ABSTRACT

Title of Dissertation: DEVELOPMENTAL PARSING AND COGNITIVE CONTROL

Zoe Loretta Ovans, Doctor of Philosophy, 2022

Dissertation directed by: Drs. Yi Ting Huang & Jared Novick, Department of Hearing and Speech Sciences, Neuroscience and Cognitive Science Program

Processing sentences incrementally entails making commitments to structure (and sometimes role assignments) before all information in a sentence is present. Children in particular have been shown to have difficulty revising the initial structural commitments they make when these turn out to be incorrect (Trueswell et al., 1999; Hurewitz et al., 2000; Weighall, 2008; Choi & Trueswell, 2010; Anderson et al., 2011). While prior research has generally ascribed this to limitations in the development of children's non-linguistic cognitive-control system, a precise account of *how* cognitive control limitations might lead to difficulty with incremental sentence processing is missing from the literature. In part, this is because existing research has focused on individual differences in children's ability to exert cognitive control over their thoughts and actions. In contrast, this dissertation makes use of within-child variation in cognitive-control engagement to provide evidence that children's domain-general cognitive-control system pushes them to rely more heavily on reliable parsing cues (and less heavily on unreliable ones) when the system is highly engaged. This conclusion brings together seemingly

disparate results from child and adult conflict adaptation studies, where adults appear to adapt to conflict but children do not. Overall, it is concluded that cognitive-control engagement leads both children and adults to re-rank parsing cues to attend more to ones that are more task-relevant, but the criteria they use to determine which cues are most relevant can change with language experience.

DEVELOPMENTAL PARSING AND COGNITIVE CONTROL

by

Zoe Loretta Ovans

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Yi Ting Huang, Chair
Professor Jared Novick, Co-Chair
Professor Colin Phillips
Professor Jan Edwards
Professor Jeffrey Lidz, Dean's Representative

Preface

The work presented in this dissertation is highly collaborative. The conception and design of the experiments presented in Chapter 2, Chapter 3, and Chapter 4, represent the joint intellectual effort of my advisors Yi Ting Huang and Jared Novick. In addition, this work has been supported by the National Science Foundation (Graduate Research Fellowship #DGE-1840340 and NRT award #DGE-1449815.

Acknowledgements

I have so many wonderful, brilliant people to thank for the work presented in this dissertation, both for their insightful thoughts and for keeping my heart full throughout my time in grad school and my academic career as a whole. Thank you for making me love what I do.

To my advisors, Jared & Yi Ting. My sincere and endless thanks to you both for embodying everything an advisor should be. Thank you for being careful scientists and mentors, for teaching me how to organize my thoughts and my data, for sharing with me your expertise in language processing and love of asking deep questions, and above all, for being kind souls.

Yi Ting, I feel like I can see your enthusiasm bubble over every time we meet. Thank you for always asking endless follow-up questions to everything I say. I've always appreciated how willing you are to both zoom in to focus on details (like exactly which word we should remove when our abstract is 501 words long), and zoom out to focus on the big picture (like figuring out whether experiments are actually answering the questions we set out to). On a personal note, thank you for remembering things about me that I don't even always remember about myself. You're one of the most thoughtful listeners I've ever met and I'm constantly blown away your ability to bring up small details about a conversation we had 4 years ago.

Jared, thank you for always being outwardly calm and collected, while also being so willing and excited to iron out gritty details and make sure we're saying exactly what we want to say. Basically every comment you've ever written on my

work has made me stop and go “Oh wow, why didn’t I think to write it like that??”. Thank you also for being the most caring advisor any graduate student could ever ask for. From insisting that I meet everyone you can possibly introduce me to at conferences, to waking up early to make sure I get one more round of feedback on a draft, to always making sure everyone has enough food. I can only hope to emulate your mentorship style for any future students I might have.

To Jeff Lidz, thank you for teaching me so much about language acquisition and, more generally, the importance of doing slow, careful research. When I was trying to figure out what I was interested in studying for grad school, for every phenomenon it seemed like you had either figured out the right answer 10 years ago or had written a completely convincing paper on why it never really existed in the first place. I’m so glad to have had the chance to learn from you.

To Colin Phillips, thank you for always being willing to question basic assumptions and never letting me get complacent. When I first came to Maryland you gave an address warning new students not to be like your cat: eager to go outside and then immediately anxious to come back in once again. I’m happy to report that I’m nothing like your cat. Happy as I was at Maryland, I’m ready and excited to apply everything you’ve taught me to future endeavors.

To Jan Edwards, thank you for many things. For being incredibly warm and welcoming, especially when I first came to UMD. For always keeping clearly in mind how you might expect my results to come out. For your valuable feedback about how to present my research and which statistical tests to run. And also for making me cry

at work that one time with a heart wrenching story about how we should all call our parents more.

To Al Kim, thank you for being a wonderful collaborator and mentor. Thank you for teaching me how to analyze EEG data, and for sharing your expertise in critically evaluating what brain signals might really mean and why we should care. I've always appreciated your careful, precise teaching style as well as how completely comfortable you are discussing any level of a project, from a particular line of analysis code to the highest-level point we're trying to make.

