Accent-Free Non-Native Speaker conditions ($p > .1$). In other words, participants treated non-native speakers with high linguistic competence like native speakers, and penalized their under-informative utterances. There was no difference in ratings for different kinds of Speakers in the other three Sentence types.

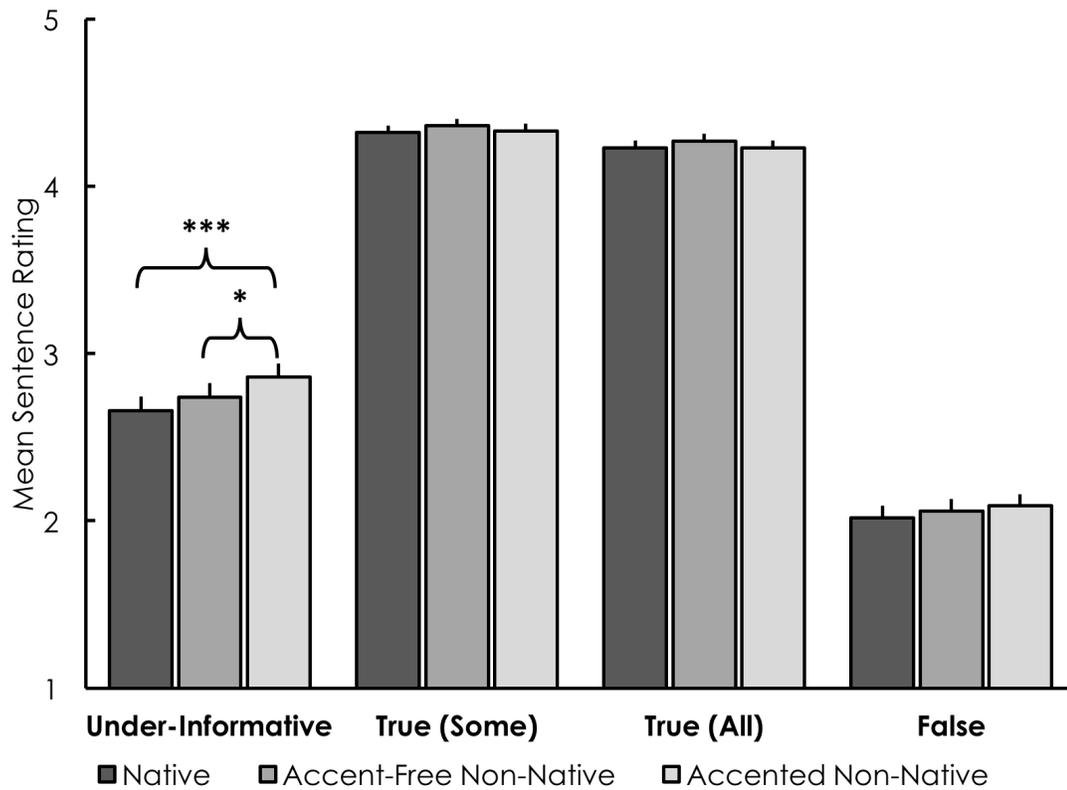


Figure 2.3: Mean Sentence Ratings by Speaker for all Sentence Types in Experiment 2. Error bars indicate +/-1 S.E.M. Asterisks denote significance as follows: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 2.4: Parameter estimates for a mixed-effects regression model predicting Sentence Ratings from Speaker and Sentence Type in Experiment 2.

Effect	β	S.E.	t	p
Intercept	3.345	0.022	150.923	< .001
Speaker (Native vs. Accented Non-Native)	-0.029	0.011	-2.685	.008
Speaker (Native vs. Accent-Free Non-Native)	-0.010	0.011	-0.928	.353
Sentence Type (Under-Inf. vs. False)	1.296	0.013	97.299	< .001
Sentence Type (Under-Inf. vs. True (Some))	-0.304	0.015	-19.755	< .001
Sentence Type (True (Some) vs. True (All))	0.600	0.013	44.998	< .001

2.2.3.2 Responder Bias Analysis

A NNS Effect score was calculated for each participant by subtracting mean Under-Informative Sentence Ratings in the Native Speaker condition from mean Under-Informative Sentence Ratings in the Accented Non-Native Speaker condition (the Accent-Free Non-Native Speaker condition was not included in calculation of the score, as these ratings did not differ significantly from the Native Speaker condition). A linear regression was performed predicting NNS Effect scores from Under-Informative Native Speaker Sentence Ratings. The analysis significantly predicted NNS Effect scores, $F(4, 178) = 19.55, p < .001, R^2 = .10$. As can be seen in Figure 2.4, the more participants adopted a pragmatic interpretation of under-informative sentences (i.e., gave them a low rating), the more lenient they were towards the same under-informative statements when they were attributed to a non-native speaker with a strong accent, $\beta = -0.177, SE = 0.040, t = -4.421, p < .001$.

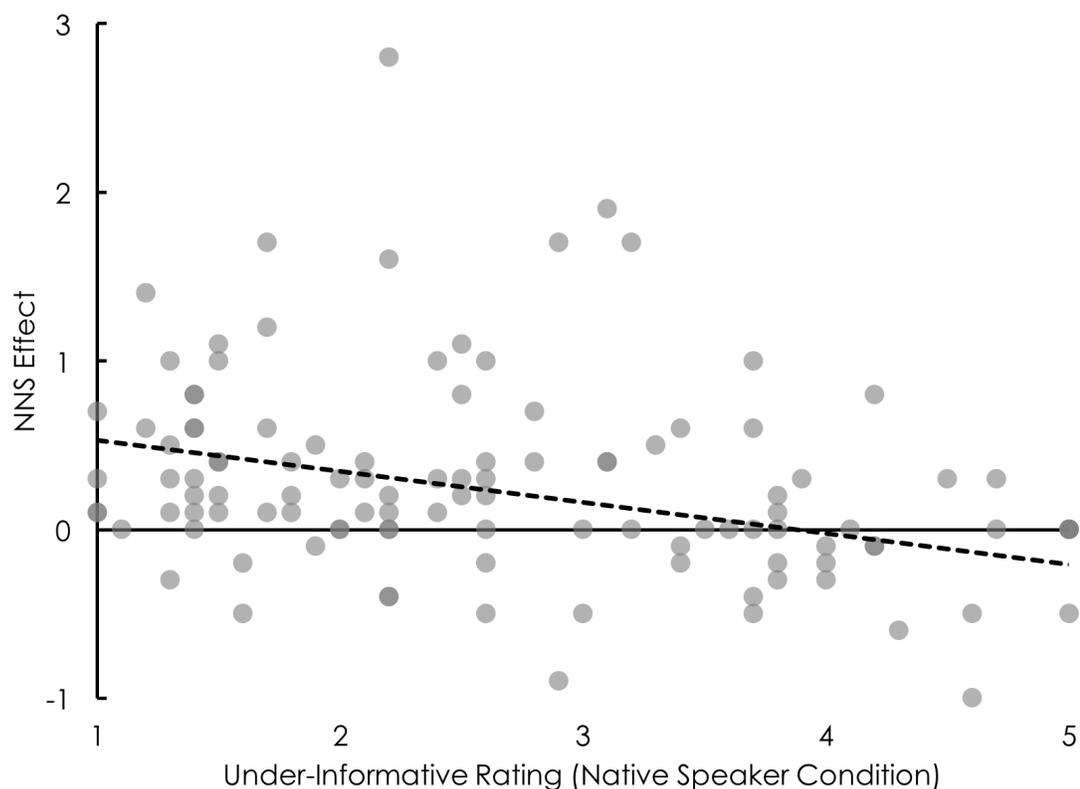


Figure 2.4: NNS Effect by Under-Informative rating in the Native Speaker condition in Experiment 2. A NNS Effect of 0 indicates no difference in ratings between speaker conditions, whereas a positive NNS Effect indicates greater lenience towards under-informativeness from Accented Non-Native Speakers as compared to Native Speakers.

2.2.4 Discussion

Experiment 2 replicated the general pattern of results in Experiment 1: regardless of whether they were attributed to a native or a non-native speaker, under-informative statements were judged as making more sense compared to completely false sentences, but less sense compared to true sentences. Additionally, Experiment 2 replicated the results of Experiment 1 by finding selectively higher ratings for under-informative statements believed to be produced by non-native compared to native

speakers. Importantly, this effect was modulated by the language proficiency of the non-native speaker: under-informative sentences were judged as making more sense when they came from an accented non-native speaker compared to a native speaker but an accent-free non-native speaker had no such advantage. Finally, we replicated the finding that this selective advantage for non-native speakers was greater in those participants who consistently derived pragmatic inferences from under-informative statements.

As with Experiment 1, our findings strongly support an Expectation-Based account of non-native speech processing: different responses to the same under-informative sentences across conditions were produced simply by altering beliefs about the language background of the speaker. Furthermore, the present data suggest that it is expectations about the language proficiency of the speaker specifically that lead to greater forgiveness of under-informativeness. We hypothesize that, given accent information alone, comprehenders make further assumptions about the non-native speakers' L2 proficiency level (cf. Kang et al., 2010) and go on to assume that only non-native speakers with a poor command of their second language should be given the benefit of the doubt when they produce under-informative statements.

2.3 Experiment 3

The under-informative sentences used in Experiments 1 and 2 (e.g., “Some people have noses with two nostrils”) relied on world knowledge (e.g., knowing that noses have two nostrils). Furthermore, the corresponding “not all” propositions (“Not all people have noses with two nostrils”) were false and unlikely to be part of speaker meaning. In Experiment 3, we sought to replicate the findings of Experiments 1 and 2 using a different set of stimuli for which judgments did not rely on evaluating

individual sentences against one's own world knowledge; furthermore, the "not all" propositions could plausibly have been intended to be part of what the speaker meant by uttering *some*. Building on materials used in a study by Bergen and Grodner (2012), we introduced three-sentence passages where a highly knowledgeable speaker used *some* in a way that was highly likely to give rise to the "not all" implicature. The "not all" implicature was either cancelled explicitly ("In fact, all...") or supported ("The rest...") in the final sentence of the passage.

In Bergen and Grodner's (2012) reading study, speaker knowledge was manipulated in a context sentence and participants showed sensitivity to speaker knowledge that manifested itself in their reading times. Participants generated stronger implicatures when the speaker was highly knowledgeable of the topic at hand ("I *meticulously compiled* the investment reports. Some of the real estate investments lost money.") and therefore were more likely to have meant that the stronger alternative *all* was false. Conversely, participants generated weaker implicatures in cases when the speaker was less knowledgeable ("I *skimmed* the investment reports. Some of the real estate investments lost money.") and therefore may not have meant that the stronger alternative was false, but simply implicated lack of knowledge about whether the stronger alternative was false. In the present experiment, we only used cases where the speaker was highly knowledgeable: the passages were preceded by information about the speaker (native vs. non-native) and participants were asked to rate the passages for meaning, as in our prior studies. We reasoned that participants would take a non-native speaker to be more likely compared to a native speaker to make (and later correct) a pragmatically under-informative statement, presumably because of poorer initial choice of words due to lower language proficiency/pragmatic competence.

2.3.1 Participants

One hundred and ten English monolinguals aged 20-42 years ($M = 28.62$, $SD = 4.34$), living in the United States were recruited from Amazon's Mechanical Turk ($N = 46$ Female, $N = 62$ Male, $N = 2$ Other/Prefer not to answer). Participants were compensated \$1.50 for the fifteen-minute study.

2.3.2 Materials and Procedure

Forty passages were created for Experiment 3 (see Appendix B), each with 4 versions (Some/All, Some/Rest, Only some/Rest, Only some/All; see Table 4). Twenty-two passages came from the stimulus list in Bergen and Grodner (2012), with minor word changes. The rest were created in the same fashion. Passages were created such that it would be believable for a college student to have produced them. All passages began with a context sentence establishing that the speaker was fully knowledgeable about the topic (e.g., "As part of my advanced accounting class, I meticulously compiled the investment reports." – see Table 2.5.)

In Some/All passages, the context sentence was followed by a critical sentence beginning with *some* meant to trigger a "not all" implicature but the final sentence cancelled the implicature ("In fact...all.") Some/Rest passages included the same critical sentence as the Some/All passages but their final sentence was consistent with the implicature ("The rest..."). The Only some/Rest passages were identical to the Some/Rest passages, except that the critical sentence began with *Only some* instead of *some*, and therefore, there was no need to calculate an implicature. The Only some/All passages were identical to the Some/All passages, except that again the critical sentence began with *Only some* instead of *some*. In this case, not only was there no

implicature to be generated but the final sentence beginning with “In fact...all” was logically inconsistent with the critical sentence (“Only some...”).

Table 2.5: Sample stimuli for Experiment 3.

Passage Type	Example
Some/All	As part of my advanced accounting class, I meticulously compiled the investment reports. Some of the investments lost money. In fact, they all did because of the recent economic downturn.
Some/Rest	As part of my advanced accounting class, I meticulously compiled the investment reports. Some of the investments lost money. The rest did not totally unfamiliar despite the recent economic downturn.
Only some/Rest	As part of my advanced accounting class, I meticulously compiled the investment reports. Only some of the investments lost money. The rest did not despite the recent economic downturn.
Only some/All	As part of my advanced accounting class, I meticulously compiled the investment reports. Only some of the investments lost money. In fact, they all did because of the recent economic downturn.

The task consisted of two blocks: a Native Speaker block and a Non-Native Speaker block (counterbalanced across participants), using the same speaker descriptions as in Experiment 1. Passages within each block were evenly distributed across the four conditions (10 of each), and were presented in a random order. Thus, both Speaker (Native, Non-Native) and Passage Type (Some/All, Some/Rest, Only some/Rest, Only some/All) were treated as within-subjects factors. At the start of each block, one of the four speaker bios appeared on the screen. Participants were instructed to read carefully in order to answer the comprehension questions that followed, and were given as much time as they needed to read the paragraph before

moving on. The speaker bio was followed by three multiple-choice questions about the speaker, presented in a random order (“Where is Emma/Yuqi from?”, “What is Emma/Yuqi majoring in?”, “What does Emma/Yuqi like to do in her spare time?”). Performance on these comprehension questions was very good (87%), indicating that participants had fully read and understood the speaker bios. Participants were then instructed that they would be reading 40 passages that were originally uttered by the person they had just read about, and that their job was to rate how “Good” each sentence was on a five-point scale where 1 is “Very bad” (makes no sense at all) and 5 is “Very good” (makes perfect sense). On each trial, a three-sentence passage appeared in the center of the screen with the ratings scale below. For half of the participants, the speaker bio was always present at the top of the screen in a muted gray color, and for the other half it was not present at the top of the screen. Results are combined for these two groups as they did not differ from one another in terms of sentence ratings for any condition.²

We predicted that Some/Rest and Only some/Rest passages would elicit high ratings, as both types of passages make sense and are pragmatically felicitous (cf. the True sentences in our previous experiments). Only some/All passages should elicit the lowest ratings, as the final sentence logically contradicts and corrects the critical sentence (cf. our earlier False sentences). Some/All passages should elicit higher ratings than Only some/All passages but lower ratings than the other two conditions

² A mixed ANOVA was conducted with Speaker (Native, Non-Native) and Sentence Type (Some/All, Some/Rest, Only some/All, Only some/Rest) as within-subjects factors and Experiment (Paragraph Present on every trial or Absent) as a between-subjects factor. The main effect of Experiment was not significant, nor were any interactions with Experiment (all p 's > .05).

because of the presence of an under-informative statement that is later corrected (cf. the Under-Informative sentences in our previous experiments).

For our critical Speaker manipulation, we expected participants to rate Some/All passages more highly in the Non-Native Speaker condition compared to the Native Speaker condition. This would indicate that participants would be more accepting of a Non-Native speaker inadvertently producing an utterance in which *some* is compatible with *all*, mirroring the results of Experiments 1 and 2. No such differences were expected in the other passages.

2.3.3 Results

2.3.3.1 Overall Analysis

A linear mixed-effects regression with crossed random intercepts for Participants and Items and Speaker (Native Speaker, Accent-Free Non-Native Speaker, Accented Non-Native Speaker) and Sentence Type (Some/All, Some/Rest, Only some/Rest, Only some/All) included as fixed effects was performed on participants' Sentence Ratings in Experiment 3 (see Figure 2.5). Sentence Ratings varied significantly by Speaker, $\chi^2(1) = 8.059, p = .005$, and by Sentence Type, $\chi^2(3) = 1439.242, p < .001$. Planned contrasts (Table 2.6) indicated that ratings were higher in the Non-Native Speaker ($M = 3.38, SD = 1.24$) condition as compared to the Native Speaker ($M = 3.26, SD = 1.22$) condition ($p = .004$). Furthermore, Some/All ($M = 2.76, SD = 1.80$) passages were rated higher than Only some/All ($M = 2.40, SD = 1.31$) passages but lower than Some/Rest ($M = 4.02, SD = 1.17$) passages (both p 's $< .001$). Only some/Rest ($M = 4.05, SD = 1.23$) passages were rated higher than Some/Rest passages ($p < .001$).

These main effects were qualified by a significant interaction between Speaker and Passage Type, $\chi^2(3) = 27.622, p < .001$. Post-hoc tests indicated that ratings of Some/All passages were higher for Non-Native speaker trials ($M = 2.98, SD = 1.83$) than for Native speaker ($M = 2.53, SD = 1.30$) trials ($p < .001$). Ratings of Some/Rest, Only some/Rest, and Only some/All passages did not differ by Speaker (all p 's $> .1$).

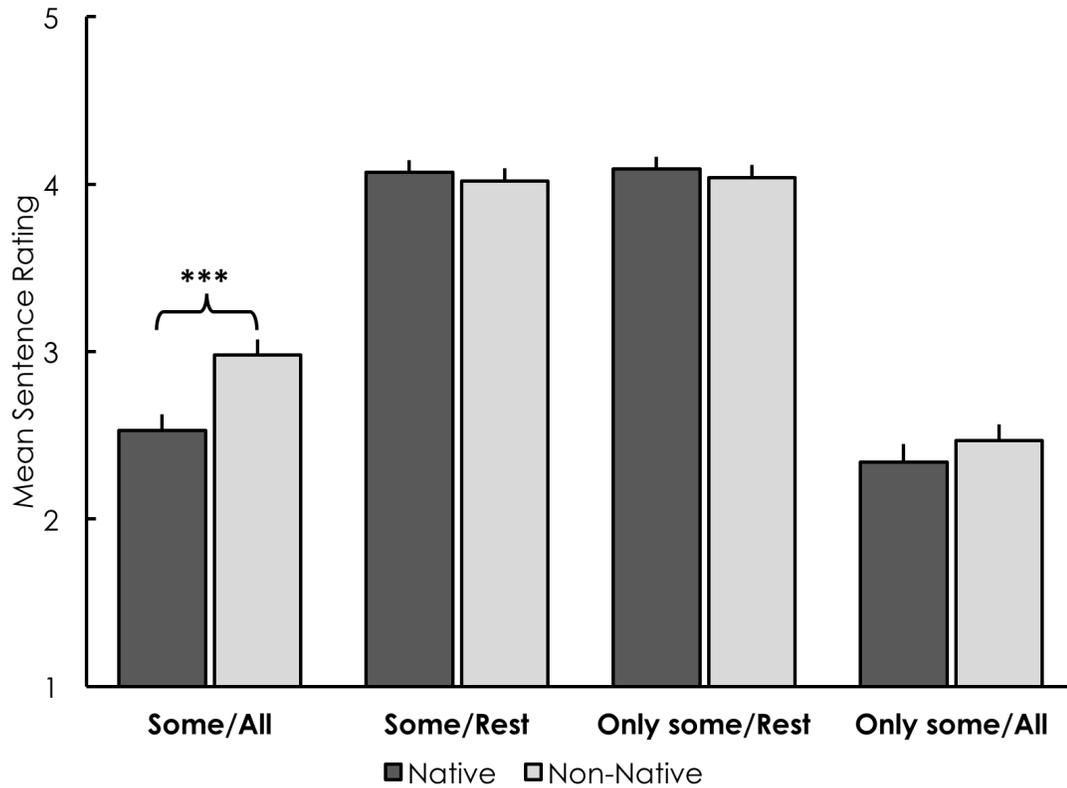


Figure 2.5: Mean Sentence Ratings by Speaker for all Passage Types in Experiment 3. Error bars indicate ± 1 S.E.M. Asterisks denote significance as follows: $*p < .05, **p < .01, ***p < .001$.

Table 2.6: Parameter estimates for a mixed-effects regression model predicting Sentence Ratings from Speaker and Sentence Type in Experiment 3.

Effect	β	S.E.	t	p
Intercept	3.271	0.038	86.978	< .001
Speaker (Adult vs. Child)	0.050	0.018	2.847	< .001
Sentence Type (Some/All vs. Only/All)	0.901	0.030	29.639	< .001
Sentence Type (Some/All vs. Some/Rest)	-1.444	0.035	-41.125	< .001
Sentence Type (Some/Rest vs. Only/Rest)	-0.730	0.030	-24.326	< .001

2.3.3.2 Responder Bias Analysis

A NNS Effect score was calculated for each participant by subtracting mean Some/All ratings in the Native Speaker condition from mean Some/All ratings in the Non-Native Speaker condition. A linear regression was then performed predicting NNS Effect scores from mean Some/All ratings in the Native Speaker condition, as in the previous experiments: the analysis yielded a significant result, $F(1, 108) = 18.03$, $p < .001$, $R^2 = .14$. As Figure 2.6 demonstrates, the worse an individual rated a passage in which *some* was compatible with *all* (i.e., made a pragmatic judgment rather than a logical one), the more lenient they were towards such passages when produced by a non-native speaker, $\beta = -0.251$, $SE = 0.059$, $t = -4.246$, $p < .001$.

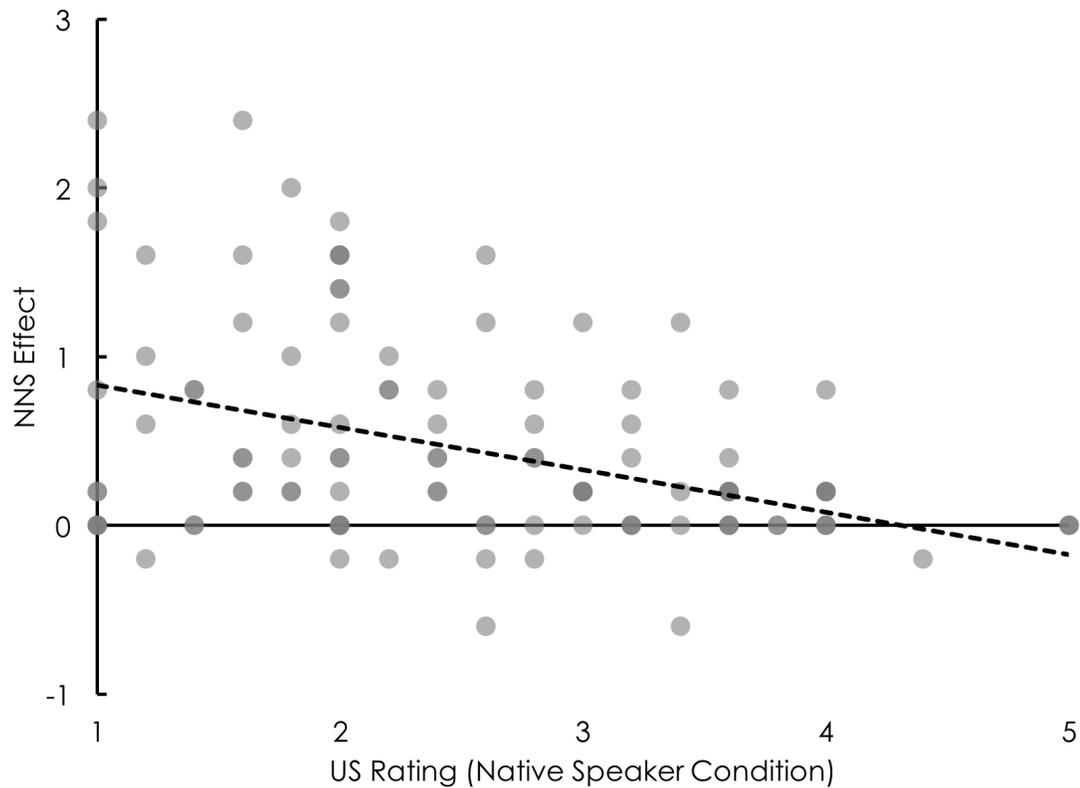


Figure 2.6: NNS Effect by Some/All rating in the Native Speaker condition in Experiment 3. A NNS Effect of 0 indicates no difference in ratings between speaker conditions, whereas a positive NNS Effect indicates greater tolerance of under-informativeness from Accented Non-Native Speakers as compared to Native Speakers.

2.3.4 Discussion

In line with the pattern of results previously observed, under-informative passages with *some* followed by *all* were treated as making more sense than logically inconsistent passages in which *only some* was followed by *all*, but less sense than passages in which *some* or *only some* was followed by *the rest*. Importantly, comprehenders judged that these under-informative passages made more sense when they believed them to have come from a non-native compared to a native speaker of

English. As in Experiments 1 and 2, this effect of speaker identity was selective to Some/All passages and did not extend to any of the other conditions, and was greatest in participants who tended to give lower meaning ratings to the Some/All statements (i.e., participants who had adopted a pragmatic interpretation of *some* upon first reading it). Experiment 3 therefore fully replicated our earlier findings with a new set of stimuli, showing that forgiveness of non-native speakers' under-informativeness is robust, generalizable, and does not rely on world knowledge.

