Effects of Working Memory Load and Speaker Reliability on Contrastive Inference and Quantifier Processing

A dissertation presented
by
Laine Stranahan
to
The Department of Linguistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Linguistics

Harvard University
Cambridge, Massachusetts

May, 2018
Effects of Working Memory Load and Speaker Reliability on Contrastive Inference and Quantifier Processing

Abstract

This dissertation investigates the processing of two distinct linguistic phenomena: Quantifier interpretation and contrastive inference. Research has indicated that non-literal, or pragmatic interpretation, e.g. scalar implicature, is slower and more effortful than truth-conditional, or semantic, interpretation alone. Consequently, much research has been guided by the assumption that easy or fast processes are likely to be semantic in nature, whereas slow or effortful ones are likely to be pragmatic. In two eye-tracking experiments investigating the interpretation of scalar quantifiers and numerals, literal interpretation of "all" was impaired by verbal working memory load, violating the broad generalization that semantics is easy and pragmatics is hard. Exact interpretation of numerals was unaffected, suggesting numeral upper bounds are easy to compute. In two further experiments, contrastive inferences were easy to generate but difficult to suppress, a result not predicted by theories in which listeners generate such inferences by modeling speakers as rational agents. The effects of working memory load and speaker reliability on adjective interpretation confirmed instead the predictions of automatic egocentric processing. Taken together, these results challenge the division of linguistic processing into two homogeneous classes distinguishable by differences in cognitive implementation.
# Table of Contents

Chapter 1: Introduction..............................................................................................................7
Chapter 2: Effects of Working Memory Load on the Processing of Numerals and Scalar Quantifiers...13
    Introduction..........................................................................................................................13
    Scalar Implicature Online.................................................................................................14
    Scalar Implicature and Working Memory........................................................................17
Experiment 1................................................................................................................................19
    Methods...............................................................................................................................21
    Participants.........................................................................................................................21
    Procedure............................................................................................................................21
    Materials.............................................................................................................................24
    Results.................................................................................................................................27
    Evidence of Scalar Implicature.........................................................................................30
    Effects of Working Memory Load...................................................................................32
    Discussion.........................................................................................................................35
Experiment 2................................................................................................................................38
    Methods...............................................................................................................................39
    Participants.........................................................................................................................39
    Procedure............................................................................................................................40
    Materials.............................................................................................................................40
    Results.................................................................................................................................40
    Effects of Working Memory Load...................................................................................42
    Discussion.........................................................................................................................44
    General Discussion..........................................................................................................45
    Conclusions.......................................................................................................................46
Chapter 3: Effects of Working Memory Load on Contrastive Inference....................................48
    Introduction..........................................................................................................................48
    Contrastive Inference.........................................................................................................52
    Contrastive Inference Processing.....................................................................................55
    Working Memory Load......................................................................................................58
Experiment 3................................................................................................................................60
    Methods...............................................................................................................................60
    Participants.........................................................................................................................60
    Materials.............................................................................................................................61
    Visual Displays..................................................................................................................61
    Auditory Commands.........................................................................................................64
    Working Memory Task.......................................................................................................65
    Procedure............................................................................................................................66
    Results.................................................................................................................................69
    Eye-tracking Analysis: Critical Region.............................................................................73
    Post Hoc Eye-tracking Analyses.....................................................................................74
    Discussion.........................................................................................................................76
Chapter 4: Effects of Working Memory Load and Speaker Reliability on Contrastive Inference......80
    Introduction.........................................................................................................................80
    Predictions..........................................................................................................................83
Experiment 4........................................................................................................................................89
  Methods...........................................................................................................................................90
    Participants.....................................................................................................................................90
    Procedure........................................................................................................................................90
    Materials.........................................................................................................................................92
      Visual Displays.........................................................................................................................92
      Auditory Commands................................................................................................................93
      Working Memory Task...........................................................................................................93
  Results..............................................................................................................................................94
    Eye-Tracking Analysis: Critical Region.......................................................................................95
    Revisiting the Effect of Working Memory Load...........................................................................95
    Overall (Omnibus) Analyses.......................................................................................................96
    Analyses by Load Condition......................................................................................................99
    Eye-tracking Analysis: 0-100ms Post-Adjective-Offset.................................................................100
  Discussion.........................................................................................................................................101
    Contrastive Inference as a Self-Oriented Process.......................................................................102
    Perspective-Taking......................................................................................................................104
  Conclusions.......................................................................................................................................108

Chapter 5: Conclusions....................................................................................................................111

References..........................................................................................................................................115

Appendix A: Experiment 1 & 2 Items...............................................................................................122
Appendix B: Experiment 1 & 2 Letter Sequences...........................................................................123
Appendix C: Experiment 3 Displays...............................................................................................124
  Critical Displays...........................................................................................................................124
  Filler Displays...............................................................................................................................125
Appendix D: Experiment 3 Letter Sequences..................................................................................126
Appendix E: Experiment 3 & 4 Working Memory Capacity Assessment.........................................128
  Materials..........................................................................................................................................128
Appendix F: Chapter 4 Displays.......................................................................................................129
  Critical Displays...........................................................................................................................129
  Filler Displays...............................................................................................................................130
Appendix G: Experiment 4 Commands.............................................................................................131
  Filler...............................................................................................................................................132
Appendix H: Experiment 3 & 4 Letter Sequences............................................................................134
Acknowledgments

Qingqing Wu and Manizeh Khan assisted with study design for Chapter 2. Study design for Chapter 3 was done in collaboration with Dylan Hardenbergh. Data collection for Chapters 2 through 4 was performed with the assistance of Sandy Kim, Chen Zhou, Adaugo Ugocha, Dylan Hardenbergh, Joe Palana, Zheng Zhang, Yiping Li, Sam Benkelman, Anna Schuliger, Emily Moya, Josh Lipson, Evgeniia Diachek, Janae Hughes, Maya Saupe, and Heyang Yin. Experiments in Chapters 3 and 4 were partially funded by a grant from the Mind, Brain & Behavior Interfaculty Initiative. I am indebted to audiences at the Snedeker Lab, the Meaning and Modality Lab, the 88th Annual Meeting of the LSA, Sinn und Bedeutung 18, ESSLLI 25, AMLaP 20 & 22, SNEWS 2016 @ Brown, BLS 43, CUNY 2017, the Stanford CSLI Workshop, ICLC 14, and many more for their feedback. Last but not least, I am grateful to my advisor, committee members, teachers, collaborators, colleagues, departmental administrators, research assistants, lab managers, students, friends, and family for their mentorship, guidance, and support.
Chapter 1: Introduction

Understanding a linguistic utterance is more than just deriving its literal meaning. Suppose a friend visiting you at home utters the sentence in (1.1) while sitting next to an open window:

(1.1) It's freezing in here!

It literally means that the temperature of the speaker's location is very cold. But by politely mentioning the cold without overtly complaining or making a demand, your friend is most likely making an indirect request that you close the window.

Literal, or semantic, meaning pertains to the truth or falsity of a sentence: It either is or isn't freezing. In order to successfully communicate, we must figure out what a speaker intends by making an utterance: Its intended, or pragmatic, meaning. In this case, we reason that the speaker wants to avoid confrontation while still getting us to close the window. By most accounts, we must derive semantic meaning in order to make the often complex leap to what the speaker actually meant. By decoding the truth-conditions of (1.1) we reason that despite its out-of-the-blue character, it must be relevant somehow, prompting us to look for a reason the speaker chose this strategy.

While semantically interpreting (1.1) is relatively straightforward, interpreting it pragmatically is a somewhat more complex and potentially unbounded task, requiring the listener to appeal to facts about social status, politeness, world knowledge, and so on. How language users manage to reliably infer speakers' intended meanings from their literal statements is a rich and open-ended question. One of the more challenging tasks of linguistics is to uncover the mechanisms by which speakers and listeners systematically relate the two.
With the implicit goal of learning how listeners are able to reliably derive speakers' intended meaning from the literal meanings of their utterances, this dissertation investigates two simple cases of listeners systematically adding to or expanding on the semantic aspects of what speakers say, i.e., two cases of pragmatic enrichment. Since pragmatic enrichment is so variable and complex, I have chosen two types of pragmatic enrichment which appear relatively constant across speakers and contexts: Scalar implicature and contrastive inference. The intended meanings of the utterances that trigger them are, in general, regular and predictable.

Investigating scalar implicature and contrastive inference, among the wide variety of pragmatic inferences language users make on a daily basis, has the potential to reveal aspects of the cognitive mechanisms underlying pragmatic enrichment more generally. Although these two inferences need not share underlying cognitive mechanics \textit{a priori}, they share essential properties which suggest a common explanatory basis. In particular, they both appear to be derived from regularities in language use concerning how much or how little information is conveyed during a typical exchange.

Scalar implicature is derivable from the fact that speakers generally provide as much information as they can in a cooperative linguistic exchange. When my friend tells me, "I ate some of the cookies," I don't typically end with the literal conclusion that they ate more than none. I reason that if they had eaten all of the cookies, they would have said so. Since they didn't, I infer that they ate some, \textit{but not all}, of the cookies.

Conversely, contrastive inference is derivable from speakers' tendency to provide as much information as necessary \textit{and no more}. If my friend asks me to "Pass the white ramekin," even if there are many white objects nearby and I have never heard the word "ramekin" before, I can search for two items differing in color, pick the white one, and likely have successfully executed the request. Without
needing to distinguish between items in the environment which are of the same type but different colors, it would have been redundant for my friend to use the modifier "white." Given such a pair nearby, I pragmatically infer that they want the white one, without even needing to hear the noun.

The fact that scalar implicature and contrastive inference both closely parallel principles of optimal informativeness suggests but does not necessitate that they are computed in similar or analogous ways. In this dissertation, I set out to investigate the cognitive resources underlying these two inferences in the hope that we may find regularities in the mental implementation of informativity-based pragmatic processes.

Luckily, we are not starting from scratch. A recurring finding in investigations of pragmatic inferences is that they tend to be time-consuming---pragmatically enriched meanings tend to take longer to reach than semantic meanings---and effortful---where semantic meaning is generally not interrupted by cognitive load, pragmatic processing often is. Online studies of scalar implicature, for example, show that the literal or lower-bounded meaning of "some" (some and possibly all) is almost instantaneous, while the pragmatic or upper-bounded meaning (some but not all) takes time to reach (Bott & Noveck, 2004; Huang & Snedeker, 2009a). Meanwhile, pragmatically-enriched upper-bounded meanings of "some," for example, are less frequent under cognitive load than they would be otherwise (Marty, et al., 2013).

This data pattern paints a picture of language comprehension in which semantics is fast and easy while pragmatics is slow and effortful. But unsurprisingly, the more data we accumulate, the more this appears to be an oversimplification. In this dissertation I contribute two key discoveries which confirm that this comfortable dichotomy, while often accurate, is not universal.
In Chapter 2, I present a study which attempts to flesh out the effects of working memory load on scalar implicature by embedding a dual task study in a visual world eye-tracking paradigm. By exploring the timecourse of the impairment of cognitive load on upper-bound interpretations of "some" discovered by Marty, et al. (2013), I sought a finer-grained picture of how working memory depletion impacts pragmatic processing. Instead, I ended up discovering that cognitive load impacts semantic processing, too: In particular, I found the interpretation of quantifiers like "all" which involve no pragmatic enrichment to be impaired under high degrees of working memory load. Already, the clean dichotomy between "easy" semantics and "hard" pragmatics has begun to fall apart: Not all semantic processes are easy, and cognitive load does not only affect pragmatic processes, as many dual-task studies have implicitly or explicitly concluded (De Neys & Schaeken, 2007; Dieussaert, et al., 2011; Marty, et al., 2013).

In Chapter 3, I set out to investigate the processing of contrastive inference, a lesser-known pragmatic process which is both less regular and more context-dependent than scalar implicature, but nonetheless appears to be based on the same principles of optimal informativeness in communication. Despite its greater context-dependence and ad-hoc nature, contrastive inference is generally computed faster than scalar implicature. But, like scalar implicature, its sensitivity to speakers' conformity to optimal informativeness (Grodner & Sedivy, 2011) suggests that to two are computed in similar ways. In particular, contrastive inference could be explained by listeners reasoning about interlocutors as rational agents who choose among potential utterances in order to maximize informativity while avoiding inefficient redundancy. In Chapter 3 I used contrastive inference to test the predictions of the easy-semantics vs. hard-pragmatics dichotomy by observing whether, as such a story predicts, it was impaired under working memory load. The results, however, were inconclusive and led me to a less-ambiguous test of the mechanism underlying contrastive inference.
In Chapter 4, I present an experiment to more thoroughly differentiate between the remaining possibilities: If contrastive inference is effortful as predicted by the simple dichotomy, it should be impaired under simultaneous cognitive load and deviation from norms of optimally informative communication. On the other hand, if contrastive inference is automatic and difficult to override, counter to all previous accounts, then cognitive load should in fact prevent us from inhibiting it when it is no longer helpful in assisting with comprehension (as when a speaker deviates from informativity norms). Surprisingly, I confirmed the predictions of the latter account, again challenging the simple view of semantics-as-easy and pragmatics-as-hard.

Finally, in Chapter 5, I synthesize the results of the three studies presented in the dissertation, concluding that while semantic processes are often easier than pragmatic ones, either type of process can be implemented as a deliberate, complex process involving the integration of information from multiple sources, or as a simple, default-like behavior which is difficult to interrupt.

Before I go on, a note about context. Almost anything could be considered part of an utterance's context: the weather that day, the astrological sign of the speaker, the number and type of referents in the room, what has been said before the utterance in question. Many of these things may turn out to have systematic and predictable effects on language processing, but for present purposes, I limit myself to consideration of just the most concrete nonlinguistic aspects of context which have been shown to affect processing. This includes number and type of referents in the immediate environment (e.g., on the table between interlocutors, on the computer screen a listener is looking at), and the behavior of the speaker (e.g., whether they conform to conversational norms). In some cases, the environment is manipulated so that the context can be controlled tightly, for example, putting participants in an experimental setting where they are looking at a computer screen with four objects on it while listening to one sentence at a time. It is my hope that the regularities we observe in these controlled settings will
extend into less controlled ones, and that eventually we will be able to understand more complex interactions that take place in noisy, uncontrolled environments with many more contextual attributes to take into consideration.
Chapter 2: Effects of Working Memory Load on the Processing of Numerals and Scalar Quantifiers

Introduction

Language users routinely make inferences beyond the literal meanings of utterances they hear. For example, although (2.1a) is logically true whenever I did all of my homework, you're likely to infer (2.1b) even if I don't say it out loud:

(2.1)  
   a. "I did some of my homework."
   
   b. I did not do all of my homework

Grice (1975) noted that listeners typically infer from the use of a weak utterance (e.g. "I did some of my homework") that a related stronger one (e.g. "I did all of my homework") must not hold. Calling this implicature, he attributed it to the fact that interlocutors are generally cooperative. In particular, speakers try to be optimally informative, i.e., to use the most informative utterance they can (lest the communication fail) but no more (lest time or energy be wasted). Listeners assuming their interlocutor is cooperative in this way thus infer from the speaker's choice of a weak utterance that they were not in a position to use a stronger one, probably because it is not true.

Horn (1972) refined the analysis of implicatures like (1.1) by noting that the alternative utterances could be derived by replacing the weak term ("some") with another term that falls alongside it in an ordered scale:

(2.2) <some, all>
Thus scalar implicature can be viewed as the procedure of identifying a weak utterance, retrieving stronger scalemates of the weak term in question, and negating the set of sentences that results from the replacement of the weak term with a stronger one. Thus "I did some of my homework" implicates that it is not the case that "I did all of my homework." Lexical scales similar to (2.2) have been noted across many syntactic categories, including verbs (<might, must>), adjectives (<warm, hot>), and logical connectives (<or, and>). While scalar implicature is quite frequent in adult language use, it doesn't happen one hundred per cent of the time: Studies have reported nontrivial rates of logical interpretations of weak utterances such as (2.1a) ("I did some and possibly all of my homework") in addition to pragmatic interpretations featuring the implicature.

In sentence processing, there is evidence that language comprehenders begin with the logical interpretation when faced with weak scalar terms, and only with time and effort reach the pragmatic one. For example, Huang & Snedeker (2009a) found that the eye gaze of participants faced with two scenes, one compatible with both interpretations of "some" and another compatible only the pragmatic one, tended to equivocate between the two for about 800ms after hearing the quantifier until converging on the pragmatic one. Other research has largely confirmed this pattern (Bott & Noveck, 2004; Huang & Snedeker, 2009b; Huang & Snedeker, 2011; Bott, et al., 2012; but see Breheny, et al., 2006; Grodner, et al., 2010). This suggests that listeners go through a logical phase of interpretation lasting a little less than a second before completing a scalar implicature and finally arriving at the final, pragmatic, interpretation.

Scalar Implicature Online

Huang & Snedeker (2009a) investigated the time course of scalar implicature using the visual world paradigm (Tanenhaus, et al., 1995), an experimental setup in which listeners' visual fixations within a
A constrained referential environment are used to infer aspects of their linguistic interpretation as they listen to a sentence. Participants were shown displays featuring a a girl with a proper subset of items (e.g. a girl with some but not all of the socks on the screen) and another girl with a total set of items (e.g. a girl with all of the soccer balls on the screen) (Figure 2.1). Their eye movements were monitored while they were verbally instructed to "Point to the girl that has some/all of the socks/soccer balls." Crucially, when they feature the quantifier "some," the instructions are logically ambiguous until the second syllable of the noun phrase (e.g. "Point to the girl that has some of the soc-") could refer to either girl since both have some objects starting with "soc-". But if a scalar implicature is computed from "some," only the girl with some of the socks is a plausible target. Huang & Snedeker (2009a) found that participants looked equally frequently at both girls for about 800ms after hearing the quantifier "some," suggesting a period of logical interpretation preceding the computation of scalar implicature. Meanwhile, participants instructed to "Point to the girl that has all of the soccer balls" looked at the girl with the total set almost immediately after hearing the quantifier.

Figure 2.1. A display from Huang & Snedeker 2009a.
Other studies have revealed similar patterns (Rips, 1975; Bott & Noveck, 2004; Feeney, et al., 2004; Panizza, et al., 2009; De Neys & Schaeken, 2007; Huang & Snedeker, 2011; Huang & Gordon, 2011; Bott, et al., 2012). For example, Bott & Noveck (2004) found that participants who spontaneously interpreted "some" with an upper bound ("some but not all") took longer than those who spontaneously interpreted it with no upper bound. (But see Breheny, et al., 2006, and Grodner, et al., 2010, for some evidence of fast scalar implicatures.)

Furthermore, offline studies, i.e., studies whose measures are not time-locked to behavior, have suggested that when under cognitive load, e.g., memorizing dot-patterns or letters, or under time pressure, people are less likely to compute scalar implicatures (Rips, 1975; Bott & Noveck, 2004; Feeney, et al., 2004; De Neys & Schaeken, 2007; Dieussaert, et al., 2011; Marty, et al., 2013; Marty & Chemla, 2013). When tasked with remembering letter sequences or visual dot-arrays, participants are more likely to judge sentences true or false based on a logical interpretation than on an interpretation with a scalar implicature (Marty, et al., 2013). Why is this? Does summoning the alternative utterances the speaker could have said ("I did all of my homework") and maintaining them in activation require working memory? Does modeling the speaker's mental state require WM resources? A common hypothesis is that the activation and maintenance of alternatives in working memory is crucial for scalar implicature computation, so depletion of working memory is responsible for the observed deficits (Marty, et al., 2013). The studies presented here utilize the visual world paradigm setup from Huang and Snedeker (2009a) to observe participants in the process of enriching the logical interpretation of "some." By putting them under differing degrees of working memory load, we can test for differences in the timing of their visual fixations that may reveal when and how working memory load interferes with scalar implicature. This, in turn, will help narrow the possible source of the effect.
by providing the first time-course information about the interpretation of scalar and numeral quantifiers under working memory load.

Scalar Implicature and Working Memory

A handful of studies indicate that working memory plays a role in listeners deriving their final, enriched interpretation of weak scalar terms (Dieussaert, et al., 2011; De Neys & Schaeken, 2007; Marty, et al., 2013). For example, Marty and colleagues (Marty, et al., 2013; Marty & Chemla, 2013) found that when made to memorize letter sequences in conjunction with a simple scalar implicature task, fewer implicatures were calculated the longer the letter strings were. But the dependent measure used in this and other working memory studies was offline acceptability rating, i.e., and end-of-the-day measure which does not reveal time-course. While suggesting that some aspect(s) of pragmatically evaluating a sentence like (1a) require working memory resources, the results therefore tell us little about the exact role of working memory in the cognitive processes underlying scalar implicature. The studies in this chapter ask specifically about the temporal nature of the role of working memory in online processing: When and how does working memory depletion affect the derivation of pragmatic interpretations of scalar quantifiers?

An early clue that depletion of cognitive resources interferes with scalar implicatures was noted by Bott & Noveck (2004). They found that when participants were either explicitly instructed to interpret "some" as "some but not all" (i.e., made to compute a scalar implicature) or spontaneously did so took longer than when they either spontaneously interpreted or were instructed to interpret "some" as "some and possibly all" (~3.3s vs. ~2.7s in a truth-value judgment task). They also found that participants forced to perform a truth value judgment within a shorter time window (900ms) responded with fewer "false" judgments to sentences like "Some elephants are mammals" (28%) than participants given
longer to respond (3s, 44%). Since time pressure reduces the cognitive resources available to perform a task, the fact that judgments based on scalar implicatures take longer than non-scalar-implicature judgments suggests that scalar implicatures are cognitively demanding.

Using a more direct manipulation of working memory, De Neys & Schaeken (2007) found that participants who memorized more complex visual dot-arrays judged sentences like "Some elephants are mammals" to be false (reflecting scalar implicature) more often (73.2%) than participants who memorized less complex ones (78.9%; load between-subjects). Dieussaert, et al. (2011) found the same effect, but only among participants in the bottom tertile of working memory capacity. Participants with low working memory capacity, as measured by a group version of the Operation Span Task (La Pointe & Engle, 1990) translated in to Dutch (GOSPAN; De Neys, d’Ydewalle, Schaeken, & Vohs, 2002), judged sentences like "Some elephants are mammals" to be false (reflecting scalar implicature) less often when concurrently memorizing more complex visual dot-arrays (68%) than when memorizing simpler ones (78%).