To other Maryland faculty who have given me great advice during lab meetings, LSLTs, colloquia, classes, or projects we worked on together: Ellen Lau, for always asking questions that seem innocuous but really get right at the heart of the matter. Naomi Feldman, for being a careful collaborator and staunch student advocate. And to Tonia Bleam, Valentine Hacquard, Bob Slevc, Matt Goupell, Nan Ratner, Stefanie Kuchinsky, and Rochelle Newman for insightful conversations and valuable feedback.

To Tara Mease, thank you for always being an awesome BU travel buddy, and for making sure I didn't accidentally mess up the ICS database. To Shevaun Lewis, thank you for helping with so many things, and always encouraging students to think deeply about what they want to do with their lives. And to Jasper Lewis, thank you for being a good boy and for never biting me. To Caitlin Eaves, for always being patient with me when I took forever to find receipts, and to Tess Wood, for somehow answering every email before I sent it.

To Pam Komarek, who does so many things that are absolutely essential to keep the spirit of the NACS program alive and thriving. As one small example: we actually have to kick students off the admissions committee after 3 years because it's so much fun. This is all because of Pam, who, along with being a delight to work with, really does the difficult logistical work, leaving students free to have a good time and get to know prospective students, and this leads to a more cohesive student body as a whole. As another example: I served as the student representative on the NACS colloquium committee for several years, and Pam was tireless and firm in making sure my voice was heard and in advocating for speakers students suggested. For these and many more things (including having the organizational skills of Marie Kondo), Pam will always have my gratitude and admiration.

To my previous advisors, mentors, and teachers who taught me why Cognitive Science is so fascinating and inspired me to pursue a career in it: Barbara Landau, Mike McCloskey, Emma Gregory, Paul Smolensky, Géraldine Legendre, and Justin Halberda. To Colin Wilson for introducing me to phonology and teaching me that linguistics is basically just solving fun puzzles all day. And especially to Akira Omaki. For your candid advising. For inspiring me and Tyler to go to UMD in the first place. For teaching me a great deal about syntax, language acquisition, and psycholinguistics. I'll always wish we could have just one more conversation.

To my NACS cohort, especially Uzma Javed, Chelsea Haakenson, and Diana Alkire, I'll miss our many Board & Brew dates where we never played and games because we were too busy catching up and chatting about our lives. And also to Shakiba Rafiee, the Kevins (Armengol & Schneider), Ta-wen Ho, Wanyi Liu, for

teaching me about walking, rats, mice, and that ferrets honestly think it's fun if you bowl them down a hallway.

To my HESP cohort, Julianne Garbarino, Allie Johnson, and Michelle Erskine. Thank you for teaching me about speech-language pathology and for making me feel at home when I first came to UMD. Michelle, thank you for being a dissertation writing pal and for always addressing emails to me "Dear Zoe" even though we're friends. Allie, thanks for convincing me to go to spin class and to go to trivia. Julianne, thank you for helping me wade through many stats classes, and, in my mind, for always being the voice of reason in any situation.

To the other NACS, HESP, and Ling students who made my time at Maryland joyful, even when we weren't allowed to see each other for over a year because of a global pandemic. To Laurel Perkins, thank you for being someone I look up to and try to model my behavior after every day, even if I fall far short. And for providing some tough competition during all of our cross-country game nights! To Kathleen Oppenheimer, thank you for being the best kind of person to work with, an incredible baker, and a great friend. Your passion for making sure everyone (especially clinicians!) understands the value of evidence inspires me, and I can't wait to see what you go on to accomplish. Thanks for always being willing to nerd-out with me. To Alex Oppenheimer, thanks for always looking incredibly cute and eventually learning my name. To Lauren Salig, for being a wonderful lab-mate and an absolute cornhole shark. To Zach Maher, for also being a great lab-mate and a fearsome trivia competitor. To Anna Tinnemore, for always asking insightful questions and for being a stellar virtual trivia host (I know everyone has opinions about who should take over

hosting Jeopardy but I think you could give them all a run for their money). To Amritha Mallikarjun and Chris Heffner, thank you for being role models to me, and thanks for the company on our trips to Psychonomics (including that time we got to hold a baby alligator!). To Erika Exton, for also being a great Psychonomics travel companion, and for also being upset that the museum of Montreal never told us why the beaver trade collapsed. To Julie Cohen, thanks for being so welcoming when I first came to UMD and for being a fellow lover of the underappreciated Casey's Coffee! To my former lab-mates, Alix Kowalski and Rachel Adler, thanks for showing me the ropes when I came to Maryland and for always being around to talk to (well, until you graduated. So selfish.). To Nina Hsu, thank you for being an amazing role model and for being willing to walk me through your data analysis pipeline so many times. To Adam Liter, thank you for the fun game nights, and for introducing me to Chessle even though it's really hard if you don't have a bunch of chess openings memorized. I'm sorry this isn't written in LaTeX. And also to Hanna Muller and Laura Bailey, letting us stay over and teaching us about salmonoids (even if they're properly called salmonids). To Mina Hirzel, Anouk Dieuleveut, Aaron Doliana, Rodrigo Ranero and Sigwan Thivierge, for being great cake bakers (and eaters), great friends, and great linguists. To Yu'an Yang, thank you for the glorious hotpot adventures. And to other linguists, SLPs, and cognitive scientists who were always a joy to talk to and who never failed to brighten my day: Paulina Lyskawa, Maxime Papillon, Phoebe Gaston, Christina Blomquist, Madison Buntrock, Arynn Byrd, Elizabeth Kolberg and my newest lab-mates, Sophie Domanski, Kelly Marshall, Tal Ness, and Val Langlois. I know you'll all do great things.