2.4 General Discussion

2.4.1 Theories of Non-Native Language Processing

Interacting with non-native speakers poses specific challenges for language processing, with prior research indicating that neural and behavioral responses to native and non-native language errors differ (Hanulikova et al., 2012; Gibson et al., 2017; Goslin et al., 2012; Romero-Rivas et al., 2015), and that listeners judge non-native speakers to be less trustworthy and more vague (Lev-Ari & Keysar, 2012). Intelligibility-Based accounts argue that non-native speech is processed differently to the extent that it requires more processing resources. Expectation-Based theories argue that differences in online and offline processing of non-native speech stem from the different expectations that listeners hold about non-native speakers (e.g., that non-native speakers have lower language proficiency, or that their speech stream will be more variable). In practice, the role of expectations is hard to disentangle from the intelligibility costs incurred by a foreign accent. Here we presented a strong test of the role of expectations in the absence of any actual intelligibility-related costs by comparing ratings of sentences presented in the written modality but believed to have

been produced by either a native or a non-native speaker. Furthermore, we broadened the empirical scope of prior work on the comprehension of non-native speech by focusing on the domain of pragmatics.

Across three experiments, we found that knowledge about the language background of the speaker affected pragmatic interpretation even in the absence of actual exposure to a foreign accent. Specifically, comprehenders rated pragmatically under-informative sentences (e.g., “Some people have noses with two nostrils”) as more meaningful when they believed that these sentences were produced by a non-native speaker as compared to a native speaker (Experiment 1). The effect was present for non-native speakers with lower second language proficiency but not for highly proficient (non-accented) non-native speakers (Experiment 2). Furthermore, the native vs. non-native speaker difference extended to an additional set of stimuli that did not rely on world knowledge (Experiment 3). Throughout these experiments, speaker sensitivity was related to individual judgment preferences: individuals who tended to respond to the pragmatic meaning of under-informative statements when these statements were attributed to a native speaker (and hence based their judgment on the presence of a more informative alternative) were most forgiving of non-native speakers’ under-informativeness - presumably because these speakers’ access or ability to evaluate alternatives was impaired; by contrast, individuals who tended to respond to the logical meaning of under-informative statements when statements belonged to a native speaker (and hence did not focus on the presence of better linguistic alternatives) were less likely to adjust their ratings for non-native speakers. Together, our findings strongly support Expectation-Based accounts of non-native language processing, even in the absence of intelligibility factors. Our results are in

line with previous studies that suggest a role for speaker identity expectations in non-native speech processing (Hanulikova et al., 2012; Goslin et al., 2012; Lev-Ari, 2015; Romero-Rivas et al., 2015).

Naturally, the present data are entirely compatible with the idea that a foreign accent introduces noise to the linguistic signal and increases processing effort. In spoken communication, non-native input to language comprehension is often an errorful or corrupted signal on multiple levels, and it as such can incur intelligibility-based cost. Communicating with non-native speakers typically makes use of both general expectations about the language and error patterns of non-native speakers of the kind discussed here, as well as situation-specific experiences with and adaptations to actual error patterns in the speech of the particular individual one is communicating with (on such situation-specific factors, see Gibson et al., 2017).

2.4.2 The Pragmatics of Accent

How can the increased tendency to forgive sins of information omission in non-native speakers be reconciled with the finding that non-native speakers' utterances are often judged to be less trustworthy and more vague compared to those of native speakers (e.g., Lev-Ari & Keysar, 2010)? One might expect an under-informative utterance coming from a non-native speaker to be judged less charitably or corrected more often by native comprehenders. We want to note that there have been cases where the linguistic instability inherent in much of non-native speech has been found to have some advantages: as mentioned already, syntactically errorful utterances are less likely to elicit surprise (Hanulikova et al., 2012) and more likely to be reinterpreted when produced by non-native speakers (Gibson et al., 2017). For our data, we believe that non-native speakers are penalized less because they (are

perceived to) have reasonable grounds for selecting a less-than-optimal linguistic stimulus – namely, they are linguistically less competent.

The computations leading to the pragmatic lenience effect are worth discussing in some detail. Recall that pragmatic aspects of meaning are driven by expectations about how rational communication works. For instance, communicators expect speakers to strive to offer sentences that are informative to the degree required by the goals of the conversation (Maxim of Quantity; Grice, 1975). When speakers fail to be as informative as required, hearers are justified to engage in further inferences to understand the reasons behind this failure. In some cases, listeners derive a scalar implicature, inferring that the speaker meant that a more informative statement would not be true. In other cases, under-informative statements give rise to the inference that the speaker was unable to commit to the stronger term because of lack of information (see also Bonnefon, Feeney, & Villejoubert, 2009; Geurts, 2010; Sperber & Wilson, 1995, for additional possibilities).

In Experiments 1 and 2, statements such as “Some humans have noses with two nostrils” violate the Maxim of Quantity: the speaker used the weaker term in a logical scale (‘some’) when she could have used a stronger, more informative scalar term (‘all’). Furthermore, the statements can be potentially misleading because they can give rise to a scalar inference corresponding to a patently false proposition (“Not all humans have noses with two nostrils”). After presumably detecting the violation, comprehenders judge under-informative sentences as “making less sense” when attributed to a native speaker because it is hard to perceive what the speaker could have meant (i.e., what the grounds for under-informativeness could be given what is known about Emma’s abilities and preferences). For a non-native speaker,

comprehenders' judgments are more charitable (even though not completely positive) since the infelicitous scalar choice could be attributed to lack of proficiency in English. (The same considerations apply to Experiment 3: here the speaker explicitly corrects 'some' to 'all' precisely because the earlier choice of quantifier was likely to lead to a misleading inference.) Comprehenders are therefore more likely to forgive non-native speakers for sins of information omission (for these speakers know not what they do.) As our data consistently show, the tendency to forgive is stronger in participants who tend to adopt a pragmatic final interpretation – probably because these participants base their rating on the alternatives that the speaker could have used but did not. Notice that these comprehenders compute the pragmatic content derivable from the speaker's utterance (“Not all...”) even though they are unlikely to believe it themselves – and may not assume that it was meant to be communicated by the speaker (cf. Mazzarella, 2015; Sperber et al., 2010).³

The present evidence for pragmatic lenience towards non-native speakers comes from an offline, metalinguistic task. Such tasks have proven very useful as a means of investigating pragmatic intuitions in both adults and young children (see

³ The present perspective differs from (and is orthogonal to) the notion of pragmatic tolerance developed by Katsos and Bishop (2011) to account for the fact that, unlike adults who penalize under-informative utterances, young children seem to find them acceptable when given a binary response scale (when given a 3-point scale, the difference disappears, and both age groups give under-informative utterances intermediate ratings). According to Katsos and Bishop, young children detect under-informativeness in binary tasks but – unlike adults - do not deem it serious enough to warrant a negative judgment. Thus in their account, the notion of tolerance is meant to explain task-specific behavior, i.e., children's apparently logical responses to under-informativeness *within a binary judgment task*. In the present data, lenience towards non-native speakers leads to *higher*, not lower, ratings for under-informativeness in adults. Crucially, these higher ratings are not taken to reflect task-specific reasoning (or a difference in how task demands are interpreted in the Native vs. Non-Native speaker conditions) but rather specific inferences about the grounds of under-informativeness in speakers of different linguistic backgrounds. Finally, and relatedly, such inferences are not meant to be limited to metalinguistic contexts but should arise spontaneously when people process non-native speech that falls short of informativeness expectations (see Chapter 3.).

Noveck, 2001). Nevertheless, we anticipate that comprehenders beyond the present context make spontaneous assumptions about why people are less informative than expected and, as part of these computations, attribute different grounds for under-informativeness to native and non-native speakers in a variety of tasks. The next chapter specifically addresses this prediction.

2.4.3 Extensions and Future Directions

Our data suggest several possibilities for future research. First, all of our experiments compared native speakers of English to native speakers of Chinese who spoke English as a second language. It remains open whether (descriptions of) different types of accents are equally likely to induce adjustments in pragmatic processing. Relatedly, it remains to be seen whether such adjustments emerge regardless of the comprehenders' specific language background (if so, the reported effect would be replicated in reverse with participants recruited in China). Versions of the present experiments could pursue these questions by varying both the language of the comprehenders and the language background (and level of proficiency) of the presumed non-native speakers. From a broader perspective, it is intriguing to explore whether selective pragmatic lenience of the kind discussed here would generalize to tokens produced by other populations whose linguistic knowledge or use is developing, atypical or otherwise limited (examples include children acquiring their first language, or aphasic patients).

Second, effects of (non)native speaker status in our data were observed selectively in under-informative sentences but not in true (and informative) or false sentences. We suspect that this selectivity is due to the fact that non-native speakers in our studies were introduced as highly educated, already living abroad and being

members of a university community; furthermore, the experimental sentences were fairly sophisticated (especially in Experiment 3) and contained no grammatical errors. Any differences between the two groups of speakers was therefore limited to relatively nuanced aspects of communication. This conclusion is reinforced by the fact that the two groups of speakers did not differ in their perceived (Experiment 1) or stated (Experiment 3) familiarity with the sentence topics so there was no basis for assuming that they differed in their ability to judge a test sentence as factually true or not. It remains possible that, in populations of non-native speakers with less secure knowledge of the mechanics of their second language, the observed lenience might extend to semantic errors (see Goslin et al., 2012).

Third, the present results cohere with a broader perspective according to which accents are not just physical features of a linguistic stimulus but sources of psychological attributions. Accents can form the basis of assumptions about the speaker's epistemic state, cultural beliefs, experience with food, music and the environment, and several other attributes beyond language. Depending on the topic, accented speakers may be considered more, not less knowledgeable than native speakers and these epistemic assumptions can themselves bear on utterance interpretation. For instance, if a Chinese-accented person used a scalar utterance in discussing Chinese politics (e.g., “Some Chinese families follow the one-child policy”), the listener's comprehension of her utterance would probably be affected by her presumed expertise (for instance, the listener may conclude that the speaker meant that not all families follow the policy, whereas the same utterance from a native speaker of English might be taken to convey lack of knowledge about the situation in all families in China). In the present work, knowledge of the topics in test sentences

was comparable across native and non-native speakers. Future work could fruitfully investigate how situational knowledge or cultural attitudes interact with the non-native speaker effect we observed here.

Chapter 3

SPEAKER EFFECTS ON PRAGMATIC MEANING: JUSTIFYING UNDER-INFORMATIVENESS IN NATIVE VS. NON-NATIVE SPEAKERS

As noted in Chapter 2, being a non-native speaker of a language presents some disadvantages. Children prefer to learn from, be friends with, and share resources with native over non-native speakers (Kinzler, Dupoux, & Spelke, 2007). Adults judge non-native speakers as being less trustworthy (Lev-Ari & Keysar, 2010) and more vague (Lev-Ari & Keysar, 2012), and non-native speakers face social discrimination (Kalin & Rayko, 1978; Gluszek & Dovidio, 2010; Hosoda & Stone-Romero, 2010). Furthermore, native comprehenders process non-native speech differently from native speech. Syntactic errors like “She *mow the lawn” typically elicit a P600 component in event-related potential studies, but this neural response is attenuated when ungrammatical sentences are spoken by a non-native speaker, suggesting that listeners expect non-native speakers to make syntactic errors (Hanulíková, van Alphen, van Goch, & Weber, 2012; Grey & Van Hell, 2017). However, recall that in Chapter 2 we reported a systematic difference in the way comprehenders interpret what native vs. non-native speakers say (and leave unsaid) that creates a bias *in favor* of non-native speakers: under-informative statements were judged as better when attributed to a non-native as compared to a native speaker of English. In Chapter 3 we demonstrate an additional pragmatic bias in favor of non-native speakers that has implications for our findings in Chapter 2.

The focus of the present chapter is under-informativeness at a broader scale than in Chapter 2, but the same theoretical framework is relevant. Listeners expect speakers to offer utterances that are sufficiently informative (Grice, 1975), causing adults to disprefer under-informative statements such as “Some dogs are mammals” in sentence rating studies (e.g., Noveck, 2001; Papafragou & Musolino, 2003; Bott & Noveck, 2004; Guasti et al., 2005). Relatedly, children avoid learning new information from ‘teachers’ with a history of under-informativeness (Gweon, Pelton, Konopka, & Schulz, 2014). In such developmental work, an informative teacher provides information about all features of a new toy, while an under-informative teacher leaves information out. In the present chapter we pursue the possibility that listeners respond to an under-informative friend or teacher differently depending on the speaker’s native language. This is a context likely to occur – and affect learning – in the real world (as, e.g., the number of foreign-born faculty at universities in the U.S. is increasing; Marvatsi, 2005).

When speakers fail to be fully informative, hearers may engage in further inferences to understand the reasons behind the failure. A first broad class of such inferences involves the speaker’s inability to offer the required information. For instance, the sentence “Some of Jane’s friends are vegetarian” may give rise to the inference that the speaker does not know whether the stronger statement (“All of Jane’s friends...”) is true; alternatively, if the speaker is assumed to be well-informed about Jane’s friends, the statement may lead to the assumption that the speaker knows for a fact that not all of Jane’s friends are vegetarian (Grice, 1975; Horn, 1972; Sauerland, 2012; Carston 1998). In a second broad class of cases, under-informativeness is attributed to the speaker’s unwillingness to communicate additional

information out of politeness (“Some people hated your poem”), desire to mislead (“A few of my projects have failed”), or some other reason (Grice, 1975; Geurts, 2010). Inability or unwillingness inferences may be computed on the basis of what the speaker said but whether they are taken as part of what the speaker intended to convey depends on how much the listener trusts the speaker, and what they know about the speaker’s preferences (Sperber & Wilson, 1995).

Within the Gricean framework, instances where the speaker is unwilling to offer relevant information generally lead to communication breakdowns (but see also Geurts, 2010; Sperber & Wilson, 1995). As a result, unwillingness to be informative is likely to be penalized more heavily than inability to offer required information. This asymmetry is confirmed in studies of intentional action understanding: infants react with more impatience (e.g., reaching, looking away) when an adult interacting with them was unwilling to give them a toy than when the adult was simply unable to give them the toy (Behne, Carpenter & Tomasello, 2005). Similar results have been observed with non-human primates (Call, Hare, Carpenter & Tomasello, 2005; Phillips et al., 2009; Canteloupe & Meunier, 2017).

Most investigations of under-informativeness have focused on a narrow set of possible speaker motivations (but see Bonnefon, Feeney, & Villejoubert, 2009; Mazzarella, 2015). Here we build on pragmatic theory to explore a fuller range of explanations of under-informativeness and their cognitive consequences. Crucially, we propose that such explanations vary systematically depending on speaker identity (native vs. non-native). Because the speech of non-native speakers is expected to be more error-prone and less controlled by the speaker’s intent as compared to native speakers’ production, we hypothesize that under-informativeness is more likely to be

attributed to *inability* (rather than *unwillingness*) in non-native as compared to native speakers. This asymmetry is likely to affect further behavior. Despite the tendency to avoid learning from under-informative individuals (Gweon et al., 2014), participants may be more willing to give the benefit of the doubt to an under-informative non-native speaker as compared to an under-informative native speaker, considering the lower social penalties associated with being an unable as opposed to an unwilling social partner (cf. Behne et al., 2005). We test these hypotheses in a series of four experiments.

3.1 Experiment 4

Experiment 4 investigated perceptions of under-informativeness in a simple, everyday context. As in Experiments 1-3, we manipulated the identity of the speaker (native vs. non-native) in the written modality, so as to keep all other properties of the linguistic stimulus identical between conditions.

3.1.1 Participants

One hundred twenty-six monolingual English speakers aged 18-47 ($M = 29.64$, $SD = 4.61$) living in the United States, 62 of whom were female, were recruited from Amazon's Mechanical Turk to participate. Participants were compensated \$0.30 for the 3-minute study.

3.1.2 Materials and Procedure

Participants saw a picture of a young Asian-looking woman and underneath it a description of either a native speaker, Emma (Native Speaker condition, $N = 63$ participants) or a non-native speaker, Yuqi (Non-Native Speaker condition, $N = 63$ participants). The description read: "This is Emma/Yuqi. Emma/Yuqi is a college

student at the University of Delaware, majoring in history. She moved to Delaware from Boston/China three years ago and still has a strong Boston/Chinese accent. In her spare time, Emma/Yuqi likes to paint and play the piano.” A comprehension question followed, which asked participants to indicate where Emma/Yuqi was from. Accuracy was high (90%). Then the picture of the woman reappeared next to a picture of a refrigerator that contained bananas, apples, and pears. The following text accompanied the picture: “Emma’s/Yuqi’s friend asks for a snack. Emma/Yuqi looks in the refrigerator and says: ‘There are bananas and apples.’”

Participants were then asked: “Why didn’t Emma/Yuqi say that there were bananas, apples, and pears in the refrigerator?” and were instructed to write in their own response.

3.1.3 Results

Responses were coded as involving inability or unwillingness, each with several sub-types (see Table 3.1). Three responses that included both types of justification (e.g., “She didn’t know the word or she didn’t want to tell her friend there were pears”) were removed. Inability for native speakers was mostly associated with difficulty of perceiving, identifying or remembering the unmentioned object (31.74% of responses); for non-native speakers, inability was again associated with perceptual or cognitive difficulty (34.93%) but also with problems with naming the omitted object (41.27%). Within the unwillingness class, a frequent explanation for native speakers was deception (26.98%) and social considerations such as politeness towards others or saving face for one’s own sake (25.40% combined); for non-native speakers, deception was the predominant sub-type but was only half as frequent as for native speakers (12.70%).

A binary logistic regression was then performed with Speaker (Native, Non-Native) as the independent variable and Justification (Inability, Unwillingness) as the dependent variable. The model accounted for a significant amount of variance, $\chi^2(1) = 17.69$, $p < .001$, R^2 (Hosmer-Lemeshow) = 0.10. As can be seen in Table 3.2 and Figure 3.1, Justification varied by Speaker, such that the odds of a participant believing that the speaker was unable – rather than unwilling – to be fully informative were 4.86 times greater in the Non-Native Speaker ($M = 76.20$, $SD = 0.43$) condition as compared to the Native Speaker ($M = 39.68$, $SD = 0.49$) condition. The proportion of Inability justifications was not significantly different from chance in the Native Speaker condition ($p = .130$), but differed from chance in the Non-Native Speaker condition ($p < .001$).

Table 3.1: Breakdown of justifications given in Experiment 4 by Speaker and Type.

Justification Type	% Native Responses	% Non-Native Responses	Example
Inability	39.68%	76.20%	
Linguistic difficulty	7.94%	41.27%	She didn't know the word for pears.
Perceptual or cognitive difficulty	31.74%	34.93%	She didn't see the pears./ She forgot about the pears./She thought the pears were apples.
Unwillingness	60.32%	23.80%	
Deception	26.98%	12.70%	She wanted to keep the pears. She knew her friend didn't like pears, so she only offered her fruit she liked.
Politeness	19.05%	3.17%	She is embarrassed of her accent.
Saving face	6.35%	3.17%	It was her choice.
Other	7.94%	4.76%	

Table 3.2: Results of the binary logistic regression model for Experiment 4.

	β	SE	Odds Ratio	p	95% Confidence Intervals	
					Lower	Upper
(Intercept)	-0.42	0.26	0.66	.104	-0.93	0.08
Speaker	1.58	0.39	4.86	<.001	0.83	2.37

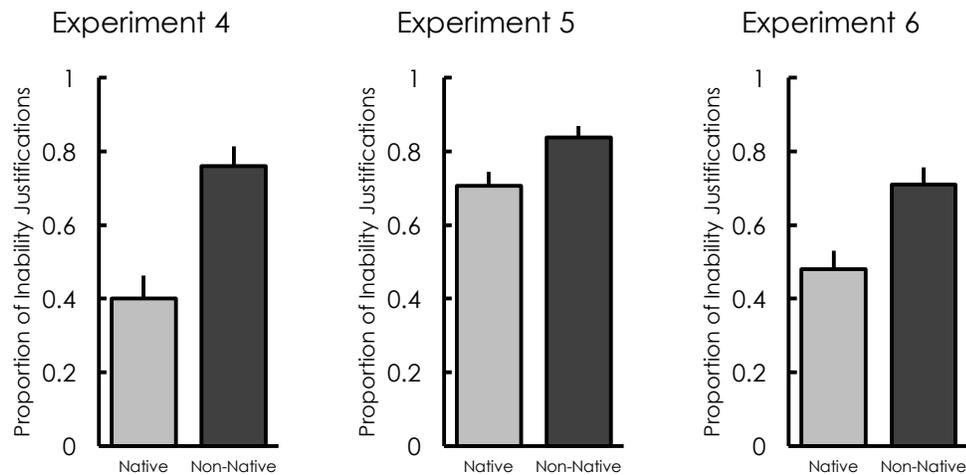


Figure 3.1: Proportion of Inability Justifications in Experiments 4 (left), 5 (center) and 6 (right). Error bars represent ± 1 S.E.M.

3.1.4 Discussion

In Experiment 4, under-informativeness elicited various interpretations as anticipated by pragmatic theory (Geurts, 2010; Sperber & Wilson, 1995). However, these interpretations differed for native and non-native speakers: leaving out information was more likely to be attributed to inability (as opposed to unwillingness) to say more in non-native compared to native speakers.

3.2 Experiment 5

In Experiment 5, participants chose between an Inability (cognitive difficulty) and an Unwillingness (deception) explanation for a speaker's under-informativeness to test whether the asymmetrical appeal to inability for native vs. non-native speakers would emerge even when a speaker's linguistic difficulty was set aside as a potential explanation.

3.2.1 Participants

Two hundred seventy-eight monolingual English speakers aged 19-67 ($M = 29.54$, $SD = 5.83$) living in the United States, 148 of whom were female, were recruited from Amazon's Mechanical Turk. Participants were compensated \$0.20 for the 2-minute study.

3.2.2 Materials and Procedure

The procedure was identical to Experiment 4, except that participants were provided with an Inability justification ("She forgot to mention the pears") and an Unwillingness justification ("She doesn't want her friend to know there are pears") to choose from. There were 136 participants in the Native Speaker condition and 142 in the Non-Native Speaker condition.

3.2.3 Results

A binary logistic regression was performed with Speaker (Native, Non-Native) as the independent variable and Justification (Inability, Unwillingness) as the dependent variable. The model accounted for a significant amount of variance, $\chi^2(1) = 6.98$, $p = .008$, R^2 (Hosmer-Lemeshow) = 0.02. As can be seen in Table 3.3 and Figure 3.1, Justification varied by Speaker, such that the odds of a participant believing that

the speaker left out a piece of information due to their inability were 2.16 times greater in the Non-Native Speaker condition ($M = .84, SD = 0.37$) as compared to the Native Speaker condition ($M = .71, SD = 0.46$). Even though Inability justifications occurred more often in the Non-Native speaker condition, in both conditions they were the preferred choice, with their proportions significantly different from chance (both p 's < .001).

Table 3.3: Results of the binary logistic regression model for Experiment 5.

	β	SE	Odds Ratio	p	95% Confidence Intervals	
					Lower	Upper
(Intercept)	0.88	0.19	2.40	<.001	0.51	1.25
Speaker	0.77	0.30	2.16	.009	0.20	1.36

3.2.4 Discussion

Experiment 5 showed that under-informativeness was more likely to be attributed to inability in a non-native compared to a native speaker. Even though the Inability choice did not involve language, the bias to treat under-informativeness as unintentional in non-native speakers produced this result.

3.3 Experiment 6

Experiment 6 sought to replicate Experiment 5 in a novel context. An inventor who was either a native or a non-native speaker taught others about her invention but omitted one feature. Participants chose between two explanations (inability vs. unwillingness) for the omission. Unlike Experiment 5, the omitted information was

clearly known to the speaker and was highly relevant for the listeners (science fair visitors trying to learn about new inventions). Additionally, the context made the two explanations more specific and plausible: the speaker had many other inventions and might have forgotten the feature ('inability'); furthermore, the omitted feature was a limitation that the speaker might have wanted to downplay ('unwillingness').

3.3.1 Participants

Two hundred monolingual English speakers aged 18-49 ($M = 29.77$, $SD = 5.72$) living in the United States, 111 of whom were female, were recruited from Amazon's Mechanical Turk. Participants were compensated \$0.20 for the 2-minute study.

3.3.2 Materials and Procedure

Participants were first presented with the image of a novel object and its three functions in Figure 3.2 and told: "Read about this Zeg and learn what it is. Try to remember what it does as best you can. Once you're done reading all of the information on this page, click the NEXT button." To encourage thorough examination of the object, the experiment would not advance until the participant had spent 30 seconds on the page (the NEXT button was not available for 30 seconds, and a countdown timer appeared in the corner of the page). Participants were then presented with descriptions of the inventor of the Zeg (see Figure 3.3) - either Emma Smith with a strong Boston accent (Native Speaker condition, $N = 100$ participants), or Yuqi Chen with a strong Chinese accent (Non-Native Speaker condition, $N = 100$ participants). The description included the information that the Zeg was one of their many inventions. Comprehension questions followed the inventor description ("Where

is Emma/Yuqi from?”, “What instrument does Emma/Yuqi play?”). These were answered with generally high accuracy (94% and 76% respectively). On the next screen, participants read the following text: “Emma/Yuqi is sharing her invention, the Zeg, at the amateur inventor club's annual public science fair. This is what Emma/Yuqi says about the Zeg to people who visit her display.” This text was accompanied by a picture of a young Asian-looking woman next to a picture of the Zeg, with a speech bubble coming from the woman that contained the following description: “The Zeg cuts dough into noodles and separates them.”