Scalar implicature thus appears costly in terms of time and cognitive resources. But all the studies above depleted working memory either indirectly by limiting time, or by asking participants to complete a visual task. Marty, et al (2013) and Marty & Chemla (2013) extended these results in two ways. First, they observed the same reduction in scalar implicature frequency with a concurrent verbal working memory task. Therefore, working memory depletion in general, not just visuo-spatial depletion, reduces scalar implicature frequency. In a dual-task experiment participants were required to memorize and then reverse short (two letters) or long (four letters) letter sequences while rating the appropriateness of a sentence like "some of the dots are red" as a description of a display full of red dots. Participants under high working memory load were less likely to rate the descriptions as
inappropriate than participants under low load (81% vs. 89%), suggesting they interpreted the quantifier without an upper bound.

Second, Marty, et al. (2013) extended the class of quantifiers studied to include cardinal quantifiers such as "two" and "three" in addition to non-cardinal (proportional) quantifiers like "some" and "all" previously studied. Strikingly, in the same dual-task paradigm involving a graded sentence-picture matching task under high and low working memory load, they found that participants under high load were less, not more, likely to judge descriptions such as "three of the dots are red" as appropriate when they all are red, as compared with participants under low load (38% vs. 54%). If the pattern observed with "some" and "all" suggests that scalar implicature requires working memory resources to compute, then this pattern suggests the basic interpretation of numerals is upper-bounded ("exactly three") and that it takes some effort to reach a non-upper-bounded interpretation ("three or more"). Are numerals somehow different from "some" and "all" in that they don't require scalar implicature to become upper-bounded, or that they require some cognitively effortful process to remove an inherent lower bound? In Experiment 2, I follow up on this question by examining the time-course of this effect of load on numerals in an eye-tracking task. First, I review the unexpected results of the experiment designed to investigate the effects of working memory load on scalar implicature from "some."

**Experiment 1**

In Experiment 1, we investigated the effects of working memory load on the time-course of scalar implicature from the weak scalar term "some." By varying the difficulty of a memory task between participants (no load, low load, and high load) and then comparing participants’ eye gaze to objects compatible or incompatible with the upper bound after hearing the quantifier, we tested whether cognitive load had an effect on the speed of upper bound computation.
We presented participants with short stories while they looked at displays featuring two sets of objects with overlapping names (e.g. "birthday cakes" and "birthday cards") distributed to two sets of boys and girls: While one set of four objects was evenly split between two characters, the other set of three was always given to one character:

![Example of a display featuring object distributions.](image)

Figure 2.2: Example of a display featuring object distributions.

After hearing a story, participants' eye movements were monitored as they listened to an instruction sentence featuring either the weak scalar term "some" (e.g. "Point to the girl that has some of the birthday cakes") or the strong scalar term "all" (e.g. "Point to the girl that has some of the birthday cards"). In "all" trials, we expected looks to the target quadrant (bottom left) to rise significantly above chance immediately after participants heard the quantifier. Critically, in “some” trials, the information after the quantifier (“Point to the girl that has some of the birthday…” and before the final syllables of the target noun phrase (“cakes”), is referentially ambiguous between the girl with the proper subset of the set of four items (e.g. the girl with exactly two birthday cakes) and the girl with the total set of three items (e.g. three birthday cards). Since the quantifier is semantically compatible with either the total set or the proper subset of items, participants' looks should be evenly distributed between the two female characters during this period as long as they are entertaining a lower-bounded interpretation of the
quantifier. Once an upper bound is computed, participants should tend to look significantly more at the female character with the proper subset of items. Thus, by observing the timing of the convergence of gaze on the target character in "some" trials, we can observe the timing of upper bound computation, i.e., scalar implicature. Then, by comparing upper bound computation across the three memory load conditions, we can learn whether working memory load delays scalar implicature.

Methods

Participants

Ninety-four undergraduate students enrolled at Harvard University and adults from the Cambridge, Massachusetts, area participated in this study. They received either course credit or US $10 for their participation. All participants were native English speakers with normal or corrected-to-normal vision who reported no delays in language development. Fourteen additional participants were excluded due to non-native English speaker status (12 participants) or developmental delay (2 participants).

Procedure

Participants were seated in front of a Tobii T60 remote eye-tracker outfitted with a touchscreen and connected to a USB keyboard via a laptop. Each session began with instructions followed by three practice trials.

Participants were randomly assigned to one of three working memory load conditions: No load (24 participants), low load (42 participants), or high load (28 participants). In the no-load condition, all trials began with a one-second white screen followed by a display featuring four characters: Two boys on the left and two girls on the right. The characters were shown on a white background, separated by black lines into quadrants (Figure 2.3).
As soon as the characters appeared, participants heard a pre-recorded story about them (for example, attending a birthday party). Then, a set of items mentioned in the story appeared simultaneously alongside the characters in the top two quadrants (Figure 2.4). (For example, two birthday cakes appeared to Mike’s right and two birthday cakes to Julie’s left.) As the objects appeared, participants heard a sentence describing their distribution. Next, an analogous sentence described the distribution of a different set of objects (e.g., birthday cakes) as they appeared alongside one of the two bottom characters.

[Four characters appear] “The boys and girls were going to a birthday party. Julie and Mike remembered to bring birthday cakes. [Four birthday cakes appear on the top of the screen, two each in Julie and Mike’s quadrants.] Sarah remembered to bring birthday cards, but Phil forgot to bring his. [Three birthday cards appear in Sarah’s quadrant on the lower right.]”
Once the object distribution was complete, participants were instructed to "point to" one of the characters using the touchscreen. Target characters were identified using definite descriptions indicating their gender and the quantity and type of objects they had (e.g. “Point to the girl that has some of the birthday cards”). Once the participant touched the screen, the next trial began after a brief delay.

Whereas previous studies used item pairs whose overlap was on average one syllable (e.g. Huang & Snedeker, 2009a), compound nouns and other polysyllabic noun phrases with at least one and sometimes two overlapping syllables (“birthday cakes”/“birthday cards”) were used in order to make the ambiguous window as long in duration as possible. (See Appendix A for full list of item pairs.)

For participants in the low load and high load conditions, the trial structure above was sandwiched between the two stages of a memory task (Figure 2.5). The first stage consisted of the presentation of a sequence of letters (two letters for low load and four for high), displayed one at a time for one second each with a one-second white screen in between. After character selection, a prompt to enter the letter sequence in reverse order appeared. After the participant entered the sequence, a feedback slide appeared, followed by a one-second white screen, and the next trial began.
Figure 2.5. Trial structure in low load and high load conditions. (Graphic: Qingqing Wu)

Materials

Trials in the low load condition featured letter strings of length two and those in the high condition strings of length four. To build these strings, eight letters were chosen from the English alphabet such that they would represent maximally phonologically dissimilar sounds (B, H, F, J, L, R, M and X). Thus, the number of letters more closely corresponds to the number of "slots" occupied in verbal working memory (Conrad and Hull, 1964; Baddeley, 1966; Salame & Baddeley, 1982). Letters were randomly permuted to create sixteen distinct two-letter strings and sixteen distinct four-letter strings. (See Appendix B for the exact strings used.) Each participant saw each letter string in their condition exactly once.
In all trials and conditions, the position of the four characters on the screen was the same: Mike and Julie on the top left and right, and Phil and Sarah on the bottom left and right, respectively. Thus the vertically adjacent characters always matched in gender and the horizontally adjacent characters never did. Sixteen different stories were constructed in which two types of objects were introduced and distributed to the boys and girls. The stories were always such that one pair of horizontally adjacent characters (e.g. Mike and Julie) split a set of four objects of a given type equally (two objects each), and one of the remaining characters (e.g. Phil) received all three objects of the other type, leaving the final character (e.g. Sarah) with nothing.

The background story helped to provide a context within which the quantifiers we used (“some” and “all”) would be interpreted with respect to the particular finite sets of objects on the screen. Stories referred to the objects either using bare noun phrases (“birthday cards”) or possessively modified noun phrases (“her birthday cakes”); quantifiers were carefully avoided so as not to prime participants to associate particular object sets with particular quantifiers. All audio stimuli were recorded by a female native English speaker instructed to place prosodic stress on the noun phrase to avoid contrastive emphasis on the quantifier.
Table 2.1. Set of four instruction sentences for an item pair with a female target.

Four lists of 16 story-command-display trios were created in order to counterbalance for three factors: First, which object type (e.g., birthday cakes or birthday cards) the requested character had, second, the strength of the quantifier used to make the request (the weak scalar term “some” or the strong scalar term “all”), and third, the location of the character requested. Each list of 16 critical trials contained exactly eight trials featuring a weak scalar term and eight featuring a strong one, and each item pair appeared exactly once. A given object type of the two featured in each item (e.g., birthday cakes from the item featuring birthday cakes and birthday cards) was requested in exactly half of the lists, and each object type was paired with a weak scalar term in half of the lists and a strong one in the other half. For each item pair (e.g. birthday cards and birthday cakes), a set of four instruction sentences, to be paired with one of two visual displays, was thus recorded (Table 2.1). Half the item pairs had female targets and half had male targets.

Instruction sentences were divided into five regions: The region from the onset of the first phoneme to the onset of the gender word (sentence start), the region from the onset of the gender word to the onset of the quantifier (gender), the region from the onset of the quantifier to the onset of the noun phrase (quantifier), the region from the onset of the noun phrase to the disambiguation of the noun phrase (noun start), and finally the region from the onset of disambiguation to the offset of the final syllable of
the sentence. The ambiguous period---that is the period during which the listener has heard the quantifier but cannot yet tell which object will be mentioned, quantifier plus noun start regions---during which we expected to see evidence of upper bounding from "some" lasted 928ms on average.

Results

The gender of the character requested in the pointing instructions was predictable based on the distribution of objects during the introductory story: If no objects were given to one of the boys, then a girl was requested, and vice versa. It is thus possible that participants were able to predict which vertically-adjacent pair of characters was likely to contain the target character and focus their looks there from the moment the distribution became apparent. In order to correct for this we used as our dependent measure the ratio of looking time to the target character to the combined looking time to the target character and its vertically-adjacent counterpart. Since this measure is independent of any preference for one gender (i.e. lateral side of the screen) over the other, it only reflects the disambiguation of reference between the two characters of the target gender.

For each participant and trial, eye gaze location during the instruction sentence was recorded at 60Hz by a Tobii T60 remote eyetracker. Samples in which either the eyetracker lost track of one or both eyes, or in which neither eye had a score in the top half of the eyetracker's validity range, were eliminated (22.1% of samples). Trials with fewer than half of the expected number of samples remaining during the critical region (the time from quantifier onset to the end of noun phrase ambiguity) were eliminated (79 trials: 24 in the No Load condition, 56 in the low load condition, and 54 in the High Load condition). Participants who did not have 4 or more trials remaining in both "some" and "all" conditions were excluded from analysis (3 participants, 2 in the high load condition and one in the low load condition).
Samples gathered during the instruction sentence were sorted into the five regions detailed above. For each trial and each region, we then derived a continuous measure of target preference consisting of the ratio of samples in which the participant was looking at the target (e.g. for a trial whose command was "Point to the girl that has some of the socks," the girl with two of four socks) to the total number of samples in which they were looking at the target or the distractor (the other girl). This ratio indicates the degree to which a participant is visually fixated on the target: If their only looks to characters that match the gender of the command are to the target, it will be equal to one. If all looks to gender-matching characters are to the distractor, it will be zero. If there were no looks to gender-matched characters (either the target or the distractor), that participant's looks during that region of that trial were excluded from analysis.

Figure 2.6: Mean target Preference by region across participants and items

Even before the sentences diverged in content (e.g., during "Point to the ..."), participants displayed a preference to look at the character with the proper subset of items (Figure 2.6). Since direct
comparisons of mean target preference in a given region would thus fail to distinguish between this perceptual bias and quantifier-induced target preference, we investigated interpretation of "some" and "all" by examining the increase in target preference induced by the quantifier rather than the average target preference in a given region. If a quantifier strongly cues the listener to the identity of the target object, target preference should increase substantially, i.e., the slope of the target preference line should be steep. If a quantifier only weakly cues the listener to the target, or does not cue at all, target preference should not increase much, i.e., the slope should be flat(ter).

We implemented such a slope-based analysis by restricting analysis to samples obtained during the gender and noun start regions (the two regions directly adjacent to the quantifier), and adding region (pre-quantifier vs. post-quantifier) to our models. A significant effect of region on target preference would indicate successful use of the quantifier for target identification. An interaction between region and quantifier strength, on the other hand, would indicate a difference in how useful the quantifiers were.

For the purposes of our analyses, the onset of each sentence region was shifted ahead 200ms to account for the time it takes to program a saccadic eye movement (Allopenna, et al., 1998; Matin, et al., 1993; Hallett, 1986). For each participant and each trial, we converted the eye gaze samples during each of the two critical regions, pre-quantifier (gender) and post-quantifier (noun start), which indicated which quadrant of the screen the participant was looking at (e.g., target, competitor, adjacent distractor, non-adjacent distractor) at a given moment into a measure of target preference. For a given region, target preference was computed in two steps. First, we calculated the ratio of target looks (samples in which the participant was looking at the target quadrant) to combined target-and-distractor looks (samples in which they were looking either at the target or the gender-matched distractor). Trials in which participants were looking at neither the target nor the distractor in one or both regions were excluded.
from analysis. Since eye movements are characterized by brief, singular fixations broken up by
saccades or quick jerks of the eyes, the distribution of the resulting difference score across participants
was bimodal, with peaks at one and zero. We therefore binarized each difference score in order to
perform a logistic analysis: Ratios greater than 0.5 were coded as 1 (target preference), differences less
than 0.5 were coded as -1 (distractor preference), and differences of exactly 0.5 were excluded from
analysis.

Using the glmer function from the lme4 package in the R statistical programming language (Bates,
Maechler, Bolker, & Walker, 2014), we built logistic mixed effects models with region (pre- vs. post-
quantifier) and quantifier strength ("some" vs. "all") and either load presence (high load vs. no load) or
load magnitude (high load vs. low load) as fixed effects and participant and item as random effects. We
planned three sets of analyses: First, among no-load participants, a test for differences between "some"
and "all" with regard to pre- to post-quantifier target preference increase, which would indicate
replication of previous observations of scalar implicature (e.g., Huang & Snedeker, 2009a); Second,
across both quantifiers, a test for effects of load presence on pre- to post-quantifier target preference
increase; and third, across both quantifiers, a test for effects of load magnitude on pre- to post-
quantifier target preference increase.

Evidence of Scalar Implicature

First, we tested for a replication of the findings of Huang & Snedeker (2009a) that the strong quantifier
"all" triggers rapid target identification while participants who hear "some" remain ambivalent between
the two characters compatible with a "some (and possibly all)" interpretation several hundred
millseconds before computing an upper bound and eventually fixating on the character with some and
not all of the target objects. A logistic mixed-effects model of no-load data with region and quantifier
strength as fixed effects and participant and item as random effects revealed main effects of both region ($\beta = 0.51$, SE = 0.09, $z = 5.80$, $p < 0.005$) and quantifier type ($\beta = -0.25$, SE = 0.06, $z = -4.0$, $p < 0.001$) such that participants looked more at the target in the post-quantifier region than the pre-quantifier region, and looked more at the target when it was a subset character. There was a significant interaction between region and quantifier strength ($\beta = 0.26$, SE = 0.09, $z = 2.91$, $p = 0.004$), but critically, there was no main effect of region among the "some" trials alone ($\beta = -0.03$, SE = 0.23, $z = -0.13$, $p = 0.90$). In other words, there was no significant increase in target preference after hearing the quantifier "some," indicating that listeners did not use it to identify the target (Figure 2.7). This suggests that participants did not compute upper bounds before disambiguation, i.e., did not perform scalar implicatures. Meanwhile, there was a significant effect of region on target preference among "all" data alone ($\beta = 0.74$, SE = 0.22, $z = 3.37$, $p = 0.0007$), confirming the effectiveness of the paradigm (Figure 2.7).
Effects of Working Memory Load

We performed two tests for effects of working memory load on target preference. First, we examined whether the presence of load had an effect on the increase in target looks after hearing the quantifier by comparing no-load participants' performance with high-load participants' performance. (Low-load participants were excluded from this analysis.) A logistic mixed effects model with load presence, quantifier strength, and region as fixed effects revealed main effects of both region ($\beta = 0.47, SE = 0.95, z = 4.92, p < 0.001$) and strength ($\beta = -0.27, SE = 0.07, z = -3.97, p < 0.001$) such that participants' target preference was substantially higher after the quantifier, and they preferred to look at the character with the proper subset of objects. While an interaction between region and quantifier...
strength was significant ($\beta = 0.29, \text{SE} = 0.09, z = 3.06, p = 0.002$), there was no main effect of load presence ($\beta = 0.06, \text{SE} = 0.08, z = 0.60, p = 0.56$) nor any other interactions (all $|\beta|$s $<$ 0.04 and all $p$s $>$ 0.2).

Figure 2.8: Target preference during pre- and post-quantifier regions by load presence and quantifier ("some" and "all")

Second, we examined whether, among participants under some degree of load, the magnitude of that load had an effect on increase in target looks after hearing the quantifier. To do this, we compared low-load to high-load participants (no-load participants were excluded). A logistic mixed-effects model with load magnitude, quantifier strength, and region as fixed effects revealed main effects of both region ($\beta$
such that participants again looked at the target substantially more after hearing the quantifier, and again preferred to look at the character with the proper subset of items even before hearing the quantifier. There was no main effect of load magnitude ($\beta = 0.02$, SE = 0.09, $z = 0.24$, $p = 0.81$), nor an interaction between region and magnitude ($\beta = -0.11$, SE = 0.11, $z = -1.03$, $p = 0.31$). There was, however, a significant interaction between quantifier type and region ($\beta = 0.26$, SE = 0.09, $z = 2.92$, $p = 0.004$), and a marginal three-way interaction between quantifier type, load magnitude, and region ($\beta = -0.19$, SE = 0.11, $z = -1.79$, $p = 0.07$) such that high memory load negatively impacted post-quantifier increase in target looks to a greater degree in "all" trials than in "some" trials. In other words, "all" trials showed a greater decrease in slope from low to high load than "some" trials (Figure 2.9).
Discussion

We measured the difference in participants' visual target preference before and after hearing the quantifiers "some" and "all" under three degrees of working memory load. We predicted that since interpretation of "all" does not require pragmatic enrichment, target preference should increase rapidly, and the increase should be robust under working memory load. On the other hand, since interpretation of "some" requires upper-bound computation, participants' target preference after hearing "some"
should increase more slowly. Further, if working memory load selectively impairs pragmatic upper bound computation as suggested by Marty, et al. (2013), the increase in target preference after "some" should be weakened by load.

Two notable findings emerged. First, we found no evidence that participants were computing upper bounds from "some" in our task. In other words, we did not observe scalar implicature. This is apparent in the lack of a significant difference in pre- and post-quantifier target preference among no-load participants, who acted as a control for our working memory manipulation. In the absence of any evidence of upper-bounding, we are unable to detect modulatory effects of working memory load on scalar implicature computation. Second, we found that when participants were under working memory load, the load's magnitude inversely affected the interpretation of the strong scalar term "all." In other words, among those under load, participants under high load were less able to use "all" to disambiguate the target than participants under low load. I will address these two findings in turn.

Why weren't upper bounds computed? A plausible explanation for the larger pattern of divergence in upper-bound computation speeds among scalar implicature studies in general, namely that participants exposed to a variety of labels for subsets and total sets take longer than those exposed to a single labeling schema, does not apply here. While participants exposed to more than one label each for subsets and total sets (e.g., "some" and "two" for subsets, "all" and "three" for total sets) take longer to compute upper bounds on "some" than participants exposed solely to "some" and "all" (Huang & Snedeker, 2018), our study included these two quantifiers alone with the specific aim of increasing the speed of upper-bound computation. Furthermore, participants in our study had an on-average 928ms-long region during which information from the quantifier "some" was available prior to noun phrase disambiguation, substantially longer than the ~200ms within which participants have been observed to compute upper bounds in some single-labeling-schema studies (Grodner, et al., 2010).
There is, however, great variation in upper-bound computing times across studies of scalar implicature in different modalities. For example, reading time studies have found evidence of upper-bounding 1800-2400ms post-quantifier (Breheny, et al., 2006; Bergen & Grodner, 2012), and ERP studies 1300-1700ms post-quantifier (Nieuwland, et al., 2010). When researchers shorten the time available for participants to compute upper bounds, they fail to find evidence of such computation even within 900ms of the quantifier (Hartshorne & Snedeker, 2014; Hartshorne, Azar, Snedeker, & Kim, 2015).

Upper bounding is this a process which varies substantially across tasks and experimental settings. Could upper-bounding also vary from person to person? Aspects of cognition such as working memory capacity and executive function display tremendous diversity, as do components of language processing (Kidd, et al., 2018). Most eye-tracking studies of scalar implicature cited in this chapter were conducted using populations consisting entirely of college students at high-ranking universities, whose cognitive and linguistic ability is likely to be (1) above average and (2) low-variance. The current experiment, however, was conducted using a diverse population of community members recruited online from all over the Cambridge, MA, region in addition to college students, who constituted only 31 of the 92 people we tested. Thus our participants likely (1) had lower average cognitive and linguistic ability than participants in other studies, meaning that it takes them longer, on average, to compute a scalar implicature, and (2) varied more widely in their abilities, adding noise to our data and reducing the likelihood that we would observe upper-bound computations should they be there.

While we were not able to examine the online effects of working memory load on upper-bound computation from "some," we did find that the interpretation of "all," which does not involve pragmatic enrichment, was impaired under high load. Where a simple account of semantic and pragmatic processing on which the former is easy (fast, cognitively undemanding) and the later difficult (slow,
cognitively demanding) predicted memory load to selectively impair pragmatic processes, here we find that semantic processing is impaired, too. I will discuss this finding, which challenges the view of semantic and pragmatic processing as two uniform categories, and its implications further in the General Discussion.

Crucially, while the effect of load on "all" was significant, it was not part of our hypothesis space and lacks corroborating evidence in other research. Before concluding that "all" is effortful to interpret, this pattern should be replicated, ideally in a number of paradigms. One data point alone is suggestive, but not proof, of the claim that "all" (and perhaps semantic interpretation in general) is demanding to process. Future work should investigate quantifier processing with an eye toward establishing firmer ground for the claim that "all" is difficult to process. The fact that load magnitude but not load presence impaired processing also requires further investigation. Until then, I adopt the tentative claim that "all" is cognitively demanding to interpret, and constitutes a counterexample to the simple generalization that semantics is easy.