To the fabulous lab managers who have helped me to no end: Kerianna Fredrick, Allesandra Sanchez, and Rhosean Asmah. To the multitudes of undergraduate research assistants who have done hours upon hours of eye-coding, and corpus work, including Jenna Nelson, Abby Rosler, Alex Peller, Alexandra Heyl, Claire Crossman, Daniella Teixeira, Felicity Vasek, Gammon Gresham, Josie Black, Juliana Camponeschi, Kara Schmidt, Madison Lenhart, Michael Fein, Razan Ahmed, Emma Abid, Sarah Gagné, Emily Thomas and Jessica Contreras. Thank you very very much. This work couldn't have been done without your careful help.

To my lovely parents, Andrea and Donald Ovans. Thank you for teaching me how to learn. Dad, for your sage advice, and for always believing in me. And for always being willing to drive me anywhere I needed to go (in both senses). Mom, for always telling me to question every assumption and find the exception to every rule. And for reading (and, of course, editing) everything I send you with glee. I know you're reading this whole dissertation right now and will tell me all of your favorite parts. I love you very much for it. To my other parents, Stephanie and Calvin, Orsula, and Jeff, as well as my sisters Linnea and Becky. Thank you all for being endlessly supportive of all my endeavors. To my nieces and nephews, Katie, Charlie, Henry, Lola, and Jack, who are all more charming and fun to talk to every time I see you. Thank you all for making home home.

And finally, to Tyler. What can I possibly say? Thank you for reading and deeply improving every draft. For not just knowing how to phrase that email but also knowing exactly what I'm trying to say. For always being ready to deliver a full report of the funniest things that happened to you that day. For being perfectly willing

to lose sleep debating a paper for hours. For also being perfectly happy to lose sleep playing the most complicated board games we can find. For your boundless *joie de vivre*. For reminding me of what matters, and for being my best friend. I couldn't have done it without you and I certainly wouldn't have wanted to. Here's to a lifetime of curiosity, rousing banter, and new adventures together.

When you begin your Ph.D. many people will warn you to be prepared for what you're getting yourself into, but who could have known how much of an absolute blast it would be. Thank you all. You made it so much fun.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Children often make early commitments to sentence structure and sometimes fail to revise them, even after encountering late-arriving information that conflicts with their initial parse (Trueswell et al., 1999; Weighall, 2008; Choi & Trueswell, 2010; Huang et al., 2013; Omaki et al., 2014; Lassotta, Omaki & Franck, 2016). Prior research broadly ascribes this to children's limited cognitive-control development (Woodard, Pozzan & Trueswell, 2016; Choi & Trueswell, 2010; Mazuka et al., 2009), since children around the age of 5 consistently show delays in general tests of executive function (Diamond et al.,2007; Bunge et al., 2002; Davidson et al., 2006; De Luca & Leventer, 2010; Zelazo et al., 2015). However, there is now mounting evidence that children's ability to exert control over their own thoughts and actions varies with context (Larson et al., 2012; Iani, Stella, & Rubichi 2014; Ambrosi, Lemaire, & Blaye, 2016). Moment-to-moment changes in children's cognitive state can influence their ability to override a prepotent response to focus on a goal. Relatively little is known about *how* children's cognitive-control engagement state influences sentence processing, and the goal of this dissertation is to determine what that effect is.

To that end, this dissertation contrasts two hypotheses for the nature of how children's cognitive control system interacts with their language processing system. One possibility is that children have limited, easily-depleted cognitive-control resources, akin to classic models of working memory (Baddeley & Hitch, 1974; Anguera et al., 2012). This represents a standard view of why children have difficulty

revising their parsing decisions: When children fail to revise their initial commitments, it is because their cognitive-control system, subserved by underdeveloped frontal lobes that undergo protracted maturation, runs out of the necessary resources quickly (e.g. Woodard, Pozzan & Trueswell, 2016; Qi, Love & Fisher, 2020; Powell & Carey, 2017; Wehbe et al., 2020; Ryskin, Levy, & Fedorenko, 2020).

Under this "Depletion" hypothesis, encountering conflict is costly. This allows us to make a clear prediction about the consequences of encountering successive instances of conflict during information-processing. Namely, navigating conflict once (e.g. on a Stroop-like task) should have a negative impact when children encounter any subsequent instances of conflict (e.g. on a sentence-processing task), since some quantity of their conflict-resolution resource will be exhausted. Under this hypothesis, when children encounter Stroop-conflict before having to revise initial interpretations of language input, this should lead to worse revision as children face the task of navigating the conflict between their initially-built and final parses with a relatively shallower pool of control resources. Consistent with this, children who perform worse on some Stroop-like tasks have more difficulty revising temporarily ambiguous or "garden-path" sentences (Woodard, Pozzan & Trueswell, 2016; Huang et al., 2016; Qi, Love & Fisher, 2020; c.f. Huang & Hollister 2019). Other support for the Depletion view comes from the broader child development literature on children's Theory of Mind. In one case, for example, it has been found that "depleting" children's Executive Function resources by having them conduct a delay-of-gratification task leads to reduced performance on a later false-belief task (Powell &

Carey, 2017). That is, children who were asked to sit in a room and wait to play with a toy were subsequently worse at recognizing that a character in a story held a false belief. It should be noted, though, that these tasks are quite different from traditional tests of cognitive control, and may rely on other aspects of executive function such as working memory or attention span. The false belief task results also represent children's scores taken at a particular snapshot in time, and as such, these do not capture the real-time trial-by-trial dynamics of cognitive-control engagement that may be necessary to argue in favor of the Depletion hypothesis.