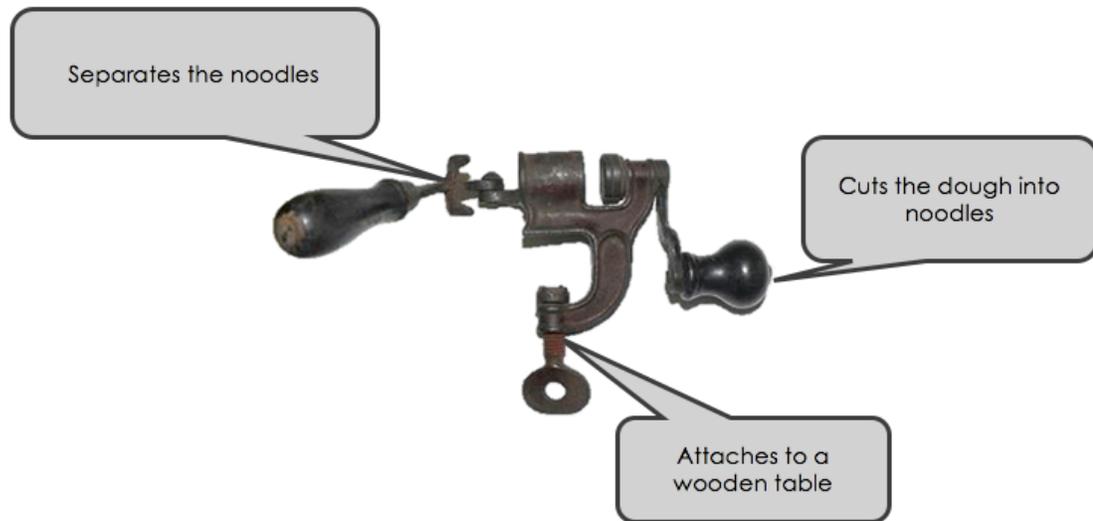


Figure 3.2: Novel object used in Experiment 6 (“Zeg”) and its three functions.

<p>The Zeg is one of Emma Smith's many inventions. She likes to create and build new tools, and is a member of an amateur inventor club. Emma also likes to hike and play the piano. She is originally from Boston but recently moved to Delaware and her neighbors say that she still has a strong Boston accent.</p>	OR	<p>The Zeg is one of Yuqi Chen's many inventions. She likes to create and build new tools, and is a member of an amateur inventor club. Yuqi also likes to hike and play the piano. She is originally from China but recently moved to Delaware and her neighbors say that she still has a strong Chinese accent.</p>
--	----	---

Figure 3.3: Speaker descriptions used in Experiment 6.

Participants then responded to the question: “Why didn’t Emma/Yuqi say that the Zeg attaches to a wooden table?” Two options were provided, an Inability justification (“She forgot because she has so many inventions”) and an Unwillingness justification (“She didn’t want people to know that it doesn’t attach to other surfaces”).

3.3.3 Results

A binary logistic regression was performed with Speaker (Native, Non-Native) as the independent variable and Justification (Inability, Unwillingness) as the dependent variable. The model accounted for a significant amount of variance, $\chi^2(1) = 11.10$, $p = .001$, R^2 (Hosmer-Lemeshow) = 0.04. As can be seen in Table 3.4 and Figure 3.1, Justification varied by Speaker, such that the odds of a participant believing that the speaker left out a piece of information due to inability were 2.65 times greater in the Non-Native Speaker ($M = 71.00$, $SD = 0.50$) condition as compared to the Native Speaker ($M = 48.00$, $SD = 0.46$) condition. The proportion of Inability justifications did not differ significantly from chance in the Native Speaker condition ($p = .764$), but did so in the Non-Native Speaker condition ($p < .001$).

Table 3.4: Results of the binary logistic regression model for Experiment 6.

	β	SE	Odds Ratio	p	95% Confidence Intervals	
					Lower	Upper
(Intercept)	-0.08	0.20	0.92	.689	-0.47	0.31
Speaker	0.98	0.29	2.65	.001	0.40	1.57

3.3.4 Discussion

We found that participants were more likely to link under-informativeness to inability in non-native compared to native speakers. In fact, for non-native speakers inability was the leading explanation for under-informativeness, whereas for native speakers explanations were split between inability and unwillingness.

3.4 Experiment 7

In Experiment 7, we investigated how perceived reasons for under-informativeness impact future learning by asking participants to decide whether or not to learn from an under-informative inventor again. We expected that participants would avoid learning from an under-informative speaker (as in Gweon et al., 2014), but that this effect would vary depending on the inventor's linguistic background. If under-informativeness is more likely to be attributed to inability instead of unwillingness in non-native vs. native speakers, participants should be more likely to choose to learn from an under-informative inventor who is a non-native as opposed to a native speaker.

3.4.1 Participants

Four hundred new monolingual English speakers aged 18-59 ($M = 30.13$, $SD = 6.26$) living in the United States, 193 of whom were female, were recruited from Amazon's Mechanical Turk. Participants were compensated \$0.30 for the 3-minute study.

3.4.2 Materials and Procedure

Procedure was similar to Experiment 6. Participants were first presented with the novel object and facts in Figure 3.2 (except that "Attaches to a wooden table" was

changed to “Attaches to a table” and the speaker descriptions began with “Emma Smith/Yuqi Chen is the inventor of the Zeg” instead of “The Zeg is one of Emma Smith’s/Yuqi Chen’s many inventions” to provide a more neutral context). Then a description of one of the inventors was presented (see Figure 3.3). The same comprehension questions as in Experiment 6 followed and were answered generally accurately (94%, 76%).

As in Experiment 6, participants next read the following text: “Emma/Yuqi is sharing her invention, the Zeg, at the amateur inventor club's annual public science fair. This is what Emma/Yuqi says about the Zeg to people who visit her display.” This text was accompanied by a picture of a young Asian-looking woman next to a picture of the Zeg, with a speech bubble that contained one of the following descriptions: “The Zeg attaches to a table, cuts dough into noodles, and separates the noodles” (Informative condition), or “The Zeg attaches to a table” (Under-Informative condition). Unlike Experiment 6, only one of the three functions was mentioned in the Under-Informative condition to increase the severity of information omission. Speaker (Native, Non-Native) and Informativeness (Informative, Under-Informative) were manipulated between-subjects in a latin-square design, with equal numbers of participants ($N = 100$) in each condition.

Participants next responded to a Helpfulness question (“How helpful was Emma/Yuqi to people who visited her display to learn about the Zeg?”) using a scale ranging from 1 (not helpful) to 5 (helpful). The Helpfulness Rating served as a check that the Informativeness manipulation was effective: the Under-Informative inventors should elicit lower ratings than the Informative inventors (as in Gweon et al., 2014). Then participants read that “Emma/Yuqi is developing a new tool called the Plib” and

saw a second novel object. Participants were asked: “How would you like to learn about the Plib?”, and had to click either on the picture of the previous inventor (with the name Emma/Yuqi mentioned underneath the picture as a reminder) or the picture of a new Asian-looking woman without any details given about her (except for the name Sue/Su – depending on Speaker condition - mentioned underneath her picture). The binary variable Teacher Choice (Same – Emma/Yuqi, New – Sue/Su) was our main dependent variable of interest.

3.4.3 Results

A 2 (Speaker: Native, Non-Native) by 2 (Informativeness: Informative, Under-Informative) factorial ANOVA was performed on participants’ mean Helpfulness Ratings (see Figure 3.4). As predicted, perceptions of helpfulness were influenced by the Informativeness of the speaker, $F(1, 396) = 114.66, p < .001, \eta_p^2 = .22$, such that Informative inventors ($M = 3.47, SD = 1.00$) were judged as more helpful than Under-Informative inventors ($M = 1.87, SD = 1.22$). Helpfulness ratings did not vary by Speaker, $F(1, 396) = 0.58, p = .448, \eta_p^2 < .01$, and the Speaker by Informativeness interaction was not significant, $F(1, 396) = 0.65, p = .420, \eta_p^2 < .01$. Thus people linked the helpfulness of a speaker simply to the informational content of their utterance.

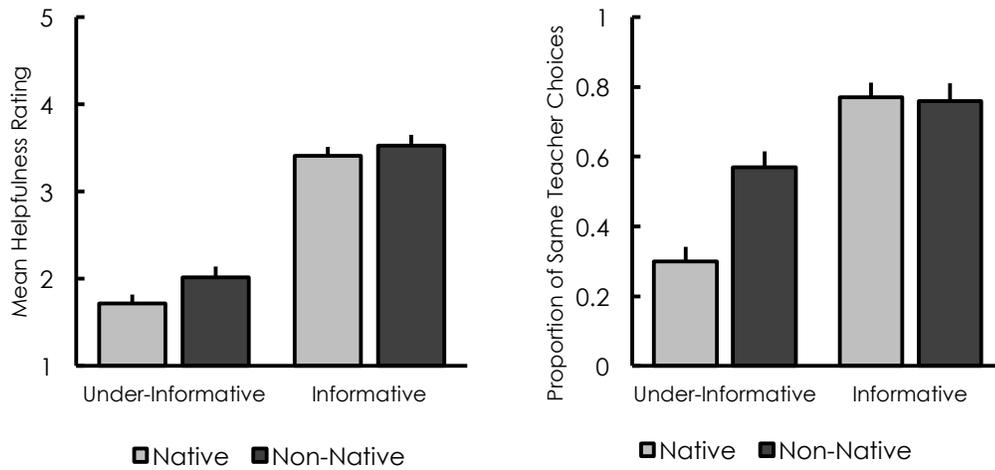


Figure 3.4: Mean Helpfulness Ratings (left) and the proportion of Same Teacher Choices (right) in Experiment 7. Error bars represent 95% confidence intervals.

To determine whether future learning behavior is affected by previous demonstrations of under-informativeness and non-native speaker status, a binary logistic regression was performed with Speaker and Informativeness as independent variables and Teacher Choice (Same, New) as the dependent variable (Figure 3.4 and Table 3.5). The model accounted for a significant amount of variance, $\chi^2(3) = 61.50, p < .001, R^2$ (Hosmer-Lemeshow) = 0.11. Teacher Choice varied by Informativeness, such that the odds of a participant choosing the Same teacher were 0.13 times lower in the Under-Informative condition than in the Informative condition. The main effect of Speaker was not significant, but there was a significant Informativeness by Speaker interaction. Specifically, post-hoc tests (Tukey) indicated that participants were more likely to choose to learn again from an Under-Informative Non-Native Speaker ($M = 0.57, SD = 0.50$) than an Under-Informative Native Speaker ($M = 0.30, SD = 0.46$), $p = .017$. The likelihood of choosing to learn again from an Informative Native ($M =$

0.77, $SD = 0.42$) vs. Non-Native ($M = 0.76$, $SD = 0.43$) Speaker did not differ significantly, $p = .999$.

Table 3.5: Results of the binary logistic regression model for Experiment 7.

	β	SE	Odds Ratio	p	95% Confidence Intervals	
					Lower	Upper
(Intercept)	1.21	0.24	3.35	< .001	0.76	1.70
Informativeness	-2.06	0.03	0.13	< .001	-2.71	-1.44
Speaker	-0.06	0.33	0.95	.868	-0.72	0.60
Informativeness*Speaker	1.18	0.45	3.27	.008	0.31	2.07

3.4.4 Discussion

As expected (see Gweon et al., 2014), under-informative speakers were deemed to be less helpful than fully informative speakers and were dispreferred as sources of further learning. However, under-informative speakers were more likely to be given a second chance if they were non-native speakers (57% vs. 30% for native speakers). The present experiment extended our previous work by showing that such explanations are computed spontaneously when needed and affect future behavior.

3.5 General Discussion

Recent research demonstrates negative social biases towards non-native speakers (Kinzler et al., 2007; Lev-Ari & Keysar, 2010, 2012, a.o.) and differences in the comprehension of native and non-native speech (Hanulíková et al., 2012; Gibson et al., 2017; Grey & Van Hell, 2017). Here we bridged and extended these two strands of research by investigating how pragmatic and further social inferences might differ

regarding native and non-native speakers. We focused on under-informativeness, a phenomenon that has received considerable attention in recent psycholinguistics work (Noveck, 2001; Papafragou & Musolino, 2003; Bott & Noveck, 2004; Guasti et al., 2005).

Overall, native comprehenders treated under-informative utterances differently depending on the identity of the speaker. In Experiment 4, comprehenders were more likely to explain under-informativeness as the result of inability to give sufficient information – rather than unwillingness to do so – for non-native as compared to native speakers. Furthermore, inability explanations for native speakers mostly invoked difficulty with seeing, recognizing or remembering the unmentioned object but for non-native speakers they also often invoked difficulty with naming the object. Unwillingness explanations mostly invoked deception or conflict with one’s own or the hearer’s social preferences, and both sub-types were numerically less frequent for non-native speakers. In Experiments 5 and 6, the inability vs. unwillingness asymmetry was replicated in a forced-choice task, even though the provided inability explanation was non-linguistic in nature. We hypothesize that this result was due to a general bias to consider failed communicative behavior as less voluntary for non-native speakers. Finally, in Experiment 7, participants were more likely to choose to learn from an under-informative non-native speaker than an under-informative native speaker, presumably because they gave the benefit of the doubt to the speaker whose prior under-informativeness was less likely to have been voluntary.

The present results provide novel evidence for the complexity of inferences underlying the processing of under-informativeness. Specifically, they show that listeners consider both the speaker’s abilities and preferences when making

conversational inferences in ways that had been outlined by linguistic theories (e.g., Grice, 1975; Sperber & Wilson, 1995; Geurts, 2011, among many others) but not yet fully confirmed experimentally.

The present data also contribute to a growing body of work demonstrating differences in how native and non-native speech is comprehended across multiple linguistic domains (Hanulíková et al., 2012; Grey & Van Hell, 2017; Gibson et al., 2017). In the domain of pragmatics, they show that, when a speaker is under-informative, listeners consider (native/non-native) speaker identity when making inferences about why the speaker failed to say more; furthermore, these inferences have further consequences for social cognition and future behavior. The current data comport with – and explain - recent findings showing that listeners are more accepting of under-informative statements when those are attributed to non-native as compared to native speakers (Chapter 2): listeners are likely to attribute under-informativeness to ineptness, not willful choice of stimulus, in non-native speakers. More broadly, our data support the position that foreign accents are not just physical features of a linguistic stimulus but sources of mental-state information about the speaker involving knowledge, preferences, and intentions.

As in Chapter 2, we observed speaker-specific differences in processing linguistic stimuli that emerged in the absence of actual encounters with foreign accents. Thus a simple, top-down manipulation of speaker identity mobilized expectations about non-native speech that subsequently affected social-pragmatic inferences across seven experiments. While this speaks to the role of expectations in language processing, future work should ask whether different types of accents are equally likely to induce changes in pragmatic processing, how comprehenders'

specific language backgrounds might affect the results, and how the present findings generalize to actually perceived (as opposed to imagined) accents.

Our conclusions may appear puzzling given prior research showing negative biases towards non-native speakers (e.g., Lev-Ari & Keysar, 2010, 2012) that seem to begin in infancy (Kinzler et al., 2007). Nevertheless, they are consistent with recent results suggesting that the error-prone nature of non-native speech can have some advantages: syntactically errorful utterances are less likely to elicit surprise (Hanulikova et al., 2012) and implausible messages more likely to be reinterpreted (Gibson et al., 2017) when produced by non-native compared to native speakers. Our data do suggest a cost for non-native speakers such that their communicative contributions are perceived as less likely to be the product of willful choice. Overall, however, being perceived as a somewhat inept communicator can have unexpected social benefits.

Chapter 4

LISTENER EFFECTS ON PRAGMATIC MEANING: INDIVIDUAL DIFFERENCES IN SCALAR IMPLICATURE AND OTHER PRAGMATIC DOMAINS

People do not always say exactly what they mean; for stylistic reasons, to mislead a listener, to keep a secret, and so forth. Despite this, the vast majority of daily communication proceeds smoothly, made possible by our ability to process meaning at two levels. Under the classic view, listeners first process the literal semantic meaning of an utterance, and then “read between the lines” to pragmatically enrich the meaning with inferences about the speaker’s intended meaning. These inferences come from listeners’ strong expectations that the speaker will be a cooperative interlocutor – informative, truthful, relevant, and concise (Grice, 1975). When any of these expectations appear to be violated by the speaker, the listener attempts to reconcile this by making an inference about the speaker’s true, cooperative, meaning. For example, when Jane says Mary is a “night owl,” it appears to be untruthful – clearly she is not a literal bird. The listener expects so strongly that Jane will be cooperative, however, that she infers that Jane really meant that Mary tends to stay up late.

The focus of the present chapter is the case of scalar implicature, in which given an under-informative statement like “Some giraffes have long necks,” the listener assumes that the presumably cooperative speaker intended *not all* giraffes have long necks, or they would have said otherwise. Such statements are false if one derives the scalar implicature, although the literal, semantic meaning is true – some

giraffes have long necks, *in fact they all do*. Listeners vary in the extent to which they adopt the literal or pragmatic meaning of such under-informative utterances, but the underlying reason for this variation is unknown.

4.1 Executive Function and Scalar Implicature

Currently, empirical evidence for the specific role of EF in scalar implicature computation is unsettled. A study by De Neys and Schaeken (2007) found support for a role of EF – and working memory specifically – in implicature computation: participants were more likely to accept under-informative statements such as “Some dogs are mammals” when they were under greater cognitive load (e.g., when they were concurrently memorizing complex as opposed to simple dot patterns; see also Marty & Chemla, 2013). Similarly, Antoniou et al. (2016) found that working memory capacity within EF predicted the rate at which adults rejected under-informative sentences used to describe visual scenes (“There are hearts on some of the cards”). Furthermore, neuroimaging studies have demonstrated that frontal cortical regions associated with EF are activated during scalar implicature computation (Shetreet, Chierchia, & Gaab, 2014; Politzer-Ahles & Gwilliams, 2015). However, because many of these studies focused specifically on EF and either did not include other variables in the design (De Neys & Schaeken, 2007; Marty & Chemla, 2013) or did not analyze other variables in detail (Antoniou et al., 2016), it is possible that the observed effects of EF were due to third factors. For instance, involvement of EF in scalar implicature could be tied to engagement of Theory of Mind (e.g., Apperly, 2012, and next section).

Adding to the complicated picture, a large-scale study with Dutch students by Heyman and Schaeken (2015) failed to find relationships between EF abilities and

scalar implicature computation. Furthermore, a very large body of individual-differences work examining the contributions of ToM and (different aspects of) EF to pragmatic abilities in both typically and atypically developing children has failed to yield converging evidence for reliable associations (see Matthews, Biney, & Abbot-Smith, in press, for a detailed review).

4.2 Theory of Mind and Scalar Implicature

Scalar implicature computation has been investigated in adolescents and adults with Autism Spectrum Disorder (ASD), a group known to have ToM deficits but to also involve individuals with widely different cognitive and linguistic profiles (Newschaffer et al., 2007). In one study, ASD participants (who scored poorly on ToM measures of false belief) adopted pragmatically under-informative interpretations of stories (e.g., that *all* of the children were in a pool when only 2 out of 3 children were, as the story read and as the accompanying picture showed) more often than neurotypical controls, as anticipated by the hypothesis that ToM is involved in scalar implicature computation (Noveck, Guelminger, Georgieff, & Labruyere, 2007). Other studies have shown that ASD adolescents (Pijnacker et al., 2009; Hochstein, Bale, & Barner, 2017) and adults (Chevallier Wilson, Happé, & Noveck, 2010) are not impaired in their ability to compute implicatures. These last studies, however, either did not measure the ToM abilities of the participants (Pijnacker et al., 2009; Chevallier et al., 2010), or reported a (relatively high⁴) ToM score only for the ASD group and

⁴ Hochstein et al. (2017) reported that their participants with Autism gave mental state justifications 72% of the time on a version of Happé (1994)'s Strange Stories task. For comparison, using Hochstein et al.'s coding scheme, our participants in Experiment 1 gave mental state justifications 78% of the time.

not the control group (Hochstein et al., 2017). Hochstein et al. (2017) argued that ToM is not required for scalar implicature, but both Pijnacker et al. (2009) and Chevallier et al. (2010) concluded that high-functioning individuals with Autism, such as their participants, could have basic ToM skills that are sufficient for the computation of some pragmatic inferences.

While the relationship between ToM and pragmatic inference follows straightforwardly from our knowledge of how scalar implicatures are calculated, researchers have yet to explicitly use ToM to explain variability in the pragmatic skills of neurotypical adults. This is perhaps due in part to a perception that because adults have a fully developed ToM, there is insufficient variation to test such a relationship. Instead, prior work has used related terminology such as “social-communicative skills” primarily measured using the Autism Quotient questionnaire (AQ; Baron-Cohen et al., 2001). The AQ – originally developed as a self-assessment tool for adults that could potentially be used to screen for Autism spectrum conditions – measures the number of Autistic traits a person possesses. On the Communicative Subscale of the AQ, which is often used in studies of pragmatics (e.g., Nieuwland, Ditman, & Kuperberg, 2010; Zhaio, Liu, Chen, & Chen, 2015), examples of such Autistic traits include being slow to understand a joke and having difficulties with turn-taking in telephone conversations. Thus, the AQ can be seen as a proxy for ToM, which is presumably why it has been used in investigations of scalar implicature. It is an imperfect proxy, however, because it is highly metacognitive, requiring participants to reason about their own social skills. It does not directly measure an individual’s ability to, for example, understand a joke or have a telephone conversation, and certainly does not directly measure the ability to reason about others’ mental states. The

relationship between AQ scores and pragmatic judgments is inconsistent, with our own prior work (among others) failing to find evidence of a link (Heyman & Schaeken, 2015; Antoniou et al., 2016; Barbet & Thierry, 2016; Chapter 2, but see Nieuwland et al.; Zhao et al., 2015 for evidence that neural responses to under-informative statements differ between high-AQ and low-AQ groups).

Outside of work on scalar implicature, there is neural evidence linking other pragmatic abilities with ToM. Engaging in ToM activates a network of cortical regions, most notably the right Temporo-Parietal Junction (rTPJ; Saxe & Kanwisher, 2003). Such regions are also activated when processing metaphors (Prat, Mason, & Just, 2012), indirect requests (Van Ackeren et al., 2012), irony (Eviatar & Just, 2006), and jokes (Feng, Ye, Mao, & Yue, 2014; Kline, Gallee, Balewski, & Fedorenko, Submitted). Furthermore, patients with lesions in these ToM areas have impairments in processing metaphors (Champagne-Lavau & Joannette, 2009) and jokes (Winner et al., 1998).

4.3 The Present Study

The evidence reviewed so far reveals several limitations in the current literature that preclude firm conclusions about the potential importance of EF and ToM in explaining individual differences in adults' scalar implicature computation. Perhaps the most striking limitation is that the two abilities have yet to be investigated in a single group of individuals. It is important to do so to tease apart the unique influence of each factor since ToM and EF are often correlated with one another (Carlson & Moses, 2001; Carlson, Moses, & Breton, 2002; Hughes & Ensor, 2007; Bull, Phillips, & Conway, 2008; Apperly, 2012). Relatedly, best practices in individual differences research (e.g., Cronbach, 1957; Miyake et al., 2000) require

large sample sizes, multiple measures, and proof of replicability. Most of the studies reviewed above are relatively small-scale and fail to fulfill these methodological requirements (see also Matthews et al., in press, for similar issues with developmental evidence).

Viewed more broadly, the studies reviewed above raise the question whether and how individual differences documented in scalar implicature computation relate to other pragmatic phenomena. If the computation of pragmatic meaning in general involves cognitive cost that results from holding and manipulating representations in working memory (cf. De Neys and Schaeken, 2007; Antoniou et al., 2016, among others), further types of pragmatically enriched meaning should also be expected to incur similar costs and be associated with EF (working memory) abilities (see Chiappe & Chiappe, 2007; and Mashal, 2013; for evidence from metaphor comprehension). Similarly, if pragmatic computation is a species of intention recognition (Grice, 1975; Sperber & Wilson, 1986), the involvement of ToM should be fairly general across pragmatic phenomena. In support of this hypothesis, neuroimaging research has shown that cortical regions, most notably the right Temporo-Parietal Junction (rTPJ) known to engage in ToM (Saxe & Kanwisher, 2003), are also activated when processing metaphors (Prat, Mason, & Just, 2012), indirect requests (Van Ackeren et al., 2012), irony (Eviatar & Just, 2006), and jokes (Feng, Ye, Mao, & Yue, 2014; Kline, Gallee, Balewski, & Fedorenko, submitted). Furthermore, patients with lesions in these ToM areas have impairments in processing metaphors (Champagne-Lavau & Joannette, 2009) and jokes (Winner, Brownell, Happé, Blum, & Pincus, 1998). Interestingly, in some of this patient work, the relationship between ToM and pragmatic abilities holds even when EF is intact (Champagne-Lavau & Joannette,

2009). However, at present, we lack evidence connecting individual differences in EF or ToM to performance across several pragmatic phenomena.