In Experiment 1, we found an effect of working memory load on tasks involving interpreting quantifiers to identify reference, even when quantifier interpretation involves no scalar implicature. Are numerals somehow different from "some" and "all" in that they don't require scalar implicature to become upper-bounded, or that they require some cognitively effortful process to remove an inherent lower bound? In Experiment 2, I examine the time-course of this effect of load on numerals.

Experiment 2

In Experiment 1 we learned that even scalar quantifiers which require no pragmatic enrichment like "all" are nevertheless more difficult to interpret while under working memory load. Do numerals'
upper-bounded meanings, which arise just as quickly as the meaning of "all" (Huang & Snedeker, 2009a), also require working memory to compute?

Marty, et al. (2013) found that upper-bounded readings of the weak scalar item "some" were less frequent under working memory load, suggesting that upper-bounding is a working-memory-intensive process. Since numerals have both upper- and lower-bounded interpretations which parallel those of "some" in many (but not all) ways, we might predict that upper-bounded readings of numerals should be less frequent under load. Surprisingly, Marty, et al. (2013) found just the opposite: Upper-bounded readings of numerals were actually more frequent under working memory load.

We thus have reasons to believe that upper-bounded interpretations of numerals could be easier or harder under working memory load. We investigated this question by observing the timecourse of the effect of working memory load on numeral upper-bounding. In an experiment exactly the same as Experiment 1, but with the numeral quantifiers "two" and "three" replacing the scalars "some" and "all," respectively, we observed participants' ability to incorporate information from quantifiers while under differing degrees of working memory load.

Methods

Participants

Ninety-one undergraduate students enrolled at Harvard University and adults from the Cambridge, Massachusetts, area participated in this study. They received either course credit or $10 for their participation. All participants were native English speakers with normal or corrected-to-normal vision who reported no delays in language development. Eighteen additional participants were excluded due to later-reported non-native English speaker status (16 participants) or developmental delay (2 participants).
Procedure

The procedure in Experiment 2 was identical to that of Experiment 1, except that instructions featured the cardinal quantifiers "two" and "three" in place of "some" and "all," respectively. (For example, "Point to the girl that has some of the socks" becomes "Point to the girl that has two of the socks," and "Point to the girl that has all of the soccer balls" becomes "Point to the girl that has three of the soccer balls.")

The working memory manipulation was carried out identically to that of Experiment 1. Thirty-four participants were assigned to the No Load condition, 30 to the Low Load condition, and 27 to the High Load condition.

Materials

The materials for Experiment 2 were exactly the same as those for Experiment 2 with the exception of the recorded audio instructions. All instructions across both experiments were recorded on the same day by the same speaker to minimize variation. A voice actor was instructed to read the cardinal sentences with prosody as identical to the non-cardinal ones as possible. All visual displays and letter strings were exactly as in Experiment 1.

Results

Eye gaze was sampled and filtered exactly as in Experiment 1. Samples in which the eyetracker lost track of one or both eyes, or for which the validity score was in the bottom half of the validity range for both eyes, were eliminated (21.1% of samples). Trials with fewer than half the expected number of samples remaining during the critical regions were eliminated (81 trials). Participants with fewer than 4
trials left in either quantifier condition, "two" or "three," were excluded from analysis (three participants, one in the high load condition and two in the no load condition).

Once again, we found a preference for participants to look at the character with the proper subset of objects, likely rooted in perceptual bias, suggesting as before that the increase in target preference after hearing the quantifier is a better measure of quantifier interpretation than comparison of mean target preference during a given time region.

A binary measure of target preference was derived in exactly the same manner as Experiment 1. It indicated whether the participant was looking more at the target or the distractor during a given region. (Regions in which the participant was looking at neither, or equally at both, were excluded from analysis.) We again constructed logistic mixed-effects models with region (pre- vs. post-quantifier), quantifier strength ("two" vs. "three"), and load presence (high load vs. no load) or load magnitude (high load vs. low load) as fixed effects and item and participant as random effects.

As in Experiment 1, we built logistic mixed effects models with region (pre- vs. post-quantifier) and quantifier strength ("two" vs. "three") and either load presence (high load vs. no load) or load magnitude (high load vs. low load) as fixed effects and participant and item as random effects. We planned two sets of analyses: First, across both quantifiers, a test for effects of load presence on pre- to post-quantifier target preference increase; and second, across both quantifiers, a test for effects of load magnitude on pre- to post-quantifier target preference increase.
Effects of Working Memory Load

We performed two tests of the effect of working memory load on target preference. First, we examined whether the presence of load had an effect on the increase in target looks after hearing the quantifier by comparing no-load participants' performance with high-load participants' performance. (Low-load participants were excluded from this analysis.) A logistic mixed effects model with load presence, quantifier type, and region as fixed effects revealed a marginal effect of load presence ($\beta = -0.14, SE = 0.07, z = -1.84, p = 0.07$) and a main effect of quantifier type ($\beta = -0.24, SE = 0.07, z = -3.42, p = 0.0006$) such that participants' target preference was lessened overall under load, and they preferred to look at the character with the proper subset of objects. No other effects interactions were found.

Figure 2.10: Target preference during pre- and post-quantifier regions by load presence and numeral ("two" and "three")
Second, we examined whether, among participants under some degree of load, the magnitude of that load had an effect on increase in target looks after hearing the numeral. To do this, we compared low-load to high-load participants (no-load participants were excluded). A logistic mixed-effects model with load magnitude, quantifier strength, and region as fixed effects revealed a main effect of quantifier type ($\beta = -0.20$, $SE = 0.06$, $z = -3.09$, $p < 0.002$) such that participants again preferred to look at the character with the proper subset of items even before hearing the numeral. There were no other effects or interactions.
Discussion

We measured the difference in participants’ visual target preference before and after hearing the numerals "two" and "three" under three degrees of working memory load. We predicted that since upper-bound interpretation of numerals is generally rapid and does not appear to be cognitively effortful, neither numeral effect on target identification should be impaired by working memory load. In fact, if the findings of Marty, et al. (2013) are reliable, lower-bounded interpretations of numerals should be cognitively effortful, and if anything, we would expect target identification to be easier under working memory load. In fact, we found neither effect, observing instead that while load impaired target identification across the board, it did not change how numerals were interpreted.

This is particularly notable in light of the findings from Experiment 1 that the purely semantic interpretation of the scalar quantifier "all" was impaired by working memory load. I argued that this finding alone cast doubt on the unitary treatment of semantic processing by suggesting that even interpretations which require no pragmatic enrichment at all apparently do require working memory resources, just as pragmatic enrichments do (Marty, et al., 2013). This could have been explained by a universal working-memory demand by all quantifiers which had simply gone previously undetected. But the lack of an effect of load on numerals in exactly the same paradigm indicates that not all quantifiers are alike in their resource-intensiveness.

While the results of Experiment 2 prima facie suggest that numeral upper bounds are not effortful to compute, lending support to theories in which numeral upper bounds are available very early and easily
in processing, a caveat is in order. Given that Experiment 1 was unable to test the effects of working memory load on upper bound computation in the visual world paradigm, we don't yet know how such effects would manifest. How much of a delay would our manipulation cause? Under what circumstances? What kinds of analyses are appropriate to detect them? In light of this uncertainty, the null findings of Experiment 2 must be interpreted with caution. While the effect on "all" to which the null results here were compared was found using the same paradigm, it only emerged in one of two analyses, casting doubt on its reliability. This in turn casts doubt on the validity of the comparison.

Furthermore, since we don't know exactly what we are looking for when we test for an effect of load on upper bound computation, the fact that our analyses revealed no effect of load does not mean that no such effect exists.

General Discussion

In two visual world eye-tracking experiments investigating the effects of working memory load on the timecourse of the interpretation of scalar quantifiers ("some" and "all") and numerals ("two" and "three"), we sought to test the predictions of the view suggested by Marty, et al. (2013) that semantics is easy and pragmatics is hard. Following up on results indicating that purely truth-conditional interpretation of quantifiers is always rapid while pragmatically enriched interpretations take extra time (e.g., Huang & Snedeker, 2009a), Marty, et al. (2013) found that semantic or lower-bounded readings of "some" were more frequent under high working memory load. Interpretations of "all," on the other hand, were unaffected. In online processing, we observed only lower-bounded interpretations of "some," precluding us from learning about the effects of load on upper-bounding of "some" in moment-to-moment processing. We did, however, find that participants under working memory load displayed online effects of impaired interpretation of "all," counter to the predictions of the semantics-is-easy, pragmatics-is-hard view. While there are reasons to hesitate before making strong conclusions from
these results, this is especially striking in light of the fact that load did not impair online interpretation of numerals, suggesting that not only are some semantic processes hard in the sense that Marty, et al. (2013) found pragmatic implicature to be hard, but that quantifiers also differ in how hard they are to interpret.

The fact that numeral interpretation was unaffected by working memory load---more precisely, that upper-bounded interpretations of numerals were reached just as rapidly under high working memory load as under no load at all---comports with a large body of data indicating divergence between numerals and scalar quantifiers with respect to how their upper bounds are computed. For example, upper-bounded readings of numerals tend to be accessed more quickly than upper-bounded readings of "some" (Huang & Snedeker, 2009a; Panizza, et al., 2009), and children tend to interpret numerals as upper-bounded well before they do so with scalars like "some" (Braine & Rumain, 1981; Noveck, 2001; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004, Hurewitz, et al., 2006; Pouscoulous, et al., 2007; Guasti, et al., 2005; Huang & Snedeker, 2009b; Katsos & Bishop, 2011).

While we did not find divergence between numeral and scalar upper-bounding per se, the fact that load did not impair numeral processing even while it seemed to impair that of "all" is consistent with divergence in the literature. While we cannot decisively interpret the null results of Experiment 2 since the nature of the effect of load on upper bounding in this paradigm is not yet clear, our findings contribute to this field of results by indicating, alongside Marty, et al. (2013), that upper-bounded interpretations of numerals are likely to be cognitively undemanding.

Conclusions

Deriving the meaning of a linguistic utterance involves first decoding its literal meaning---what would have to be the case for it to be true?---and then reasoning about its use by a particular speaker in a
particular context to figure out what exactly was intended. Research in many domains has indicated that the former component of linguistic interpretation, semantics, is both faster and easier than the latter type, pragmatics. Studies of scalar implicature in particular find listeners rapidly reaching the literal meaning of quantifiers like "some" and only after a few hundred milliseconds achieving the pragmatically enriched meaning that most speakers typically intend to convey. Moreover, investigations of comprehension under cognitive load have revealed that semantic interpretations can become more frequent when comprehenders' working memory is taxed, suggesting that semantics is not only faster, but also easier than pragmatics, which appears to actively involve working memory resources.

The findings, detailed in this chapter, that even purely semantic interpretations appear to actively recruit working memory resources pose a possible challenge to this simplistic view of linguistic comprehension. Yes, pragmatics can be hard, but so, it seems, can semantics. Moreover, it appears that all semantic processes recruit working memory to the same degree: One type of quantifier (numerals) can be immune to working memory depletion even as another (the strong scalar quantifier "all") is impaired by it.
Chapter 3: Effects of Working Memory Load on Contrastive Inference

Introduction

Understanding an utterance typically requires access to the context in which it was said. The referents of phrases like "my house" or "this table" depend on who produces them and what the environment contains. The fact that we use context to interpret language has been understood for centuries (see Parret, 1976, for reference to ancient grammarians; Reichenbach, 1947, for the earliest influential account in modern philosophy of language). Exactly how we do so is not obvious. What kinds of contextual information matter? Does context influence syntactic, semantic, and pragmatic processes in the same way? At what point during an utterance do we begin to use context in comprehension?

Take, for example, the case of someone interpreting the phrase "The wooden spoon" in the setting depicted in Figure 3.1:

Figure 3.1: Visual scene with two metal pots, a metal pan, a wooden cutting board, a wooden spool, and two spoons, one wooden and one metal.
Even before the noun phrase is complete, it is possible to derive information about which object in the utterance's environment it will refer to. Listeners can iteratively use the set of objects in the context to constrain hypotheses about the interpretation of the noun phrase based on partial utterances as in Table 3.1:

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;The...&quot;</td>
<td>Referent must be unique (not one of the pots)</td>
</tr>
<tr>
<td>&quot;The wooden...&quot;</td>
<td>Referent must be one of the wooden objects</td>
</tr>
<tr>
<td>&quot;The wooden spoo...&quot;</td>
<td>Referent must be the spool or the wooden spoon</td>
</tr>
<tr>
<td>&quot;The wooden spoon&quot;</td>
<td>The referent must be the wooden spoon</td>
</tr>
</tbody>
</table>

Table 3.1: Hypothetical stages of interpretation of "The wooden spoon"

Indeed, language processing proceeds incrementally at phrasal, lexical, and sub-lexical scales (Marslen-Wilson, 1987; Eberhard, et al., 1995). But the incorporation of context is more than simply ruling out referents incompatible with the literal meaning of the current speech increment. Relationships between referents and reasoning about the speaker's behavior and intentions can help listeners to narrow interpretation beyond literal meaning. For example, given that speakers generally use modifiers like "wooden" only when they are necessary to disambiguate between similar referents (e.g. a wooden spoon and a metal spoon), a listener might infer from the presence of such a modifier that the target object is one of a set of objects of the same type. Since "wooden" would be redundant in an utterance referring to the spool, after hearing "wooden" they might guess that the speaker is referring to the only wooden object which has a counterpart made of something different: The spoon.

The ability to anticipate the intended referent of an unfinished referential expression on the basis of a modifier when the referential environment features contrast has been observed in a number of
psycholinguistic studies (e.g., Eberhard, et al., 1995; Sedivy, et al, 1999; Sedivy, 2003; Heller, et al., 2008; Grodner & Sedivy, 2011). Listeners’ tendency to look more at the member of a contrast pair which is compatible with the incoming adjective (e.g. the wooden spoon) than a non-contrastive alternative (the spool) is even taken for granted in studies investigating other phenomena (Keysar, Barr, Balin & Brauner, 2000; Nadig & Sedivy, 2002; Brown-Schmidt, et al., 2008, i.a.). But how is contrastive inference computed? What mental representations and processes give rise to such early referent identification in contrastive contexts?

One salient possibility is that listeners compute Gricean quantity inferences (Grice, 1975; Horn, 1984) on the basis of beliefs about the speaker's intentions, the set of referents in the context, and the utterances the speaker could have made (Grodner & Sedivy, 2011; Sedivy, 2003). While incrementally analyzing speech, the listener could establish relationships between the environment and possible utterances (e.g. "it would be redundant to say the wooden spool since there's only one, but it would be helpful to say the wooden spoon since there are two"). Then, assuming the speaker intends to be helpful, the listener can narrow the set of possible referents. Indeed, the fact that referent anticipation is impaired when the person making the description is unreliable (Grodner & Sedivy, 2011) suggests that contrastive inference is computed this way.

On the other hand, listeners could reach narrowed interpretations entirely without the help of speaker representations. For instance, we may automatically conceptualize a scene like that in Figure 3.1, resulting in the activation of associated lexemes which can then be linked to incoming speech. Unprompted mental activation of nominal labels for everyday objects in noncommunicative contexts has been observed in both infants and adults (Zelinsky & Murphy, 2000; Mani & Plunkett, 2010). Since unambiguously committing objects to memory requires encoding not just object type but also contrast between objects of the same type, listeners would activate the label "wooden spoon" for the wooden
spoon, "silver spoon" for the silver one, "spool" for the lone spool, and so on. After hearing "wooden," they need only map the speech stream to their implicit labeling schema to narrow the set of possible referents to the wooden spoon. Because the demands of successful memory encoding are similar to the demands of unambiguous speech, this procedure permits listeners to link informativity (the presence of a modifier) to reference in a way that appears to involve speaker representations but in fact does not (cf. Barr, 2008; Barr, 2014).

Identifying and characterizing the cognitive representations and processes involved in pragmatic inferences is essential not only for interpreting the large and growing body of data in experimental pragmatics but also for constraining theories of linguistic competence. But the data concerning contrastive inference processing are inconclusive. In order to differentiate between the two hypotheses above, we tested for effects of working memory load on rates of contrastive inference. If maintaining and engaging speaker representations is required to achieve early contrastive interpretations of modifiers, participants under high working memory load should exhibit lower rates of contrastive inference compared to those under low load.

In this chapter, I review past experimental evidence for contrastive inference, noting that many researchers assume a speaker modeling account of the phenomenon in order to explain the negative effects of speaker unreliability on inference magnitude. After exploring alternative explanations consistent with the contrastive inference data, I then present an experiment in which participants were observed computing contrastive inferences under different degrees of verbal working memory load. I present the results, which reveal no reliable difference in contrastive inference magnitude between the two load conditions, and suggest two possible interpretations of this outcome. First, the early referential narrowing which characterizes contrastive inference may be the result of a cognitive process less working-memory-intensive than speaker modeling; for instance, conceptual pre-encoding with lexical
activation. Second, in light of post hoc tests suggesting that load actually does impair contrastive interpretation, but in a time window other than the one we initially set out to examine, it may be the case that contrastive inference is computed via speaker modeling, but that our study was underpowered to detect the predicted impairment.

**Contrastive Inference**

While early theories of sentence processing emphasized modularity and the late role of context (e.g., Ferreira & Clifton, 1986; Frazier & Fodor, 1978), psycholinguistic research has revealed comprehension to be an incremental process guided by context from its outset. Phonological, syntactic, and semantic aspects of a sentence are not decoded independently: Structures at each level inform each other as they are simultaneously constructed (Kuramada, et al., 2014; Snedeker & Trueswell, 2004; Yee & Sedivy, 2006; etc.). Rather than incorporating contextual information after a sentence has been fully analyzed, context guides this interactive construction as soon as the sentence begins (Tanenhaus, et al., 1995; Niewland & van Berkum, 2006; Spivey, et al., 2002; etc.).

The diversity of contextual factors rapidly integrated into online processing is striking. For example, eye movements during temporarily ambiguous sentences like "Put the apple on the napkin in the box" reveal an early preference to parse the prepositional phrase "on the napkin" as VP-attached or NP-attached depending on the number of apples in the scene (Tanenhaus, et al., 1995; Spivey, et al., 2002). Representations of the speaker also appear to play a role in early sentence processing. For instance, participants in an interactive object selection task look substantially more at an empty glass visible both to them and their interlocutor after hearing "the empty martini glass" than one visible only to themselves (Hanna, et al., 2003), suggesting that representations of our interlocutor's perspective rapidly constrain our interpretation of potentially-ambiguous noun phrases.
Some of the more sophisticated context-sensitive incremental interpretation we perform is related to the fact that speakers tend to provide prenominal modifiers only when there is a contrastive relationship between two or more objects in the immediate environment. For example, adjectives like "wooden" are typically produced to distinguish between referents of the same kind (e.g. a wooden spoon and a metal spoon) (Arnold, 2010; Brown-Schmidt & Tanenhaus, 2006; Gundel, et al., 1993; Ariel, 1990; Levelt, 1989; Sridhar, 1988; Osgood, 1971; Chafe, 1976; but see Engelhardt, et al., 2006). The amount of information speakers include in referential constructions thus varies with the presence of closely-related objects in the environment, and listeners are accordingly able to make predictions about what the referent of such a construction will be before the speaker is even finished uttering it (Sedivy, et al., 1994; Eberhard, et al., 1995; Sedivy, et al., 1999; Sedivy, 2003; Heller, et al., 2008; Grodner & Sedivy, 2011; Huang & Snedeker, 2013; Kronmüller, et al., 2014).

In work investigating the effect of prosodic stress on contrastive interpretation of modifiers, Sedivy, et al. (1994, reported in Eberhard, et al., 1995) found that listeners who heard the stressed modifier "large" in "touch the LARGE blue square" were quicker than listeners who heard the unstressed version to look at a large blue square which appeared with a corresponding small one when there was no other contrast pair present. But when another contrast pair was present whose members differed along the same dimension (e.g. a large and a small yellow circle), these predictive looks disappeared, suggesting listeners expect information conveyed by optional prenominal modifiers to perform critical disambiguation between similar referential candidates.

In a more direct investigation of this effect, Sedivy, et al. (1999) found that participants rapidly used the presence of adjectives like "tall" to narrow their visual attention in a display featuring a contrast pair (e.g. a tall and a short glass), but not in a display in which no tall object had a short counterpart. Using head-mounted eye-tracking, researchers observed the eye movements of participants seated in front of a
table while they listened to auditory instructions like "Pick up the tall glass." Every table display featured at least two objects which had the property designated by the adjective (e.g. a tall glass and a tall pitcher), one of which was designated as the target (e.g. the tall glass). Half of the displays additionally featured an object of the same type as the target but differing in the dimension of the adjective (e.g. a short glass). The other half featured an unrelated distractor. After hearing "tall," listeners looked more at the tall glass than a competing tall object such as a pitcher when the short glass was present (proportion of target looks minus proportion of competitor looks $\approx 0.11$), but this effect disappeared when the short glass was replaced with a second distractor such as a book ($T - C \approx -0.13$).

Figure 3.2: (A) Display with contrast item. (B) Display with no contrast item (Adapted from Sedivy, et al., 1999)

Further investigations of contrastive inference extended this pattern of findings to other adjective classes (Sedivy, 2003) and populations (Huang & Snedeker, 2013; Kronmüller, et al., 2014). While the magnitude and timing of the effect vary depending on details of the experimental task and variable(s) measured, adults typically display a substantial preference for the contrast item within a few hundred milliseconds of the optional modifier. Furthermore, the ability to link informativity of speech with
contrast in the environment appears to begin developing as early as age five (Huang & Snedeker, 2013), but does not reach full maturity until after age ten (Kronmüller, et al., 2014).

Contrastive Inference Processing

Contrastive inference thus has different signatures in different experimental contexts and seems to emerge gradually during development. By the time they are adults, language users routinely rapidly incorporate representations of referential context, in particular, representations of contrast among objects of the same type, into incremental interpretation of referential expressions. But how exactly does environmental contrast influence modifier interpretation in real time? The account most frequently suggested in the literature appeals to listeners' ability to model the speaker's intentions and choices.