Contra the Depletion view, a second hypothesis is that children's cognitive control is best described not as a pool of resources, but as a dynamic system that can change to bias children to attend to task-relevant cues, as it does for adults. The difference between children and adults' performance, then, lies in the particular specification of which cues are relevant to the task at hand. This "Cognitive biasing" hypothesis makes a prediction that directly contrasts with that of the Depletion view: encountering information-processing conflict once can lead to improved performance when more conflict is encountered subsequently (e.g. during a ambiguous sentence-processing task). If, when children's cognitive control system is in an upregulated state it acts to boost the relative weight placed on task-relevant (reliable) cues, and these cues are also cues to revision, the Engagement view predicts that garden-path revision can be improved. Initial support for this hypothesis comes from work with adults, where cognitive-control engagement during parsing seems to help comprehenders attend to revision cues and revise subsequent garden-path sentences (Hsu & Novick, 2016; Hsu, Kuchinsky & Novick, 2021). Essentially, the Cognitive

Biasing view is that children's cognitive-control system can assist in sentence processing in much the same way that adults' can, allowing children to attend more to task-relevant cues when the system is more engaged, improving comprehension.

Further support for this view can be found in work on older children as well: Stroop tasks show that cognitive-control engagement temporarily makes the same child *better* at disregarding dominant but task-irrelevant cues (e.g., the word form in a Stroop task) in favor of task-relevant ones (e.g., ink color in a Stroop task) (Larson et al., 2012; Iani, Stella, & Rubichi, 2014, Ambrosi, Lemaire, & Blaye, 2016). While traditional approaches in support of the Depletion view have relied on relating individual differences in cognitive control capacity between children, to test the Cognitive Biasing view it will be necessary to manipulate the relative engagement status of children's cognitive control system, and observe the consequences of this manipulation on sentence processing within the same child.

To be explicit, the Depletion and Cognitive Biasing views make divergent predictions about the consequences of children's cognitive control system being in varying states. Under the Depletion account, children's cognitive control system is resource limited and is in a more depleted state when information conflict must be navigated several times in a row (e.g. by experiencing two incongruent trials of a Stroop task). Under the Cognitive Biasing account, children's cognitive control system serves to bias processing toward task-relevant cues, and is in a more engaged state after information-conflict is encountered. As a general method of distinguishing these two accounts, the predictions they make can be observed to come apart in situations where conflict is encountered twice in rapid succession. In order to bring

about just such a condition, the studies in this dissertation will make use of a "conflict adaptation" paradigm, wherein high- and low-conflict trials from tasks in disparate domains are interleaved (Gratton, G., Coles, M. G., & Donchin, E., 1992); Clayson & Larson, 2011).

To distinguish the two accounts, it will be necessary to tightly control the timing of cognitive control depletion/engagement before children conduct sentence-processing tasks. Finding worse performance at sentence revision may (potentially) be taken as evidence for the Depletion view, while finding improved performance at sentence revision when cues to revision are task-relevant will be taken as evidence for the Cognitive Biasing view.

Chapter 2 of this dissertation uses visual-world eye-tracking to test 5-year-old children's comprehension of active and passive sentences (e.g. "The cat will be quickly chas(ing/ed by) the dog"). In Experiment 1, these sentences will be temporarily-ambiguous as verbs will reliably signal role assignment, but occur late. In contrast, early word-order cues (i.e., children's bias to assume the first NP in an utterance will be agentive) imply correct roles for actives but not for passives (Huang & Arnold, 2016; Abbot-Smith et al., 2017; Huang, Leech & Rowe, 2017). Across trials, these sentences are interleaved with a child-friendly Stroop task. Based on prior work (Hsu & Novick, 2016; Huang et al., 2016), incongruent Stroop trials engage cognitive-control more than congruent Stroop trials do. If children's cognitive-control resources are easily depleted, prior engagement should *hinder* comprehension of passives, but not actives. If prior engagement helps children ignore unreliable parsing cues, it should *improve* comprehension of passives that generate an agent-first bias.

The agent-first bias was chosen as the cue to initial structure building since while children reliably use it, it has been shown that children as young as 3 are able to overcome it to interpret passive sentences (Abbot-Smith et al., 2017). In contrast, the morphosyntactic cues provided by the verb in this study always provide unambiguous evidence to argument role mapping. Experiment 2 attempts to replicate these results, but while Experiment 1 uses a novel word identification paradigm, this follow-up uses known nouns to reduce children's working memory burden during the task. Experiment 3 attempts to isolate baseline effects of cognitive-control engagement. It uses similar sentences to Experiment 1 but with materials known *not* to lead children to rely on an agent-first bias. Across these studies, this work aims to show that cognitive-control engagement does not uniformly make children worse at navigating sentential ambiguity, but rather engages a system that can help them ignore unreliable parsing cues, as it does for adults.