Here we present two experiments designed to address these crucial gaps in the literature. In Experiment 8, we seek to determine the unique contributions of EF and ToM to scalar implicature, a domain known for variation in judgments and a central topic in the study of pragmatic processing and development. In Experiment 9, we investigate the influence of EF and ToM on metaphor and indirect request comprehension, in addition to scalar implicature. In these experiments we go beyond prior work that typically investigated only one factor at a time in the context of scalar implicature derivation. Furthermore, we try to implement best practices in individual differences research by including large sample sizes (approximately 200 participants for each experiment), multiple measures (where possible) of pragmatic ability, EF, and ToM, as well as evidence of replicability. Finally, we investigate the mechanisms underlying individual variation in pragmatic computations in domains other than scalar implicature.

4.4 Experiment 8

We investigated scalar implicature computation using a dual-task paradigm as well as a number of individual differences measures. In the Dual Task, participants judged the extent to which under-informative (and other types of) sentences made sense while holding simple (Control trials) or complex (Load trials) dot patterns in memory. The goal of this Dual Task was to assess the contribution of EF to scalar implicature by comparing Control and Load memory conditions. The prediction from De Neys and Schaeken (2007) is that we under-informative statements should be judged as making more sense on Load trials as compared to Control trials, in keeping

with their finding that individuals are more logical (i.e., less likely to compute an implicature) under heavy cognitive load. This prediction rests on the assumption that EF is recruited in order to compute a scalar implicature. Ratings for the control sentences, where no implicature is involved, should not differ by Cognitive Load. In the Simple Task, we assessed scalar implicature computations when people were not engaged in a dual task paradigm. We chose a straightforward binary judgment task that has been widely used in the literature on scalar implicature for this purpose (e.g., Bott & Noveck, 2004; De Neys & Schaeken, 2007; Pijnacker et al. 2009; Slabakova, 2010; Bott, Bailey, & Grodner, 2012). Participants were asked to judge whether under-informative and fully informative statements made sense or not, and we related performance on this Simple Task to measures of EF and ToM to investigate the unique contribution of each to pragmatic inference.

4.4.1 Participants

Two hundred monolingual English speakers aged 18-47 ($M = 28.68$, $SD = 4.84$) living in the United States, 89 of whom were female, were recruited from Amazon's Mechanical Turk to participate. Participants were compensated \$1.50 for the 15-minute study. Data from 22 individuals who reported to be bilingual and/or diagnosed with Autism Spectrum Disorder were excluded, leaving 179 participants for analysis.

4.4.2 Materials and Procedure

Participants completed five tasks in the order described below: Dual Scalar Implicature Task, Auditory Digit Span Task (as a measure of EF), Simple Scalar Implicature Task, and the Mind in the Eyes and Strange Stories Tasks (as measures of

Theory of Mind). Stimuli for the Dual Scalar Implicature Task and Simple Scalar Implicature Task can be found in Appendices C and D, respectively.

4.4.2.1 Dual Scalar Implicature Task

The Dual Scalar Implicature Task was based on De Neys and Schaeken (2007). At the beginning of each trial, participants were presented with a pattern of 3 or 4 dots on a 3x3 grid, and were instructed to remember the pattern. There were two types of patterns which represented two Cognitive Load conditions (following Bethell-Fox & Shepard, 1988 and De Neys & Schaeken, 2007): Control patterns that were simple to remember, in which there were three dots in a horizontal, vertical, or diagonal row (a “1-piece” arrangement) and Load patterns designed to increase cognitive demands, in which there were four dots arranged in a “3-piece” pattern. Examples of dot patterns are presented in Figure 4.1. On the next screen, participants read a sentence and were asked to rate each of it on a scale from 1 (Very Bad – Doesn’t make sense) to 5 (Very Good – Makes perfect sense). Sentences of four types (see Table 4.1) used in our previous studies (Fairchild & Papafragou, October 2017) were presented to participants: True but Under-Informative sentences beginning with *some* (henceforth Under-Informative), True and Felicitous sentences beginning with *some* (henceforth, True (Some)), True and Felicitous sentences beginning with *all* (henceforth, True (All)), and False sentences beginning with *all* (henceforth, False). The Under-Informative sentences served as the critical trials, and the other three Sentence Types were treated as control sentences. The four Sentence Types did not differ from one another in sentence length as measured in either words or syllables (all p 's > .1). After the rating the sentence, participants were asked to recreate the dot pattern by clicking on the appropriate squares in the grid. There were 80 total trials in the task presented

in a random order for each participant: 20 trials of each Sentence Type, half with Control patterns and half with Load patterns (counterbalanced across participants resulting in two lists).

Of interest was to compare Control and Load memory conditions to assess the contribution of EF to scalar implicature. The prediction from De Neys and Schaeken (2007) is that we should find a Cognitive Load by Sentence Type interaction such that Under-Informative ratings should be higher on Load trials as compared to Control trials. Ratings for other Sentence Types should not differ by Cognitive Load, as no potentially cognitively costly inferences are required to process the meaning of these sentences.

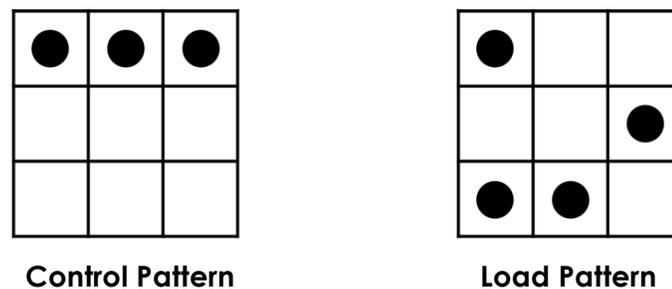


Figure 4.1: Examples of Control (left) and Load (right) patterns used in Experiment 8.

Table 4.1: Examples of stimuli used in Experiment 8.

Sentence Type	Example
Under-informative	Some people have noses.
True (Some)	Some people have pets.
True (All)	All snow is cold.
False	All women are doctors.

4.4.2.2 Auditory Digit Span Task

This task provided an EF measure. It consisted of 12 trials. On each trial, participants heard a computerized voice (Apple’s “Samantha”) utter 5, 6, 7, or 8 digits

between 1 and 9. Then, participants were asked to enter the string of digits in reverse. . We measured working memory instead of other types of EF, such as inhibition, for its demonstrated relationship with higher-order cognition in general (Engle, Tuholski, Laughlin, & Conway, 1999; Kane & Engle, 2002; Kane et al., 2004) and scalar implicature specifically (De Neys & Schaeken, 2007; Marty & Chemla, 2013). This reverse digit span was chosen specifically as it provides a truer measure of working memory than, e.g., a forward digit span task because participants are required to perform operations on stored material (Diamond, 2013). A memory span score representing participants' EF was calculated for each participant by taking the number of digits in the longest span correctly recalled (maximum possible score of 8).

4.4.2.3 Simple Scalar Implicature Task

Ten sentences borrowed from Bott and Noveck (2004) were presented individually in a random order, and participants judged whether they were “Good” or “Bad.” Participants were instructed that a Good sentence is one that makes sense, and a Bad sentence is one that does not make sense. Participants were instructed that a Good sentence is one that makes sense, and a Bad sentence is one that does not make sense. Five of the sentences were Under-Informative (“Some dogs are mammals”) and five were Informative (“Some fish are tuna”). The Informative sentences served as control trials, and were designed to be consistently judged as “Good.” The critical Under-Informative sentences could also be judged as “Good” if the participant adopted the literal interpretation of the statement (e.g., “Some *and possibly all* dogs are mammals”), or they could be judged as “Bad” if the participants derived a scalar implicature (e.g., “*Not all* dogs are mammals”). Results from this task were related to EF and ToM measures.

4.4.2.4 Mind in the Eyes and Strange Stories Task

These tasks were designed to be advanced tests of ToM and had previously been used with adults (Baron-Cohen et al., 1997; Joliffe & Baron-Cohen, 1999; Baron-Cohen et al., 2001). Additionally, both tasks required sophisticated linguistic abilities, and therefore were more likely to relate to the linguistic judgments on the scalar implicature tasks. We used an abridged 12-trial version of the Mind in the Eyes Task (Baron-Cohen et al., 1997; Baron-Cohen et al., 2001). Participants were presented with pictures of eyes and were asked to choose one word out of four choices that best described what the person was thinking or feeling (see Figure 4.2 for an example). The 12 trials were presented randomly, and the locations of the four answer choices were also randomized.

We also used an abridged version of the Strange Stories Task (Happé, 1994) consisting of 7 experimental and 3 control trials. Each of the experimental trials featured a short story describing a situation that involved pretend play, joking, white lies, figures of speech, irony, misunderstanding, or forgetting. Here is an example of such a story: “Katie and Emma are playing in the house. Emma picks up a banana from the fruit bowl and holds it up to her ear. She says to Katie, ‘Look! This banana is a telephone!’” Participants were asked to explain why Emma said this, and the number of mental state (e.g., “She is pretending”) and physical state (e.g., “A banana looks like a telephone”) responses were tallied (coding was verified by a second rater, $IRR = .91$). Control stories did not involve mental states. A composite ToM measure was calculated by taking the mean number of correct Mind in the Eyes trials and Strange Stories mental state justifications.

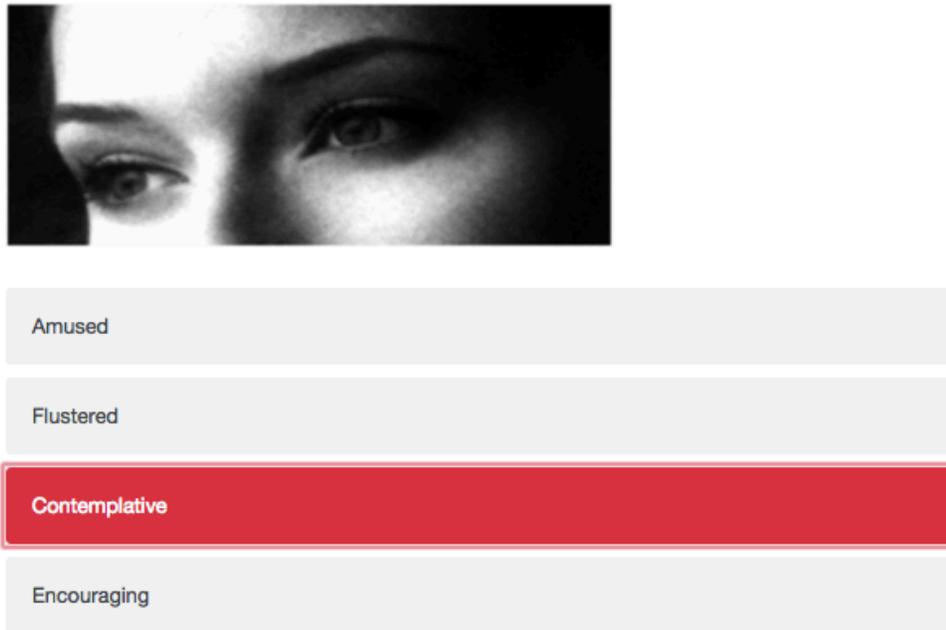


Figure 4.2: Example trial from the Mind in the Eyes Task (correct answer highlighted).

4.4.3 Results

4.4.3.1 Dual Scalar Implicature Task

Results are presented in Figure 4.3. A linear mixed-effects regression was performed on Sentence Rating data (excluding incorrect memory trials) using the *nlme* package (Pinheiro et al., 2017) for the R Project for Statistical Computing v3.2.2 (R Core Team, 2015). Cognitive Load (Control, Load), Sentence Type (Under-Informative, True (Some), True (All), False), and the interaction between the two were included in the model as fixed effects, with crossed random intercepts for Participants and Items. Sentence Ratings differed significantly across Sentence Types, $\chi^2(3) = 6953.639, p < .001$. Planned contrasts (presented in Table 4.2) indicated that Under-Informative ($M = 2.89, SD = 1.59$) sentences were rated higher than False ($M = 2.18,$

$SD = 1.43$) sentences, $p < .001$, but worse than True (Some) ($M = 4.51$, $SD = 0.83$) sentences, $p < .001$. True (All) ($M = 4.34$, $SD = 1.05$) sentences were rated lower than True (Some) sentences, $p < .001$ (perhaps because participants thought that the quantifier *all* was superfluous). Importantly for present purposes, Sentence Ratings did not differ significantly between Cognitive Load conditions, $\chi^2(1) = 0.898$, $p = .343$, and the interaction between Cognitive Load and Sentence Type did not reach significance, $\chi^2(3) = 6.859$, $p = .077$.

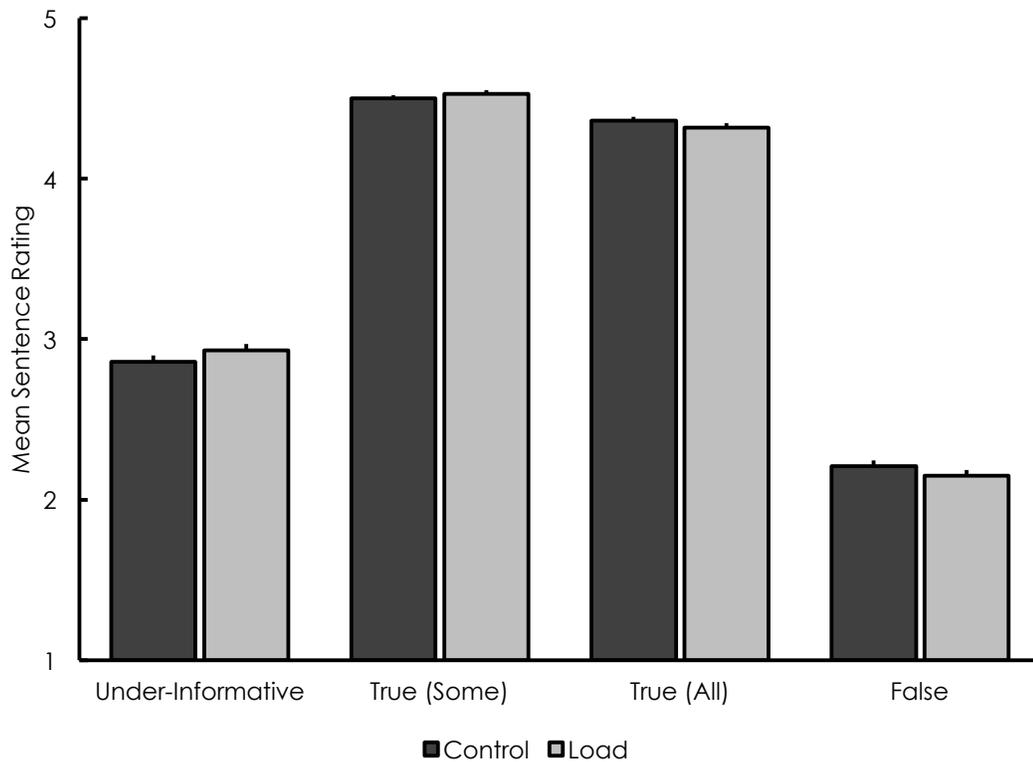


Figure 4.3: Mean Sentence Ratings by Sentence Type and Cognitive Load condition in the Dual Scalar Implicature Task in Experiment 8. Error bars represent +1 S.E.M.

Table 4.2: Linear mixed-effects regression for the Dual Scalar Implicature Task results of Experiment 8, with Cognitive Load, Sentence Type, and their interaction included as fixed effects and crossed random intercepts for Participant and Item.

Effect	β	S.E.	t	p
Intercept	3.475	0.039	88.269	< .001
Cognitive Load (Control vs. Load)	-0.002	0.010	-0.220	.826
Sentence Type (Under-Inf. vs. False)	1.305	0.018	73.894	< .001
Sentence Type (Under-Inf. vs. True (Some))	-0.275	0.020	-13.415	< .001
Sentence Type (True (Some) vs. True (All))	0.586	0.018	32.802	< .001

4.4.3.2 Auditory Digit Span, Mind in the Eyes, and Strange Stories Tasks

Results of all EF and ToM tasks are presented in Table 4.3.

Table 4.3: Scores on all EF and ToM tasks in Experiment 8.

Task	Mean	SD	Min	Max
Digit Span Task				
Memory Span	5.66	1.31	3	7
Mind in the Eyes Task				
# Correct	8.19	2.60	1	12
Strange Stories Task				
Mental State Justifications	10.06	3.10	0	14
Physical State Justifications	0.93	1.23	0	6
Control Justifications	5.34	1.28	0	6
Composite Scores				
EF	39.73	6.67	15	51
ToM	18.25	4.50	3	25

4.4.3.3 Simple Scalar Implicature Task

Performance on the Simple Scalar Implicature Task is shown in Figure 4.4. As expected, participants were more likely to give “Bad” ratings to Under-Informative (M

= 3.32, $SD = 1.96$) sentences than Under-Informative ($M = 0.39$, $SD = 0.88$) ones, $t(177) = 17.74$, $p < .001$.⁵

We next related the number of “Bad” ratings of Under-Informative sentences on this task (henceforth, the *Pragmatic Score*) to performance on Executive Function and Theory of Mind Tasks. As a first step, we investigated the extent to which an individual’s Pragmatic Score was correlated with their EF and ToM abilities. Pragmatic Score was significantly positively correlated with EF (Digit Span) scores, $\tau_b(176) = 0.171$, $p = .006$, and composite ToM scores, $\tau_b(176) = 0.207$, $p < .003$. These analyses align with prior work in demonstrating a role of each ability in scalar implicature, but do not inform our understanding of the relative contribution of EF and ToM.

To tease apart the *unique* roles of EF and ToM in explaining variation in judgments in the Simple Scalar Implicature task, a multiple linear regression was conducted with the EF and composite ToM scores as independent variables and Pragmatic Score as the dependent variable. EF and ToM scores were significantly correlated in the present data set, $\tau_b(176) = 0.226$, $p < .001$, a common finding in the

⁵ To investigate the extent to which scalar implicature judgments are consistent within a single individual, a correlation analysis was performed between scores on the Simple Scalar Implicature task (number of pragmatic responses out of 5) and mean Under-Informative sentence ratings in the Dual Scalar Implicature task (all trials in the low Cognitive Load control condition). A Kendall’s tau analysis was chosen to account for the positively skewed scores on the simple task, which represented the fact that people tended to respond to the pragmatically-enriched meaning of the utterances. When participants were given a 5-point scale for responses, scores were fairly evenly distributed across the entire range of possible values. Despite these task differences in overall data patterns, scores on the two tasks were significantly negatively correlated, $\tau_b(176) = -0.292$, $p < .001$, suggesting a general tendency within an individual to respond either pragmatically or logically. However, the correlation is only a moderate one, indicating that that individuals do vary somewhat in the way they respond to the pragmatic meaning of an utterance across tasks (for similar observations, see Tavano & Kaiser, 2010; Degen & Tanenhaus, 2015).

literature (e.g., Carlson et al., 2002; Bull et al., 2008) that further demonstrates the need to determine the unique impact of each factor on pragmatic ability, but the moderate correlation did not raise any issues of multicollinearity (all VIF values <1.5, all Tolerance values < 1) so the regression was performed as planned. The model accounted for a significant amount of variance, $F(2, 175) = 8.805, p < .001$. As can be seen in Table 4, EF was not significantly associated with Pragmatic Score, $p = .256$. ToM was significantly positively associated with Pragmatic Score, $p = .001$. In other words, participants who performed better on Theory of Mind tasks penalized Under-Informative sentences more on the Simple Scalar Implicature task, and this relationship held even when controlling for EF. In contrast, EF had no unique impact on the responses on the Simple Scalar Implicature task.⁶

For our last analysis, we subtracted the number of times Informative sentences were judged as “Bad” by an individual participant from the number of times Under-Informative sentences were judged as “Bad”. Calculating this *Pragmatic Difference Score* (PDS) allowed us to account for individual responding preferences (i.e., the general likelihood that a participant would judge a sentence as “Bad”), and thus gave a more sensitive picture of pragmatic sensitivity. The maximum PDS of 5 represents a completely pragmatic participant who would always reject Under-Informative statements and accept Informative statements. A score of 0 represents a completely logical participant who would always accept both Under-Informative and Informative statements. Finally, a highly unlikely negative score would be indicative of a

⁶ We repeated this analysis using ratings of Under-Informative sentences in the Dual Scalar Implicature Task (all trials) as the dependent variable and found the same results. The model was significant overall, $F(2, 175) = 7.065, p = .001$, and ToM ($\beta = -0.069, t = -3.640, p < .001$) but not EF ($\beta = 0.002, t = 0.152, p = .880$) was associated with sentence ratings.

participant who would judge Informative statements to be “Bad” more often than Under-Informative statements.

We repeated the previous analyses using PDS as the dependent variable instead of Pragmatic Score. PDS was significantly positively correlated with EF (Digit Span) scores, $\tau_b(176) = 0.209, p < .001$, and composite ToM scores, $\tau_b(176) = 0.263, p < .001$. To investigate the *unique* roles of EF and ToM in explaining variation in judgments in the Simple Scalar Implicature task, a multiple linear regression was conducted with the EF and composite ToM scores as independent variables and PDS as the dependent variable. The model accounted for a significant amount of variance, $F(2, 175) = 16.110, p < .001$. EF was not significantly associated with PDS, $\beta = 0.151, t = 1.205, p = .230$. ToM was significantly positively associated with PDS, $\beta = 0.173, t = 4.723, p = .001$. Thus, our finding that participants who performed better on Theory of Mind – but not EF – tasks behaved more pragmatically on the Simple Scalar Implicature task was demonstrated again even when controlling for response preferences by using PDS.⁷

⁷ To enrich our assessments of working memory, we repeated the two regressions using a composite EF score made up of scores on the Digit Span Task and the number of correct responses to Load trials on the memory portion of the Dual Scalar Implicature Task. The results did not change. The model predicting Pragmatic Bias accounted for a significant amount of variance, $F(2, 175) = 8.99, p < .001, R^2 = .09$, as did the model predicting Pragmatic Responding, $F(2, 175) = 17.04, p < .001, R^2 = .16$. In both analyses, ToM was significantly positively associated with the outcome variable (p 's $< .001$) while EF was not (p 's $> .05$).

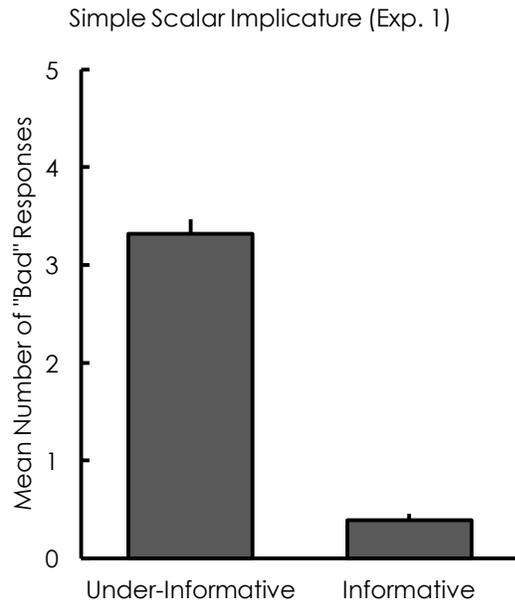


Figure 4.4: Results of the Simple Scalar Implicature Task in Experiment 8. Error bars represent +1 S.E.M.

Table 4.4: Multiple linear regression predicting Pragmatic Bias (the number of “Bad” ratings of Under-Informative sentences in the simple Scalar Implicature task) in Experiment 8 from composite ToM and EF scores.

Effect	β	S.E.	t	p
Intercept	4.903	0.921	5.323	< .001
EF	-0.028	0.022	-1.280	.202
ToM	-0.115	0.033	-3.500	< .001

4.4.4 Discussion

The main goal of Experiment 8 was to test the contributions of EF and ToM in explaining variability in scalar implicature judgments. As a first step towards this goal, in a Dual Scalar Implicature Task we attempted to replicate the previous finding that individuals make more literal judgments under greater cognitive load (De Neys &

Schaeken, 2007) due to the recruitment of EF in scalar implicature computation. We did not observe such a pattern of results in the present study, with judgments of Under-Informative sentences not differing significantly between high memory load and control conditions.

Our next step involved relating individual scores on EF and ToM tasks to judgments in a Simple Scalar Implicature task. This step went beyond much prior work on pragmatics, including the De Neys and Schaeken (2007) study, that typically investigated only one ability – EF, in this case – at a time. In line with the Dual task results, EF did not have a unique influence on pragmatic judgments. ToM, however, was significantly associated with scalar judgments: participants who performed better on the ToM tasks were less literal in their interpretation of Under-Informative utterances, more often making judgments consistent with the calculation of a (potential) scalar implicature. This association is predicted by a wide class of pragmatic accounts inspired by Grice (1975) which require a listener to reason about the intentions of the speaker in order to successfully compute the implicature.

Why might our results differ from the De Neys and Schaeken study? Marty and Chemla (2013) conducted an exact replication of the study and also found that participants were more logical under heavy cognitive load, making it unlikely that the original results were due to error. As our study was a conceptual replication, methodological differences (which included differences in the sentence stimuli, number of trials, and response scale) may have contributed to our finding that participants' pragmatic judgments did not differ between high and low cognitive load conditions, but we do not discuss these differences in detail here for two reasons. First, our data demonstrate the task-sensitivity of the effect observed by previous researchers

(De Neys & Schaeken, 2007; Marty & Chemla, 2013): individuals are not *always* more logical under greater cognitive load. Furthermore, and more importantly, we show that even when a relationship between EF and scalar implicature is found – such as in our correlation analyses – this relationship disappears when ToM is accounted for. Previous work, including the comparatively smaller-scale studies conducted by De Neys and Schaeken (2007) and Marty and Chemla (2013) did not measure ToM and therefore missed a critical contributor to pragmatic inference.