Like many pragmatic implicatures, contrastive inferences are argued to be the result of computations which directly recruit representations of the speaker (Sedivy, 2003; Grodner & Sedivy, 2011). On this account, listeners assume speakers are subjected to conflicting pressures to be (1) as efficient as possible, i.e., not to waste time or energy with unnecessary material, and (2) as informative as necessary to succeed in the goals of communication (in this case, referent identification). Listeners then consider the speaker's awareness of the referential environment and the set of relevant utterances available to them (e.g., "Pick up the (tall) glass," "Pick up the (tall) pitcher," etc.). After comparing the relative informativeness of each utterance in the speech context, the listener can identify the optimal utterance given a referential target. For example, for a speaker referring to the tall glass in Figure 3.2 (A), "tall glass" is optimally informative but "glass" is underinformative. In case the speaker intends to refer to the pitcher, "tall pitcher" is overinformative but "pitcher" optimal. Since the tall glass is the only referent which requires a modifier for disambiguation, listeners who believe their interlocutors are optimally informative can thus conclude the target is the tall glass the moment they hear "tall."
There is experimental evidence in support of such a speaker modeling account for contrastive inference. Grodner and Sedivy (2011) had participants perform a contrastive inference task like that in Sedivy, et al. (1999), but manipulated between-subjects whether the person giving the instructions conformed to or routinely violated optimal informativity expectations, for example, by providing overinformative utterances like "Pick up the plastic spoon" when there was only one spoon present. While participants exposed to the reliable speaker showed a tendency to look at the tall glass more than the tall pitcher after hearing "tall" (T - C ≈ 0.11), participants exposed to the unreliable speaker did not (T - C ≈ -0.05).

On a speaker-modeling account, this is easy to explain: Since unreliable speakers provide listeners with evidence that they do not, in general, produce optimally-informative referential utterances, a listener's speaker model is missing the crucial premise that the speaker will not choose over- or underinformative descriptions. This prevents them from computing the inference linking "tall" with the tall glass. Thus it is due to accurate speaker modeling that listeners cease to expect optional modifiers to convey contrastive meaning.

But the early target disambiguation that characterizes contrastive inference is not necessarily the result of a computation involving representations of the speaker's intentions and choices. There are ways in which listeners could link the presence of a modifier with contrast in the environment that do not involve integrating speaker representations with linguistic ones. One extreme possibility is that listeners have learned a simple association between the presence of modifiers in referential phrases and the presence of contrast in the environment. (See Arnold, et al., 2007 for a similar proposal linking disfluency in the speech stream with novelty in the referential environment.) But there is no obvious reason that simple association should be modulated by the speaker's reliability.
A more plausible account of contrastive inference processing without speaker representations is that listeners typically mentally encode visual stimuli in their environment, taking contrast into account, and activation spreads to the lexical labels associated with each object such that only members of contrast pairs are mentally associated with modifiers. For instance, a listener viewing the four objects in Figure 3.2 (A) automatically conceptualizes these objects for the purposes of both comprehending and remembering the scene, resulting in the activation of concepts for both object types (GLASS, PITCHER) and for properties which disambiguate otherwise identically encoded objects (TALL + GLASS, SHORT + GLASS). As activation spreads, a concept implicitly triggers the lexeme associated with it (e.g. GLASS -> "glass", TALL -> "tall"). Evidence for rapid unconscious activation of single-noun labels for objects in the visual environment has been found in both infants (Mani & Plunkett, 2010; 2011) and adults (Zelinsky & Murphy, 2000; Yee & Sedivy, 2006), and while to the best of the author's knowledge, no studies have tested for activation of adjectives or adjective-noun combinations, the same mechanisms could result in such activations.

For someone viewing the scene in Figure 3.2 A, this process results in the association of the four objects with the implicit labels "tall glass," "short glass," "pitcher" and "key." Then, as verbal instructions unfold (e.g. "Pick up the tall glass"), they align incoming linguistic material to their internal labeling schema. Since the only object implicitly labeled "tall" is the tall glass, as soon as the adjective is detected in the speech stream, they can link the utterance to the tall glass.

In this way, unambiguously conceptualizing visual stimuli results in precisely the same pattern of modifier-to-contrast correspondences that speaker modeling does, but without appealing to the speaker (cf. Barr, 2008; Barr, 2014). When the listener notices the speaker's unreliability, without drawing on a fine-grained model of the speaker's intentions and choices, pre-encoding is suspended, eliminating the one-to-one mapping between the tall glass and the label "tall." For instance, the listener might adjust
their pre-encoding by suppressing their own internal conceptualization of the scene to make room for
the speaker's (potentially different) perspective (cf. Nadig & Sedivy, 2002; Hanna, et al., 2003; Hanna
& Tanenhaus, 2004; Brown-Schmidt, et al., 2008; Heller, et al., 2008). Note that on this account,
listeners maintain and update representations of the speaker, using them to guide high-level decisions
about whether or not to pre-encode, or with respect to which context to do it. But individual instances
of contrastive inference, that is, context-sensitive anticipatory looks to contrastive targets, are
performed egocentrically, without the guidance of fine-grained models of the speaker and their
intentions and choices.

Working Memory Load

Given that both speaker-modeling and self-oriented processing explanations are consistent with the
data, how can we test whether speaker modeling underlies the capacity for contrastive inference? There
are several reasons to think that working memory might provide a key.

First, working memory load appears to selectively interfere with interpretations involving scalar
implicature, another quantity inference frequently suggested to involve speaker modeling of precisely
the sort detailed above (De Neys & Schaeken, 2007; Dieusaaert, et al., 2011; Marty, et al., 2013; Marty
& Chemla, 2013). If working memory resources are required to maintain representations of, e.g., the
speaker's alternative utterances or the relationships between these alternatives and potential referents in
the environment, depletion of working memory stores should specifically interfere with a speaker-
modeling implementation of contrastive inference.¹

¹ Note that it is not necessarily the case that scalar implicature, despite sharing aspects of Gricean
quantity implicatures with contrastive inference, is processed in exactly the same way. Perhaps the two
inferences involve different mechanisms and representations despite appearing superficially similar.
Moreover, recent data suggests that scalar implicature, while often computed by means of speaker
modeling, may arise through conceptual pre-encoding in certain contexts (Huang & Snedeker, 2018).
While it is not clear whether participants in studies which showed negative effects of working memory
Second, working memory is implicated in perspective-taking and theory-of-mind capacities. Working memory load impairs performance on tasks involving interpreting speech with respect to an interlocutor's perspective (Lin, et al., 2010), and working memory capacity is correlated with performance on such tasks (Brown-Schmidt, 2009). Since speaker modeling involves listeners taking interlocutors' intentions and choices into account with respect to the visual environment they see, contrastive inference should be impaired under working memory load if it is computed via speaker modeling.

Third, working memory is also implicated in the integration of information from multiple sources. Cognitive processes involving fewer types of information are less susceptible to interference from working memory load than tasks which involve many types of information. Since speaker modeling requires integrating information about the speaker with linguistic information and information about the environment, it should require more working memory resources than self-oriented conceptualization.

Furthermore, while some degree of working memory is required for any act of linguistic comprehension, there is evidence that pre-encoding of the sort which may explain self-oriented processing of contrastive inferences is not particularly working-memory-intensive. Mani & Plunkett (2010; 2011) found that preverbal infants, who have minimal working memory capacity, exhibited signs of spontaneous lexical activation from simple depictions of everyday objects in noncommunicative contexts.

In light of these facts, if listeners derive contrastive interpretations of adjectives by performing a series of inferences from models of their interlocutor, contrastive inference behaviors should be weaker under load on scalar implicature were computing these inferences using speaker modeling, there is no evidence that they were not. It is not possible to decisively link working memory to speaker modeling on the basis of this data alone, but one must take it into consideration alongside other factors.
high working memory load than under low load. In order to test this hypothesis, we presented participants with a contrastive inference task similar to that of Sedivy, et al. (1999) embedded within a verbal working memory task modeled after Marty, et al. (2013). Using the visual world eye-tracking paradigm, we measured the difference in the strength of contrastive inference effects between participants under low and high verbal working memory load.

Experiment 3

To investigate whether contrastive inference is computed by speaker modeling, we designed an experiment to test whether working memory load impaired the predictive looking behaviors associated with contrastive inference. We had participants hold in memory sequences of letters, manipulating their lengths following Marty et al. (2013), while they performed contrastive inference-eliciting task modeled after Sedivy, et al. (1999). If contrastive inference is performed by active speaker modeling, listeners should display a smaller preference to look at contrastive target items over competitors when memorizing more letters.

Methods

We conducted a dual-task study using the visual world paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, Sedivy, 1995). Participants memorized either a long (3-6 letters) or short (one letter) sequence of letters while we used a remote eye-tracker to observe their eye movements to a visual display featuring pairs of contrast objects during instructions that featured an optional modifier.

Participants

Sixty undergraduate students at Harvard University with normal or corrected-to-normal vision participated in this study and were compensated with either course credit or US $10. Participants were
randomly assigned to the high working memory load condition (N = 30) or the low working memory load condition (N = 30).

Materials

Visual Displays

We created 40 physical displays featuring four everyday objects positioned on a 2x2 wooden platform lined with white shelf-liner (Figure 3.3). (See Appendix C for a list of all display objects.)

![Figure 3.3: An example of a visual display.](image)

Sixteen of the displays (the critical displays) had two versions: a contrast version including a pair of objects of the same type, and a non-contrast version with an unrelated distractor object in place of one object in the pair (Table 3.2). For example, the contrast version of the critical display featuring a short candle included a tall candle, a notepad, and a shot glass; the non-contrast version included a stapler, a notepad, and a shot glass. The object in the contrasting pair which appeared in both contrast and non-contrast versions of the display (e.g., the short candle) was the target object (the object participants would be directed to click on during the first part of the trial; see Procedure). The contrasting object of the same type (the contrast object, e.g., the tall candle) only appeared in the contrast version of the display. In both versions of every display, one of the remaining two objects, the competitor, was always
an exemplar of the property distinguishing the target (e.g., a short shot glass). Finally, in non-contrast versions, the contrast item was replaced with a second unrelated distractor (e.g., a stapler).

Table 3.2: Example of the contrast version (A) and the no-contrast version (B) of a critical display. The critical audio command is “Click on the short candle.” The contrast version features a pair of objects differing in one dimension (height), while the no-contrast version features a distractor in place of one of the objects.

Contrast pairs (consisting of the target object and the contrast object) were constructed exclusively of objects differing in size. Ten of the sixteen contrast pairs consisted of objects differing in overall height and had targets characterized by the adjectives "small" (6) or "large" (4). Three of the pairs differed in vertical height and had targets described as "tall" (2) or "short" (1). Two pairs differed in horizontal length and targets were both described as "long," and one pair differed in width and the target was described as "thin." (See Appendix C for a full list of the display objects.) So that participants would not develop a general bias for either the positive (large/tall/long/wide) or the negative end (small/short/short/thin) of the scales along which the contrast pairs differed, target objects were chosen

54
to fall equally often on the positive end (4 large, 2 tall, and 2 long objects) and the negative end (6 small, 1 short, and 1 thin object) of their scales.

We chose to use exclusively scalar adjectives like "tall" instead of a mix of scalar and material adjectives (like "wooden," cf. Grodner & Sedivy, 2011) after pilot testing revealed a substantial difference in the size of contrastive inferences induced by the two adjective classes. Since participants displayed greater target preference after hearing scalar adjectives than after hearing material ones, likely because size is much more visually salient in our stimuli than material, we used exclusively scalar adjectives in the present study to maximize the baseline contrastive inference effect on which we were testing for modulatory effects of working memory load.

We presented participants with photographs of physical displays instead of using digitally-rendered images in order to increase the visual salience of the contrast properties. Since previous demonstrations of contrastive inference involved participants viewing and manipulating actual physical objects on a table (Sedivy, et al., 1995; Sedivy, et al. 1999; Grodner & Sedivy, 2011), or shelf (Keysar, Barr, Balin & Brauner, 2000; Nadig & Sedivy, 2002; Brown-Schmidt, et al., 2008), we reasoned that if all four objects were presented together in a visually coherent scene with cues for depth, texture, shape and size, participants would be more likely to accurately encode the objects as being related to each other in these dimensions than if they were presented as isolated photographs of objects not occupying the same physical space, or as separate cropped images on a neutral background.

In addition to the sixteen pairs of critical displays, 24 single filler displays were constructed. Sixteen featured four unrelated objects (e.g., a pen, a roll of toilet paper, a pair of tongs, and a styrofoam cup). The remaining eight contained a contrast pair and two unrelated distractors and were used to prevent participants from developing a bias to look toward the contrast objects whenever a contrast pair was
present. (The eye gaze pattern predicted by a contingency between the presence of a contrast pair and the target being a member of the pair is identical to the pattern characteristic of contrastive inference.) Whereas every contrast version of a critical display was accompanied by initial instructions to choose a member of the contrast pair, these eight filler displays had initial instructions to choose a distractor instead (see Procedure for other instructions accompanying both critical and filler displays).

Auditory Commands

For each display we recorded two verbal commands directing participants to click on one of the four objects, adding to a total of 80 sentences of the form "Click on the <NP>." (The two versions of a given critical display were accompanied by the same pair of sentences.) Commands were recorded by an adult male native English speaker instructed to pronounce all sentences with primary stress on the noun (e.g., "Click on the CANDLE" or "Click on the short CANDLE"). This was in order to prevent the placement of contrastive stress on the adjective modifying the target noun, which is known to increase pre-noun looks to the target object when the object is a member of a contrast pair in the absence of other factors (Sedivy, et al, 1999).

The first command accompanying either version of a critical display picked out the target object (e.g., "Click on the short candle"), and included a modifier ("short") to uniquely identify it. (In the contrast version of the display, the modifier was necessary for disambiguation. In the no-contrast version, it was redundant.) The second command always picked out the distractor (e.g., "Click on the notepad"). For five of the eight filler displays which included a contrast pair, the first command picked out a distractor and the second command picked out a member of the contrast pair using a modified noun phrase. (These latter commands were included to decrease the ratio of overinformative to optimally-informative modifiers so that participants would be less likely to deem the speaker unreliable.) For the other three fillers with contrast pairs, as well as for the remaining 16 fillers, two commands were
recorded picking out two different distractors. The commands were such that no object in a given
display was referred to more than once.

Working Memory Task

Five sets of 40 letter sequences were designed for the working memory task, which was administered
simultaneously with the contrastive inference task. Since serial letter recall can be more difficult among
letters with phonologically similar names, the thirteen letters from which the sequences were built (B,
F, H, I, J, L, M, O, Q, R, U, X, and Y) were chosen to maximize the phonological difference between
members of any given subset. This way, the number of letters more closely corresponds to the number
of "slots" occupied in verbal working memory (Conrad and Hull, 1964; Baddeley, 1966; Salame &
Baddeley, 1982).

To minimize load while maintaining comparability within the dual-task paradigm², participants in the
low load condition were given only one letter to recall per trial. A set of forty single letters was
constructed such that no single letter occurred substantially more often than any other. Participants in
the high load condition memorized strings between three and six letters long. The exact length was
determined by the participant’s score on the working memory capacity assessment (see Procedure). Sets
of 40 strings of lengths three, four, five, and six were constructed to meet the following conditions: (1)
no single string contained a given letter more than once, (2) no string was repeated in the experiment.

---

² Participants in the low load condition were asked to memorize one letter rather than none in order to
maintain maximal comparability between the two conditions. The primary differences between one-
letter and multi-letter conditions are (1) the number of items to be recalled and (2) the difficulty of
reversal (trivial in the case of one letter, non-trivial in the case of many). Thus any differences in
performance between the two groups are primarily attributable to these factors. While zero-letter and
multi-letter conditions also differ in these respects, they additionally differ in that the former is a single-
task condition and the latter a dual-task one. Differences in performance between these two groups thus
are not exclusively attributable to (1) and (2), but also to the task-switching demands present in the
dual-task but not the single-task condition.
and (3) no single letter occurred substantially more often than any other over the course of the experiment. (See Appendix D for an exhaustive list of the strings used.)

Procedure

Prior to beginning the experiment, each participant's working memory capacity was assessed using a backwards digit span test. A participant's score was the greatest length at which they were able to correctly recall and reverse a number sequence given two attempts (See Appendix E for the full assessment procedure and materials.) The experiment was then administered using a Tobii T60 remote eye tracker sampling gaze at 60Hz. Character entry during the working memory task was conducted using a USB keyboard and selection of objects on the screen during the contrastive inference task was conducted using a USB mouse.

Participants were randomly assigned to either the high or low load condition. In both conditions, the contrastive inference task (detailed below) was sandwiched between the two stages of the reverse letter sequence recall task, which were as follows. First, participants were shown a series letters, displayed one at a time. Letters were shown in black on a white background for one second, separated by 500ms of a blank white screen. Next, the contrastive inference task was administered. Finally, a screen appeared asking participants to enter the letter sequence they saw, in reverse order. After they pressed <ENTER>, a feedback screen appeared which signaled that the response was either correct or incorrect.

Participants in the low load condition were always given one letter per trial to memorize, no matter the results of their working memory capacity assessment. In the high memory load conditions, participants were given strings two letters shorter than their score on the reverse digit span test. Someone who scored six on the assessment was given letter strings of length four, someone who scored eight would...
see strings six letters long, and so on. The maximum available string length was six and the minimum
three, therefore anyone scoring below five or higher than eight was given strings of length three and
six, respectively. The minimum was established to ensure that even low-scoring participants in the high
memory load condition were challenged substantially more than low memory load participants. The
maximum was established to ensure that that high-scoring participants in the high load condition were
not challenged so much they would be unable to successfully complete the memory task in most trials.

The contrastive inference task sandwiched between the two stages of the working memory task
featured a photo of four everyday objects accompanied by two audio commands. Each command
prompted the participant to select an object. (Object selection was performed using the mouse.) There
were 40 displays and 80 commands total. No two commands picked out the same object in a given
display. In both versions of the 16 critical trials, the first command picked out the target object and
featured a scalar adjective like "short" in “click on the short candle.” A given participant saw the
contrast version of eight of the 16 critical displays (e.g. tall candle, short candle, notebook, shot glass)
and the no-contrast version of the remaining eight (e.g. stapler, short candle, notebook, shot glass).
Presentation was counterbalanced so that for each display, the same number of participants saw its
contrast version as saw its no-contrast version. In both versions of critical displays, the second
command accompanying the display featured an unmodified noun phrase (e.g., "click on the notepad")
which picked out the distractor object.

The 24 filler trials were the same across all participants. Eight fillers had displays featuring a contrast
pair, five of which were accompanied by instructions first to click on the distractor and second to click
on a member of the contrast pair. This was in order to prevent a contingency between the presence of a
contrast pair in a display and the first of the two commands identifying an object in the contrast pair.
The first of these commands always featured an unmodified noun phrase (e.g., "click on the key") and
the second always featured a modified one (e.g., "click on the metal slinky"). The other three contrast displays were accompanied by two commands, both featuring unmodified noun phrases, and both picking out distractors (i.e. the two objects not part of the contrast pair). This was in order to prevent a contingency between the presence of a contrast pair and the identification of one of its members at any point in the trial. The remaining 16 filler trials had displays consisting of four unrelated distractors, and were accompanied by two commands with unmodified noun phrases identifying one and then another of the distractors.

Across the experiment, a given participant thus saw 16 displays with contrast pairs and 24 without. Of the 32 commands accompanying the 16 contrast displays, a given participant was asked to click on a member of the contrast pair a total of 13 times, eight of which occurred as the first command in the trial and five as the second. Of the 48 commands accompanying the 24 displays with no contrast pair, eight commands featured an unnecessary modifier (e.g., "click on the short candle" when no tall candle was present); this was unavoidable if linguistic input was to be controlled for. The other 40 commands featured unmodified noun phrases. In total, of the 80 commands a participant heard, 31 were modified (of which eight were unnecessarily so), and the remaining 49 were unmodified.

The displays were counterbalanced across four lists such that no quadrant (e.g., upper left, lower right) was referred to more than any other across the experiment. With the exception of the first three trials, which were always the same three fillers and served as practice trials to acclimate participants to the nature of the task, trials were presented in randomized order such that no two participants saw the displays in the same order.
Results

Mean object selection task accuracy was 90.38% in the low load condition and 90.07% in the high load condition. Mean letter recall accuracy was 90.5% among low and 68.17% among high load participants. Mean score on the working memory capacity test (backwards digit span) was 6.3 numbers (6 among participants then assigned to the low load condition and 6.6 among those assigned to high load). Thus, the average length of the letter sequences memorized by high load participants was 4.6, while all low load participants memorized one letter at a time.

In order to measure contrastive inference effects under the two load conditions, we analyzed the proportion of participants' fixations on the target object during a fixed time window, compared to fixations on the competitor, as a function of contrast in the referential environment and the degree of working memory load the participant was under. The location of participants' eye gaze was sampled at 60 Hz during the command portion of the trial (i.e., during the instructions "Click on the short candle").

This window was then broken down into three regions: Critical (immediately after adjective offset), pre-critical, and post-critical. Each sample was automatically assigned a validity score by the Tobii, indicating the degree of certainty of that measurement; We eliminated all samples for which neither eye had a score within the top half of the validity range. Trials which had fewer than 4 remaining samples within the critical region were removed from analysis. Forty-six trials, or 4.8% of trials, were removed this way.
Figure 3.4: Average proportion of looks to target, contrast, competitor, and distractor objects across all subjects and items in contrast trials. 0ms on the y-axis corresponds to the 200ms-shifted modifier offset.
Figure 3.5: Average proportion of looks to target, contrast, competitor, and distractor objects across all subjects and items in no-contrast trials. 0ms on the y-axis corresponds to the 200ms-shifted modifier offset.

In order to best capture a possible contrastive inference effect using a fixation proportion analysis, we defined the critical region for our analysis to include samples between zero and 500ms after (200ms-adjusted) adjective offset.\(^3\) In fixation proportion analyses, regions of analysis are typically offset

---

3 In contrastive inference studies measuring fixation latency, target fixations prompted by modifiers in contrastive environments have been reported anywhere from 217ms to 852ms after the onset of the disambiguating word, e.g. "candle" (Sedivy, et al., 1999; Eberhard, et al., 1995; Hanna, et al., 2003). In studies measuring fixation proportions in a given time window, divergence in preferential target looking between environments with and without contrast has been reported during the 500ms region.
200ms after the relevant linguistic cue to account for the time it takes to program and launch an eye-movement (Allopenna, et al., 1998; Matin, et al., 1993; Hallett, 1986). A region described as beginning at adjective offset and ending 500ms after offset thus actually begins 200ms after and ends 700ms after offset. The pre-critical region included samples between the onset of the first word of the sentence and the offset of the adjective, and the post-critical region included samples from adjective offset through the end of the sentence. Only observations from the first command of the 16 critical trials were analyzed, meaning that utterances always featured a prenominal adjective, but whether there was a contrast object or a second distractor present varied between subjects.