Chapter 3 expands on the findings of the first three studies. It attempts to define more specifically what sentence-processing cues children are up-weighting when their cognitive control system is more highly activated. In Experiment 4, children are presented with sentences that vary only in the extent to which verbs are strong predictors of upcoming sentence structure. Following cognitive-control engagement, children are more likely to parse sentences according to the cues from the more strongly-biased verbs, indicating that cognitive-control engagement increases use of a parsing heuristic that increases reliance on processing cues that are themselves more reliable predictors of upcoming structure. Experiment 5 further zooms in on the results of Experiment 4 by determining whether children are

attending more to the bias of individual verbs or the bias that most English verbs follow, more generally. Children were presented with modifier-biased verbs, which differ from the overall bias of verbs in English, to determine whether cognitive control engagement pushes them to use the word-specific bias or the category-general one. As these studies make use of a novel "virtual-world" eye-tracking paradigm, Experiments 6 & 7 are methodical validation studies, replicating well-known eye-tracking effects using this method in an effort to demonstrate that it differs little from lab-based eye-tracking.

Chapter 4 consists of two corpus analyses of child-directed speech with the aim of empirically quantifying "reliability" as it's used in the prior studies. The goal of Experiment 8 is to measure the reliability of each verb in Experiments 4 & 5, based on the input that children have heard, as opposed to relying on adult Cloze task norming data. Each verb was coded for the percentage of times it predicts an upcoming with-phrase to attach to the VP or NP. If children are using a reliability heuristic, this predicts that the more consistent the verb, based on this measurement, the more likely children will be to rely on this verb's bias following cognitive-control engagement. Further, if the strength of the bias in child-directed speech correlates with the extent to which children rely on the verb bias following cognitive control engagement in Chapter 3, this will provide further evidence that children rely more strongly on information from verbs when their cognitive-control system is upregulated because they are reliable predictors of the sentence structure they are about to hear. Experiment 9 is a second corpus analysis of child-directed speech: Prior work has shown that for "put" imperatives with PP attachment ambiguity (e.g.

"Put the frog on the napkin into the box"), cognitive-control engagement impairs children's online processing abilities (Huang et al., 2016). This may be because children regard the initial verb as a highly reliable cue that the first PP will attach to the VP and will be a location for the putting event. Thus, children may show depletion-like results but for a different reason: when their cognitive control system is engaged they commit to the parse suggested by "Put" because it's an ordinarily reliable cue. The goal of Experiment 9 is to compare, given children's input, the relative reliability of the verb "put" in predicting a location for the putting event in the proximate PP, versus the reliability of children's agent-first bias. If "Put" is a significantly more reliable parsing cue than the expectation that initial NPs will be agents, this will lend credence to the claim that cognitive-control engagement does indeed lead children to rely more on cues that are more reliable, and less on cues that are less predictive of upcoming structure.

1.1: Overview

To address the question of how children's cognitive control system interacts with the developing parser, prior research has targeted an area of sentence processing where cognitive control is likely to be most obviously necessary: the interpretation of ambiguous sentences. As Sections 1.2 and 1.3 outline, both adults and children process sentences incrementally, meaning that they make commitments to structure and/or role assignments before an entire utterance is heard. When this commitment turns out to be incorrect, a potential conflict signal is generated. In order to correctly interpret the sentence, the "initial guess" representation the listener built must be suppressed, and they must instead build a new representation of the sentence based

upon the late-arriving information. To the extent that they differ, these early and late versions of the sentence may generate representational conflict that requires mediation by the comprehender's nonlinguistic cognitive control system.

As Section 1.3.2 details, a spate of evidence demonstrates that children have relative difficulty exercising their cognitive control system in non-linguistic tasks when one mental representation has to be ignored in favor of a different, task-relevant one. How then, does this difficulty in exercising cognitive control affect real-time sentence processing? Two possibilities will be discussed, each presenting a different picture of how children's cognitive control system itself is structured: under the Depletion View, the cognitive control system is resource-limited and conflict mediation during sentence processing involves using up some portion of this resource, leaving less for subsequent mediation of representational conflict. For children, then, successful execution of cognitive control is difficult because they have less of this resource to begin with. One issue with this hypothesis is that it presents a disconnect between how the child and adult cognitive-control systems function. In contrast, under the Cognitive Biasing view, the cognitive-control system is not resource-limited, and conflict mediation instead involves putting the system in a particular state where sources of evidence that are goal-relevant are more highly attended to. Under this view, successful execution of cognitive control is difficult for children when they either fail to reach this state or fail to choose the cue that is truly most task-relevant at the time.

1.2: Incremental sentence processing and cognitive control in adults

In this section, relevant literature establishing both incremental sentence processing and cognitive control in adults is introduced in subsections 1.2.1 and 1.2.2, respectively. Subsection 1.2.3, then presents evidence that domain-general models of cognitive control, such as the Botvinick (2001) model of Conflict Monitoring in the Stroop task, can be applied to linguistic tasks that require cognitive control, despite these tasks making use of representations in separate domains.