Given that we did observe significant positive bivariate correlations between EF and both scalar implicature tasks, but that these relationships were overridden by ToM, we conclude that the most likely role for EF in scalar implicature computation is most likely during engagement of ToM reasoning. In support of this conclusion, many researchers have argued that it is effortful to engage in ToM and that it thus recruits EF (Keysar et al., 2000; Apperly, Samson, Chiavarino, & Humphreys, 2004; Apperly, Back, Samson, & France, 2008; Lin, Keysar, & Epley, 2010). For example, adults are slower to select an appropriate referent when they need to incorporate another person's perspective to do so (Keysar et al., 2000; Epley, Morewedge, & Keysar, 2004). Given this evidence, previously observed associations between EF and scalar implicature may have been actually indicative of the link between EF and ToM. For instance, individuals under heavy cognitive load in previous work may have had fewer cognitive resources available to engage in ToM and reason about the intended meaning of the under-informative sentences, and thus responded more often to the literal meanings. An additional possibility – that is not mutually exclusive – is that EF is required in scalar implicature to hold the ToM reasoning in memory during implicature computation, and combine it with other information such as knowledge of

the speaker's abilities and preferences (Antoniou et al., 2016). Thus, when listeners' EF was taxed in prior work, they may have had difficulties not only in engaging in ToM, but also in comparing the result of this reasoning – that the speaker is fully knowledgeable and would have used the stronger quantifier *all* if it had been true – with the activation of that stronger alternative utterance (e.g., “*All* giraffes have long necks”). We return to this possibility in the General Discussion below.

4.5 Experiment 9

In Experiment 8, we found that ToM – but not EF – was important in making pragmatic judgments. To test whether this association is specific to scalar implicature or generalizes to other pragmatic domains, in Experiment 9 we investigated the contributions of EF and ToM to metaphor and indirect request, as well as scalar implicature (to replicate the findings of Experiment 8).

Metaphors (“I’m a night owl”) require the listener to infer that the speaker did not intend to convey the literal meaning of the utterance (as it violates the Maxim of Quality) but rather some kind of similarity (staying up late; Grice, 1975). As common metaphors become familiar/conventionalized, the pragmatic meaning may become automatically available (Glucksberg, 2003). Indirect requests (“It’s hot in here”) require the listener to infer that the speaker did not intend to convey simply the literal meaning of the utterance (since this would violate the Maxim of Quantity – this information is obvious and known to both speaker and listener; Grice, 1975); the purpose of the utterance is to indirectly ask the hearer to perform an action that he/she is willing and able to do (e.g., open the window). There is evidence that both EF and ToM are recruited in both metaphors and indirect requests. Behavioral work measuring individual differences in working memory has demonstrated an association

between metaphor comprehension and EF (Chiappe & Chiappe, 2007; Mashal, 2013), and ERP research suggests that processing indirect requests increases higher memory retrieval demands as compared to processing non-request utterances (Coulson & Lovett, 2010). Furthermore, metaphor comprehension has been associated with ToM skills (Happé, 1993; but see Norbury, 2005), and engagement of ToM regions of the brain has been reported for both metaphor (Prat et al., 2012) and indirect request comprehension (Van Ackeren et al., 2012). However, the specific contribution of EF and ToM to pragmatic variation across individuals has not been tested yet.

To assess metaphor comprehension, we borrowed a task from Jankowiak, Rataj, and Naskręcki (2017). That study investigated the neural correlates of processing novel metaphors (“to harvest courage”), conventional metaphors (“to gather courage”), literal phrases (“to experience courage”) and anomalous phrases (“to move courage”) in Polish-English bilinguals’ first and second languages. (For ease of presentation, we focus here on the first-language results.) The authors reasoned that novel metaphors would require more cognitive resources to be processed than conventional metaphors because understanding a novel metaphor requires a comparison between the literal and the intended meaning (Bowdle & Gentner, 2005), whereas understanding more familiar, conventional metaphors does not (Glucksberg, 2003). In line with this prediction, novel metaphors elicited greater N400 components than conventional metaphors, literal phrases, and anomalous phrases, indicative of the increased semantic processing necessary to comprehend them. Interestingly, novel metaphors were also judged as less meaningful than conventional metaphors, and those metaphors as less meaningful than literal phrases (anomalous phrases were the least meaningful items). Here we adopted the behavioral paradigm of Jankowiak et al.

(2017) and asked participants to judge the meaningfulness of metaphorical and literal phrases, expecting to replicate the finding that novel metaphors are judged as less meaningful than literal phrases overall. Of interest was whether the gap between literal and metaphorical phrases would be narrower for participants with higher ToM (indicating successful comprehension of the pragmatic meaning of metaphors).

To test indirect request comprehension, we used a paradigm originally developed by Van Ackeren et al. (2017) to investigate the activation of ToM regions of the brain during the processing of indirect requests. In this task, participants were presented with picture-sentence pairs that either suggested that the speaker was making a request (e.g., a picture of a closed window paired with the sentence “It is very hot here”) or not (e.g., a picture of a desert paired with the sentence “It is very nice here”). Compared to control trials, request trials elicited greater activation of cortical regions associated with ToM. Furthermore, behavioral judgments indicated that participants were more likely to feel that the speaker was making a request in the critical indirect request condition as compared to control trials. Here we administered a behavioral adaptation of the task using the same stimuli, and asked participants to rate how strongly they felt that the speaker was making a request. We expected such ratings to be higher for request trials as compared to control trials. Of interest was whether better ToM (after controlling for EF) would enhance this pattern.

4.5.1 Participants

Two hundred monolingual English speakers aged 19-68 ($M = 29.38$, $SD = 5.66$) living in the United States, 91 of whom were female, were recruited from Amazon’s Mechanical Turk to participate. Participants were compensated \$1.50 for the 15-minute study. Data from 26 individuals who reported to be bilingual and/or

diagnosed with Autism Spectrum Disorder were excluded, leaving 174 participants for analysis.

4.5.2 Materials and Procedure

Participants completed five tasks in the order presented below: three pragmatic tasks (Metaphor, Indirect Request, Scalar Implicature), an Auditory Digit Span Task (as a measure of EF), and the Mind in the Eyes and Strange Stories Tasks (as ToM measures). Stimuli for the Metaphor, Indirect Request, and Scalar Implicature tasks can be found in Appendices E, F, and C, respectively.

4.5.2.1 Metaphor Task

Participants were presented with twenty verb phrases individually, and were asked to rate on a 5-point scale how meaningful each one was. Ten phrases were novel metaphors (“to harvest courage”) and 10 literal phrases (“to feel anger”). The materials were borrowed from a set of stimuli used in Jankowiak et al. (2017) that had been extensively normed on word length, concreteness, metaphoricity, etc. The task was chosen in part due to the careful creation of the novel stimuli, as well as its prior use with adults. Higher meaningfulness ratings for novel metaphors should indicate better metaphor comprehension; the literal phrases served as a control condition.

4.5.2.2 Indirect Request Task

This task was borrowed from Van Ackeren et al. (2012). Thirty-six picture-sentence pairs were presented to participants, evenly divided over four conditions (see Table 4.5). On the critical Indirect Request trials, the sentences accompanying the pictures (e.g., the sentence “It is very hot here” paired with a picture of a closed window) could be used to imply that the speaker wished the listener to do something

(e.g., open the window). The Picture Control, Utterance Control, and Picture-Utterance Control trials were designed in such a way as to preclude a request interpretation: Picture Control trials featured the same picture as the Indirect Request condition but paired it with a sentence that was not a request (e.g., “It is very nice here” paired with the same picture of a closed window); Utterance Control trials consisted of the same sentence as the Indirect Request condition paired with a different picture that did not lead to a request (e.g., “It is very hot here” paired with a picture of a desert); and Picture-Utterance Control trials consisted of the Picture Control picture paired with the Utterance Control sentence, which combined did not suggest a request (e.g., “It is very nice here” paired with a picture of a desert). On each trial, the sentence appeared beneath the picture and participants were asked “How much do you feel that the speaker wants something from you?” They were given a 5-point scale to input their responses, and were instructed that a higher rating meant that they felt strongly that the speaker wanted something from them. Trials were presented in a random order for each participant, and stimuli were fully rotated such that only one version of a given item was presented to each participant. Higher ratings on Indirect Request trials indicate pragmatic (request) interpretations. We expected the three Control trials to elicit low (literal, non-request) ratings.

Table 4.5: Examples of stimuli used in the indirect request comprehension task in Experiment 2 (borrowed from Van Ackeren et al., 2012).

Picture	Sentence	Condition
	It is very hot here.	Indirect Request
	It is very nice here.	Picture Control
	It is very hot here.	Utterance Control
	It is very nice here.	Picture-Utterance Control

4.5.2.3 Simple Scalar Implicature Task

Participants completed the same simple, binary task administered in Experiment 8.

4.5.2.4 Auditory Digit Span Task

This EF measure was identical to the task administered in Experiment 8.

4.5.2.5 Mind in the Eyes and Strange Stories Tasks

These ToM measures were identical to those administered in Experiment 1. Coding for the Strange Stories Task was verified by a second rater ($IRR = .89$).

4.5.3 Results

For each pragmatic task, we analyze results separately and relate them to EF and ToM measures. Overall results from EF/ToM tasks are given in Table 4.6. The three pragmatic tasks are compared in the final subsection.

Table 4.6: Scores on all EF and ToM tasks in Experiment 9.

Task	Mean	SD	Min	Max
Digit Span Task Memory Span	5.55	1.53	3	7
Mind in the Eyes Task # Correct	8.77	2.47	0	12
Strange Stories Task Mental State Justifications	10.57	2.80	0	14
Physical State Justifications	1.17	1.54	0	6
Control Justifications	5.45	1.17	0	6
Composite Scores EF	5.55	1.53	3	7
ToM	19.34	4.47	3	26

4.5.3.1 Metaphor Task

Performance on the Metaphor task is shown in Figure 4.5. Participants judged literal control phrases ($M = 3.48$, $SD = 0.69$) as being more meaningful than the critical metaphorical statements ($M = 2.92$, $SD = 0.50$), $t(173) = 8.846$, $p < .001$. As before, we first investigated the extent to which an individual's responses were

correlated with their EF and ToM abilities. Meaningfulness ratings on critical trials of the Metaphor Task were not significantly correlated with EF scores, $\tau_b(173) = -0.019$, $p = .743$, but were significantly negatively correlated with ToM scores, $\tau_b(173) = -0.120$, $p = .026$.

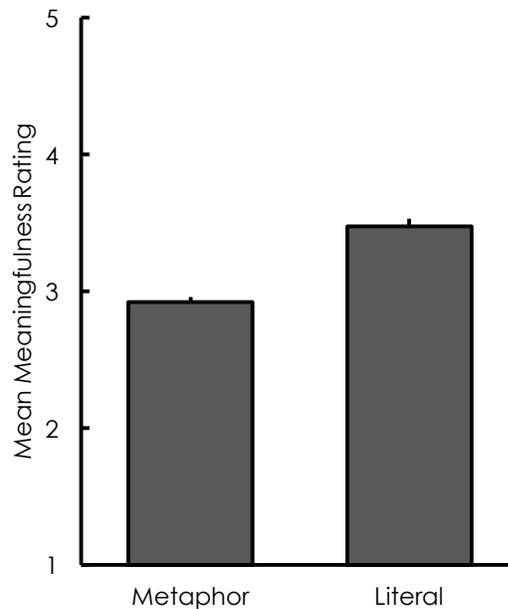


Figure 4.5: Results of the Metaphor Task in Experiment 9. Error bars represent +1 S.E.M.

We then conducted a regression analysis predicting pragmatic ability in the Metaphor task from EF and ToM, to determine the unique contribution of each. As in Experiment 8, to make results easier to compare across tasks and to account for general response preferences, the regression analysis was based on a difference score that sought to measure the level of pragmatic responding tendencies given responses to non-pragmatic trials. Specifically, for each participant, we used the mean metaphor phrase ratings minus the mean control (literal) phrase ratings to construct a Pragmatic

Difference Score (PDS): more negative values represent more logical responding (i.e., treating literal phrases as more meaningful than metaphorical ones) whereas values closer to zero indicate that the metaphors were judged more closely to literal statements in terms of meaningfulness, presumably because their enriched, non-literal meanings were available to the comprehender. A PDS score of 0 represented an individual who took metaphorical statements to make as much sense as literal ones.

A multiple linear regression with EF and composite ToM scores as predictor variables and PDS as the outcome variable predicted a significant amount of variance, $F(2, 171) = 3.152, p = .045, R^2 = 0.04$ (Table 4.7). EF was not significantly associated with PDS on the Metaphor Task, $p = .716$. Surprisingly, better scores on ToM tasks were negatively associated with PDS, such that individuals who performed better on the ToM tasks tended to judge metaphorical phrases as less meaningful than literal phrases, $p = .015$.

Table 4.7: Multiple linear regression analyses predicting scores on metaphor, indirect request and scalar implicature tasks in Experiment 9 from EF and ToM scores.

Effect	Metaphor		Indirect Request		Scalar Implicature	
	β (S.E.)	p	β (S.E.)	p	β (S.E.)	p
Intercept	0.062 (0.310)	.841	-0.706 (0.263)	.008	-0.794 (0.745)	.288
EF	0.016 (0.044)	.716	0.022 (0.037)	.545	0.057 (0.105)	.585
ToM	-0.037 (0.015)	.015	0.068 (0.013)	<.001	0.176 (0.036)	<.001

4.5.3.2 Indirect Request Task

Performance on the Indirect Request Task is shown in Figure 4.6. A linear mixed-effects regression performed on responses in the Indirect Request Task with

Trial Type (Indirect Request, Picture Control, Utterance Control, Picture-Utterance Control) included in the model as fixed effects and crossed random intercepts for Participants and Items indicated that the extent to which participant felt that a request was being made differed significantly across Trial Types, $\chi^2(3) = 254.061, p < .001$. Post-hoc tests indicated that Indirect Request ($M = 3.04, SD = 0.69$) trials were rated higher than Picture Control ($M = 2.18, SD = 0.83$), Utterance Control ($M = 2.50, SD = 0.76$), and Picture-Utterance Control ($M = 2.28, SD = 0.77$) trials, all p 's $< .001$. Picture Control trials were rated lower than Utterance Control and Picture-Utterance Control trials, both p 's $< .001$. Picture-Utterance Control and Utterance Control trials did not differ significantly from one another, $p = .060$. Request ratings on the critical trials of the Indirect Request Task were significantly positively correlated with EF scores, $\tau_b(173) = 0.123, p = .031$, and ToM scores, $\tau_b(173) = 0.258, p < .001$.

As with metaphor, we created a Pragmatic Difference Score (PDS) that corresponded to the mean request rating in the critical request trials minus the mean request rating for the three types of control trials combined for each participant. The maximum PDS of 4 represented perfect identification of indirect requests and non-requests, whereas smaller values indicated an inability to detect indirect requests. A multiple linear regression predicting PDS from EF and composite ToM scores was significant, $F(2, 171) = 17.910, p < .001, R^2 = 0.17$ (Table 4.7). The association between EF and PDS was not significant, $p = .545$. ToM was positively associated with PDS, $p < .001$, such that participants who performed better on the ToM tasks were also more accurate at identifying an indirect request.

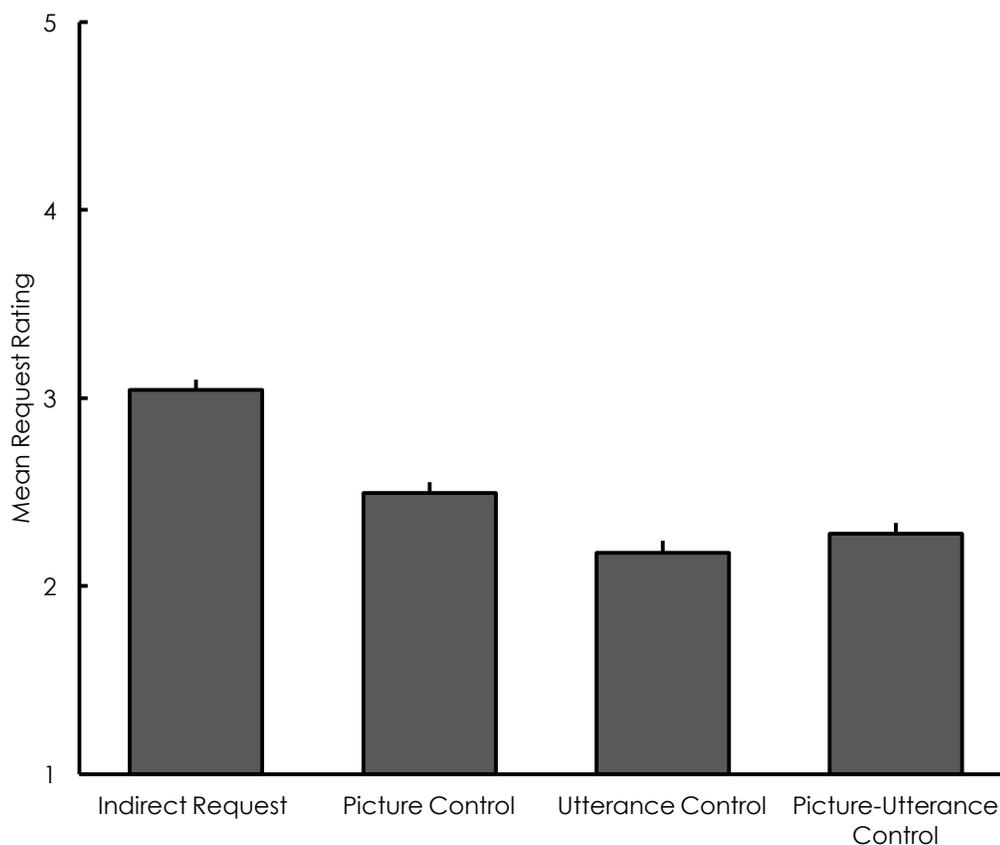


Figure 4.6: Results of the Indirect Request Task in Experiment 9. Error bars represent +1 S.E.M.

4.5.3.3 Simple Scalar Implicature Task

Performance on the Scalar Implicature Task is shown in Figure 4.7. As expected, participants were more likely to give “Bad” ratings to Under-Informative ($M = 3.29, SD = 1.81$) sentences than Under-Informative ($M = 0.36, SD = 0.89$) ones $t(173) = 18.149, p < .001$. The number of “Bad” ratings to Under-Informative sentences was significantly positively correlated with EF scores, $\tau_b(173) = 0.140, p = .023$, and ToM scores, $\tau_b(173) = 0.259, p < .001$, in line with our findings in Experiment 8.

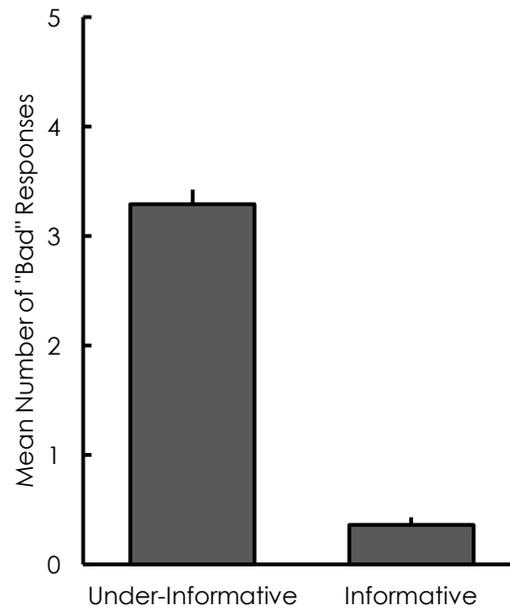


Figure 4.7: Results of the Simple Scalar Implicature Task in Experiment 9. Error bars represent +1 S.E.M.

We computed a PDS for this task corresponding to the number of “Bad” ratings for Under-Informative sentences minus the number of such ratings for Informative sentences for each participant. As in Experiment 8, the maximum PDS of 5 represented a perfectly pragmatic responder who always highly penalized Under-Informative sentences but never Informative sentences. A multiple linear regression was then performed with EF and composite ToM scores as predictor variables and PDS as the outcome variable to determine the unique influence of each predictor on scalar implicature judgments. Overall, the analysis predicted a significant amount of variance on the Scalar Implicature Task, $F(2, 171) = 15.010, p < .001, R^2 = 0.15$ (Table 4.7). The association between EF and PDS was not significant, $p = .585$. ToM was positively associated with PDS, $p < .001$.

4.5.3.4 Correlations Among Tasks

To directly test the relationships among metaphor, indirect request, and scalar implicature understanding, a series of correlation analyses was performed with the pragmatic difference scores used in the preceding regression analyses (see Figure 4.8). Metaphor PDSs were not significantly correlated with either Indirect Request, $\tau_b(171) = -0.070, p = .174$, or Scalar Implicature PDSs, $\tau_b(171) = -0.043, p = .444$. PDSs on the Scalar Implicature Task were significantly positively correlated with Indirect Request PDSs, $\tau_b(171) = 0.175, p = .002$.

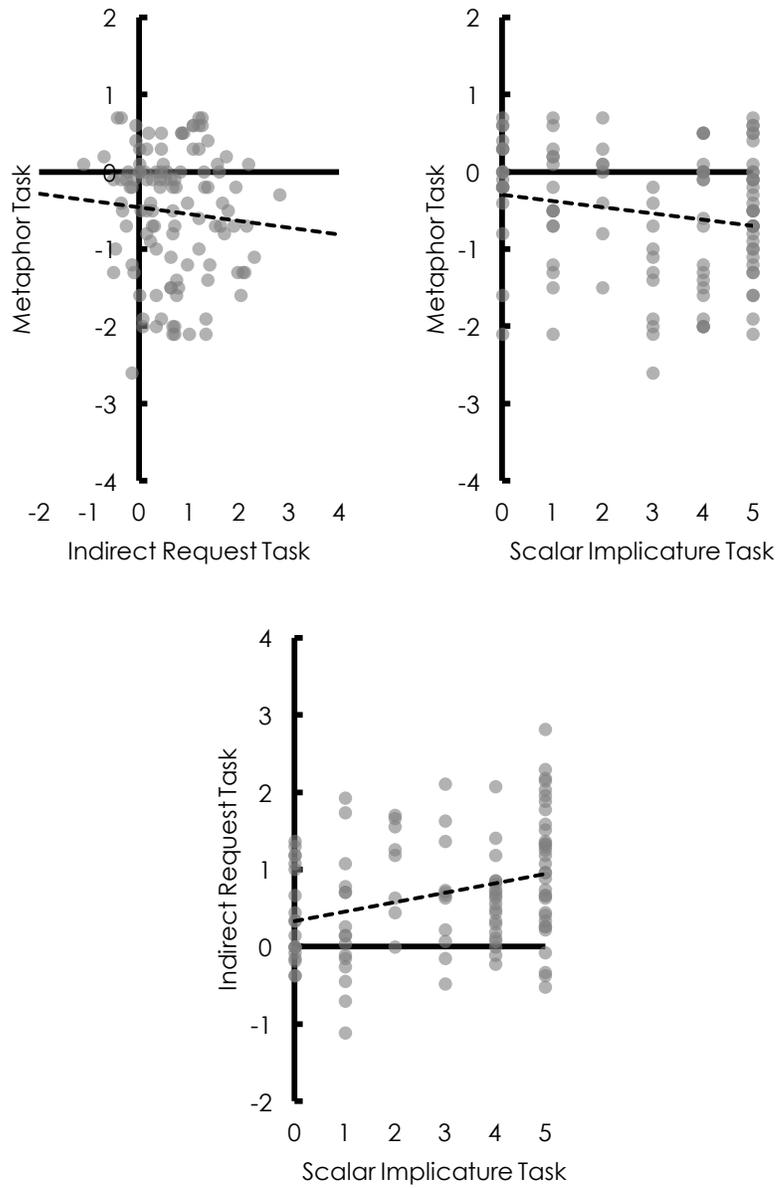


Figure 4.8: Correlations among pragmatic tasks in Experiment 9 (Top Left: Indirect Request and Metaphor, Top Right: Scalar Impicature and Metaphor, Bottom: Scalar Impicature and Indirect Request).

4.5.4 Discussion

In Experiment 9, our primary aim was to investigate the relationships between performance on a broad array of pragmatic tasks (scalar implicature, metaphor, and indirect request judgments) and measurements of EF and ToM within individuals. ToM – but not EF – was significantly positively associated with scalar implicature and indirect request judgments (as captured by PDSs), replicating and extending Experiment 8. EF was not predictive of participants' judgments of metaphorical statements, but ToM was negatively associated with pragmatic responding on the metaphor task. In line with these findings, scalar implicature and indirect request judgments were correlated with one another but not with responses on the metaphor task.

Our findings from the metaphor task appear particularly counterintuitive given the link observed here and elsewhere between better ToM skills and enhanced pragmatic abilities (Happé, 1993; Chiappe & Chiappe, 2007; Mashal, 2013; even though see Norbury, 2005). Why should participants with better ToM rate metaphorical phrases as being less meaningful than the literal phrases, as compared to participants with worse ToM? Notice that, in both Jankowiak et al. (2017) and our own data, metaphorical phrases overall were judged as less meaningful than literal ones. This is unsurprising since such phrases had both literal (false) and metaphorical (true) meanings, unlike literal phrases that only had one true meaning (cf. also under-informative statements that are typically rated higher than completely false statements but lower than completely true statements, as listeners take into account both the false semantic and true pragmatic meanings – see Experiment 8, Dual Scalar Implicature Task; cf. Fairchild & Papafragou, 2017; Katsos & Bishop, 2011). What requires an

explanation is the fact that in the present study the literal-metaphorical ratings gap became *wider* with increased ToM sophistication.