For each participant and each trial, we converted the eye gaze samples during the critical region, which indicated which quadrant of the screen the participant was looking at (e.g., target, contrast object, competitor, distractor) at a given moment into a measure of target preference. Target preference for a given trial was computed in two steps. First, competitor looks (the proportion of samples in the critical region during which a participant was looking at the competitor) were subtracted from target looks (the proportion of samples in which they were looking at the target). Since eye movements are characterized by brief, singular fixations broken up by saccades or quick jerks of the eyes, the distribution of the resulting difference score across participants was bimodal, with peaks at one and zero. We therefore binarized each difference score in order to perform a logistic analysis: Differences greater than zero were coded as 1 (target preference), differences less than zero were coded as -1 (competitor preference), and differences of exactly zero (mostly trials in which the participant was looking neither at the target nor the competitor during the critical region) were excluded from analysis. Trials were thus excluded for one of only two reasons: (1) The number of samples in which the participant was looking beginning at adjective offset (Grodner & Sedivy, 2011), the 667ms region beginning at noun onset (Huang & Snedeker, 2013), the 600ms region beginning 200ms after noun onset (Hanna, et al., 2003), the region beginning 200ms before noun onset and ending 200ms after noun onset (Heller, et al., 2008), and the region beginning at adjective onset and ending 600ms after noun onset (Ryskin, et al., 2015).
at the target was equal to the number of samples in which they were looking at the competitor during
the critical region, and (2) there were no samples during the critical region in which the participant was
looking at either the target or the competitor.

Eye-tracking Analysis: Critical Region

Mean target preference during the critical region in the low load condition was 58.9% in no-contrast
trials and 77.5% in contrast trials. A logistic mixed effects model with contrast as a fixed effect and
participant and item as random effects showed a main effect of contrast among the low load
participants ($\beta = 0.52$, $SE = 0.12$, $z = 4.117$, $p < 0.001$). In the high load condition, target preference
among no-contrast trials was 56.8% and among contrast trials was 65.4%. An analogous model showed
a main effect of contrast among high load participants, albeit a smaller one ($\beta = 0.27$, $SE = 0.13$, $z =
2.12$, $p < 0.05$).

Among all data, a logistic mixed effects model with contrast and load as fixed effects and participant
and item as fixed effects revealed a significant main effect of contrast ($\beta = 0.39$, $SE = 0.09$, $z = 4.42$, $p
< 0.001$), a main effect of load on target preference ($\beta = -0.24$, $SE = 0.1$, $z = -2.41$, $p < 0.05$), but no
significant interaction between the two factors ($\beta = -0.14$, $SE = 0.09$, $z = 1.55$, $p = 0.12$).
Figure 3.6: Proportion of trials with more looks to target than competitor object in contrast vs. no-contrast trials during the critical region 0-500ms after adjective offset.

Post Hoc Eye-tracking Analyses

Because the variety of time frames in which contrastive inference has been observed (See Footnote 2), we also performed the same analyses on the five 100ms sub-regions of the critical region (Table 3.2). Significant main effects of contrast were found in all five regions. Main effects of load emerged only in the fourth (300-400ms post-adjective-offset) and fifth (400-500ms post-adjective-offset) sub-regions, and a significant interaction between contrast and load was present in the first sub-region (0-100ms post-adjective-offset), but absent thereafter.
<table>
<thead>
<tr>
<th>Time Window</th>
<th>Contrast (β, SE, z, p)</th>
<th>Load (β, SE, z, p)</th>
<th>Contrast * Load (β, SE, z, p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-100ms</td>
<td>Contrasts 0.29 (0.1), z 2.94, p 0.003&lt;br&gt;Loads 0.03 (0.10), z 0.28, p 0.78&lt;br&gt;Contrast * Load -0.2 (0.1), z -2.01, p 0.045</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100-200ms</td>
<td>Contrasts 0.35 (0.1), z 3.6, p 0.0004&lt;br&gt;Loads -0.11 (0.11), z -1.02, p 0.31&lt;br&gt;Contrast * Load -0.09 (0.1), z -0.93, p 0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200-300ms</td>
<td>Contrasts 0.45 (0.1), z 4.5, p 0.000005&lt;br&gt;Loads -0.11 (0.12), z -0.99, p 0.32&lt;br&gt;Contrast * Load -0.002 (0.1), z -0.03, p 0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300-400ms</td>
<td>Contrasts 0.29 (0.1), z 2.99, p 0.003&lt;br&gt;Loads -0.29 (0.10), z -2.82, p 0.005&lt;br&gt;Contrast * Load -0.01 (0.1), z -0.10, p 0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>400-500ms</td>
<td>Contrasts 0.25 (0.1), z 2.47, p 0.01&lt;br&gt;Loads -0.33 (0.10), z -3.2, p 0.001&lt;br&gt;Contrast * Load -0.1 (0.1), z -0.97, p 0.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Effects of load and contrast on target preference in five 100ms time windows after adjective offset from a linear mixed effects model.

Visual inspection of gaze trajectories suggests that the interaction is driven by the fact that target preference is both earlier and longer-lasting in the low-load condition than in the high-load condition. Whereas low-load target preference begins well before adjective offset and appears to continue past the end of the critical region, target preference in high load participants did not emerge until ~200ms after adjective offset, and lasted only a few hundred milliseconds.
Discussion

To test whether contrastive inference is computed by speaker modeling, we had participants perform a contrastive inference task under high and low memory load. Reduction or elimination of contrastive inference under high load would suggest memory-intensive speaker modeling. We observed a main effect of contrast in both load conditions: Participants looked significantly more at the target object after hearing a contrastive adjective when a contrastive item was present vs. absent no matter the degree of memory depletion. However, we found no strong evidence that contrastive inference was reduced under high load: No significant interaction between contrast and working memory load was observed during the primary region of interest. On the one hand, this suggests that contrastive inference is not computed by speaker modeling. On the other, it is consistent with the possibility that load does impair contrastive inference but our study was underpowered to detect it. I will address each of these possibilities in turn.
The fact that our findings did not provide clear evidence for an effect of working memory load on contrastive inference is perhaps because there simply is no such effect: Contrastive inference is computed by a process less resource-intensive than speaker modeling. What could this look like? Any process wherein the link between the modifier (e.g. "short") and the presence of contrast in the referential environment is derived in a self-oriented way is a candidate. For instance, perhaps listeners conceptually pre-encode the objects in our displays, taking into account contrast such that each object is given a unique mental representation to facilitate identification and recall. Then, as activation spreads from the concepts to their associated lexemes, participants end up with labeling schemata that happen to conform to Gricean informativity expectations. With mental labels like "tall candle," "short candle", "shot glass" and "notepad" activated, they then check incoming speech against their internal list. Referents whose internal labels match the speech stream are entertained as referents, while those whose labels do not are excluded. Since members of contrast pairs are the only objects whose conceptualizations require encoding of properties like size and height to uniquely identify them, prenominal modifiers in the speech stream immediately cue participants to members of contrast pairs. Though the internal labeling schema matches the labels a Gricean speaker would generate under communicative pressure, it is constructed without direct reference to communication: The two are similar because the demands of successful encoding and recall are similar to the demands of successful and efficient communication. If conceptual pre-encoding is a fast and automatic process, as priming research in infants and adults suggests (Zelinsky & Murphy, 2000; Mani & Plunkett, 2010; 2011; Mani, et al., 2012), it should be difficult to impair even under high degrees of working memory load.

On the other hand, the fact that our findings did not provide clear evidence for an effect of working memory load on contrastive inference may be because our study design was underpowered to detect such an effect, which nevertheless might be real. If there is a real effect of working memory load on
contrastive inference lurking underneath our results, then contrastive inference may indeed be computed by speaker modeling. Weakness in our working memory manipulation or our analysis strategy may have obscured our ability to observe this.

There are several reasons for us to hesitate before concluding that there is no effect of working memory load on contrastive inference. First, strong claims (such as that contrastive inference is not computed by speaker modeling) cannot be made from null effects in general: In order to decisively make such a conclusion, we would want, for instance, to have a direct comparison between contrastive inference and some other phenomenon in which depletion affected the latter but not the former.

Second, our best guess as to the expected size of the reduction of contrastive inference by working memory load, based on the size of the effect of memory depletion on scalar implicature (Marty, et al., 2013), was roughly half. In other words, if scalar implicature is computed by speaker modeling, we expected an approximately 50% reduction in the rate of inference computation from low to high load participants if contrastive inference is computed in an analogous way. Given that the baseline target preference observed in the low load condition is about 20.3%, a reduction in inference by half translates to a target preference of 10.15% in the high load condition. In fact, we observed a 10.3% target preference in the high working memory load condition, suggesting that our study may have been statistically underpowered to detect the predicted effect.

Additionally, the load manipulation we chose may not have been strong enough to thoroughly tax participants' working memory in the high load condition. The population we studied had high working memory capacity overall, and may not have been challenged sufficiently by our task. While six letters may have occupied the working memory of a participant who scored eight on the capacity test enough to interfere with memory-intensive processes, participants who scored nine, ten or eleven may have had...
plenty of resources left over to perform memory-intensive processes. A stronger manipulation in which participants assigned to the high load condition were given longer strings more closely titrated to their working memory capacity might do a better job of ensuring individuals are taxed to the point of observable deficit.

Finally, post-hoc analyses of 100ms time windows suggest that perhaps we are looking at the wrong time region when analyzing for modulatory effects on contrastive inference. The significant interaction between load and contrast in the 0-100ms time window suggests that the primary deficits caused by working memory load are in the latency and duration of target preference, and future analyses should focus on the short time window immediately after adjective-offset rather than the 0-500ms post-adjective-offset region used by Grodner & Sedivy (2011) based on previous research (Sedivy, et al., 1999; Eberhard, et al., 1995). By including samples from outside the region in which we are most likely to observe effects of working memory load, we may have reduced our power even further.

In sum, our findings are consistent with two possibilities. I have briefly outlined both, but without further experimentation we cannot distinguish between them once and for all. One of the chief aims of Experiment 4, detailed in the next chapter, is to re-test the above hypothesis using stronger manipulations and better-motivated analyses. A version of this study with greater power is more likely to yield interpretable results.
Chapter 4: Effects of Working Memory Load and Speaker Reliability on Contrastive Inference

Introduction

In Chapter 3, we set out to test whether contrastive inference is computed by speaker modeling, as Grodner & Sedivy (2011) suggested. Since speaker modeling involves building, maintaining, and manipulating representations of the speaker during comprehension, we expect contrastive inferences to be impaired by working memory load if they are computed this way. Self-oriented contrastive inference, on the other hand, is less likely to exhibit vulnerability to load interference. Spontaneous, contrast-sensitive scene conceptualization with automatic lexical activation, for instance, involves fewer steps and less integration of information from diverse sources than speaker modeling. In Experiment 3, an eye-tracking study examining contrastive inference rates under high and low working memory load, we found no evidence that load impairs contrastive inference. On the one hand, this suggests that it is computed not via speaker modeling but by a less resource-intensive self-oriented process. On the other hand, post hoc tests suggesting that our analysis region was misplaced, in conjunction with the relative ease of the memory task and the fact that the effect size we observed was roughly what we expected on a speaker modeling account (Marty, et al., 2013), prevent us from drawing strong conclusions from Experiment 3.

One of the aims of the experiment presented in the current chapter is to re-test the same hypothesis using a stronger version of the working memory manipulation and a better-motivated analysis strategy. To that end, we once again monitored participants' contrastive-inference-making under high and low memory load, but using a more challenging memory task, new audio stimuli, and a modified analysis
plan. With these modifications, we are in a better position to detect whether load affects contrastive inference, and to learn about the mechanism behind it.

But while an effect of load on contrastive inference would suggest speaker modeling over self-oriented processing, a null result would be difficult to interpret: Is the lack of effect due to the real-life absence of effects of load on inferencing? Or is it because the study lacks power or the memory task is too easy? We faced this interpretation challenge in Chapter 3 and do our best here to minimize it. But there is another, more serious difficulty in interpreting the outcome of this study. While we predict speaker modeling to tax cognitive resources more than self-oriented processing, both presumably involve some degree of working memory: Object identification, contrast encoding, and implicit labeling could not occur without some involvement of working memory. Therefore, both theories are consistent with an effect of working memory load on contrastive inference. Finding a reliable effect of load could at most tell us how severely memory depletion affects inference.

In light of this, the current study design was modified to provide a more conclusive test of contrastive inference processing. We added a second factor, speaker reliability, which is predicted to interact with working memory in completely different ways by the two hypotheses under consideration. Following Grodner & Sedivy (2011), we randomly exposed half of the participants to an unreliable speaker after assigning them to the high or low memory load conditions. If contrastive inference is the result of speaker modeling, we predict it will be impaired by each of the two factors independently, and a fortiori by both factors in conjunction. An effortful and accurate modeling procedure which generates inferences under typical circumstances should be impaired due to lack of resources by working memory load, resulting in the reduction or disappearance of contrastive inference. It should continue running when the speaker is unreliable, but the model should not yield inferences since the speaker's behavior does not warrant them. With both factors present, the modeling process should be impaired by
load (leading to reduced inferences) and missing the critical premises the model needs to generate inferences due to unreliability. If any working memory is left to run speaker modeling, it will still be missing the premises crucial to generate contrastive inference, so inferences should be impaired just as much or more than under each factor alone.

Recall from Chapter 3 that contrastive inference can just as easily be explained as a self-oriented process, for example spontaneous contrast-sensitive conceptualization of the objects in the visual display followed by automatic lexical activation which rapidly interacts with incoming speech. On such an account, listeners could inhibit their self-oriented conceptualization when presented with evidence that the speaker deviates from conversational norms (Nadig & Sedivy, 2002; Hanna, et al., 2003; Hanna & Tanenhaus, 2004; Brown-Schmidt, et al., 2008; Heller, et al., 2008). If contrastive inference is the outcome of this type of process, it should be impaired under speaker unreliability. Since it involves working memory to a lesser degree than speaker modeling, it should be minimally impaired by load alone. But critically, under both factors simultaneously, contrastive inference should appear stronger than under unreliability alone: While load depletes working memory resources required for inhibiting the spontaneous contrast-sensitive construal underlying the lexical activation that leads to contrastive inference, it leaves the components of contrastive inference themselves relatively unaffected.

In this chapter, I briefly review the effects of speaker unreliability on contrastive inference observed by Grodner & Sedivy (2011). I then detail the predictions of each of the two hypotheses introduced in Chapter 3 with respect to working memory load and speaker unreliability. I review an extension of Experiment 3 designed to more thoroughly probe the mechanism underlying contrastive inference by investigating the effects of working memory load and speaker unreliability together. Using eye-tracking, we monitored participants' contrastive interpretation of prenominal modifiers under all four combinations of load (high and low) and reliability (reliable and unreliable). Contrastive inference was
substantially impaired by speaker unreliability as predicted by Grodner & Sedivy (2011), but largely unaffected by working memory load, which merely slowed verbal processing. Notably, participants under high memory load computed robust contrastive inferences even under unreliable speakers, suggesting that suppressing contrastive inference is cognitively demanding, while the inference itself is relatively easy. I propose a detailed account of contrastive inference processing which explains this counterintuitive result, drawing on research in implicit labeling and perspective-taking in language use.

**Predictions**

Speaker-modeling accounts of pragmatic inference have gained support primarily from studies investigating scalar implicature processing. Some of the strongest evidence that listeners actively recruit speaker representations in computing pragmatic inferences comes from studies that manipulate what the listener knows about the speaker. For example, Bergen & Grodner (2012) found that participants were less likely to interpret "some" as "some (but not all)" when the context suggested the speaker might have incomplete knowledge. This lends support to an account on which listeners deploy representations of the particular speaker they are listening to each time they hear an underinformative scalar term like "some." Under typical circumstances, listeners reason that since the speaker both knows whether "some" and "all" apply and strives to be maximally informative, they would have said "all" if it had been applicable, allowing them to conclude that it must not be and yielding the upper-bounded interpretation "some (but not all)." Similar findings lend further support to this account (Goodman & Stuhlmüller, 2013; Hochstein, et al., 2014).

While this speaker-modeling mechanism can also explain why listeners compute contrastive inferences (Sedivy, et al., 1999; Sedivy, 2003), only one published study directly investigated the role of speaker representations in contrastive inference processing. Grodner & Sedivy (2011) found that listeners
exposed to speakers who used modifiers redundantly, among other "unreliable" behaviors, computed fewer contrastive inferences than listeners exposed to speakers who did not, paralleling the findings on scalar implicature. In a head-mounted eye-tracking study, participants' eye movements were observed as they followed instructions like "Pick up the tall glass" while seated at a table. Among other objects on the table were the target (a tall glass), a competitor which was also accurately described by the target adjective (a tall pitcher), and either a contrast object of the same nominal category as the target but differing along the dimension of the adjective (a short glass) or an unrelated distractor object. When the speaker was reliable, listeners' tendency to look more at the target object than the competitor in the 500 milliseconds after the end of "tall" was greater in trials with a contrast object (e.g., a short glass) than an unrelated distractor. In other words, even though "tall" applied equally well to the tall glass and the pitcher, listeners rapidly focused on the glass when it had a short partner from whom the adjective was necessary to distinguish it. But this contrastive inference effect disappeared when speakers appeared with four key modifications meant to indicate to participants their unreliability with respect to informative language use.

The four modifications were as follows: (1) Instructions at the beginning of the experiment told the participant that the person who recorded their instructions had "an impairment that caused language and social problems" (Grodner & Sedivy, 2011, p. 250); (2) Five of the several hundred referential descriptions uttered over the course of the experiment contained semantically-constrained lexical errors (e.g., a toothbrush was called a "hairbrush"); (3) Three of the location descriptions referred to nonexistent locations (e.g., instead of saying "above A and below B," the speaker said "above B and below A" when no such location existed); and (4) Most referential descriptions were more informative than they needed to be (297), either as unreduced noun phrases when pronouns were clearly licensed (37 instances) or noun phrases with unnecessary modifiers (197 instances). In a follow-up study which
removed the first of these modifications from the unreliable speaker condition but kept the other three, listeners displayed the same contrast-sensitive target preference when listening to unreliable speakers as reliable ones, suggesting that observing speakers' unreliable behavior instance-by-instance is not enough to trigger contrastive inference suspension, but rather that top-down cues to the speaker's unreliability are crucial.

The sensitivity of listeners' contrastive inference rates to the speaker's perceived conformity to conversational norms has been used to argue that listeners deploy Gricean inferences in real time whenever they interpret modified referential descriptions (Grodner & Sedivy, 2011). But, as noted above, the predictive target looks prompted by speakers' adjective use that characterize contrastive inference need not be the result of reasoning about the speaker. Listeners may instead employ a self-oriented process to rapidly link modifiers in incoming speech to contrast in the environment without appealing to the speaker's choices and intentions. For example, upon viewing a display of four objects, two of which differ only in size (e.g., a short candle, a tall candle, a shot glass, and a notepad), listeners might spontaneously conceptualize the objects for their own unambiguous identification and recall. Since the concept CANDLE does not uniquely pick out either candle, they are encoded with salient distinctive properties: TALL CANDLE and SHORT CANDLE. (Meanwhile, SHOT GLASS and NOTEPAD suffice to pick out the other two.) Conceptual activation could then spread to the lexemes associated with these concepts, leaving the listener with implicit labels for the four objects. These labels are active in the listener's mind when they begin to parse incoming speech. Upon hearing "short," the label "short candle" is triggered, resulting in looks to the object associated with it.

How do each of these theories of contrastive inference—speaker modeling and self-oriented processing—account for the findings of Grodner & Sedivy (2011)? On the speaker modeling theory, listeners build, maintain, and manipulate representations of their interlocutor which they employ from
the first moments of comprehension. When these representations are of a speaker who uses modifiers informatively, listeners can conclude that "short" in an unfolding referential expression indicates contrast with similar but non-short object. But when these representations are of a speaker who does not use modifiers informatively, listeners can't conclude anything in particular after hearing "short," resulting in no contrastive inference.

On the self-oriented processing theory, listeners can't help but conceptually encode and implicitly label the world around them, whether their interlocutor is reliable or not. No matter the status of the speaker, unambiguous encoding of contrast pairs results in modified implicit labels which permit rapid target identification upon encountering a modifier. When the speaker is reliable, this rapid identification proceeds unhindered. But when cued to the speaker's unreliability, the listener actively inhibits their internal labeling schema, preventing it from guiding moment-to-moment interpretation of referential noun phrases. This results in literal processing with no contrastive inference under unreliable speakers. (It is important to note that this theory does not claim that listeners do not maintain or use representations of the speaker at all: It is clear that listeners at least attend and respond to high-level speaker attributes like unreliability. Rather, this theory simply claims that fine-grained representations of the speaker such as their intentions, possible utterances, and choices do not play a role in listeners' computation of contrastive meaning from optional modifiers.)

While both speaker-modeling and self-oriented accounts of contrastive inference predict sensitivity to speaker reliability, they begin to diverge in their predictions when it comes to listener behavior under working memory load. A speaker-modeled contrastive inference is computed from a model of the speaker that includes attributes like the speaker's intention to use language informatively, the set of utterances available to them in a given context, and the relative informativeness of each of those utterances in context. Furthermore, the listener must perform, upon hearing a prenominal modifier like
"short" and based on the model, a counterfactual inference to rule out the competitor referent. For example, they must reason as follows: "If the speaker had meant to refer to the shot glass, to say "short shot glass" would have been overinformative; therefore they must mean to refer to the other short object." Speaker modeling involves building, maintaining, and manipulating representations of the speaker and their choices and intentions, and integrating these representations via counterfactual reasoning with representations of the context and of incoming linguistic material. The complexity of this process suggests that it should be working-memory-intensive, predicting that contrastive inference should be impaired under load.

On the other hand, the spontaneous and contrast-sensitive encoding of the visual world predicted by the self-oriented theory predicts that contrastive inference should have relatively low working memory demands. While object recognition and unambiguous encoding, followed by lexical activation and implicit label maintenance, surely require some working memory resources, these processes are spontaneous and, being self-oriented, do not involve integration of information from the variety of sources speaker modeling does. Thus, while self-oriented contrastive inference may be impaired by load, it is less likely than speaker-modeled inference to be strongly affected or eliminated.

But while the two theories predict different degrees of working memory involvement in contrastive inference, evidence for a negative effect of load on inference rates could be interpreted in two ways. First, it could indicate that working-memory-intensive speaker modeling is responsible for contrastive inference. Depletion of working memory resources impairs modeling, leading to fewer or weaker contrastive inferences in the high load condition. Second, it could indicate that relatively easy self-oriented processes which nevertheless require working memory lead to contrastive inference. Depletion impairs the aspects of self-oriented inferences that depend on working memory, again resulting in fewer or weaker inferences under high load. While the size of the effect we observe might
nudge us one way or the other, the outcome of the working memory manipulation alone cannot rule out either theory.