1.2.1: Incremental sentence processing in adults

A wealth of evidence over the last several decades has shown that adults process sentences incrementally (Altmann & Steedman, 1988; MacDonald, Pearlmutter, & Seidenberg, 1994; Altmann & Kamide, 1999; Kamide, Altmann & Haywood, 2003, *inter alia*). We construct likely parses for the sentences we hear as we take in each new word, instead of hedging our bets and waiting until they're over. Evidence for this comes from studies of sentences that violate these rapidly-built parses: longer reading times are found for "garden-path" sentences that violate comprehenders' syntactic expectations than for ones that follow them, particularly at the point where the sentence is disambiguated and the true structure is revealed. For example, Frazier & Rayner, (1982) presented participants with sentences like "Since Jay always jogs a mile seems like a very short distance to him." If listeners used a non-parallel parsing strategy, e.g. they opted to attach incoming words to the phrase they're currently processing, they would initially interpret "a mile" to be the direct object of the verb "jogs." Under this interpretation, the following word "seems" is

incongruous. To correctly interpret the sentence, listeners would then have to reanalyze the NP "a mile" as the subject of the next clause. They observed longer reading times at the point of disambiguation ("seems") for garden-path sentences, compared to minimally different sentences that contained no ambiguity (e.g. "Since Jay always jogs a mile *this* seems like a very short distance to him"). Here, if comprehenders were to wait until the end of a sentence to commit to its structure, no such slow-downs would be predicted at the point of disambiguation when an initially-likely structure is ruled out. Thus, these results have been taken as a clear demonstration that listeners make temporary commitments to sentence structure online, as they're hearing a sentence, even at the risk of these guesses being incorrect and having to be revised.

Subsequent work using visual-world eye-tracking has demonstrated that when adults process temporarily ambiguous sentences, they commit to a particular semantic analysis of the sentences they hear, and must revise this initial commitment if late-arriving disambiguating information proves it to be incorrect. Tanenhaus et al. (1995) demonstrated this by presenting participants with garden-path sentences that contained prepositional-phrase attachment ambiguity such as "Put the apple on the towel in the box," and minimally different unambiguous phrases that contained an overt complementizer, e.g. "Put the apple that's on the towel in the box." They simultaneously presented listeners with corresponding visual scenes, such as one containing an apple on a towel, an empty towel, a box, and a pencil. An analysis of participants' eye-movements as they heard the sentences revealed that upon hearing the ambiguous phrase "on the towel," participants in the ambiguous condition looked

to the empty towel 55% of the time, but participants in the unambiguous condition rarely looked at the empty towel. This demonstrated that listeners make quick assumptions about phrase attachment in real-time, as they process the sentences they hear.

Expanding on this work, Pickering & Traxler (1998) demonstrated that plausibility of the initial analysis plays a role in how strongly comprehenders commit to it. They presented participants with sentences that contained subordinate clause ambiguities, such as "As the woman edited/sailed the magazine about fishing amused the reporters." Sentences were constructed so that the noun following the initial verb was either a plausible match or mismatch as a direct object for that verb. These sentences were then compared to identical ones that contained a comma after the initial verb, and were therefore unambiguous. They tracked participants' word-by-word reading time and found slow-downs following the disambiguating verb (e.g. "amused") regardless of plausibility, but also found that this slow-down was greater in the plausible condition, where the initial mis-analysis was more tempting. These results provide further evidence that comprehenders parse sentences incrementally. They moreover show that while plausibility has an effect on the relative strength of the commitment to an incorrect parse, the pressure to parse in an incremental manner is great, even causing adults to trundle through implausible scenarios to do so.

The existence of temporary commitments to sentence structure that must be revised when they are incorrect presents a clear need for a mental system that is able to revoke activation of an initially promising representation in favor of another one that is more relevant for the task at hand. As the next section outlines, this is precisely

the job that is thought to be carried out by the sub-system within executive function known as the cognitive control system. It is reasonable, then, to assume that the process of mediating between conflicting sentence representations is done by the general system that performs cognitive control in other domains, such as visual processing. The question of domain generality of the system that mediates conflict in language has been the topic of some debate though, and will be returned to in Section 1.2.3.

1.2.2: The adult cognitive control system

Cognitive control has long been the name attributed to the mental system that mediates disputes between conflicting mental representations. Hammond & Summers (1972) introduced this system and provided a framework for how this system might be organized. They characterized cognitive control as "the extent to which the subject controls the execution of his knowledge" and took it to be a measure of how predictably an individual would make a particular response Y when given a particular cue X, irrespective of (or assuming perfect knowledge of) the parameters of the task. This defined a role for cognitive control as separate from other executive functions like memory, but fell short of specifying a mechanistic model for how it might work. Norman & Shallice (1986) proposed an account of what they called "attentional control" in which two tasks might interfere with each other in two different ways. One way was through "structural interference," wherein two tasks require the same processing structures, and the other they called "attentional interference," when two tasks compete for the same attentional resources. Baddeley & Della Salla (1996) characterized cognitive control as living within the central executive component of

Baddeley & Hitch (1974)'s well-known model of working memory. While these studies offer a picture of what the components of cognitive control might be and how it might fit into executive function more broadly, a common limitation of them is that they do not specify how control might be used to overcome interference within a domain or on any particular tasks.