The hypothesis that follows most closely from the original study is that the meaningfulness judgments in the metaphor task reflected perceived effort of processing. Jankowiak et al. (2017) demonstrated that the novel metaphors required a greater amount of cognitive resources than the literal phrases. Participants with better ToM who may have invested a greater amount of resources into generating potential metaphorical meanings may have given lower ratings to phrases that required this extra effort to become meaningful, as compared to literal phrases that are already meaningful at the semantic level. Consequently, if meaningfulness judgments were based purely on the perceived level of effort, it is reasonable that the low-ToM participants, who have put less effort into processing potential additional meanings of the metaphors, judged the metaphorical phrases to be as meaningful as the literal phrases.

Alternatively, in line with our predictions, participants with better ToM may have been more likely to infer metaphorical meanings but gave lower meaningfulness ratings of the metaphorical phrases because they judged the metaphors to be unsuccessful. For every metaphor in this task (e.g., “to harvest courage”; see Appendix C), there is a related conventional metaphor (e.g., “to gather courage”; included as a control in Jankowiak et al., 2017). It is possible that high ToM- individuals who processed the metaphorical meanings reasoned that these novel phrases (understood to be less familiar than both the conventional metaphors and the literal phrases; Jankowiak et al., 2017) were odd ways of conveying the speaker’s intended message, especially given the availability of the corresponding conventional

metaphors in the language. According to this possibility, high-ToM comprehenders would rely on the assumption that speakers should strive to offer messages that are clear and concise (see the Manner maxim in Grice, 1975) and would penalize these novel metaphors for being more convoluted than necessary (see also Sperber & Wilson, 1986). Notice that all phrases in the metaphor task appeared in the infinitival form (“to do X”), a fact that may have made participants more likely to focus on them as linguistic tokens and judge them in terms of the Manner maxim (this is in contrast to our Indirect Request Task, for example, where participants were presented with whole utterances and explicitly asked to reason about the speaker’s desires).

While we leave open these possibilities concerning the metaphor task, the results of Experiment 9 overall highlight the role of ToM in pragmatic processing. The fact that we observed very similar results for scalar implicature and indirect request comprehension suggests that scalar implicature is indeed a suitable tool for evaluating pragmatic competence and shares underlying (ToM) mechanisms with other pragmatic phenomena.

4.6 General Discussion

Adults are often used as the benchmark for pragmatic computation in both developmental studies (Noveck, 2001; Papafragou & Musolino, 2003; Katsos & Bishop, 2011) and comparisons with atypical populations (Noveck et al., 2007; Pijnacker et al., 2009, Chevallier et al., 2010; Hochstein et al., 2017), yet there is tremendous variation in (neurotypical) adults’ pragmatic judgments. Here we explored the roots of this variation focusing on the well-studied case of scalar implicature. Many psycholinguistic investigations of this phenomenon have taken into account individual differences when interpreting adults’ pragmatic performance (Noveck &

Posada, 2003; Bott & Noveck, 2004; Politzer-Ahles et al., 2013; Heyman & Schaeken, 2015; Degen & Tanenhaus, 2015; Fairchild & Papafragou, 2017); nevertheless, the underlying cause for such individual differences is not clear. Furthermore, it is unknown whether an individual's ability to compute scalar implicatures is linked to other pragmatic abilities. Our goal in the present paper was to address these gaps in the literature.

Drawing on the theoretical assumption that deriving a scalar implicature can be cognitively costly (e.g., Grice, 1975; Sperber & Wilson, 1986; Sauerland, 2004; Geurts, 2010), and prior empirical findings (De Neys & Schaeken, 2007; Antoniou et al., 2016), we tested the hypothesis that EF – specifically working memory - might be associated with scalar implicature calculations. Similarly, drawing on theoretical assumptions about the process of scalar implicature computation (Grice, 1975; Sperber & Wilson, 1986; Sauerland, 2004) and previous experimentation (Noveck et al., 2007, among others), we tested the hypothesis that ToM might be associated with scalar pragmatic judgments. In Experiment 8, we presented the first investigation to incorporate measures of both EF and ToM in an attempt to tease apart the unique impact of the two factors on scalar implicature calculations within an individual. In Experiment 9, we extended our investigation to metaphor and indirect request in an effort to understand whether the mechanisms underlying individual differences in scalar implicature computation extend to other types of pragmatic phenomena.

Across the two experiments, we found that the better a participant performed on tasks measuring ToM, the more likely they were to respond to the pragmatic meaning (as opposed to the literal meaning) of an under-informative utterance such as “Some dogs are mammals.” Importantly, this association represented the unique

influence of ToM on scalar implicature judgments, controlling for EF. In addition to the strong and replicable association between ToM and scalar implicature judgments, we found that participants with stronger ToM were better able to identify when a speaker was making an indirect request. Interestingly, we saw a moderate association in the opposite direction for metaphor, such that participants with better ToM displayed higher meaningfulness ratings for *literal* compared to metaphorical sentences. Further investigation showed that performance on the scalar implicature task was correlated with the indirect request task but that performance on the metaphor task was not correlated with either of the two other tasks. In terms of measures of EF (working memory), we failed to find a unique relationship between EF and pragmatic judgments of any type (including scalar implicature, metaphor, and indirect request), suggesting that – at least for adults – the strength of one’s cognitive resources is not uniquely important for making pragmatic judgments.

4.6.1 Executive Function and Pragmatic Judgments

In both of our experiments there were significant positive correlations between EF and performance on scalar implicature and indirect request tasks, but these relationships disappeared after factoring in ToM. We are not suggesting here that EF is not required in any way for pragmatic judgments (or in other words, that scalar implicatures are computed by default as has been argued by Levinson, 2000). There is a wealth of prior evidence demonstrating that processing scalar implicatures is more cognitively costly than processing the fully informative equivalent statements (e.g., Breheny et al., 2006; Huang & Snedeker, 2009). Instead, we conclude that EF is recruited to the extent that it is required for engaging ToM reasoning.

There are two specific, not mutually exclusive possibilities about the role of EF in ToM reasoning during pragmatic interpretation. One possibility is that EF is recruited during ToM computations because these computations are inherently effortful (Keysar et al., 2000; Apperly, et al., 2004; Apperly, et al., 2008; Lin, Keysar, & Epley, 2010). Some evidence supporting this possibility comes from referential communication paradigms in which adults are slower to select an appropriate referent when they need to incorporate another person's perspective (Keysar et al., 2000; Epley, Morewedge, & Keysar, 2004). Additionally, dual-task studies demonstrate that adults have difficulty selecting the appropriate visual perspective in a perspective-taking task (but did not affect simple level-1 visual perspective calculations; Qureshi, Apperly, & Samson, 2010⁸) and completing the very same Mind in the Eyes and Strange Stories tasks we administered here (Bull, et al., 2008) when executive function demands are high. According to this possibility, previously observed associations between EF and scalar implicature (De Neys & Schaeken, 2007; Antoniou et al., 2016) may have been actually indicative of the link between EF and ToM. For instance, individuals under heavy cognitive load in previous work may have had fewer cognitive resources available to engage in ToM and reason about the intended meaning of the under-informative sentences, and thus responded more often to the literal meanings.

Alternatively, EF might be required to hold the products of ToM reasoning in memory and combine them with other information such as linguistic context during

⁸ This finding, along with evidence that even infants have some basic ToM abilities (e.g., Onishi & Baillargeon, 2005; Soutgate, Senju, & Csibra, 2007), suggests that EF is needed to complete the types of complex instances of mental-state reasoning that adults frequently engage in (and would have to engage in for pragmatic inferences), but may not be a necessity for simpler ToM processes.

the computation of pragmatic meaning (cf. Antoniou et al., 2016). In the specific case of scalar implicature in the present chapter and in previous studies, participants needed to combine their reasoning about what a cooperative, knowledgeable interlocutor would say with their knowledge of the world (e.g., that dogs are mammals, and that there are no exceptions). Thus, participants who failed to derive a scalar implicature may have done so because they did not have the EF resources to engage in ToM and/or connect that ToM reasoning with their own world knowledge.

Given our findings, the specific subcomponent(s) of EF involved in pragmatic inference should be the topic of further investigation. There is some disagreement as to whether working memory is synonymous with EF or is simply one subcomponent, along with inhibition and task-switching (Miyake, 2000). Some of this disagreement may be attributed to the wide variety of tasks used to measure EF, which often suffer from issues of reliability and validity (Duckworth & Kern, 2011). We specifically chose the backwards digit span task as our measure of EF because it has good test-retest (Wechsler, 1981) and split-half (Waters & Caplan, 2003) reliability and is thought to measure working memory more accurately than other tasks as it requires participants to actively work with stored information (Diamond, 2013). Future research could include multiple tasks of each EF subcomponent and create composite scores from them to counteract methodological concerns about individual tasks.

4.6.2 Theory of Mind and Pragmatic Judgments

A key aspect of our findings is that scalar implicature and indirect request calculations in neurotypical adults (drawn from the general population of participants in online studies) are related to these adults' ToM skills, even when controlling for EF. This finding is in line with previous research reporting impaired pragmatic reasoning

in patients with cortical lesions to ToM areas (Champagne-Lavau & Joannette, 2009; Spoto et al., 2012) and increased neural activation of ToM areas in healthy adults when processing indirect requests (Van Ackeren et al., 2012). Taken together, this evidence supports the claim that understanding the pragmatic meaning of an utterance requires actively thinking about what the speaker intended to say – a claim with a long history within the linguistic and philosophical literature on pragmatics (Grice, 1975; Sperber & Wilson, 1986; Sauerland, 2004; Geurts, 2010). (On the somewhat different pattern of results on metaphor, see Section 4.5.4.)

Our findings confirm that there is substantial variation across individuals in reasoning about others' minds. It is important to consider, however, that, much like EF, ToM is a broad term for a collection of abilities relating to thinking about other people, including perspective-taking, false belief understanding, emotion recognition, and so forth. The tasks we administered to measure ToM both required a high degree of linguistic ability – vocabulary knowledge for the Mind in the Eyes Task and the ability to synthesize multiple pieces of information embedded in a story and generate a response based on this information for the Strange Stories Task. We chose these tasks because they are more complex and particularly well-suited for use with adults (Happé, 1994; Jolliffe & Baron-Cohen, 1999; Baron-Cohen et al., 2001). It would be interesting to extend our approach to include established ToM tasks with reduced linguistic demands (see, e.g., Samson, Apperly, Braithwaite, & Andrews, 2010).

Our results leave it open whether ToM is recruited online during the early stages of pragmatic inference, or at a later stage of reasoning before providing a judgment. Online measures of scalar implicature computation (e.g., Breheny et al., 2006; Huang & Snedeker, 2009; Niewland et al., 2010), EF (e.g., Eriksen & Eriksen,

1974), and ToM (e.g., Dumontheil, Apperly, & Blakemore, 2010) can be integrated into future work to help tease these two possibilities apart (cf. Breheny et al., 2013, for evidence that speaker knowledge is consulted early during the processing of scalar terms).

4.6.3 Pragmatic Judgments Within and Across Tasks

The central motivation for the present paper was the observation that adults differ in how consistently they make pragmatic inferences. This variation has been dealt with in the past by splitting participants into Logical and Pragmatic Responders (e.g., Bott & Noveck, 2004). Here, we chose to treat pragmatic responding as a continuum rather than a dichotomy given inconsistencies within individuals observed in previous research (Degen & Tanenhaus, 2015; Fairchild & Papafragou, 2017), as well as in the present data, where judgments spanned the entire scale of possible responses. Regardless of how variation in pragmatic inference is treated, it is important to consider the issue of stability of responding – whether “pragmatic” responders generally behave pragmatically, for example, across contexts.

Our findings point to differences in adults’ pragmatic response patterns based on (presumably stable) cognitive characteristics. Additionally, there was a significant degree of response consistency across scalar implicature and indirect request judgments (Experiment 9). Taken together, these findings suggest that stable, participant-internal characteristics are important influences on pragmatic calculations, with the tendency to be a “pragmatic” responder being relatively stable (and related specifically to the individual’s ToM).

Nevertheless, as the metaphor data show, this picture is far from simple. Other data in the present study suggest that participant-external factors – especially, the

nature of the task - are likely to shift participant-internal biases. First, judgments on the two scalar implicature tasks were only moderately correlated in Experiment 8 (see footnote 5). Second, judgments on scalar implicature and indirect request tasks were similarly moderately correlated in Experiment 9. These results are reminiscent of studies highlighting the influence of context and task demands on scalar implicature computation (Breheny et al., 2006; Bonnefon et al., 2009; Bergen & Grodner, 2012; Huang & Snedeker, 2018) and point to the need to integrate participant-driven and task-driven factors in understanding variation in adults' pragmatic responding within and across tasks (cf. Degen & Tanenhaus, 2015).

4.6.4 Future Directions

Large-scale work on individual differences in (neurotypical) adults' pragmatic communication is only beginning. One important next step is to understand when ToM is recruited during pragmatic inference, either online during comprehension or at a later stage of reasoning before providing a judgment. Our results do not differentiate between these two explanations because our tasks were all untimed. As such, our results may reflect an association between slower, reflective mental state reasoning and, final interpretations of linguistic material after all online processing is complete. Online measures of scalar implicature computation (e.g., Breheny et al., 2006; Huang & Snedeker, 2009; Niewland et al., 2010), EF (e.g., Eriksen & Eriksen, 1974), and ToM (e.g., Dumontheil et al., 2010) are well-established and can be integrated into future work to help tease these two possibilities apart. Converging evidence from neuroimaging data focusing on ToM regions such as the rTPJ would also help to shed light on ToM recruitment during pragmatic inference. Finally, we hope that the present study encourages future researchers to incorporate multiple pragmatic

phenomena – beyond what we have done here – into research on pragmatic competence. Developmental research has shown that some pragmatic tasks are more difficult for children than others, for example, irony requires 2nd order false-belief reasoning abilities while metaphor requires only 1st order (Happè, 1993). Additionally, research in referential communication underscores the importance of fine-grained, theoretically-motivated investigations into relationships between specific cognitive abilities and pragmatic phenomena. Classic pragmatic theories of referential communication assume that the ability to interpret referring expressions such as “Pick up the tall glass” (in a situation where the speaker may or may not be able to see the tallest glass) depends on the ability to take the speaker’s perspective (Clark & Marshall, 1978). In support of this connection, the EF component responsible for inhibitory control has been associated with success with understanding referential expressions when the speaker’s and the addressee’s viewpoints conflict (see Nilsen & Graham, 2009 for developmental data; and Brown-Schmidt, 2009 for adult evidence; but see Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015). Additional comprehensive, theoretically-motivated research – similar to what we have presented here – on irony, humor, sarcasm, and other pragmatic phenomena is needed to specify the precise mechanisms underlying adults’ processing of pragmatic meanings and thereby gain a more complete understanding of language comprehension.

Chapter 5

SUMMARY & CONCLUSIONS

The uniquely human ability of language is perhaps most useful to us when it is used as a tool to facilitate interpersonal communication: for example, when it is used for transmission of knowledge and cultural norms, planning, or storytelling.

Understanding language in such social contexts requires comprehending the semantic meaning of utterances, but importantly also requires listeners to consider the speaker's knowledge, intentions, and preferences to understand meaning at the pragmatic level.

Pragmatic inferences vary widely across contexts and individuals, leading to many potential instances of miscommunication. Accordingly, the central goal of this dissertation was to systematically address variation in pragmatic inference from two angles; variation due to speaker characteristics and variation due to listener characteristics. Previous research suggests that both sides are important: pragmatic inferences are more likely to be made for more knowledgeable speakers (Bergen & Grodner, 2012), and neurotypical adult listeners are more likely to make pragmatic inferences than children (Papafragou & Musolino, 2003) or individuals with Autism (Noveck et al., 2007). However, important open questions remain as to what specific speaker and listener characteristics affect pragmatic inference, and whether they affect all types of pragmatic inference equally.

We addressed these open questions in three independent studies investigating how the language background of the speaker (Chapters 2 and 3) and the cognitive abilities of the listener (Chapter 4) affect scalar implicature computation – a specific

type of pragmatic inference that is widely studied in the literature. Furthermore, we investigated the extent to which scalar implicature computation relates to other types of pragmatic inference – namely, metaphor and indirect request comprehension.

In doing so, we made several important contributions. First, variation in performance is acknowledged in the majority of investigations of pragmatic inference (Noveck, 2001; Noveck & Posada, 2003; Bott & Noveck, 2004; Guasti et al., 2005; Politzer-Ahles, Fiorentino, Jiang, & Zhou, 2011; Heyman & Schaeken, 2015; Degen & Tanenhaus, 2015) but its cause is unclear. Our findings shed light on this variation, and thus the theories concerning the mechanisms underlying pragmatic processing. Second, adults are very often used as the benchmark for studies of pragmatics involving children and clinical populations. We demonstrated the importance of considering even neurotypical adults' pragmatic abilities as being on a spectrum. Finally, by manipulating non-native speaker status our findings contribute to our understanding of the social biases faced by non-native speakers as well as our understanding of theories of non-native speech processing.

The two sections that follow summarize the findings of Chapters 2, 3, and 4 as they relate to speaker and listener effects on the processing of pragmatic meaning. Finally, we consider the broader implications of these findings for theories of native and non-native speech processing, real-world interactions between native and non-native speakers, and future work in pragmatics.

5.1 Speaker Effects on Pragmatic Processing

Previous research has found that a speaker's situational knowledge of the context at hand affects the listener's pragmatic inferences (Bergen & Grodner, 2012). Specifically, scalar implicatures are less likely to be computed when the speaker only

has partial knowledge of the topic of their utterance (as compared to a fully knowledgeable speaker). It was unknown whether stable speaker characteristics – such as gender, age, or in this case, language background – affected scalar implicature in a similar way. In Chapter 2 of the present work, we addressed this question by comparing judgments of under-informative sentences (e.g., “Some giraffes have long necks”) when these are uttered by native speakers of English vs. non-native, Chinese-accented speakers. Native comprehenders consistently judged such sentences as better when they were attributed to non-native as compared to native speakers. Importantly, this was specific to non-native speakers with lower English language proficiency, and did not extend to highly proficient non-native speakers. In Chapter 3, we further demonstrated that native listeners are more forgiving of non-native speakers’ than native speakers’ under-informativeness, and that this leniency is specifically due to perceptions of inability on the part of the non-native speaker. Additionally, we found evidence that these differing perceptions of native and non-native speakers’ under-informativeness affected learning behavior, with native listeners being more likely to choose to learn from an under-informative non-native speaker (who is probably perceived to have omitted information unintentionally) than an under-informative native speaker (who was more likely to be perceived as unwilling to say more).

Taken together, these findings demonstrate a strong and replicable effect of stable speaker identity on pragmatic judgments: listeners take into account the native language and second language proficiency level of the speaker when making judgments of under-informative utterances. In connection with previous findings demonstrating the effect of situational knowledge on scalar implicature, our results further suggest that listeners integrate multiple attributes of the speaker into their

understanding of an utterance. Importantly, we also show that listeners use these attributes to make additional inferences that the speaker did not likely intend, in our case about the cause of under-informativeness. This happens spontaneously during conversation and affects behavior towards the speaker, as we demonstrated in Experiment 7.

5.2 Listener Effects on Pragmatic Processing

Individuals vary in how likely they are to take the pragmatic interpretation of a sentence like “Some dogs are mammals” (e.g., Bott & Noveck, 2004). Some previous research has suggested that this variation is influenced by ToM (Noveck et al., 2007), while others have shown evidence for an influence of EF on scalar implicature (e.g., De Neys & Schaeken, 2007). Previous findings have resulted in unanswered questions concerning individual variation in scalar implicature and the mechanisms underlying pragmatic inference; a synthesis is obscured further by the fact that ToM and EF are often correlated. Additionally, to the extent that ToM and/or EF is involved, it is unclear whether this relationship generalizes to other types of pragmatic inference. In Chapter 4 we aimed to resolve these mixed findings in a pair of experiments in which we measured ToM, EF, and multiple types of pragmatic judgments within the same participants.

Across both experiments, we found that individuals with better ToM gave more pragmatic responses on scalar implicature tasks. This relationship persisted even when controlling for EF, which had no unique impact on pragmatic judgments. Furthermore, better ToM (but not EF) was also associated with better identification of indirect requests. (Data from metaphor suggested a potentially different outcome, perhaps because of task demands.)

These results align nicely with theories of the mechanisms behind pragmatic inference, according to which we arrive at the pragmatic meaning of an utterance by reasoning about the speaker's intention (e.g., Grice, 1975; Sperber & Wilson, 1995). The better an individual is at reasoning about others' minds, the better able they are to use this skill in conversation to uncover additional meanings of a statement. Our results also potentially suggest that prior work demonstrating an influence of EF on scalar implicature may be due in part to the relationship between EF and ToM.

5.3 Broader Implications

In addition to addressing open questions concerning speaker and listener variation in pragmatic inference, the interdisciplinary nature of this dissertation makes it relevant to a number of other strands of research. Before touching on these lines of work, it is worth noting the real-world application of our findings. Perhaps most importantly, we showed that native listeners tend to feel that non-native speakers are less competent than their native speaker counterparts, both for linguistic (e.g., the speaker could not retrieve the necessary words for an utterance) and cognitive (e.g., the speaker forgot to mention omitted information) reasons. While this bias leads to more lenience with non-native speakers in our results, it may also explain the negative social consequences experienced by non-native speakers inside (Kinzler et al., 2007) and outside of the lab. Children (Kinzler, Corriveau, & Harris, 2011) and adults (Lev-Ari & Keysar, 2010) who find native speakers more trustworthy than non-native speakers may base their judgments on a feeling that non-native speakers are incompetent, and therefore their information should be trusted less. If these experimental findings generalize to the real world, a perception that non-native

speakers are incompetent could contribute to the reasons that employers are less likely to hire non-native speakers (Hosoda & Stone-Romero, 2010), and other similar facts.

The distinction between inability and unwillingness to be under-informative also bears on discussions in the broader literature on pragmatics. Previous theorists have noted that there are multiple reasons why a speaker might provide incomplete information, which can be broadly classified into explanations involving the speaker's inability to say more or unwillingness to do so (Grice, 1975; Sperber & Wilson, 1995; Geurts, 2011). Here we experimentally confirm that these two categories accurately describe listeners' justification of under-informativeness, and further show that such justifications of under-informativeness vary across individual speaker identities. The specific contexts that we investigated in Chapter 3 made it unlikely that the unwillingness or inability to be fully informative was intentionally communicated by the speaker, but such inferences were made freely by the listener. It remains to be seen whether these inferences about the reasons behind under-informativeness also vary across listeners.

Our findings also have implications for fields outside of pragmatics, most notably for theories of non-native speech processing. As discussed in Chapter 2, the predominant theories either posit that non-native speech is processed qualitatively differently than native speech due to expectations about the knowledge and abilities of non-native speakers (Expectation-Based Accounts; Niedzielski, 1999; Lev-Ari, 2015), or that speech processing differs by speaker only to the extent that processing non-native speech is cognitively difficult (Intelligibility-Based Accounts; Davis et al., 2005; Floccia, Goslin, Girard, & Konopczynski, 2006). We do not rule out the possibility that processing costs play a role in influencing non-native speech

processing, but importantly we do show that even when processing costs are not involved, expectations alone affect linguistic judgments of native vs. non-native speech.

Our data leave open whether non-native speaker status is integrated on-line during the earliest stages of sentence processing to guide pragmatic inference or affects later stages of processing. There is evidence that other properties of the speaker such as speaker knowledge of the situation at hand are integrated early during sentence processing (Bergen & Grodner, 2012; Breheny et al., 2013). Electrophysiological studies indicate that the non-native status of the speaker affects syntactic processing online (Goslin et al., 2012; Hanulikova et al., 2012; Romero-Rivas et al., 2015; Grey & Van Hell, 2016), but it remains to be seen whether the same is true in the domain of pragmatics.

Viewed most broadly, our findings contribute to a long line of evidence demonstrating that both speaker and listener identity strongly modulate language processing (e.g., Just & Carpenter, 1992; Prat, Keller, & Just, 2007; van Berkum et al., 2008; Nakano, Saron, & Swaab, 2010; Regel, Coulson, & Gunter, 2010; Prat & Just, 2011; Boudewyn, Long, & Swaab, 2012; Kamide, 2012; Tanner & Van Hell, 2014). For example, van Berkum et al. (2008) found that speaker identity affected online semantic integration: well-formed sentences such as “Every evening I drink wine before I go to bed” with no apparent semantic violations led to increased N400 responses when produced by an unlikely speaker given one’s world knowledge (e.g., a young child). Later related work has shown that the ability to integrate such world knowledge during sentence processing varies with a listener’s cognitive abilities, such as working memory capacity (e.g., Nakano et al., 2010). While many models of

language comprehension take into account individual variation in listener characteristics (Just & Carpenter, 1992; Gernsbacher, Varner, & Faust, 1990), the strong evidence for speaker sensitivity reported here supports models which can also account for the use of speaker properties in language processing (e.g., Nadig & Sedivy, 2002; Sedivy, 2007; Brennan & Hanna, 2009; Pickering & Garrod, 2013). Additionally, speaker characteristics such as gender, age (e.g., van Berkum et al., 2008), and emotion (Nygaard & Lunders 2002) and listener characteristics such as EF (King & Just, 1991), bilingual language experience (Hahne & Friederici, 2001), and socio-economic status (Fernald & Marchman, 2013), have received a considerable amount of attention in the literature on syntactic and semantic processing, but such work is perhaps even more imperative in the domain of pragmatics in order to understand how language comprehension varies in increasingly diverse social contexts.