However, since the two theories make strongly divergent predictions about the effortfulness of eliminating contrastive inferences under an unreliable speaker, observing the effects of working memory load and speaker reliability together can tell us more about the nature of contrastive inference than studying the effects of working memory alone. Since the two theories predict different patterns of contrastive inference rates under each of the four combinations of load (high vs. low) and speaker type (reliable vs. unreliable), observing contrastive inference under all four conditions has the potential to shed light on the underlying mechanism.

What are the crucial differences among the predictions of speaker modeling and self-oriented processing? Both theories predict strong contrastive inferences under a reliable speaker and low load, if by different mechanisms. Both theories are consistent with weaker contrastive inferences under a reliable speaker and high load, although speaker modeling arguably predicts weaker inferences than self-oriented processing. Both theories predict weak or absent contrastive inferences under low load and an unreliable speaker, but for very different reasons. Speaker modeling predicts that listeners will run the same inference based on the same process of speaker modeling that resulted in inferences from the reliable speaker, but that the unreliable speaker model differs crucially in that its parameters do not license the reasoning that results in contrastive inference. Self-oriented processing predicts that listeners spontaneously generate an implicit labeling schema which incorporates contrast in both speaker conditions, but that they attempt to inhibit it or suppress its effects after the fact when the speaker is unreliable.
Critically, suspending inferences on the self-oriented account is more effortful than letting them run, while suspending and computing inferences are equally resource-intensive processes on the speaker-modeling theory. If contrastive inferences are computed by speaker modeling, participants under high working memory load should thus display the same rates of inference regardless of speaker reliability. Under the reliable speaker, load should impair partially or fully participants’ ability to compute contrastive inferences. If load fully impairs modeling, resulting in no inference at all under the reliable speaker, then participants should also be unable to model the unreliable speaker, again resulting in no inference. If load only partially impairs modeling, resulting in reduced inference under the reliable speaker, then whatever working memory is left to model the speaker in the unreliable condition will be deployed on a speaker who doesn't license inferences, resulting in no inference under the unreliable speaker. Thus speaker modeling predicts no contrastive inference in the high-load, unreliable-speaker condition.

If contrastive inferences are the result of self-oriented processing, then suspending them under an unreliable speaker is more effortful than letting them run under a reliable one. Under high load, participants’ ability to perform effortful suspension of inference should be impaired, resulting in at least partially unsuspended contrastive inferences in the high-load, unreliable-speaker condition. In other words, while the speaker-modeling theory predicts no contrastive inference under simultaneous high load and unreliability, the self-oriented theory predicts inferences in this condition.

**Experiment 4**

In this study, which is based on Experiment 3, we manipulate two independent variables: Working memory load and speaker reliability. As in Experiment 3, working memory load was manipulated by providing participants with long or short strings of letters to memorize in a dual task paradigm, but this
time titrated exactly to participants’ working memory capacity as indicated by a backwards digit span test. Speaker reliability was implemented following Grodner & Sedivy (2011) by creating two sets of auditory verbal stimuli, one from a "reliable" speaker who always provides sufficient information for reference identification and no more, and one from an "unreliable" speaker who is reported to have a language disorder and frequently provides too much, insufficient, or erroneous information.

Methods

Participants

One hundred thirty-six undergraduate students at Harvard University with normal or corrected-to-normal vision participated in this study and were compensated with either course credit or US $10. Eleven others participated but were excluded from analysis: Five due to experimenter error or software malfunction, two due to ineligibility (dyslexia and age outside the intended range) and four due to track loss (see Results).

During the same session, some participants engaged in another eye-tracking study testing for differences in eye movement patterns between sentences with unaccusative and unergative verbs; this study was performed first and took about 30 minutes to complete, and participants were given a break in between the two studies. The first study consisted of passive listening and looking with a small number of comprehension questions. None of the objects in the first study's visual stimuli appeared in the current study.

Procedure

As in Experiment 3, participants were given a working memory capacity test prior to beginning the experiment. The experimenter assessed participants' backwards digit span by reading aloud sequences
of numbers of increasing length and asking participants to repeat back these sequences in reverse order. A participant's score was the greatest length at which they were able to correctly recite the reverse sequence given two attempts.

Participants were randomly assigned to one of two working memory load conditions (low or high) and one of two speaker conditions (reliable or unreliable). All participants were given a cover story in which the study's purpose was to investigate how successful the referential descriptions of different speakers were in directing listeners quickly to one of several objects in a display. They were told that the verbal instructions they would hear were recorded by a fellow participant in an earlier stage of the study, and shown an example of that hypothetical participant's task. The hypothetical task was to sit in front of a screen with the same visual displays the participant was about to see, wait for a red box to appear around one of the four depicted objects, and tell a hypothetical listener to click on that object. Participants in the reliable speaker condition were not given any further information; participants in the unreliable speaker condition were told that the participant whose instructions they were about to hear had been diagnosed with a disorder affecting their language and social behavior. Participants in the unreliable condition were told to do their best in selecting the referent they thought the speaker meant to indicate, even if it was sometimes difficult.

Trial structure and the number of trials were the same as in Experiment 3. Over the course of the experiment, participants saw 40 displays, each featuring four everyday objects and accompanied by two verbal commands. The two commands were sandwiched between the two stages of a working memory task identical to that in Experiment 3: First, participants saw letters displayed one at a time, and last, they had to type the letters in reverse order. Of the 80 commands presented, 16 critical commands were designed to elicit contrastive inference when presented with a display featuring a contrast pair (e.g. "Click on the short candle" with a tall and a short candle). In the unreliable speaker
condition, 49 of the remaining 64 filler commands provided too much, too little, or erroneous information (see Materials). In the reliable speaker condition, all 64 fillers were optimally informative.

Materials

In order to increase the likelihood that we would observe an effect of working memory load on contrastive inference should one exist, the stimuli from Experiment 3 were modified slightly. First, the auditory commands were re-recorded by a female native English speaker. Next, the visual displays were edited to be more clear. Finally, the upper limit on the working memory task was increased from six to nine, and participants were given strings of length equal to their score on the assessment rather than their score minus two letters. For example, a participant in the high load condition who scored eight on the assessment would have memorized six-letter strings in Experiment 3, but eight-letter strings in Experiment 4. Participants scoring nine or above would be assigned strings nine letters long, and participants scoring three or below would be assigned strings of three letters. As in Experiment 3, participants in the low load condition were given only one letter to memorize at a time regardless of their score.

Visual Displays

The visual displays were the same as in Experiment 3, but the digital photographs were edited in two ways. In Experiment 3, the objects were closer to the center of the screen than to the edges, potentially obscuring the distinction between looks to objects in different quadrants. For the current study, photographs were cropped as closely as possible while maintaining four-way symmetry among the quadrants. Additionally, objects in some images from Experiment 3 were more difficult to see than others due to lighting or object surface attributes; The brightness and contrast of these images were manipulated to maximize clarity and consistency across the experiment.
Auditory Commands

The auditory stimuli for Experiment 4 were recorded by a female native English speaker instructed to place prosodic stress on the head of each noun phrase (e.g., "Click on the CANDLE," "Click on the square plastic CLOCK"). All stimuli were recorded during the same session to maximize prosodic consistency, in particular pitch and speech rate. The distribution of verbal commands was analogous to that of Experiment 3: Two commands per display, adding up to a total of 80. In the reliable speaker condition, the content of all 80 commands was identical to that of the commands in Experiment 3. In the unreliable condition, while the 16 critical commands and two of the fillers were unaltered, noun phrases in the sixty-two remaining fillers were changed in one of three ways. Fifty-two noun phrases were modified so that they became overinformative (e.g., "Click on the square plastic clock" when only one clock was present), five became underinformative (e.g., "Click on the glove" when two gloves were present), and eleven contained lexical errors (e.g., "Click on the knife" when only a fork was present). (See Appendix G for a complete list.) Overall, 65 of the 80 total commands in the unreliable speaker condition, or about 81%, were non-optimal in some way. In contrast, only eight of the 80 commands (10%) in the reliable speaker condition were overinformative. (This was inevitable: Due to the design of the contrastive inference task, each time a participant was presented with the no-contrast version of a display accompanying a critical command, they were necessarily exposed to an overinformative utterance.)

Working Memory Task

The working memory task was identical to that in Experiment 3, except that the maximum difficulty was increased from six to nine letters. The same sets of letter sequences were used, but with the addition of three new sets: One set each for sequences seven, eight, and nine letters long. Just as in the
sets of shorter sequences, the new sets were constructed so that no letter was repeated within a given sequence and no one letter occurred substantially more often than another across the entire set.

Results

Across all participants, mean object selection task accuracy was 89% (see Table 4.1 for breakdown by condition) and mean letter recall accuracy 78% (see Table 4.2 for breakdown by condition).

<table>
<thead>
<tr>
<th></th>
<th>Reliable Speaker</th>
<th>Unreliable Speaker</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Load</td>
<td>91%</td>
<td>85%</td>
<td>88%</td>
</tr>
<tr>
<td>High Load</td>
<td>92%</td>
<td>88%</td>
<td>90%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>91%</strong></td>
<td><strong>87%</strong></td>
<td><strong>89%</strong></td>
</tr>
</tbody>
</table>

Table 4.1: Experiment 4 object selection accuracy by condition.

<table>
<thead>
<tr>
<th></th>
<th>Reliable Speaker</th>
<th>Unreliable Speaker</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Load</td>
<td>96%</td>
<td>95%</td>
<td>96%</td>
</tr>
<tr>
<td>High Load</td>
<td>67%</td>
<td>57%</td>
<td>62%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>82%</strong></td>
<td><strong>74%</strong></td>
<td><strong>78%</strong></td>
</tr>
</tbody>
</table>

Table 4.2: Experiment 4 letter recall accuracy by condition.

Mean score on the working memory capacity test (backwards digit span) was 5.73 numbers (5.70 among participants then assigned to the low load condition and 5.74 among those assigned to high load). Thus, the average length of the letter sequences memorized by high load participants was 5.74, while all low load participants memorized only one letter at a time.

Eye-tracking was performed in the same manner as in Experiment 3. Samples were collected during the 500ms window, offset by 200ms, after the end of the adjective. All samples for which neither eye had a score within the top half of the validity range were eliminated, and trials with fewer than three remaining samples during the critical region were removed from analysis. One hundred eighty-one
trials, or 9.3% of trials, were removed in this way. Subjects who did not have at least three trials left in each condition were removed from analysis; data from four subjects were excluded in this way.

In light of the results of post-hoc tests in Experiment 3, which revealed the strongest effects of working memory load on contrastive inference to be in the 0-100ms post-adjective-offset time window, we planned two sets of logistic mixed effects regressions. First, we examined the original region of interest defined as 0-500ms after adjective offset. Second, we performed the same analyses on data from the 0-100ms time window, reasoning that this may be the real locus of any modulatory effects of load.

As in Experiment 3, only observations from the first command of the 16 critical trials were analyzed. Gaze was converted into a binarized measure of target preference exactly as in Experiment 3. The resulting target preference measure indicated whether, for a given participant and a given trial, they were looking more at the target (e.g., short candle) than the competitor (e.g., shot glass) during the region of interest.

Eye-Tracking Analysis: Critical Region

Revisiting the Effect of Working Memory Load

In order to test for effects of working memory load on contrastive inference, which we failed to find in Chapter 3, we examined data from the reliable speaker condition in isolation. A logistic mixed effects regression with contrast and load as fixed effects and participant and item as random effects revealed a main effect of contrast ($\beta = 0.30$, SE = 0.09, $z = 3.249$, $p = 0.001159$), a main effect of load ($\beta = -0.33$, SE = 0.09, $z = -3.512$, $p = 0.000445$), and a marginal interaction of load and contrast ($\beta = -0.16$, SE = 0.09, $z = -1.698$, $p = 0.089474$) such that the magnitude of contrastive inference under low load (75.33% - 56.20% = 19.13%) was greater than under high load (60.49% - 53.75% = 6.74%; Figure 4.1). This is consistent with working memory load impairing verbal processing overall, delaying
phoneme and word recognition across all conditions. Contrastive inferences were thus likely delayed partially past the end of the analysis window in the high load condition, leading to a marginal interaction of load and contrast.

![Figure 4.1: Average target preference (proportion of trials in which there were more target looks than competitor looks) by contrast and load during the 0-500ms region post-adjective-offset.](image)

**Overall (Omnibus) Analyses**

Among all participants, mean target preference during the critical region was 56.70% in no-contrast trials and 66.04% in contrast trials (Figures 4.2 and 4.3). A logistic mixed effects model with contrast, load and reliability as fixed effects and participant and item as random effects showed a main effect of contrast ($\beta = 0.25$, $SE = 0.65$, $z = 3.837$, $p = 0.000125$), indicating that participants were robustly computing contrastive inferences.
Figure 4.2: Average proportion of looks to target, contrast, competitor, and distractor objects across all subjects and items in no-contrast trials. 0ms on the y-axis corresponds to the 200ms-shifted modifier offset.

Figure 4.3: Average proportion of looks to target, contrast, competitor, and distractor objects across all subjects and items in contrast trials. 0ms on the y-axis corresponds to the 200ms-shifted modifier offset.
The model also showed a main effect of load such that target preference was higher overall (65.63%) in the low load condition than in the high load condition (57.37%; $\beta = -0.22$, SE = 0.65, $z = -3.456$, $p = 0.000549$), again suggesting that working memory load impaired verbal processing in general. (The absence of a significant two-way interaction of load with contrast corroborates this interpretation.)

Finally, a significant three-way interaction among contrast, load, and reliability was found ($\beta = -0.14$, SE = 0.65, $z = -2.148$, $p = 0.031724$; Figure 4.4). The model yielded no other significant main effects or interactions (all $\beta$s < 0.1, all $p$s > 0.12).

Figure 4.4: Average target preference (proportion of trials in which there were more looks to the target than to the competitor) by contrast, load and reliability during the region 0-500ms after adjective offset.
In order to explore the three-way interaction further, we separated the data by load condition and performed mixed effects regressions with contrast and reliability as fixed effects on each of the subsets. Among data from participants in the low working memory load condition, the model revealed a significant main effect of contrast ($\beta = 0.26$, $SE = 0.10$, $z = 2.715$, $p = 0.007$) such that target preference was higher in contrast trials (70%) than no-contrast trials (61%). There was also a significant interaction between contrast and reliability ($\beta = 0.2$, $SE = 0.10$, $z = 2.091$, $p = 0.04$), such that the magnitude of contrastive inference in the reliable condition (76% - 61% = 15%) was greater than that in the unreliable condition (64% - 61% = 3%). This is consistent with the findings of Grodner & Sedivy (2011) that contrastive inference is eliminated when the speaker is unreliable. (A separate mixed-effects regression on low-load, unreliable-speaker data revealed no main effect of contrast ($\beta = 0.10$, $SE = 0.14$, $z = 0.69$, $p = 0.49$), suggesting that contrastive inference was eliminated, not merely impaired, by speaker unreliability.) A main effect of reliability, such that target preference was higher in the reliable condition (68%) than the unreliable condition (62%) was only marginal ($\beta = 0.26$, $SE = 0.10$, $z = 1.903$, $p = 0.06$), and likely driven by the difference among contrast trials only.

Among participants under high load, there was a significant main effect of contrast ($\beta = 0.26$, $SE = 0.10$, $z = 2.715$, $p = 0.00664$) such that target preference was higher in contrast trials (62%) than no-contrast trials (53%). Neither a significant main effect of reliability nor an interaction between reliability and contrast were observed (all $|\beta|$s $< 0.06$, all $p$s $> 0.5$).

The lack of an interaction between reliability and contrast among high-load data suggests that participants in both the reliable- and the unreliable-speaker conditions computed contrastive inferences. In order to corroborate the finding that contrastive inferences were computed under simultaneous high memory load and unreliable speaker exposure, we further split the low-load data into reliable- and
unreliable-speaker subsets and performed mixed-effects regressions with contrast as the only fixed effect on each subset. In particular, we sought to rule out the possibility that the contrastive inference effect among high-load data was being driven by participants in the reliable speaker condition alone.

Although target preference was higher in contrast trials (60%) than no-contrast trials (54%) among participants in the high-load reliable-speaker condition, a logistic mixed-effects model showed that this difference was not reliable ($\beta = 0.16, \text{SE} = 0.13, z = 1.194, p = 0.232$). An analogous model of high-load unreliable-speaker data, on the other hand, suggested that the moderately greater difference between target preference in contrast trials (60%) and no-contrast trials (46%) in that condition was reliable ($\beta = 0.25, \text{SE} = 0.12, z = 2.053, p = 0.0401$). This confirms that the main effect of contrast seen among high-load data cannot be explained by contrastive inference behavior in the reliable speaker condition alone: The effect of the unreliable speaker, which eliminated contrastive inference on its own, was mitigated by working memory load.

Eye-tracking Analysis: 0-100ms Post-Adjective-Offset

In light of post-hoc test results from Chapter 3 which suggested that the peak modulatory effects of working memory load on contrastive inference occur during the first 100 milliseconds after the prenominal modifier, we performed separate logistic mixed-effects regressions on data from that region. Just as in the 0-500ms analysis, an omnibus model including contrast, load and reliability as fixed effects and participant and item as random effects revealed a main effect of contrast during this region ($\beta = 0.24, \text{SE} = 0.08, z = 3.098, p = 0.00195$), as well as a main effect of load ($\beta = -0.18, \text{SE} = 0.08, z = -2.295, p = 0.02172$) such that target preference overall was lower under high load than low load. While the size of the contrastive inference effect was smaller among high load (56.64% - 53.42% = 3.22%) than low load data (68.98% - 53.98% = 15.00%), the interaction between contrast and load
was only marginal ($\beta = -0.14$, SE $= 0.08$, $z = -1.842$, $p = 0.06547$). Again, this is consistent with an overall slow-down of verbal processing by working memory load pushing contrastive inferences past the end of the analysis region.

Discussion

Experiment 4 investigated the processing of contrastive inference by measuring participants’ visual target preference in contrast vs. no-contrast trials under different degrees of working memory load and speaker reliability. Three main findings emerged: First, contrastive inference was absent when the speaker was unreliable, replicating the findings of Grodner & Sedivy (2011). This confirms that contrastive inference is sensitive to attributes of the speaker, but does not tell us why. Second, working memory load slowed verbal processing overall, likely resulting in contrastive inference effects being pushed partially past the end of the window of analysis and yielding a marginal effect of load on contrast. This is consistent with the self-oriented processing account of contrastive inference, as it does not indicate a major role for working memory in contrastive inference computation. But it is also consistent with a version of the speaker-modeling hypothesis in which working memory plays only a minor role. (I will discuss this possibility below.) Finally, we confirmed the counterintuitive prediction of the self-oriented processing hypothesis that contrastive inference should be stronger under both manipulations than under unreliability alone. Contrastive inference appears to be a self-oriented process which can be inhibited when we learn that our interlocutor violates conversational norms, but which reappears under the additional burden of working memory load.

In the remainder of this section, I will do two things. First, I present a detailed proposal for how contrastive inference is computed as a self-oriented process, drawing on insights from research on implicit labeling. Second, I will examine the possibilities for the mechanism by which contrastive
inference is suspended under speaker unreliability, noting that results from investigations of perspective-taking in language processing support an account on which suspension results from effortful inhibition of one's own perspective.

Contrastive Inference as a Self-Oriented Process

What is contrastive inference? By what cognitive mechanism do listeners predictively associate prenominal modifiers with contrast among objects in the environment? In contrast to the predictions of a speaker-modeling account, our results confirmed the predictions of a self-oriented processing account on which rapid contrastive interpretation of prenominal modifiers is due to a highly-automated, listener-internal procedure. In particular, we found that contrastive inference was robust under simultaneous high memory load and speaker unreliability, a fact which is easily explained if contrastive interpretation occurs automatically but is effortful to suppress.

What exactly does this process look like? First, listeners viewing a simple display containing everyday objects recognize and conceptually encode the objects they see for unambiguous identification and recall. Someone viewing a display featuring an apple and a book recognizes these familiar objects and mentally "labels" them with the concepts into which they fit: APPLE and BOOK. Importantly, members of pairs or sets of objects of the same type (e.g., two candles or three glasses) must be encoded with an additional piece of information to distinguish them from the rest. For instance, if the same display also featured two candles, one tall and one short, two instances of the concept CANDLE would not be sufficient for disambiguation and successful recall. This problem is solved by attaching complex concepts to similar objects consisting of the object's category and its most salient distinctive feature in context. Thus, APPLE, BOOK, TALL CANDLE and SHORT CANDLE are the set of
concepts (some simple, some complex) whose activation results from viewing a display featuring these five objects.

Next, activation spreads from the object concepts to the lexemes most closely associated with them. Just as free viewing of displays featuring images of everyday objects in noncommunicative contexts results in subconscious, unintentional activation of their names in both infants and adults (Yee & Sedivy, 2006; Mani & Plunkett, 2010; 2011; Mani, et al., 2012), participants in our studies implicitly labeled the objects on the screen in front of them without realizing it. In particular, I argue, they labeled them in a way that goes beyond traditional implicit labeling in which the common noun associated with the everyday object is subconsciously activated, but in a way such that they could differentiate the two objects from each other on the basis of the label alone at a later date. While evidence exists for implicit labeling (which involves only single-noun activation), no studies to date have investigated whether adjectives, multi-word phrases, or in particular, adjective-noun combinations, are subconsciously activated by free-viewing. In order to test the self-oriented processing account of contrastive inference more thoroughly, we should look for evidence that viewing contrast pairs (tall candle/short candle) primes recognition of adjectives like "tall" and "short."

Now that the participant has an implicit mental labeling schema for the objects on the display from which they are about to be asked to choose an object, hearing a word or phoneme that singles out one of these labels triggers a chain of activation from the word to the concept and ultimately to the object itself, resulting in looks to the object. For example, in the case of the display described above, as soon as the participant hears "tall," the mental label "tall candle" is activated to the exclusion of all others, and the participant---already looking for an object on the screen to match the speaker's description---rapidly and effortlessly connects the label to the tall candle on the screen. This quick, context-sensitive contrastive interpretation of adjectives is contrastive inference.
When there are no objects of the same nominal category in the display, participants need not use complex concepts to successfully encode the scene, and thus implicit labels can all be simple nouns. Thus, in no-contrast trials, participants must wait until they hear the noun before they are able to identify the target object, and no early target looks are observed---no contrastive inferences are made.