Since then, the needle has perhaps swung too far in the opposite direction. Many subsequent theories of cognitive control have been closely, if not irrevocably, tied to particular tasks used to test it. In a notable paper, Botvinick et al. (2001) proposed detailed models of the mental processes underlying several "classic" control tasks, including the Stroop task (Stroop, 1935), Stem completion (Reicher, 1969), and the Eriksen Flanker task (Eriksen & Eriksen, 1974). Taking their model of the Stroop task (first proposed by Cohen & Huston, 1994) as an example, they characterize cognitive control (or conflict monitoring, as they call it) as a component that acts upon a task demand sub-system (Figure 1.1). Specifically, when the Stroop task is performed, the mental system that processes the relevant information contains units that encode word meanings as well as individual color meanings. These then have weighted connections to a response system that also contains multiple options corresponding to the various colors available in the response set. Once the response is called for, the dual activation of multiple options in the response set sum together to increase activation of a conflict monitoring node. Top-down control is then executed by increasing the weight of the node that corresponds to the goal of the task within the task-demand sub-system (in this case, the color of the word and not the word itself). In model simulations, they found that increasing activation of this color node

decreased interference in the model, and led to lower total activation being sent to the conflict monitoring node. In contrast, decreasing activation of the task-demand color node led to more interference and higher total activation being sent to the conflict monitoring node. Over and above previous accounts, this model provides a reasonable explanation for how conflict is generated in a Stroop-like task, and how cognitive control might influence domain-specific representations in order to overcome that conflict.

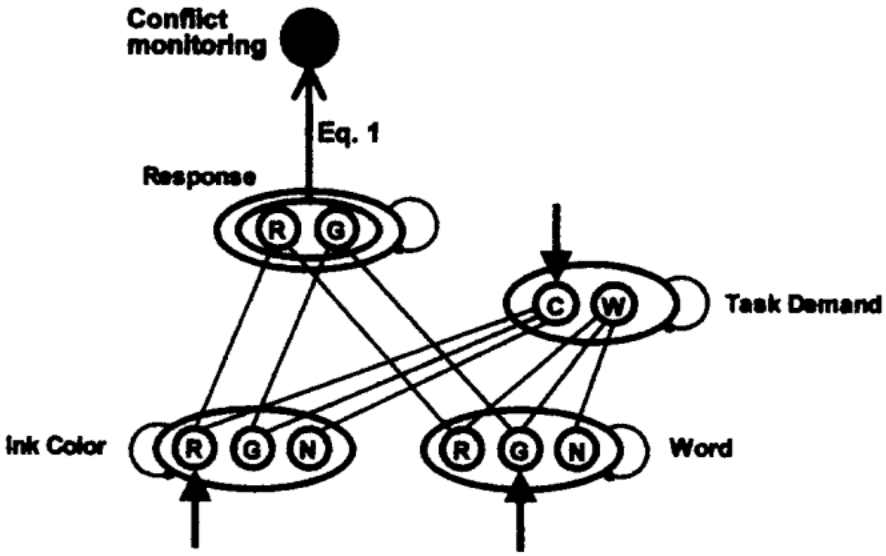

Figure 1.1: Model of conflict monitoring in the Stroop task, from Botvinick et al., 2001. Originally proposed in Cohen & Huston, 1994

Since 2001, various additional accounts of cognitive control execution have been proposed. Braver, Gray, & Burgess (2007) and Braver (2012) proposed a Dual Mechanisms of Control (DMC) framework, in which there are two different "operating modes" of cognitive control: proactive and reactive. Proactive control is the term used to describe sustained up-regulation of goal-relevant representations

within the cognitive control system in anticipation of its necessity on an upcoming task. Conversely, reactive control refers to transient activation of these representations that occurs in reaction to bottom-up stimuli. For example, in a Stroop task, participants may use proactive control to actively keep the task goal in mind (attend to stimulus color or ignore stimulus word) prior to seeing the word in question. They may use reactive control to do the same after seeing the stimulus word, but without having to actively maintain the task goal during the course of the trial.

Evidence for this division of control modes comes from neuroimaging studies that show differential prefrontal cortex activation when conflict is more or less expected. Burgess & Braver (2010) presented participants with a Recent Probes task in which, during a given trial, participants see a set of letters flash on the screen and must verify whether a new letter was in that set or not. On "recent-negative" trials, the letter in question is not in the set for that trial, but was in the set for the previous trial, creating an interference effect. Burgess & Braver varied the frequency of these recent-negative trials and found that this change affected activation patterns in the left inferior prefrontal cortex. When recent-negative trials were common (and therefore conflict between trials was more likely to occur), participants exhibited increased lateral PFC activity during the interval between the probe (the letter display) and the target letter. When recent-negative trials were rare (and therefore conflict between trials was relatively unlikely), they observed lateral PFC activity only in response to the probe. This, the authors conclude, provides evidence for both a sustained,

proactive control and a transient mode of control that responds to stimulus conflict only after it has been encountered.

Recently, other mechanisms for cognitive control have been proposed that subdivide the cognitive control system in various other ways. Koechlin, Ody, & Kouneiher (2003) separate the cognitive control system into episodic, contextual and sensory control. Van veen & Carter (2006) propose a distinction between evaluative and executive sub-systems. Badre (2008) outlines a multi-level hierarchy within the cognitive control system that incorporates Koechlin's model but adds a fourth layer dubbed "branching control."

While there is no prevailing overall model of the human cognitive control system, there is general consensus that a mental system exists that works to mediate representational conflict when it occurs. This system can be distinguished from other executive function systems such as working memory or attention, despite sometimes being characterized as a component that interacts with these other frontal systems. While various proposals have then further attempted to characterize the sub-components of cognitive control itself, it is useful to consider the aspects of the models that have been proposed that might apply to the task of sentence processing. The Botvinick model of the Stroop task, for example, provides a helpful starting point for imagining how cognitive control might play a role in navigating ambiguity during parsing. To begin with, though, it first necessary to establish that the sort of cognitive control needed for ambiguity processing does indeed make use of the same domain-general system described in the non-linguistic cognitive control literature.