REFERENCES

- Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics*, *99*, 78–95.
- Apperly, I. A. (2012). What is “theory of mind”? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825-839.
- Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults’ performance on a non-inferential theory of mind task. *Cognition*, *106*(3), 1093-1108.
- Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience*, *16*(10), 1773-1784.
- Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*(3), EL174-EL180.
- Barbet, C., & Thierry, G. (2016). Some alternatives? Event-related potential investigation of literal and pragmatic interpretations of some presented in isolation. *Frontiers in Psychology*, *7*, 1479.
- Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children’s pragmatic inference. *Cognition*, *118*(1), 84-93.
- Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry*, *38*(7), 813-822.
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5-17.

- Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: infants' understanding of intentional action. *Developmental Psychology, 41*(2), 328-37.
- Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America, 114*(3), 1600-1610.
- Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1450-1460.
- Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 14*(1), 12.
- Bialystok, E., Craik, F. I., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychology and Aging, 19*(2), 290-303.
- Bonnefon, J. F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition, 112*(2), 249-258.
- Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language, 66*(1), 123-142.
- Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language, 51*(3), 437-457.
- Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2012). Cognitive control influences the use of meaning relations during spoken sentence comprehension. *Neuropsychologia, 50*(11), 2659-2668.
- Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review, 112*(1), 193.
- Braine, M. D., & Romain, B. (1981). Development of comprehension of "or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology, 31*(1), 46-70.
- Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition, 126*(3), 423-440.

- Breheny, R., Katsos, N., & 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*(3), 434-463.
- Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274-291.
- Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*(5), 893-900.
- Bull, R., Phillips, L. H., & Conway, C. A. (2008). The role of control functions in mentalizing: Dual-task studies of theory of mind and executive function. *Cognition*, *107*(2), 663-672.
- Call, J., Hare, B., Carpenter, M., & Tomasello, M. (2004). 'Unwilling' versus 'unable': chimpanzees' understanding of human intentional action. *Developmental Science*, *7*(4), 488-498.
- Canteloup, C., & Meunier, H. (2017). 'Unwilling' versus 'unable': Tonkean macaques' understanding of human goal-directed actions. *PeerJ*, *5*, e3227.
- Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032-1053.
- Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*, *11*(2), 73-92.
- Carston, R. (1995). Quantity maxims and generalized implicature. *Lingua*, *96*, 213-244.
- Carston, R. (1998). Informativeness, relevance and scalar implicature. In Carston, R. and Uchida, S. (eds.). *Relevance theory: Applications and implications*. Amsterdam: John Benjamins.
- Champagne-Lavau, M., & Joannette, Y. (2009). Pragmatics, theory of mind and executive functions after a right-hemisphere lesion: Different patterns of deficits. *Journal of Neurolinguistics*, *22*(5), 413-426.
- Chevallier, C., Wilson, D., Happé, F., & Noveck, I. (2010). Scalar inferences in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *40*(9), 1104-1117.

- Chiappe, D. L., & Chiappe, P. (2007). The role of working memory in metaphor production and comprehension. *Journal of Memory and Language*, 56(2), 172-188.
- Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In *Proceedings of the 25th Boston University conference on language development* (pp. 157-168). Somerville, MA: Cascadilla Press.
- Clark, H. H., & Marshall, C. R. (1978). Reference diaries. In D. L. Waltz (Ed.), *TINLAP-2: Theoretical issues in natural language processing-2* (pp. 57– 63). New York, NY: Association for Computing Machinery.
- Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6), 3647-3658.
- Clyne, M. (1987). Constraints on code switching: How universal are they? *Linguistics*, 25(290), 739-764.
- Costa, A., Hernández, M., Costa-Faidella, J., & Sebastián-Gallés, N. (2009). On the bilingual advantage in conflict processing: Now you see it, now you don't. *Cognition*, 113(2), 135-149.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12(11), 671.
- Davies, C., Andrés-Roqueta, C., & Norbury, C. F. (2016). Referring expressions and structural language abilities in children with specific language impairment: A pragmatic tolerance account. *Journal of Experimental Child Psychology*, 144, 98-113.
- Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, 134(2), 222-241.
- De Bot, K., & Schreuder, R. (1993). Word production and the bilingual lexicon. In R. Schreuder & B. Weltens (Eds.), *The Bilingual Lexicon* (pp. 191-214).
- De Marchena, A., Eigsti, I. M., Worek, A., Ono, K. E., & Snedeker, J. (2011). Mutual exclusivity in autism spectrum disorders: Testing the pragmatic hypothesis. *Cognition*, 119(1), 96-113.

- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, 54(2), 128-133.
- Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature A constraint-based approach. *Cognitive Science*, 39(4), 667-710.
- Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64, 135-168.
- Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, 45(3), 259-268.
- Dumontheil, I., Apperly, I. A., & Blakemore, S. J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science*, 13(2), 331-338.
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General*, 128(3), 309.
- Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, 40(6), 760-768.
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16(1), 143-149.
- Eviatar, Z., & Just, M. A. (2006). Brain correlates of discourse processing: An fMRI investigation of irony and conventional metaphor comprehension. *Neuropsychologia*, 44(12), 2348-2359.
- Fairchild, S., & Papafragou, A. (October 2017). *Non-native speaker identity influences pragmatic judgments*. Talk given at the Meaning In Flux: Connecting Development, Variation, and Change workshop. Yale University, October.
- Fairchild, S., & Papafragou, A. (Resubmitted). Sins of omission are more likely to be forgiven in non-native speakers. *Cognition*.
- Fairchild, S., & Papafragou, A. (Submitted). Unwilling versus unable: Understanding under-informativeness in native and non-native speakers
- Fairchild, S., & Papafragou, A. (In Prep). Executive function and theory of mind in adults' pragmatic computations.

- Feng, S., Ye, X., Mao, L., & Yue, X. (2014). The activation of theory of mind network differentiates between point-to-self and point-to-other verbal jokes: an fMRI study. *Neuroscience Letters*, *564*, 32-36.
- Feeney, A., Scafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology*, *58*(2), 121.
- Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234-248.
- Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, *41*(1), 78-104.
- Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1276-1293.
- Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, 47-59.
- Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 430-445.
- Geurts, B. (2010). *Quantity implicatures*. New York: Cambridge University Press.
- Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't Underestimate the Benefits of Being Misunderstood. *Psychological Science*, *28*(6), 703-712.
- Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences*, *7*(2), 92-96.
- Gluszek, A., & Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review*, *14*(2), 214-237.
- Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and Language*, *109*(2), 55-67.
- Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language*, *122*(2), 92-102.

- Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1(02), 67-81.
- Grey, S., & van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of language comprehension. *Journal of Neurolinguistics*, 42, 93-108.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (pp. 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(1), 42-55.
- Grosjean, F. (2010). *Bilingual: Life and Reality*. Cambridge, MA: Harvard University Press.
- Gualmini, A., Crain, S., Meroni, L., Chierchia, G., & Guasti, M. T. (2001, October). At the semantics/pragmatics interface in child language. In *Semantics and Linguistic Theory* (Vol. 11, pp. 231-247).
- Guasti, M., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, 20(5), 667-696.
- Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are under-informative. *Cognition*, 132(3), 335-341.
- Hahne, A., & Friederici, A. D. (2001). Processing a second language: Late learners' comprehension mechanisms as revealed by event-related brain potentials. *Bilingualism: Language and Cognition*, 4(2), 123-141.
- Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, 24(4), 878-887.
- Happé, F. G. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2), 101-119.
- Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2), 129-154.

- Heyman, T., & Schaeken, W. (2015). Some differences in some: examining variability in the interpretation of scalars using latent class analysis. *Psychologica Belgica*, 55(1).
- Hilchey, M. D., & Klein, R. M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. *Psychonomic Bulletin & Review*, 18(4), 625-658.
- Hirschberg, J. L. B. (1985). *A theory of scalar implicature*. Philadelphia: University of Pennsylvania.
- Hochstein, L., Bale, A., & Barner, D. (2017). Scalar implicature in absence of epistemic reasoning? The case of autism spectrum disorder. *Language Learning and Development*, 1-17.
- Hochstein, L., Bale, A., Fox, D., & Barner, D. (2014). Ignorance and inference: do problems with Gricean epistemic reasoning explain children's difficulty with scalar implicature? *Journal of Semantics*, 33(1), 107-135.
- Horn, L. R. (1972). *On the semantic properties of the logical operators in English*. Doctoral diss., UCLA.
- Horn, L., R. (1984). Toward a New Taxonomy for Pragmatic Inference: Q-based and R-based Implicature. In D. Schiffrin (Ed.), *Georgetown University Round Table on Languages and Linguistics 1984*, 11– 42. Washington, D.C.: Georgetown University Press.
- Hosoda, M., & Stone-Romero, E. (2010). The effects of foreign accents on employment-related decisions. *Journal of Managerial Psychology*, 25(2), 113-132.
- Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, 58(3), 376-415.
- Huang, Y. T., & Snedeker, J. (2018). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology*, 102, 105-126.
- Hughes, C., & Ensor, R. (2007). Executive function and theory of mind: Predictive relations from ages 2 to 4. *Developmental Psychology*, 43(6), 1447.

- Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture–sentence verification. *Neuroscience Letters*, *534*, 246-251.
- Jankowiak, K., Rataj, K., & Naskręcki, R. (2017). To electrify bilingualism: Electrophysiological insights into bilingual metaphor comprehension. *PLoS One*, *12*(4), e0175578.
- Jolliffe, T., & Baron-Cohen, S. (1999). The strange stories test: A replication with high-functioning adults with autism or Asperger syndrome. *Journal of Autism and Developmental Disorders*, *29*(5), 395-406.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, *99*(1), 122-149.
- Kalin, R., & Rayko, D. S. (1978). Discrimination in evaluative judgments against foreign-accented job candidates. *Psychological Reports*, *43*(3_suppl), 1203-1209.
- Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition*, *124*(1), 66-71.
- Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637-671.
- Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189.
- Kang, O., Rubin, D. O. N., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554-566.
- Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67-81.
- Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32-38.

- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5), 580-602.
- Kinzler, K. D., Corriveau, K. H., & Harris, P. L. (2011). Children's selective trust in native-accented speakers. *Developmental Science*, 14(1), 106-111.
- Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, 104(30), 12577-12580.
- Kline, M., Gallee, J., Balewski, Z., & Fedorenko, E. (Submitted). Understanding jokes relies on the theory of mind system.
- Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33(2), 149.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621-647.
- Lev-Ari, S. (2015). Comprehending non-native speakers: theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*, 5(1546), 1-12.
- Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6), 1093-1096.
- Lev-Ari, S., & Keysar, B. (2012). Less-detailed representation of non-native language: why non-native speakers' stories seem more vague. *Discourse Processes*, 49(7), 523-538.
- Lev-Ari, S., van Heugten, M., & Peperkamp, S. (2017). Relative difficulty of understanding foreign accents as a marker of proficiency. *Cognitive Science*, 41(4), 1106-1118
- Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT press.
- Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46(3), 551-556.

- Marty, P. P., & Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Frontiers in Psychology, 4*, 403.
- Mashal, N. (2013). The role of working memory in the comprehension of unfamiliar and familiar metaphors. *Language and Cognition, 5*(4), 409-436.
- Matthews, D., Biney, H., & Abbot-Smith, K. (In Press). Individual differences in children's pragmatic ability: A review of associations with formal language, social cognition and executive functions. *Language Learning and Development*.
- Mazzarella, D. (2015). Politeness, relevance and scalar inferences. *Journal of Pragmatics, 79*, 93-106.
- Milroy, L., & Muysken, P. (1995). *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge, UK: Cambridge University Press.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49-100.
- Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and Language, 80*(2), 188-207.
- Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*(3), 289-306.
- Myers-Scotton, C., & Jake, J. L. (2000). Testing the 4-M Model: Introduction. *International Journal of Bilingualism, 4*(1), 1-8.
- Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13*(4), 329-336.
- Nakano, H., Saron, C., & Swaab, T. Y. (2010). Speech and span: Working memory capacity impacts the use of animacy but not of world knowledge during spoken sentence comprehension. *Journal of Cognitive Neuroscience, 22*(12), 2886-2898.
- Newschaffer, C. J., Croen, L. A., Daniels, J., Giarelli, E., Grether, J. K., Levy, S. E., & Reynolds, A. M. (2007). The epidemiology of autism spectrum disorders. *Annual Review of Public Health, 28*, 235-258.

- Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*(1), 62-85.
- Nilsen, E. S., & Graham, S. A. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cognitive Psychology, 58*(2), 220-249.
- Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language, 63*(3), 324-346.
- Norbury, C. F. (2005). The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British Journal of Developmental Psychology, 23*(3), 383-399.
- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition, 78*(2), 165-188.
- Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language, 85*(2), 203-210.
- Noveck, I. A., Guelminger, R., Georgieff, N., & Labruyere, N. (2007). What autism can reveal about every... not sentences. *Journal of Semantics, 24*(1), 73-90.
- Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition, 30*(4), 583-593.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255-258.
- Ozturk, O., & Papafragou, A. (2015). The acquisition of epistemic modality: From semantic meaning to pragmatic interpretation. *Language Learning and Development, 11*(3), 191-214.
- Ozturk, O., & Papafragou, A. (2016). The acquisition of evidentiality and source monitoring. *Language Learning and Development, 12*(2), 199-230.
- Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology, 66*(2), 232-258.
- Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition, 86*(3), 253-282.

- Papafragou, A., Friedberg, C., & Cohen, M. L. (In Press). The Role of Speaker Knowledge in Children's Pragmatic Inferences. *Child Development*.
- Pavlenko, A. (Ed.). (2009). *The bilingual mental lexicon: Interdisciplinary approaches*. Bristol, UK: Multilingual Matters.
- Phillips, W., Barnes, J. L., Mahajan, N., Yamaguchi, M., & Santos, L. R. (2009). 'Unwilling' versus 'unable': Capuchin monkeys' (*Cebus apella*) understanding of human intentional action. *Developmental Science*, *12*(6), 938-945.
- Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329-347.
- Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J. P., & Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, *39*(4), 607.
- Pinheiro J, Bates D, DebRoy S, Sarkar D and R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131, <https://CRAN.R-project.org/package=nlme>.
- Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*(2), 191-215.
- Politzer-Ahles, S., & Gwilliams, L. (2015). Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography. *Language, Cognition and Neuroscience*, *30*(7), 853-866.
- Politzer-Ahles, S., Fiorentino, R., Jiang, X., & Zhou, X. (2013). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research*, *1490*, 134-152.
- Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Espanol: Toward a typology of code-switching. *Linguistics*, *18*(2), 221-256.
- Prat, C. S., & Just, M. A. (2011). Exploring the neural dynamics underpinning individual differences in sentence comprehension. *Cerebral Cortex*, *21*(8), 1747-1760.
- Prat, C. S., Keller, T. A., & Just, M. A. (2007). Individual differences in sentence comprehension: a functional magnetic resonance imaging investigation of syntactic and lexical processing demands. *Journal of Cognitive Neuroscience*, *19*(12), 1950-1963.

- Pratt, C. S., Mason, R. A., & Just, M. A. (2012). An fMRI investigation of analogical mapping in metaphor comprehension: The influence of context and individual cognitive capacities on processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(2), 282.
- Prior, A., & MacWhinney, B. (2010). A bilingual advantage in task switching. *Bilingualism: Language and Cognition*, 13(02), 253-262.
- Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition*, 117(2), 230-236.
- R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing.
- Regel, S., Coulson, S., & Gunter, T. C. (2010). The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony. *Brain Research*, 1311, 121-135.
- Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience*, 9(167), 1-15.
- Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, 144(5), 898.
- Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5), 1255.
- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, 27(3), 367-391.
- Sauerland, U. (2012). The computation of scalar implicatures: Pragmatic, lexical or grammatical? *Language and Linguistics Compass*, 6(1), 36-49.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". *Neuroimage*, 19(4), 1835-1842.

- Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., & Werker, J. F. (2012). A bilingual advantage in visual language discrimination in infancy. *Psychological Science, 23*(9), 994-999.
- Sedivy, J. C. (2007). Implicature during real time conversation: A view from language processing research. *Philosophy Compass, 2*(3), 475-496.
- Shetreet, E., Chierchia, G., & Gaab, N. (2014). When some is not every: Dissociating scalar implicature generation and mismatch. *Human Brain Mapping, 35*(4), 1503-1514.
- Skordos, D., & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition, 153*, 6-18.
- Slabakova, R. (2010). Scalar implicatures in second language acquisition. *Lingua, 120*(10), 2444-2462.
- Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587-592.
- Sperber, D. & Wilson, D. (1986) *Relevance: Communication and cognition*. Blackwell.
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press. 2nd edition with Postface.
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359-393.
- Spotorno, N., Koun, E., Prado, J., Van Der Henst, J. B., & Noveck, I. A. (2012). Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage, 63*(1), 25-39.
- Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia, 56*, 289-301.
- Tavano, E., & Kaiser, E. (2010). Processing scalar implicature: What can individual differences tell us? *University of Pennsylvania Working Papers in Linguistics, 16*(1), 24.
- U.S. Census Bureau. (2011). *American Community Survey Reports*. Washington, DC: U.S. Census Bureau.

- Van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., & Rueschemeyer, S. A. (2012). Pragmatics in action: indirect requests engage theory of mind areas and the cortical motor network. *Journal of Cognitive Neuroscience*, *24*(11), 2237-2247.
- Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580-591.
- Van Heugten, M., & Johnson, E. K. (2014). Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General*, *143*(1), 340-350.
- Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 550-564.
- Wei, L. (2002). The bilingual mental lexicon and speech production process. *Brain and Language*, *81*(1), 691-707.
- Wechsler, D. (1944). *The Measurement of Adult Intelligence*. Baltimore, MD: Williams & Wilkins.
- Wechsler, D. (1981). *WAIS-R manual: Wechsler adult intelligence scale-revised*. Psychological Corporation.
- Wilson, D., & Carston, R. (2006). Metaphor, relevance and the 'emergent property' issue. *Mind & Language*, *21*(3), 404-433.
- Winner, E., Brownell, H., Happé, F., Blum, A., & Pincus, D. (1998). Distinguishing lies from jokes: Theory of mind deficits and discourse interpretation in right hemisphere brain-damaged patients. *Brain and Language*, *62*(1), 89-106.

Appendix A

STIMULI FOR EXPERIMENTS 1 AND 2

(UI = Under-Informative, TS = True (Some), TA = True (All), FA = False)

Sentence	Type	Exp.1	Exp.2
Some people have noses with two nostrils.	UI	√	√
Some dogs are mammals with four legs.	UI	√	√
Some doctors attended college to obtain a degree.	UI	√	√
Some televisions have screens on them.	UI	√	√
Some mothers have children.	UI	√	√
Some tables have legs to support them.	UI	√	√
Some fire is hot to the touch.	UI	√	√
Some eyes have pupils.	UI	√	√
Some apples have cores.	UI	√	√
Some children are under eighteen.	UI	√	√
Some rain is wet and falls from the sky.	UI	√	√
Some baseballs are round and can be thrown.	UI	√	√
Some wine is liquid and can be drunk.	UI	√	√
Some books have pages in them.	UI	√	√
Some toast is bread that is heated.	UI	√	√
Some chickens have feathers on their bodies.	UI	√	√
Some colleges have students that attend class.	UI	√	√
Some food is edible.	UI	√	√
Some restaurants serve food to their customers.	UI	√	√
Some helicopters have propellers.	UI	√	√
Some cats are mammals.	UI	√	√
Some frogs are amphibians.	UI		√
Some monkeys have two hands with thumbs.	UI		√
Some spoons are eating utensils.	UI		√
Some sailboats have sails.	UI		√
Some greenhouses grow plants.	UI		√
Some firemen extinguish fires.	UI		√
Some cars have wheels.	UI		√
Some airports have security screening.	UI		√
Some triangles have three sides.	UI		√
Some people have dogs as pets in the house.	TS	√	√
Some cookies are chocolate flavored with chocolate chips.	TS	√	√
Some shoes have high heels.	TS	√	√

Some women wear lipstick as makeup.	TS	√	√
Some students are failing the algebra class.	TS	√	√
Some plants grow flowers.	TS	√	√
Some planets have rings around them.	TS	√	√
Some hair is brown.	TS	√	√
Some food is spicy.	TS	√	√
Some lawyers drive fast cars.	TS	√	√
Some restaurants serve pizza.	TS	√	√
Some clothes are expensive.	TS	√	√
Some houses are blue.	TS	√	√
Some men carry briefcases.	TS	√	√
Some singers are popular.	TS	√	√
Some Europeans have motorcycles.	TS	√	√
Some beaches are public.	TS	√	√
Some foods are allergens.	TS	√	√
Some birds can fly.	TS	√	√
Some gyms have machines.	TS		√
Some birds can talk.	TS		√
Some women wear dresses.	TS		√
Some runners use treadmills.	TS		√
Some farmers raise cows.	TS		√
Some trees grow fruit.	TS		√
Some gardens grow vegetables.	TS		√
Some insects have wings.	TS		√
Some schools are private.	TS		√
Some artists make sculptures.	TS		√
Some men wear suits.	TS		√
All blue jays are birds.	TA	√	√
All horses have hooves.	TA	√	√
All hammers have handles.	TA	√	√
All airplanes have wings.	TA	√	√
All snow is cold.	TA	√	√
All pancakes are flat.	TA	√	√
All people have necks.	TA	√	√
All telephones make calls.	TA	√	√
All dressers have drawers.	TA	√	√
All hearts have aortas.	TA	√	√
All pumpkins have seeds.	TA	√	√
All atoms are small.	TA	√	√
All elephants have trunks.	TA	√	√
All watermelons are red inside.	TA	√	√
All chameleons have skin that changes color.	TA	√	√
All colleges have requirements for admission.	TA	√	√
All dentists recommend cleaning your teeth.	TA	√	√
All chocolate is candy.	TA	√	√
All houses have rooms.	TA	√	√
All apes have opposable thumbs.	TA	√	√

All unicycles have one wheel.	TA		√
All bears have fur.	TA		√
All nuts contain protein.	TA		√
All mountains have peaks.	TA		√
All lobsters are crustaceans.	TA		√
All bars serve alcohol.	TA		√
All toothpastes clean teeth.	TA		√
All raccoons are nocturnal.	TA		√
All skyscrapers are tall.	TA		√
All lifeguards know CPR.	TA		√
All women are doctors.	FA	√	√
All rabbits are white.	FA	√	√
All restaurants serve cakes.	FA	√	√
All ears are small.	FA	√	√
All businessmen own banks.	FA	√	√
All politicians are liberal.	FA	√	√
All cheese has holes.	FA	√	√
All mathematicians are Ukrainian.	FA	√	√
All sisters play tennis.	FA	√	√
All hunters eat lasagna.	FA	√	√
All families have twins.	FA	√	√
All animals are domesticated.	FA	√	√
All magazines have gossip.	FA	√	√
All shirts have long sleeves.	FA	√	√
All bands play jazz.	FA	√	√
All windows have curtains.	FA	√	√
All judges carry purses.	FA	√	√
All pizzas have meat.	FA	√	√
All textbooks have equations.	FA	√	√
All bottles are glass.	FA	√	√
All insects are harmful.	FA		√
All athletes are runners.	FA		√
All houses have garages.	FA		√
All shoes have laces.	FA		√
All bands have saxophonists.	FA		√
All trees have leaves.	FA		√
All kids like asparagus.	FA		√
All boxes are large.	FA		√
All people are retired.	FA		√
All fishermen catch sharks.	FA		√

Appendix B

STIMULI FOR EXPERIMENT 3

Passages were adapted from Bergen and Grodner (2012). Each passage consisted of a context sentence (which was the same across all conditions) and two continuation sentences (which differed according to Passage Type).