It is important to note that while contrastive inference appears to be sensitive to communicative pressures, communication does not play an essential role in the self-oriented processing account. In contrastive inference, modifiers in speech are linked to contrast in the referential environment when and only when they would be necessary for a cooperative speaker adhering to the Maxim of Quantity to successfully convey reference to a listener. But the demands of successful recall---which we might think about as successful communication with our future selves---are isomorphic to the demands of such communication. Conceptual modification is necessary for me to remember what I was looking at in just those cases where linguistic modification is necessary for you to understand which object I'm referring to. What was thought to be a quantity implicature computed via active, context-sensitive modeling of the speaker's intentions and choices, I argue, is actually the result of a self-driven, egocentric process of encoding the world independently of particular communicative circumstances. Contrastive inference is much easier than previously thought; suppressing it is the difficult part.

Perspective-Taking

If language users are constantly subconsciously encoding the visual world around them in a manner that both displays sensitivity to contrast among like items and results in lexical activation (implicit labeling), how do we keep this from affecting our online language comprehension when it is no longer helpful? For example, when we receive cues that our interlocutor does not conform to typical modifier usage, how do we stop interpreting the adjectives they produce contrastively? We found that
participants were generally able to eliminate early contrastive interpretation of modifiers from unreliable speakers, but once burdened with working memory load, they were no longer able to keep contrastive inferences from happening. This demonstrates that however easy contrastive inference may be (its robustness under both load and unreliability in our findings suggest that it is relatively easy), inhibiting it is cognitively effortful and, in particular, makes demands on the working memory system.

What cognitive mechanisms underlie the inhibition of contrastive inference? How exactly do listeners keep contrastive internal labeling schemata from automatically influencing their moment-to-moment interpretation of speech? Some clues---but no definite answers---come from research on perspective-taking in language comprehension, in particular, investigations of the implicit context with respect to which people interpret language.

In an object manipulation task similar to that in Sedivy, et al. (1999), Keysar, et al. (2000) arranged objects in the cubbyholes of a shelf situated between a speaker and a listener. While all of the objects were visible to the listener, some of them were blocked from the speaker's view (Figure 4.5). Critically, one of the objects visible only to the listener (i.e., in the listener's privileged ground) was of the same type as a pair of objects visible to both parties (i.e., in their common ground). For example, while the common ground contained two candles, one large and one small, the listener's privileged ground contained a third, even smaller, candle. When the speaker asked the listener to move "the small candle," the location of listeners' visual fixations shortly after "small" was used to infer whether their processing was sensitive to the difference between common and privileged ground. A listener who looks to the smaller of the two candles in common ground indicates rapid sensitivity to their interlocutor's perspective, quickly integrating it into online processing to constrain possible interpretations. On the other hand, a listener who looks to the candle only they can see indicates interference from their own
Studies using this paradigm show that listeners are able to rapidly integrate information about their interlocutor's perspective into online processing, but not perfectly. Knowledge of common ground influences early eye movements (Nadig & Sedivy, 2002; Heller, et al., 2008; Brown-Schmidt, et al., 2008), but privileged ground nevertheless interferes with target identification (Keysar, et al., 2003; Nadig & Sedivy, 2002; Hanna, et al., 2003). Comprehension appears to integrate contextual information about the difference between one's own and one's interlocutor's perspective, but nevertheless to be imperfect: Listeners will always be biased by their point of view. The idea that common-ground-related context-sensitivity requires inhibition of one's own perspective gains support from studies showing that behaviors indicative of such sensitivity are weaker or less frequent when participants have low working memory capacity or perform poorly on measures of executive function.
or inhibitory capacity (Lin, et al., 2010; Wardlow, 2013; Brown-Schmidt, 2009) or when participants are under working memory load (Lin, et al., 2010). For example, Brown-Schmidt (2009) found that participants' score on a modified Stroop task predicted their sensitivity to linguistic common ground in a contrastive inference task: Those with lower inhibitory capacity displayed greater interference from privileged ground when interpreting potentially ambiguous utterances. Lin, et al. (2010) found that participants under high working memory load made more fixations to an object in privileged ground that was compatible with a speaker's description than when they were under low load, suggesting that working memory in particular is required to inhibit one's own perspective when doing so aids in comprehension.

Just as privileged ground is difficult to inhibit when interpreting utterances from a speaker whose referential perspective clearly differs from our own, so are the self-oriented conceptualizations that lead to contrastive inference. Our spontaneous contrastive encodings of the world around us, which usually help us to more rapidly and accurately understand utterances from speakers who conform to informativity norms, are difficult to inhibit even when they are no longer helpful and may in fact be misleading (as when our interlocutor clearly violates such norms, using modifiers superfluously). In particular, when our working memory resources are depleted, we can't help but let these egocentric conceptualizations exert rapid online influence on our comprehension of referential expressions, even when we know it would be better to do without them.

Contrastive inference, a pragmatic enrichment of literal meaning that at first appeared effortful to compute, is in fact the inevitable outcome of our default understanding of the world of objects around us. It is triggered independently of communicative context, and requires active inhibition to cancel when the circumstances do not license it.
Conclusions

In this chapter I have presented the results of an experiment which investigated the effects of working memory load and speaker reliability on contrastive inference. The goals of this study were two-fold: First, to revisit the empirical question Experiment 3 from the previous chapter was meant to answer, namely, whether contrastive inference is a working-memory-intensive process, and second, to find out whether it is most likely computed by speaker modeling as many authors have suggested (e.g., Grodner & Sedivy, 2011) or an alternative, self-oriented process. By measuring contrastive inference rates under different degrees of working memory load (low vs. high) and speaker reliability (reliable vs. unreliable), we were able to differentiate between the two theories on the basis of their divergent predictions, in particular, for observations in the high-load, unreliable-speaker condition. Where contrastive inferences computed by speaker modeling should have been eliminated under simultaneous load and speaker unreliability, we found reliable evidence for robust contrastive inferences computation under these circumstances.

These results are consistent with an account on which contrastive inference is the result of a contrast-sensitive implicit labeling process which is both highly automatic and difficult to inhibit. On the basis of evidence that language users spontaneously conceptualize the world around them in a way that results in robust activation of lexemes associated with the objects they see, I suggested that contrastive inference is the result of listeners rapidly linking incoming speech with implicit labels. When contrast in the environment requires that a conceptualization include distinctive properties like TALL and SHORT to unambiguously encode objects of the same type, listeners immediately connect matching adjectives in the speech stream to contrast objects, and not other objects which happen to share the
property denoted by the adjective. When no contrast sets are present, no properties are encoded, and adjectives in incoming speech do not generate inferences.

It thus appears as though language users initially spontaneously encode their surroundings in a contrast-sensitive way no matter the communicative circumstances. But when they receive cues that their interlocutor does not conform to communicative norms, they attempt to inhibit the egocentric encoding schema with respect to which they would otherwise interpret the speaker, an effortful process which requires working memory to complete. Thus, when listeners display sensitivity to speaker attributes like adherence to Gricean maxims, it is not the result of incorporating the speaker's intentions and choices into a dynamically unfolding model which guides a rational interpretation algorithm online, but rather the result of actively suppressing a default conceptualization of the environment upon encountering evidence that the speaker themselves is not rational.

These results are striking primarily for two related reasons. First, they challenge the widely held notion that semantic processes are easy and pragmatic processes are hard. Whereas pragmatic inferences are typically taken to be---and often are---effortful, socially-sensitive processes which can be easily cancelled, our data show that language users' deployment of a sophisticated link between informativeness and environmental contrast to enrich literal meaning can be the reflection of a relatively trivial and entirely egocentric process which itself is very difficult to interrupt. Just because an interpretive process is pragmatic---i.e. it involves deriving received, not literal meaning---doesn't entail that it is more difficult to perform than something semantic---i.e. part of literal meaning.

Second, our results demonstrate that some sophisticated-looking behaviors which at first seem only explicable by speaker modeling are actually "impostors": Inevitable reflexes of how our minds work independently of communicative contexts which happen to perfectly parallel the pressures of efficient
communication (Barr, 2008). What at first seemed like the highly rational and sophisticated deployment of knowledge of the speaker's choices and intentions, and links between these and the referential environment, into counterfactual reasoning processes that allowed a listener to infer aspects of a speaker's intended message before they even completed it turned out to be a side effect of how people in general, outside of communication, perceive and encode the world around them.
Chapter 5: Conclusions

This dissertation investigated core questions about the computation of pragmatic inferences by examining the effects of internal (working memory load) and external factors (speaker reliability) on the online processing of scalar implicature and contrastive inference. By explicitly manipulating the combination of internal an external circumstances while observing participants’ contrastive inference behavior, we learned that this pragmatic process is likely a more automatic, egocentric inference than previously thought. By manipulating the internal resources available to listeners as they processed quantifier meaning, we inadvertently discovered that even semantic processes typically assumed to be fast and undemanding can make nontrivial demands on working memory, even while others remain relatively effortless.

In Chapter 2, I compared the online processing of scalar quantifiers like "some" and "all" to that of numerals like "two" and "three" under different degrees of working memory load. By extending the design of the study conducted by Marty, et al. (2013) which revealed offline effects of load on upper-bound computation, I hoped to shed light on the timecourse of the detrimental effect of cognitive load on scalar implicature processing (i.e., on the upper-bounding of semantically lower-bounded quantifiers like "some"). While I failed to observe upper-bounding of "some," I discovered a significant difference in the effect of working memory load on the processing of the scalar quantifier "all" and on the numerals "two" and "three." In particular, processing of "all" was more susceptible to impairment than that of numerals, suggesting in light of a growing body of evidence that numerals and scalar quantifiers are processed differently, and more importantly, that not all semantic processes are easy.

In Chapter 3, I compared rates of contrastive inference under high and low working memory load with an eye toward learning about the nature of the inference as either easy and automatic, or effortful and
computed with reference to the speaker's intentions and choices. Predicting that if listeners model their interlocutors as rational agents to pragmatically derive contrastive meaning from prenominal adjectives, processing signatures of contrastive inference should be impaired under working memory load, I instead found a pattern of null results open to a number of interpretations.

In Chapter 4, I presented a careful elaboration on the study presented in Chapter 3, designed to more decisively differentiate between possible accounts of contrastive inference processing. By observing participants' rates of contrastive inference under different degrees of working memory load and speaker reliability simultaneously, I was able to test the specific predictions of an automated-contrastive-inference theory according to which inference should be effortfully suppressed when speakers violate communicative norms, but under load, should reappear as listeners lack the necessary resources for inhibition. I confirmed the counterintuitive predictions of this hypothesis, demonstrating that whereas contrastive inference was thought to be an effortful process of active speaker modeling, it actually appears to be the result of a trivial and difficult-to-interrupt default encoding of the world around us. This finding fundamentally challenges the simple view of the semantics-pragmatics processing dichotomy by suggesting that some pragmatic inferences which at first appear explicable only by appeal to sophisticated social reasoning about our interlocutors' intentions and choices can actually be the automatic reflexes of simple self-oriented behavior.

While the simplistic view of the semantics-pragmatics dichotomy retains its explanatory power in many cases, this dissertation has demonstrated that it is nevertheless not universal. Semantic processes are often easy and quick, but they can also be complex and resource-intensive. In particular, quantifiers appear to vary in how much working memory we need in order to understand them. Likewise, pragmatic inferences are often complicated inferential computation that invoke social reasoning and
rich context-sensitivity, but they can also be highly automated, egocentric processes that show up in
disguise.

Despite this striking result, many important questions concerning the processing of semantic and
pragmatic meaning this dissertation began with remain unanswered. If semantic computations vary in
their complexity as results from Experiments 1 and 2 indicate, with e.g. proportional quantifiers being
more difficult to process than numerals, what exactly determines what makes a given semantic
operation more or less challenging? What exactly is the role of working memory in semantic
computations of different types? Future research should examine the working-memory demands of
different types of quantifiers, from simpler---e.g., "none" and "all"---to more complex---e.g., "most,"
"less than half," and "between five and ten."

Furthermore, now that we know that (at least some) pragmatic inferences can likely be computed
automatically, is it the case that automatic inferences are always automatic, and active ones are always
active? Or are some or all pragmatic inferences computed active or automatically depending on the
circumstances? If so, what determines when one process is employed over the other? Are some
pragmatic inferences always "impostors," egocentric but giving the impression of being sensitive to the
subtle pressures of optimal communication? Are some inferences always actively computed by speaker
modeling? Are some, as Huang & Snedeker (2018) claim, potentially computed either way, depending
on particular aspects of the circumstances? Future research should investigate whether and how various
inferences display the signatures of active or automatic processing under different circumstances, and
when and why during development these different strategies arise.

Finally, the question we were unable to address in Chapter 2 remains unanswered: What is the
timecourse of the impairment of working memory load on scalar implicature? Does load merely slow
upper-bounding of "some," or stop upper-bounding altogether? Future research into this and other aspects of the cognitive mechanisms and processes underlying the interpretation of both literal and intended meaning will help to shed light on how language works in the mind, as well as clarify and refine the connections between theory and evidence in the study of linguistic meaning.
References


## Appendix A: Experiment 1 & 2 Items

<table>
<thead>
<tr>
<th>Onset</th>
<th>Object 1 Continuation</th>
<th>Object 2 Continuation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Birthday</td>
<td>cakes</td>
<td>cards</td>
</tr>
<tr>
<td>2 Baseball</td>
<td>bats</td>
<td>gloves</td>
</tr>
<tr>
<td>3 Football</td>
<td>helmets</td>
<td>jerseys</td>
</tr>
<tr>
<td>4 Christmas</td>
<td>lights</td>
<td>trees</td>
</tr>
<tr>
<td>5 Apple</td>
<td>pies</td>
<td>sauce</td>
</tr>
<tr>
<td>6 Music</td>
<td>boxes</td>
<td>stands</td>
</tr>
<tr>
<td>7 Micro</td>
<td>phones</td>
<td>waves</td>
</tr>
<tr>
<td>8 Motor</td>
<td>boats</td>
<td>cycles</td>
</tr>
<tr>
<td>9 Hockey</td>
<td>pucks</td>
<td>sticks</td>
</tr>
<tr>
<td>10 Fire</td>
<td>crackers</td>
<td>flies</td>
</tr>
<tr>
<td>11 Butter</td>
<td>cups</td>
<td>flies</td>
</tr>
<tr>
<td>12 Toilet</td>
<td>scrubbers</td>
<td>paper</td>
</tr>
<tr>
<td>13 Coffee</td>
<td>creamers</td>
<td>makers</td>
</tr>
<tr>
<td>14 Table</td>
<td>cloths</td>
<td>spoons</td>
</tr>
<tr>
<td>15 Honey</td>
<td>bees</td>
<td>dews</td>
</tr>
<tr>
<td>16 Water</td>
<td>fountains</td>
<td>melons</td>
</tr>
</tbody>
</table>
## Appendix B: Experiment 1 & 2 Letter Sequences

Letter sequences presented to participants during the working memory portion of the trial.

<table>
<thead>
<tr>
<th>LOW LOAD</th>
<th>HIGH LOAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 BH</td>
<td>BHFJ</td>
</tr>
<tr>
<td>2 FJ</td>
<td>LRMX</td>
</tr>
<tr>
<td>3 LR</td>
<td>HLXF</td>
</tr>
<tr>
<td>4 MX</td>
<td>RHML</td>
</tr>
<tr>
<td>5 HL</td>
<td>JHFR</td>
</tr>
<tr>
<td>6 XF</td>
<td>BLJX</td>
</tr>
<tr>
<td>7 RH</td>
<td>RMHF</td>
</tr>
<tr>
<td>8 ML</td>
<td>XHLB</td>
</tr>
<tr>
<td>9 JH</td>
<td>BRFM</td>
</tr>
<tr>
<td>10 FR</td>
<td>HXRL</td>
</tr>
<tr>
<td>11 BR</td>
<td>FHXM</td>
</tr>
<tr>
<td>12 JX</td>
<td>MRHB</td>
</tr>
<tr>
<td>13 RM</td>
<td>XJHR</td>
</tr>
<tr>
<td>14 HF</td>
<td>RBFX</td>
</tr>
<tr>
<td>15 XH</td>
<td>LMJR</td>
</tr>
<tr>
<td>16 LB</td>
<td>JFMB</td>
</tr>
</tbody>
</table>
## Appendix C: Experiment 3 Displays

### Critical Displays

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Large crayon</td>
<td>Small crayon</td>
<td>Large piece of chalk</td>
<td>Shell</td>
<td>Cocktail umbrella</td>
</tr>
<tr>
<td>2</td>
<td>Large hairbrush</td>
<td>Small hairbrush</td>
<td>Large bowl</td>
<td>Playing card</td>
<td>Medal</td>
</tr>
<tr>
<td>3</td>
<td>Large flashlight</td>
<td>Small flashlight</td>
<td>Large oven mitt</td>
<td>Yo-yo</td>
<td>Orange</td>
</tr>
<tr>
<td>4</td>
<td>Large binder clip</td>
<td>Small binder clip</td>
<td>Large light bulb</td>
<td>Lego piece</td>
<td>Battery</td>
</tr>
<tr>
<td>5</td>
<td>Small envelope</td>
<td>Large envelope</td>
<td>Small cell phone</td>
<td>Wooden stick</td>
<td>Toy globe</td>
</tr>
<tr>
<td>6</td>
<td>Small whisk</td>
<td>Large whisk</td>
<td>Small rock</td>
<td>Slipper</td>
<td>Soda can</td>
</tr>
<tr>
<td>7</td>
<td>Small deodorant</td>
<td>Large deodorant</td>
<td>Small candy</td>
<td>Camera</td>
<td>Power strip</td>
</tr>
<tr>
<td>8</td>
<td>Small Post-It note pad</td>
<td>Large Post-It note pad</td>
<td>Small thumb tack</td>
<td>Ladle</td>
<td>Twine</td>
</tr>
<tr>
<td>9</td>
<td>Small funnel</td>
<td>Large funnel</td>
<td>Small Chapstick</td>
<td>Necktie</td>
<td>Squirt gun</td>
</tr>
<tr>
<td>10</td>
<td>Small leaf</td>
<td>Large leaf</td>
<td>Small pocket knife</td>
<td>Hair bow</td>
<td>Watch</td>
</tr>
<tr>
<td>11</td>
<td>Tall jar</td>
<td>Short jar</td>
<td>Tall basket</td>
<td>Flip-flop sandal</td>
<td>Necklace</td>
</tr>
<tr>
<td>12</td>
<td>Tall cup</td>
<td>Short cup</td>
<td>Tall soap dispenser</td>
<td>Highlighter</td>
<td>Hot glue gun</td>
</tr>
<tr>
<td>13</td>
<td>Short candle</td>
<td>Tall candle</td>
<td>Short shot glass</td>
<td>Legal pad</td>
<td>Stapler</td>
</tr>
<tr>
<td>14</td>
<td>Long pencil</td>
<td>Short Pencil</td>
<td>Long toothbrush</td>
<td>Sunglasses</td>
<td>Computer mouse</td>
</tr>
<tr>
<td>15</td>
<td>Long spoon</td>
<td>Short spoon</td>
<td>Long hammer</td>
<td>Clock</td>
<td>Thimble</td>
</tr>
<tr>
<td>16</td>
<td>Thin marker</td>
<td>Thick marker</td>
<td>Thin paintbrush</td>
<td>Tissue box</td>
<td>Tennis ball</td>
</tr>
</tbody>
</table>
## Filler Displays

<table>
<thead>
<tr>
<th></th>
<th>Glass mug</th>
<th>Ceramic mug</th>
<th>Nail</th>
<th>Paper bag</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>Wool glove</td>
<td>Latex glove</td>
<td>Stick of gum</td>
<td>Tupperware container</td>
</tr>
<tr>
<td></td>
<td>Plastic Slinky</td>
<td>Metal Slinky</td>
<td>Safety goggles</td>
<td>Bottle of hand sanitizer</td>
</tr>
<tr>
<td>18</td>
<td>Wide roll of tape</td>
<td>Narrow roll of tape</td>
<td>Key</td>
<td>Zipper lock bag</td>
</tr>
<tr>
<td>19</td>
<td>Short glass</td>
<td>Tall glass</td>
<td>Tea bag</td>
<td>Nail clippers</td>
</tr>
<tr>
<td>20</td>
<td>Black feather</td>
<td>White feather</td>
<td>Picture frame</td>
<td>Ice cube tray</td>
</tr>
<tr>
<td>21</td>
<td>Red towel</td>
<td>Blue towel</td>
<td>Matchstick</td>
<td>One dollar bill</td>
</tr>
<tr>
<td>22</td>
<td>Gray sock</td>
<td>White sock</td>
<td>Eyeglasses</td>
<td>Tweezers</td>
</tr>
<tr>
<td>23</td>
<td>Doll</td>
<td>Bracelet</td>
<td>StyrofoamTM ball</td>
<td>Baby bottle</td>
</tr>
<tr>
<td>24</td>
<td>Metal fork</td>
<td>Ice cream scoop</td>
<td>Mechanical pencil</td>
<td>Toy Truck</td>
</tr>
<tr>
<td>25</td>
<td>Screwdriver</td>
<td>Piggy bank</td>
<td>Clothespin</td>
<td>Car air freshener</td>
</tr>
<tr>
<td>26</td>
<td>Paper airplane</td>
<td>Tape measure</td>
<td>Headphone set</td>
<td>Wrench</td>
</tr>
<tr>
<td>27</td>
<td>Calculator</td>
<td>Hair dryer</td>
<td>Swimming goggles</td>
<td>Cotton swab</td>
</tr>
<tr>
<td>28</td>
<td>Penny</td>
<td>Scissors</td>
<td>Plastic fork</td>
<td>Safety razor</td>
</tr>
<tr>
<td>29</td>
<td>Plastic straw</td>
<td>Bottle of nail polish</td>
<td>Carabiner</td>
<td>Loofah</td>
</tr>
<tr>
<td>30</td>
<td>Wine glass</td>
<td>Eraser</td>
<td>Miniature American flag</td>
<td>Pom-pom</td>
</tr>
<tr>
<td>31</td>
<td>Bucket</td>
<td>Pizza cutter</td>
<td>Feather duster</td>
<td>Paper plate</td>
</tr>
<tr>
<td>32</td>
<td>Pen</td>
<td>Roll of toilet paper</td>
<td>Pair of tongs</td>
<td>Styrofoam cup</td>
</tr>
<tr>
<td>33</td>
<td>Tiara</td>
<td>Quarters</td>
<td>Christmas stocking</td>
<td>Flask</td>
</tr>
<tr>
<td>34</td>
<td>Bottle cap</td>
<td>Baseball</td>
<td>Lipstick</td>
<td>Medical thermometer</td>
</tr>
<tr>
<td>35</td>
<td>Pencil sharpener</td>
<td>Lighter</td>
<td>Ping pong ball</td>
<td>Shoe</td>
</tr>
<tr>
<td>36</td>
<td>Coffee filter</td>
<td>Pair of chopsticks</td>
<td>Frisbee</td>
<td>Party blower</td>
</tr>
<tr>
<td>37</td>
<td>Compact Disc</td>
<td>Wine cork</td>
<td>Rubber duck</td>
<td>Coiled rope</td>
</tr>
<tr>
<td>38</td>
<td>Travel mirror</td>
<td>Leather sandal</td>
<td>Container of dental floss</td>
<td>Screw</td>
</tr>
</tbody>
</table>

117
Appendix D: Experiment 3 Letter Sequences

Letter sequences presented to participants during the working memory task. Participants assigned to the high load condition were given sequences whose length depended on their score on the working memory capacity assessment (see Appendix E).