Context Sentence	Continuation Sentences			
	Some/All	Some/Rest	Only Some/All	Only Some/Rest
Today I graded all of the exams as part of my teaching assistant position.	Some of the students in the class passed. In fact, they all did because they had studied for many hours.	Some of the students in the class passed. The rest failed the test, despite that they had studied for many hours.	Only some of the students in the class passed. The rest failed the test, despite that they had studied for many hours.	Only some of the students in the class passed. In fact, they all did because they had studied for many hours.
While eating breakfast, I pored over the stock prices from yesterday.	Some of my stocks went up. In fact, they all did because the stock market is doing well.	Some of my stocks went up. The rest went down, despite that the stock market is doing well.	Only some of my stocks went up. The rest went down, despite that the stock market is doing well.	Only some of my stocks went up. In fact, they all did because the stock market is doing well.
Before the hurricane landed, I checked every house in town.	Some of the residents had evacuated. In fact, they all did because they know proper disaster protocol.	Some of the residents had evacuated. The rest stayed, despite that they know proper disaster protocol.	Only some of the residents had evacuated. The rest stayed, despite that they know proper disaster protocol.	Only some of the residents had evacuated. In fact, they all did because they know proper disaster protocol.
As the first to arrive on the scene of an accident, I examined all of the passengers	Some of the passengers had been injured. In fact, they all had because the bus	Some of the passengers had been injured. The rest were unharmed, despite that the bus	Only some of the passengers had been injured. The rest were unharmed, despite that the bus	Only some of the passengers had been injured. In fact, they all had because the bus

while waiting for the ambulance to arrive.	crashed into a building.	crashed into a building.	crashed into a building.	crashed into a building.
While volunteering in the veterinary clinic, I checked all of the vaccination records for a collie that was scheduled to come in the next day.	Some of the dog's vaccinations were up to date. In fact, they all were because of the owner's diligence.	Some of the dog's vaccinations were up to date. The rest were outdated, despite the owner's diligence.	Only some of the dog's vaccinations were up to date. The rest were outdated, despite the owner's diligence.	Only some of the dog's vaccinations were up to date. In fact, they all were because of the owner's diligence.
When they returned, I reviewed each of the receipts for my family's trip to Europe.	Some of the hotels were expensive. In fact, they all were because my family loves luxurious holidays.	Some of the hotels were expensive. The rest were cheap despite my family's love for luxurious holidays.	Only some of the hotels were expensive. The rest were cheap despite my family's love for luxurious holidays.	Only some of the hotels were expensive. In fact, they all were because my family loves luxurious holidays.
After the massive storm, I went to the amusement park to inspect the damage and compile a comprehensive report.	Some of the rides would have to be replaced. In fact, they all would because of being barraged by debris.	Some of the rides would have to be replaced. The rest were undamaged, despite being barraged by debris.	Only some of the rides would have to be replaced. The rest were undamaged, despite being barraged by debris.	Only some of the rides would have to be replaced. In fact, they all would because of being barraged by debris.
On a school trip to a prison, I helped the medical science professor to give each inmate a thorough medical examination.	Some of the inmates are infested with lice. In fact, they all are because of the poor conditions of the prison.	Some of the inmates are infested with lice. The rest are clean, despite the poor conditions of the prison.	Only some of the inmates are infested with lice. The rest are clean, despite the poor conditions of the prison.	Only some of the inmates are infested with lice. In fact, they all are because of the poor conditions of the prison.
My dad is the only mechanic in our town, and I had to help him inspect all of the local school buses last week.	Some of the buses will need new brakes. In fact, they all will because of being driven pretty roughly.	Some of the buses will need new brakes. The rest are running smoothly, despite being driven pretty roughly.	Only some of the buses will need new brakes. The rest are running smoothly, despite being driven pretty roughly.	Only some of the buses will need new brakes. In fact, they all will because of being driven pretty roughly.

In the middle of the night, I was woken by a commotion coming from the chicken coop and went out to inspect the damage.	Some of the chickens had been eaten. In fact, they all had because a fox got into the coop.	Some of the chickens had been eaten. The rest were safe, despite that a fox got into the coop.	Only some of the chickens had been eaten. The rest were safe, despite that a fox got into the coop.	Only some of the chickens had been eaten. In fact, they all had because a fox got into the coop.
My mom's company has been losing money so I carefully reviewed each client's invoices.	Some of our clients have been undercharged. In fact, they all have because of a mistake in the books.	Some of our clients have been undercharged. The rest have been charged accurately, despite a mistake in the books.	Only some of our clients have been undercharged. The rest have been charged accurately, despite a mistake in the books.	Only some of our clients have been undercharged. In fact, they all have because of a mistake in the books.
As the manager of a college book store, I reviewed the new shipment of text books carefully. I inspected each package of fish in my family's large deep freezer.	Some of the new textbooks had water stains. In fact, they all did because of the rainstorm. Some of the fish were rotten. In fact, they are because of the power outage.	Some of the new textbooks had water stains. The rest were fine despite the rainstorm. Some of the fish were rotten. The rest were fine despite the power outage.	Only some of the new textbooks had water stains. The rest were fine despite the rainstorm. Only some of the fish were rotten. The rest were fine despite the power outage.	Only some of the new textbooks had water stains. In fact, they all did because of the rainstorm. Only some of the fish were rotten. In fact, they are because of the power outage.
As the vice president of the student magic club I carefully checked my deck of cards before the show.	Some of my cards were bent. In fact, they all were because of my small pockets.	Some of my cards were bent. The rest were fine despite my small pockets.	Only some of my cards were bent. The rest were fine despite my small pockets.	Only some of my cards were bent. In fact, they all were because of my small pockets.
I stocked all of the shelves at my mom's business this morning.	Some of the merchandise was dusty. In fact, it all was because of the lack of customers.	Some of the merchandise was dusty. The rest was clean, despite the lack of customers.	Only some of the merchandise was dusty. The rest was clean, despite the lack of customers.	Only some of the merchandise was dusty. In fact, it all was because of the lack of customers.
This morning, I carefully checked the text messages I received over night.	Some of the messages from Jane were angry. In fact, they all	Some of messages from Jane were angry. The rest were	Only some of messages from Jane were angry. The rest were	Only some of the messages from Jane were angry. In fact, they all

	were because of our recent fight.	nice, despite our recent fight.	nice, despite our recent fight.	were because of our recent fight.
As the oldest in the family, it's my job to check all of my younger siblings' homework.	Some of my siblings understood the concepts covered in class. In fact, they all do because their teacher really cares about them.	Some of my siblings understood the concepts covered in class. The rest are confused by the material, even though their teacher really cares about them.	Only some of my siblings understood the concepts covered in class. The rest are confused by the material, even though their teacher really cares about them.	Only some of my siblings understood the concepts covered in class. In fact, they all do because their teacher really cares about them.
Before my driving test, I inspected the car in great detail.	Some of the doors have scratches on them. In fact, they all do because of the age of the car.	Some of the doors have scratches on them. The rest look new, despite the age of the car.	Only some of the doors have scratches on them. The rest look new, despite the age of the car.	Only some of the doors have scratches on them. In fact, they all do because of the age of the car.
I carefully inspected the new jewelry my sister bought.	Some of the gold watches were fakes. In fact, they all were so she is planning to return them.	Some of the gold watches were fakes. The rest were real, but she is still planning to return them.	Only some of the gold watches were fakes. The rest were real, but she is still planning to return them.	Only some of the gold watches were fakes. In fact, they all were so she is planning to return them.
This morning, I took attendance at an important meeting with the manager at work.	Some of the company's accountants were there. In fact, they all were to communicate how budget cutbacks were crippling their division.	Some of the company's accountants were there. The rest were missing because they had to audit the company's finances before the end of the quarter.	Only some of the company's accountants were there. The rest were missing because they had to audit the company's finances before the end of the quarter.	Only some of the company's accountants were there. In fact, they all were to communicate how budget cutbacks were crippling their division.
Last Saturday I took all of my younger cousins to the playground.	Some of my cousins got mosquito bites. In fact, they all did because I forgot the bug spray.	Some of my cousins got mosquito bites. The rest did not, despite that I forgot the bug spray.	Only some of my cousins got mosquito bites. The rest did not, despite that I forgot the bug spray.	Only some of my cousins got mosquito bites. In fact, they all did because I forgot the bug spray.

For my advanced accounting class, I meticulously compiled the investment report.	Some of the real estate investments lost money. In fact, they all did because of the recent economic downturn.	Some of the real estate investments lost money. The rest were successful despite the recent economic downturn.	Only some of the real estate investments lost money. The rest were successful despite the recent economic downturn.	Only some of the real estate investments lost money. In fact, they all did because of the recent economic downturn.
As the most tech-savvy employee at my job, I had to check each computer for the dangerous new virus.	Some of our computers were infected. In fact, they all were and the virus nearly destroyed the whole system.	Some of our computers were infected. The rest were clean because their owners had been very cautious.	Only some of our computers were infected. The rest were clean because their owners had been very cautious.	Only some of our computers were infected. In fact, they all were and the virus nearly destroyed the whole system.
Earlier today, I was leading a small group of prospective students around the sights down town.	Some of the prospective students got soaked by the rainstorm. In fact, they all did because they had forgotten their umbrellas.	Some of the prospective students got soaked by the rainstorm. The rest were dry because they remembered their umbrellas.	Only some of the prospective students got soaked by the rainstorm. The rest were dry because they remembered their umbrellas.	Only some of the prospective students got soaked by the rainstorm. In fact, they all did because they had forgotten their umbrellas.
After my garage sale, I cataloged all of the remaining items.	Some of the old couches had been sold. In fact, they all had since they were stylish and cheap.	Some of the old couches had been sold. The rest were going to be stored until the following summer.	Only some of the old couches had been sold. The rest were going to be stored until the following summer.	Only some of the old couches had been sold. In fact, they all had since they were stylish and cheap.
Last week, I tasted every dish at a family potluck.	Some of the dishes were spicy. In fact, they all were but fortunately I love spicy food.	Some of the dishes were spicy. The rest were mild and I found them to be bland.	Only some of the dishes were spicy. The rest were mild and I found them to be bland.	Only some of the dishes were spicy. In fact, they all were but fortunately I love spicy food.
When I entered Disney World, I asked about the status of each of the rides.	Some of my favorite rides were still running. In fact, they all were since they were still popular.	Some of my favorite rides were still running. The rest were shut down because they were no longer popular.	Only some of my favorite rides were still running. The rest were shut down because they were no longer popular.	Only some of my favorite rides were still running. In fact, they all were since they were still popular.

After my house was burglarized, I carefully inventoried my wine collection.	Some of my bottles of Chardonnay were missing. In fact, they all were even though I had secured them.	Some of my bottles of Chardonnay were missing. The rest were safe but I was still extremely upset.	Only some of my bottles of Chardonnay were missing. The rest were safe but I was still extremely upset.	Only some of my bottles of Chardonnay were missing. In fact, they all were even though I had secured them.
While volunteering in the veterinary clinic, I closely examined the mouth of a large bulldog.	Some of the dog's teeth were missing. In fact, they all were because its owner completely neglected its oral hygiene.	Some of the dog's teeth were missing. The rest were intact so it should still be able to eat solid food.	Only some of the dog's teeth were missing. The rest were intact so it should still be able to eat solid food.	Only some of the dog's teeth were missing. In fact, they all were because its owner completely neglected its oral hygiene.
In the school parking lot, I carefully inspected an old bus.	Some of its tires were flat. In fact, they all were so the cost to repair it would be enormous.	Some of its tires were flat. The others were fine so it wouldn't cost too much to fix it.	Only some of its tires were flat. The others were fine so it wouldn't cost too much to fix it.	Only some of its tires were flat. In fact, they all were so the cost to repair it would be enormous.
To check on the progress of my class research project, I meticulously recorded the results of the experiment.	Some of my predictions were correct. In fact, they all were so I should be able to publish the results.	Some of my predictions were correct. The rest were wrong so my theory must be mistaken.	Only some of my predictions were correct. The rest were wrong so my theory must be mistaken.	Only some of my predictions were correct. In fact, they all were so I should be able to publish the results.
After the babysitter left, I carefully examined my liquor collection.	Some of my new bottles of vodka were opened. In fact, they all were but I decided not to call her parents because it was so hard to find a babysitter.	Some of my new bottles of vodka were opened. The rest were untouched, but I was still concerned and decided to call her parents.	Only some of my new bottles of vodka were opened. The rest were untouched, but I was still concerned and decided to call her parents.	Only some of my new bottles of vodka were opened. In fact, they all were but I decided not to call her parents because it was so hard to find a babysitter.
I examined the damage after I dropped a bowling ball down the steps.	Some of the steps were damaged. In fact, they all were so they will require extensive repairs.	Some of the steps were damaged. The others were fine so the repairs shouldn't be too expensive.	Only some of the steps were damaged. The others were fine so the repairs	Only some of the steps were damaged. In fact, they all were so they will require extensive repairs.

			shouldn't be too expensive.	
At a friend's suggestion, I completely worked through an entire math textbook.	Some of its problems were difficult. In fact, they all were but it received a positive review anyway.	Some of its problems were difficult. The rest were straightforward and I feel like I learned a lot.	Only some of its problems were difficult. The rest were straightforward and I feel like I learned a lot.	Only some of its problems were difficult. In fact, they all were but it received a positive review anyway.
Before grocery shopping, I wrote down exactly how much of each item we had left.	Some of the condiments needed to be refilled. In fact, they all did because we cooked a lot this week.	Some of the condiments needed to be refilled. The rest were full despite that we cooked a lot this week.	Only some of the condiments needed to be refilled. The rest were full despite that we cooked a lot this week.	Only some of the condiments needed to be refilled. In fact, they all did because we cooked a lot this week.
The student government picked me to organize every award for the schoolwide assembly.	Some of the honors students received prizes. In fact, they all did because the professors didn't want anyone to feel left out.	Some of the honors students received prizes. The rest weren't invited because the professors didn't want anyone to feel left out.	Only some of the honors students received prizes. The rest weren't invited because the professors didn't want anyone to feel left out.	Only some of the honors students received prizes. In fact, they all did because the professors didn't want anyone to feel left out.
This weekend, I made it my project to catalog every book in my large collection.	Some of the dictionaries were labeled incorrectly. In fact, they all were which made my job much more difficult.	Some of the dictionaries were labeled incorrectly. The rest were labelled correctly, but a few were shelved in the wrong place.	Only some of the dictionaries were labeled incorrectly. The rest were labelled correctly, but a few were shelved in the wrong place.	Only some of the dictionaries were labeled incorrectly. In fact, they all were which made my job much more difficult.
I am a huge fan of my old high school football team and attended every game last season.	Some of their losses were close. In fact, they all were which made the games stressful to watch.	Some of their losses were close. The rest were blowouts, which made the games boring to watch.	Only some of their losses were close. The rest were blowouts, which made the games boring to watch.	Only some of their losses were close. In fact, they all were which made the games stressful to watch.
To prepare for my Spanish test, I spent hours	Some of the words sounded like they do in	Some of the words sounded like they do in	Only some of the words sounded like they do in	Only some of the words sounded like they do in

studying the new vocabulary items.	English. In fact, they all did which made the test somewhat easier.	English. The rest were totally unfamiliar which made the test somewhat challenging.	English. The rest were totally unfamiliar which made the test somewhat challenging.	English. In fact, they all did which made the test somewhat easier.
In preparation for the party, I gathered every chair in the house.	Some of the chairs are plastic. In fact, they all are which will make them easy to clean.	Some of the chairs are plastic. The rest are cloth, which will make them more difficult to clean.	Only some of the chairs are plastic. The rest are cloth, which will make them more difficult to clean.	Only some of the chairs are plastic. In fact, they all are which will make them easy to clean.

Appendix C

STIMULI FOR THE DUAL SCALAR IMPLICATURE TASK IN EXPERIMENT 8

Sentence	Type
Some noses have two nostrils.	Under-Informative
Some dogs are mammals.	Under-Informative
Some doctors attended college.	Under-Informative
Some televisions have screens.	Under-Informative
Some mothers have children.	Under-Informative
Some tables have legs.	Under-Informative
Some fire is hot to the touch.	Under-Informative
Some eyes have pupils.	Under-Informative
Some apples have cores.	Under-Informative
Some children are under eighteen.	Under-Informative
Some rain is wet.	Under-Informative
Some baseballs are round.	Under-Informative
Some wine is liquid.	Under-Informative
Some books have pages.	Under-Informative
Some toast is bread that is heated.	Under-Informative
Some chickens have feathers.	Under-Informative
Some colleges have students.	Under-Informative
Some food is edible.	Under-Informative
Some restaurants serve food.	Under-Informative
Some helicopters have propellers.	Under-Informative
Some women wear dresses.	True (Some)
Some people have dogs.	True (Some)
Some cookies are chocolate.	True (Some)
Some shoes have high heels.	True (Some)
Some women wear lipstick.	True (Some)
Some students are failing the algebra class.	True (Some)
Some plants grow flowers.	True (Some)
Some planets have rings around them.	True (Some)
Some hair is brown.	True (Some)
Some food is spicy.	True (Some)
Some lawyers drive fast cars.	True (Some)
Some restaurants serve pizza.	True (Some)
Some clothes are expensive.	True (Some)
Some houses are blue.	True (Some)
Some men carry briefcases.	True (Some)

Some singers are popular.	True (Some)
Some Europeans have motorcycles.	True (Some)
Some beaches are public.	True (Some)
Some foods are allergens.	True (Some)
Some birds can fly.	True (Some)
All blue jays are birds.	True (All)
All horses have hooves.	True (All)
All hammers have handles.	True (All)
All airplanes have wings.	True (All)
All snow is cold.	True (All)
All pancakes are flat.	True (All)
All people have necks.	True (All)
All telephones make calls.	True (All)
All dressers have drawers.	True (All)
All hearts have aortas.	True (All)
All pumpkins have seeds.	True (All)
All atoms are small.	True (All)
All elephants have trunks.	True (All)
All watermelons are red inside.	True (All)
All chameleons have skin that changes color.	True (All)
All colleges have requirements for admission.	True (All)
All dentists recommend cleaning your teeth.	True (All)
All chocolate is candy.	True (All)
All houses have rooms.	True (All)
All apes have opposable thumbs.	True (All)
All women are doctors.	False
All rabbits are white.	False
All restaurants serve cakes.	False
All ears are small.	False
All businessmen own banks.	False
All politicians are liberal.	False
All cheese has holes.	False
All mathematicians are Ukrainian.	False
All sisters play tennis.	False
All hunters eat lasagna.	False
All families have twins.	False
All animals are domesticated.	False
All magazines have gossip.	False
All shirts have long sleeves.	False
All bands play jazz.	False
All windows have curtains.	False
All judges carry purses.	False
All pizzas have meat.	False
All textbooks have equations.	False
All bottles are glass.	False

Appendix D

STIMULI FOR THE SIMPLE SCALAR IMPLICATURE TASK IN EXPERIMENTS 8 AND 9

Sentences borrowed from Bott and Noveck (2004).

Sentence	Type
Some trout are fish.	Under-Informative
Some lizards are reptiles.	Under-Informative
Some sparrows are birds.	Under-Informative
Some dogs are mammals.	Under-Informative
Some ants are insects.	Under-Informative
Some fish are tuna.	Informative
Some reptiles are alligators.	Informative
Some birds are eagles.	Informative
Some mammals are elephants.	Informative
Some insects are mosquitos.	Informative

Appendix E

STIMULI FOR THE METAPHOR TASK IN EXPERIMENT 9

Stimuli borrowed from Jankowiak, Rataj, and Naskręcki (2017).

Phrase	Type
To harvest courage	Metaphor
To store courtesies	Metaphor
To swallow defeat	Metaphor
To freeze departure	Metaphor
To smell excuses	Metaphor
To divide glory	Metaphor
To choke laughter	Metaphor
To repair legacy	Metaphor
To taste privilege	Metaphor
To kill wishes	Metaphor
To feel anger	Control
To tolerate anxiety	Control
To file assault	Control
To note attendance	Control
To lack awareness	Control
To advise caution	Control
To start charity	Control
To regulate conduct	Control
To study consciousness	Control
To sign consent	Control

Appendix F

STIMULI FOR THE INDIRECT REQUEST TASK IN EXPERIMENT 9

Stimuli adapted from Van Ackeren, Casasanto, Bekkering, Hagoort, and Rueschemeyer (2012).

Item	Picture 1 (IR & UC Conditions)	Indirect Request Sentence	Utterance Control Sentence	Picture 2 (PC & PUC Conditions)	Picture Control Sentence	Picture- Utterance Control Sentence
1	Closed window	It is very hot here.	It is very nice here.	Truck in desert	It is very hot here.	It is very nice here.
2	Candy dish	That looks delicious.	That looks colorful.	Cookbook and spoon	That looks delicious.	That looks colorful.
3	Folded blanket	It is very cold here.	It is very pretty here.	Mountain and skis	It is very cold here.	It is very pretty here.
4	Coffee maker	I am still very tired.	I am very awake.	Toaster	I am still very tired.	I am very awake.
5	Closed door	It is still closed.	That is not it.	Locksmith storefront	It is still closed.	That is not it.
6	Die on board game	I am finished now.	This is a good game.	Computer mouse	I am finished now.	This is a good game.
7	Remote control	That is in Chinese.	That is in the house.	Computer keyboard	That is in Chinese.	That is in the house.
8	Vacuum cleaner	It is very sandy here.	It is very annoying.	Sand pile	It is very sandy here.	It is very annoying.
9	Crooked painting	That is crooked.	That is unique.	Leaning tower of Pisa	That is crooked.	That is unique.
10	Plate of food	That is very tasteless.	That is very healthy.	Can of beans	That is very tasteless.	That is very healthy.
11	Window with curtains drawn	It is already light outside.	These are long curtains.	Window with open curtains	It is already light outside.	These are long curtains.
12	Umbrella	It is starting to rain.	This was a good hike.	Ski poles	It is starting to rain.	This was a good hike.
13	Snow shovel	It is very slippery here.	It is durable.	Mop	It is very slippery here.	It is durable.

14	Cheese and knife	That looks like a lot.	It is very sharp.	Knives for sale	That looks like a lot.	It is very sharp.
15	Watering can and plant	It is very dry.	It is very expensive.	Wine in glasses	It is very dry.	It is very expensive.
16	Glasses on table	I can't see very well.	This is a famous place.	Binoculars in fog	I can't see very well.	This is a famous place.
17	Light switch	It is dark here.	It is old.	Train tracks	It is dark here.	It is old.
18	Back massager	My back hurts.	I can lift it easily.	Barbell	My back hurts.	I can lift it easily.
19	Dog	He looks hungry.	He is large.	Tiger	He looks hungry.	He is large.
20	Ticket machine	I do not have train tickets.	We already have train tickets.	Closed ticket booth	I do not have train tickets.	We already have train tickets.
21	Mailbox on home	The mailman has just come.	The mailman did not come.	Public mailbox	The mailman has just come.	The mailman did not come.
22	Heater	It is too hot here.	It is nice and warm.	Boombox on beach	It is too hot here.	It is nice and warm.
23	Watch	It is now Daylight Savings time.	It is time to go.	Big Ben	It is now Daylight Savings time.	It is time to go.
24	Lamp	I think it has a brighter setting.	This is a nice lamp.	Streetlight	I think it has a brighter setting.	This is a nice lamp.
25	Document on computer	There is a comma missing.	She is a good author.	E-reader	There is a comma missing.	She is a good author.
26	Telephone	Today is Mother's Day.	Today is Independence Day.	Cake and coffee	Today is Mother's Day.	Today is Independence Day.
27	Toy car	The engine is broken.	The engine is functioning.	Plane wheels	The engine is broken.	The engine is functioning.
28	Elevator buttons	I need to go to the third floor.	This goes to the doctor's office.	Stairs behind closed door	I need to go to the third floor.	This goes to the doctor's office.
29	Television	I don't like this program.	This is a nice program.	Conference schedule	I don't like this program.	This is a nice program.
30	Stacked plates	It is almost time.	It is not time yet.	Jenga	It is almost time.	It is not time yet.
31	Camera	This is a nice location.	I just bought this.	Compass	This is a nice location.	I just bought this.

32	Service bell	There is no one here to help us.	There is someone to ring the bell.	Church bell	There is no one here to help us. It smells	There is someone to ring the bell.
33	Cookies in oven	It smells very good here.	I am full now.	Crackers in box	very good here.	I am full now.
34	December calendar	It is January now.	It is Monday again.	January calendar	It is January now.	It is Monday again.
35	Gas pump	The tank is low on gas.	The tank is full.	Plane wing	The tank is low on gas.	The tank is full.
36	Towels hanging on line	It is dry now.	The color white looks nice.	Room with painting supplies	It is dry now.	The color white looks nice.

Appendix G

IRB APPROVAL FOR HUMAN SUBJECTS RESEARCH



RESEARCH OFFICE

210 Hulihan Hall
University of Delaware
Newark, Delaware 19716-1551
Ph: 302/831-2136
Fax: 302/831-2828

DATE: February 7, 2018

TO: Anna Papafragou, PhD
FROM: University of Delaware IRB

STUDY TITLE: [312739-18] Language Acquisition: Word learning and pragmatic inference

SUBMISSION TYPE: Continuing Review/Progress Report

ACTION: APPROVED
APPROVAL DATE: February 7, 2018
EXPIRATION DATE: March 1, 2019
REVIEW TYPE: Expedited Review

REVIEW CATEGORY: Expedited review category # (7)

Thank you for your submission of Continuing Review/Progress Report materials for this research study. The University of Delaware IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Expedited Review based on the applicable federal regulation.

Please remember that informed consent is a process beginning with a description of the study and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms for this procedure. All sponsor reporting requirements should also be followed.

Please report all NON-COMPLIANCE issues or COMPLAINTS regarding this study to this office.

Please note that all research records must be retained for a minimum of three years.

Based on the risks, this project requires Continuing Review by this office on an annual basis. Please use the appropriate renewal forms for this procedure.

Appendix H

IRB APPROVAL FOR HUMAN SUBJECTS RESEARCH



RESEARCH OFFICE

210 Hullihen Hall
University of Delaware
Newark, Delaware 19716-1551
Ph: 302/831-2136
Fax: 302/831-2828

DATE: April 2, 2018

TO: Anna Papafragou
FROM: University of Delaware IRB

STUDY TITLE: [165481-19] The interface between spatial cognition and language

SUBMISSION TYPE: Amendment/Modification

ACTION: APPROVED
APPROVAL DATE: April 2, 2018
EXPIRATION DATE: April 7, 2019
REVIEW TYPE: Expedited Review

REVIEW CATEGORY: Expedited review category # (7)

Thank you for your submission of Amendment/Modification materials for this research study. The University of Delaware IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Expedited Review based on the applicable federal regulation.

Please remember that informed consent is a process beginning with a description of the study and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms for this procedure. All sponsor reporting requirements should also be followed.

Please report all NON-COMPLIANCE issues or COMPLAINTS regarding this study to this office.

Please note that all research records must be retained for a minimum of three years.

Based on the risks, this project requires Continuing Review by this office on an annual basis. Please use the appropriate renewal forms for this procedure.