<table>
<thead>
<tr>
<th>LOW LOAD</th>
<th>HIGH LOAD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1-Letter</strong></td>
<td><strong>3-Letter</strong></td>
</tr>
<tr>
<td>1</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>U</td>
</tr>
<tr>
<td>3</td>
<td>O</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
</tr>
<tr>
<td>9</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>H</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
</tr>
<tr>
<td>13</td>
<td>L</td>
</tr>
<tr>
<td>14</td>
<td>Y</td>
</tr>
<tr>
<td>15</td>
<td>X</td>
</tr>
<tr>
<td>16</td>
<td>X</td>
</tr>
<tr>
<td>17</td>
<td>I</td>
</tr>
<tr>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>19</td>
<td>Y</td>
</tr>
<tr>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>21</td>
<td>Q</td>
</tr>
<tr>
<td>22</td>
<td>R</td>
</tr>
<tr>
<td>23</td>
<td>L</td>
</tr>
<tr>
<td>24</td>
<td>R</td>
</tr>
<tr>
<td>25</td>
<td>O</td>
</tr>
<tr>
<td>26</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>27</td>
<td>J</td>
</tr>
<tr>
<td>28</td>
<td>U</td>
</tr>
<tr>
<td>29</td>
<td>I</td>
</tr>
<tr>
<td>30</td>
<td>Y</td>
</tr>
<tr>
<td>31</td>
<td>B</td>
</tr>
<tr>
<td>32</td>
<td>I</td>
</tr>
<tr>
<td>33</td>
<td>Q</td>
</tr>
<tr>
<td>34</td>
<td>L</td>
</tr>
<tr>
<td>35</td>
<td>R</td>
</tr>
<tr>
<td>36</td>
<td>Y</td>
</tr>
<tr>
<td>37</td>
<td>Q</td>
</tr>
<tr>
<td>38</td>
<td>X</td>
</tr>
<tr>
<td>39</td>
<td>Y</td>
</tr>
<tr>
<td>40</td>
<td>M</td>
</tr>
</tbody>
</table>
Appendix E: Experiment 3 & 4 Working Memory Capacity Assessment

Procedure:

- The experimenter reads a string of numbers aloud to the participant, who is asked to repeat the string of numbers aloud in reverse order.
- If the participant repeats back a given string correctly, the experimenter moves on to the first string of the next highest string length.
- If the participant fails twice at a given string length (in other words, if they fail to repeat both the first and second string correctly), the assessment ends.
  - The participant's score is the highest string length at which they repeated at least one of the two strings correctly.
  - If the participant succeeds in repeating one or more strings of length ten correctly, they are given a score of ten and the assessment ends.

Materials

<table>
<thead>
<tr>
<th>String Length</th>
<th>First String</th>
<th>Second String</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4 0 3</td>
<td>5 3 6</td>
</tr>
<tr>
<td>4</td>
<td>4 0 1 8</td>
<td>1 1 0 7</td>
</tr>
<tr>
<td>5</td>
<td>1 9 0 6 8</td>
<td>2 0 6 4 3</td>
</tr>
<tr>
<td>6</td>
<td>7 0 8 3 5 4</td>
<td>1 5 7 8 0 6</td>
</tr>
<tr>
<td>7</td>
<td>7 9 2 4 8 3 6</td>
<td>5 9 4 6 8 2 7</td>
</tr>
<tr>
<td>8</td>
<td>4 8 3 0 5 6 7 4</td>
<td>5 0 1 8 7 4 3 2</td>
</tr>
<tr>
<td>9</td>
<td>5 7 2 1 8 6 9 3 4</td>
<td>8 9 5 1 2 4 7 3 6</td>
</tr>
<tr>
<td>10</td>
<td>0 2 8 3 4 7 1 9 5 6</td>
<td>1 9 2 8 3 4 6 7 0 5</td>
</tr>
</tbody>
</table>
## Appendix F: Chapter 4 Displays

### Critical Displays

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Long pencil</td>
<td>Short pencil</td>
<td>Long toothbrush</td>
<td>Sunglasses</td>
<td>Computer mouse</td>
</tr>
<tr>
<td>Tall cup</td>
<td>Short cup</td>
<td>Tall soap dispenser</td>
<td>Highlighter</td>
<td>Hot glue gun</td>
</tr>
<tr>
<td>Small whisk</td>
<td>Large whisk</td>
<td>Small rock</td>
<td>Slipper</td>
<td>Soda</td>
</tr>
<tr>
<td>Small deodorant</td>
<td>Large deodorant</td>
<td>Small candy</td>
<td>Camera</td>
<td>Power strip</td>
</tr>
<tr>
<td>Large flashlight</td>
<td>Small flashlight</td>
<td>Large oven mitt</td>
<td>Yo-yo</td>
<td>Orange</td>
</tr>
<tr>
<td>Long spoon</td>
<td>Short spoon</td>
<td>Long hammer</td>
<td>Clock</td>
<td>Thimble</td>
</tr>
<tr>
<td>Small Post-Its</td>
<td>Large Post-Its</td>
<td>Small thumb tack</td>
<td>Peanut butter</td>
<td>Twine</td>
</tr>
<tr>
<td>Skinny marker</td>
<td>Fat marker</td>
<td>Skinny paint brush</td>
<td>Tissue box</td>
<td>Tennis ball</td>
</tr>
<tr>
<td>Tall jar</td>
<td>Short jar</td>
<td>Tall basket</td>
<td>Flip-flop</td>
<td>Beads</td>
</tr>
<tr>
<td>Small envelope</td>
<td>Large envelope</td>
<td>Small cell phone</td>
<td>Stick</td>
<td>Globe</td>
</tr>
<tr>
<td>Short candle</td>
<td>Tall candle</td>
<td>Short shot glass</td>
<td>Notepad</td>
<td>Stapler</td>
</tr>
<tr>
<td>Big crayon</td>
<td>Small crayon</td>
<td>Big chalk</td>
<td>Shell</td>
<td>Cocktail umbrella</td>
</tr>
<tr>
<td>Large hairbrush</td>
<td>Small hairbrush</td>
<td>Large bowl</td>
<td>Playing card</td>
<td>Medal</td>
</tr>
<tr>
<td>Large binder clip</td>
<td>Small binder clip</td>
<td>Large light bulb</td>
<td>Lego</td>
<td>Battery</td>
</tr>
<tr>
<td>Small funnel</td>
<td>Large funnel</td>
<td>Small Chapstick</td>
<td>Necktie</td>
<td>Squirt gun</td>
</tr>
<tr>
<td>Small leaf</td>
<td>Large leaf</td>
<td>Small snake</td>
<td>Hairbow</td>
<td>Watch</td>
</tr>
</tbody>
</table>
## Filler Displays

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>Fork</td>
<td>Mechanical pencil</td>
<td>Ice cream scoop</td>
<td>Toy truck</td>
</tr>
<tr>
<td>18</td>
<td>Screwdriver</td>
<td>Clothespin</td>
<td>Car air freshener</td>
<td>Piggy bank</td>
</tr>
<tr>
<td>19</td>
<td>Paper airplane</td>
<td>Tape measure</td>
<td>Wrench</td>
<td>Headphones</td>
</tr>
<tr>
<td>20</td>
<td>Hair dryer</td>
<td>Calculator</td>
<td>Swimming goggles</td>
<td>Q-Tip</td>
</tr>
<tr>
<td>21</td>
<td>Plastic fork</td>
<td>Penny</td>
<td>Scissors</td>
<td>Razor</td>
</tr>
<tr>
<td>22</td>
<td>Carabiner</td>
<td>Straw</td>
<td>Nail polish</td>
<td>Loofah</td>
</tr>
<tr>
<td>23</td>
<td>American flag</td>
<td>Wine glass</td>
<td>Pom-pom</td>
<td>Eraser</td>
</tr>
<tr>
<td>24</td>
<td>Pizza cutter</td>
<td>Feather duster</td>
<td>Bucket</td>
<td>Paper plate</td>
</tr>
<tr>
<td>25</td>
<td>Plate</td>
<td>Toilet paper</td>
<td>Pen</td>
<td>Styrofoam cup</td>
</tr>
<tr>
<td>26</td>
<td>Christmas stocking</td>
<td>Flask</td>
<td>Tiara</td>
<td>Quarter</td>
</tr>
<tr>
<td>27</td>
<td>Baseball</td>
<td>Lipstick</td>
<td>Thermometer</td>
<td>Bottle cap</td>
</tr>
<tr>
<td>28</td>
<td>Ping pong ball</td>
<td>Pencil sharpener</td>
<td>Shoe</td>
<td>Lighter</td>
</tr>
<tr>
<td>29</td>
<td>Rubber duck</td>
<td>Rope</td>
<td>Wine cork</td>
<td>CD</td>
</tr>
<tr>
<td>30</td>
<td>Dental floss</td>
<td>Screw</td>
<td>Sandal</td>
<td>Mirror</td>
</tr>
<tr>
<td>31</td>
<td>Gum</td>
<td>Cloth glove</td>
<td>Tupperware</td>
<td>Latex glove</td>
</tr>
<tr>
<td>32</td>
<td>Plastic Slinky</td>
<td>Safety goggles</td>
<td>Hand sanitizer</td>
<td>Metal Slinky</td>
</tr>
<tr>
<td>33</td>
<td>Zip-Loc bag</td>
<td>Narrow tape</td>
<td>Key</td>
<td>Wide tape</td>
</tr>
<tr>
<td>34</td>
<td>Picture frame</td>
<td>Black feather</td>
<td>Ice cube tray</td>
<td>White feather</td>
</tr>
<tr>
<td>35</td>
<td>Red towel</td>
<td>Blue towel</td>
<td>Dollar bill</td>
<td>Match</td>
</tr>
<tr>
<td>36</td>
<td>Tall glass</td>
<td>Nail clippers</td>
<td>Short glass</td>
<td>Tea bag</td>
</tr>
<tr>
<td>37</td>
<td>Gray sock</td>
<td>Glasses</td>
<td>White sock</td>
<td>Tweezers</td>
</tr>
<tr>
<td>38</td>
<td>Doll</td>
<td>Baby bottle</td>
<td>Styrofoam ball</td>
<td>Bracelet</td>
</tr>
<tr>
<td>39</td>
<td>Coffee filter</td>
<td>Chopsticks</td>
<td>Frisbee</td>
<td>Party blower</td>
</tr>
<tr>
<td>40</td>
<td>Paper bag</td>
<td>Nail</td>
<td>Glass mug</td>
<td>Ceramic mug</td>
</tr>
</tbody>
</table>
Appendix G: Experiment 4 Commands

Commands for critical displays. Only command 1 is critical command, and is the same in both reliable and unreliable conditions; Command 2 is a filler command and differs between the two conditions.

<table>
<thead>
<tr>
<th>Command 1</th>
<th>Command 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(same in reliable &amp; unreliable)</td>
<td><strong>Reliable</strong></td>
<td><strong>Unreliable</strong></td>
</tr>
<tr>
<td>1 Click on the long pencil.</td>
<td>Click on the sunglasses.</td>
<td>Click on the dark plastic sunglasses.</td>
</tr>
<tr>
<td>2 Click on the tall cup.</td>
<td>Click on the highlighter.</td>
<td>Click on the little shiny highlighter.</td>
</tr>
<tr>
<td>3 Click on the small whisk.</td>
<td>Click on the slipper.</td>
<td>Click on the boot.</td>
</tr>
<tr>
<td>4 Click on the small deodorant.</td>
<td>Click on the camera.</td>
<td>Click on the Nikon digital camera.</td>
</tr>
<tr>
<td>5 Click on the large flashlight.</td>
<td>Click on the yo-yo.</td>
<td>Click on the little shiny plastic yo-yo.</td>
</tr>
<tr>
<td>6 Click on the long spoon.</td>
<td>Click on the clock.</td>
<td>Click on the square plastic clock.</td>
</tr>
<tr>
<td>7 Click on the small Post-Its.</td>
<td>Click on the peanut butter.</td>
<td>Click on the big plastic peanut butter jar.</td>
</tr>
<tr>
<td>8 Click on the skinny marker.</td>
<td>Click on the tissue box.</td>
<td>Click on the toilet paper.</td>
</tr>
<tr>
<td>9 Click on the tall jar.</td>
<td>Click on the flip-flop.</td>
<td>Click on the big patterned flip-flop sandal.</td>
</tr>
<tr>
<td>10 Click on the small envelope.</td>
<td>Click on the stick.</td>
<td>Click on the long knobby stick.</td>
</tr>
<tr>
<td>11 Click on the short candle.</td>
<td>Click on the notepad.</td>
<td>Click on the flat blank notepad.</td>
</tr>
<tr>
<td>12 Click on the big crayon.</td>
<td>Click on the shell.</td>
<td>Click on the tiny speckled shell.</td>
</tr>
<tr>
<td>13 Click on the large hairbrush.</td>
<td>Click on the playing card.</td>
<td>Click on the flat playing card.</td>
</tr>
<tr>
<td>14 Click on the large binder clip.</td>
<td>Click on the Lego.</td>
<td>Click on the small two-pronged Lego piece.</td>
</tr>
<tr>
<td>15 Click on the small funnel.</td>
<td>Click on the necktie.</td>
<td>Click on the dark curled-up necktie.</td>
</tr>
<tr>
<td>16 Click on the small leaf.</td>
<td>Click on the bow.</td>
<td>Click on the bright polka-dot bow.</td>
</tr>
<tr>
<td>Reliable</td>
<td>Command 1</td>
<td>Command 2</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>17</td>
<td>Click on the fork.</td>
<td>Click on the ice cream scoop.</td>
</tr>
<tr>
<td>18</td>
<td>Click on the screwdriver.</td>
<td>Click on the piggy bank.</td>
</tr>
<tr>
<td>19</td>
<td>Click on the paper airplane.</td>
<td>Click on the tape measure.</td>
</tr>
<tr>
<td>20</td>
<td>Click on the calculator.</td>
<td>Click on the hair dryer.</td>
</tr>
<tr>
<td>21</td>
<td>Click on the scissors.</td>
<td>Click on the penny.</td>
</tr>
<tr>
<td>22</td>
<td>Click on the straw.</td>
<td>Click on the nail polish.</td>
</tr>
<tr>
<td>23</td>
<td>Click on the eraser.</td>
<td>Click on the wine glass.</td>
</tr>
<tr>
<td>24</td>
<td>Click on the bucket.</td>
<td>Click on the pizza cutter.</td>
</tr>
<tr>
<td>25</td>
<td>Click on the pen.</td>
<td>Click on the toilet paper.</td>
</tr>
<tr>
<td>26</td>
<td>Click on the tiara.</td>
<td>Click on the quarter.</td>
</tr>
<tr>
<td>27</td>
<td>Click on the bottle cap.</td>
<td>Click on the baseball.</td>
</tr>
<tr>
<td>28</td>
<td>Click on the pencil sharpener.</td>
<td>Click on the lighter.</td>
</tr>
<tr>
<td>29</td>
<td>Click on the cork.</td>
<td>Click on the CD.</td>
</tr>
<tr>
<td>30</td>
<td>Click on the mirror.</td>
<td>Click on the sandal.</td>
</tr>
<tr>
<td>31</td>
<td>Click on the gum.</td>
<td>Click on the latex glove.</td>
</tr>
<tr>
<td>32</td>
<td>Click on the safety</td>
<td>Click on the metal</td>
</tr>
<tr>
<td></td>
<td>goggles.</td>
<td>slinky.</td>
</tr>
<tr>
<td>---</td>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td>33</td>
<td>Click on the key.</td>
<td>Click on the Zip-Loc bag.</td>
</tr>
<tr>
<td>34</td>
<td>Click on the ice cube tray.</td>
<td>Click on the black feather.</td>
</tr>
<tr>
<td>35</td>
<td>Click on the match.</td>
<td>Click on the red towel.</td>
</tr>
<tr>
<td>36</td>
<td>Click on the tea bag.</td>
<td>Click on the nail clippers.</td>
</tr>
<tr>
<td>37</td>
<td>Click on the glasses.</td>
<td>Click on the white sock.</td>
</tr>
<tr>
<td>38</td>
<td>Click on the doll.</td>
<td>Click on the bracelet.</td>
</tr>
<tr>
<td>39</td>
<td>Click on the party blower.</td>
<td>Click on the chopsticks.</td>
</tr>
<tr>
<td>40</td>
<td>Click on the paper bag.</td>
<td>Click on the glass mug.</td>
</tr>
</tbody>
</table>
Appendix H: Experiment 3 & 4 Letter Sequences

Letter sequences presented to participants during the working memory task. Participants assigned to the high load condition were given sequences whose length depended on their score on the working memory capacity assessment (see Appendix E).

<table>
<thead>
<tr>
<th>LOW LOAD</th>
<th>HIGH LOAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 F RXI RXIJ RXIJL RXIJLH RXIJLHQ RXIJLHQB RXIJLHQBM</td>
<td></td>
</tr>
<tr>
<td>2 U UFX UFXY UFXYJ UFXYJO UFXYJOR UFXYJORH UFXYJORHI</td>
<td></td>
</tr>
<tr>
<td>3 O MYO MYOU MYOUR MYOURI MYOURIL MYOURILF MYOURILFH</td>
<td></td>
</tr>
<tr>
<td>4 R OIJ OIJM OIJMQ OIJMQL OIJMQLH OIJMQLHB OIJMQLHBF</td>
<td></td>
</tr>
<tr>
<td>5 X IJM IJMB IJMBX IJMBXY IJMBXYF IJMBXYFH IJMBXYFHL</td>
<td></td>
</tr>
<tr>
<td>6 B MLQ MLQB MLQBU MLQBUH MLQBUHO MLQBUHOJ MLQBUHOJI</td>
<td></td>
</tr>
<tr>
<td>7 M FXY FXYL FXYLQ FXYLQJ FXYLQJM FXYLQJMO FXYLQJMOB</td>
<td></td>
</tr>
<tr>
<td>8 F XQR XQRH XQRHF XQRHFL XQRHFLY XQRHFLYJ XQRHFLYJI</td>
<td></td>
</tr>
<tr>
<td>9 X XOM XOML XOMLR XOMLRQ XOMLRQH XOMLRQHB XOMLRQHBI</td>
<td></td>
</tr>
<tr>
<td>10 H QLY QLYJ QLYJX QLYJXR QLYJXRH QLYJXRB QLYJXRBMF</td>
<td></td>
</tr>
<tr>
<td>11 B YOB YOBF YOBFJ YOBFJH YOBFJHX YOBFJHXI YOBFJHXIM</td>
<td></td>
</tr>
<tr>
<td>12 M RYB RYBM RYBMO RYBMOF RYBMOFQ RYBMOFQH RYBMOFQL</td>
<td></td>
</tr>
<tr>
<td>13 L IHM IHMX IHMXJ IHMXJY IHMXJYB IHMXJYBQ IHMXJYBFQ</td>
<td></td>
</tr>
<tr>
<td>14 Y BIO BIOQ BIOQF BIOQFX BIOQFXL BIOQFXHL BIOQFXLHJ</td>
<td></td>
</tr>
<tr>
<td>15 X LUQ LUQI LUQIO LUQIOF LUQIOFH LUQIOFHB LUQIOFHB</td>
<td></td>
</tr>
<tr>
<td>16 X QJI QJIX QJIXM QJIXML QJIXMLR QJIXMLRO QJIXMLROH</td>
<td></td>
</tr>
<tr>
<td>17 I RIM RIMX RIMXY RIMXYH RIMXYHL RIMXYHJI RIMXYHJL</td>
<td></td>
</tr>
<tr>
<td>18 I UYJ UYJO UYJOQ UYJOQF UYJOQFB UYJOQFBH UYJOQFBHI</td>
<td></td>
</tr>
<tr>
<td>19 Y IHJ IHJM IHJMQ IHJMQR IHJMQRL IHJMQRLB IHJMQRLBF</td>
<td></td>
</tr>
<tr>
<td>20 F QUO QUOH QUOHF QUOHFB QUOHFBJ QUOHFBJL QUOHFBJML</td>
<td></td>
</tr>
<tr>
<td>21 Q HYU HYUO HYUOL HYUOLF HYUOLFJ HYUOLFJH HYUOLFJIM</td>
<td></td>
</tr>
<tr>
<td>22 R BYF BYFI BYFIL BYFILM BYFILMX BYFILMXR BYFILMXR</td>
<td></td>
</tr>
<tr>
<td>23 L LYM LYMU LYMUJ LYMUJH LYMUJHB LYMUJHBF LYMUJHBFI</td>
<td></td>
</tr>
<tr>
<td>24 R IOY IOYF IOYFL IOYFLB IOYFLBM IOYFLBMX IOYFLBMXH</td>
<td></td>
</tr>
<tr>
<td>25 O BLH BLHY BLHYF BLHYFX BLHYFXR BLHYFXRI BLHYFXRIJ</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>O</td>
</tr>
<tr>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>27</td>
<td>J</td>
</tr>
<tr>
<td>28</td>
<td>U</td>
</tr>
<tr>
<td>29</td>
<td>I</td>
</tr>
<tr>
<td>30</td>
<td>Y</td>
</tr>
<tr>
<td>31</td>
<td>B</td>
</tr>
<tr>
<td>32</td>
<td>I</td>
</tr>
<tr>
<td>33</td>
<td>Q</td>
</tr>
<tr>
<td>34</td>
<td>L</td>
</tr>
<tr>
<td>35</td>
<td>R</td>
</tr>
<tr>
<td>36</td>
<td>Y</td>
</tr>
<tr>
<td>37</td>
<td>Q</td>
</tr>
<tr>
<td>38</td>
<td>X</td>
</tr>
<tr>
<td>39</td>
<td>Y</td>
</tr>
<tr>
<td>40</td>
<td>M</td>
</tr>
</tbody>
</table>