1.2.3: Cognitive control and parsing in adults

To answer the question of how domain-general cognitive-control interacts with sentence processing, a first step has been to establish that the type of cognitive control needed for comprehension, in both adults and children, is indeed domain-general at all, and not, as has been claimed, a language-specific subsystem (Vuong & Martin, 2014; Wehbe et al., 2021; Ryskin, Levy, & Fedorenko, 2020). An initial way of showing this has been to see if performance on sentence processing tasks that might be expected to require recruitment of cognitive control correlates with performance on other tasks that we also expect to need recruitment of cognitive control, but are not linguistic in nature. When performance on these two tasks correlates, it can be taken as evidence that the same underlying system is being recruited for both. When it does not, a few studies have argued that this provides evidence that different underlying systems are being recruited.

Vuong & Martin (2014) attempted to show just such a correlation. They presented adults with garden path sentences that contained nouns with temporary subject/direct object ambiguity (e.g. "While the man coached the woman attended the party") and used plausibility ratings as a measure of how well participants had successfully recovered from the garden-path. In addition, participants were given a verbal and non-verbal Stroop task. In the verbal Stroop, the task was presented in the classic manner, with participants orally naming the ink color of a visually presented word. The non-verbal Stroop task was similar to a Simon task – participants were instructed to point in the direction of a visually-presented arrow, while ignoring the arrow's physical location on the left or right of their screen. The authors found that

performance on the verbal but not the non-verbal Stroop task correlated with garden-path recovery, leading them to conclude that the cognitive-control system implicated in sentence processing is language-specific.

More recently, Wehbe et al., (2021) sought to draw a similar conclusion using a combination of behavioral and functional neuroimaging data. They presented participants with a series of naturalistic short stories and articles. As participants read, self-paced reading, eye-tracking or fMRI data were tracked (for each method a separate participant pool was used). In particular, in the fMRI task, the authors compared the response profile when participants read these stories to areas of the brain that were more active for sentences than nonwords (as a localizer for language networks) and areas that were more active for nonwords than sentences (as a localizer for executive function networks). They found that the reading task activated areas of the brain that were also more active for sentences than nonwords, and failed to activate areas of the brain that were more activated for non-words than sentences. Moreover, behavioral reading-time slowdowns correlated with an increase of activity in the language-selective network they identified, and not in the executive function (or multiple demand) network. This, they argue, provides evidence that language comprehension does not make use of domain-general executive function at all, but rather recruits a language-specific system for executive functions like cognitive control.

While these results are bold, at various levels, these claims may be overstated. To begin with, using brain areas that are differentially more active in response to sentences than nonwords as a functional localizer will, of course, identify all areas

that underlie processing that is needed more for sentences than for nonwords. This would include areas that are required for meaning extraction, but also would include areas required for combing parts into a whole, storing information in memory, and ignoring incorrect sentence parses. Any of these processes would reasonably be expected to be recruited more for sentence than for nonword processing. There is therefore little reason to believe that the language network they identify would include only processes that are language-specific. Additionally, areas they identify as being a part of the language-specific network include those commonly thought to underlie cognitive control processing, such as the left-inferior frontal gyrus.

There are, therefore, strong reasons to doubt the conclusions that there is a language-specific sub-system within our cognitive control system that is implicated during parsing but not other tasks where information processing conflict must be mediated. More generally, it is unclear what exactly this sub-system would and wouldn't encompass. For the verbal Stroop task from Vuong & Martin, conflict arises at the level of competing representations of individual color words. It is unclear which representations this sort of conflict and garden-path recovery share that are not shared by conflict in a visuo-spatial Stroop task. Additionally, the lack of correlation with the non-verbal Stroop task may be due to a lack of power to detect a one, even if errors on both tasks are caused by a failure in the same underlying system (for a recent overview of issues with trying to observe individual differences in paradigms that weren't designed to detect them, see Hedge, Powell, & Sumner, 2018).

To skirt these empirical concerns, several studies have used a design whereby sentence processing and non-linguistic trials are interleaved in such a way that they

form pairs, and performance following incongruent trials is compared to performance following congruent ones. This paradigm has been referred to as "conflict adaptation" (Gratton, Coles, & Donchin, 1992; Botvinick et al., 2001; Ullsperger, Bylsma, & Botvinick, 2005), after the classic finding that performance on an incongruent trial generally improves following another incongruent trial, as opposed to following a congruent trial. In Botvinick et al. (2001)'s model, this finding reflects a mental biasing mechanism that, when activated, helps participants meet task demands. Importantly, the experimental manipulation of cognitive-control engagement inherent in this design also allows for a causal inference to be made: if performance on trial N-1 leads to improved mediation of conflict on trial N, the two tasks must have made use of a common mental system. To the extent that these two tasks share only a common need to ignore one already-built representation in favor of a new, more task-relevant one, this indicates that the tasks on these two trials made use of the same underlying system to resolve conflict. If two separate control systems underlie the tasks on each trial, we would not expect use of one system to affect performance on the other. Note that these analyses are always done on incongruent trials and compared to congruent conditions, to rule out the possibility that the carry-over is due to low-level attentional or motivational processes (such as engagement with the task).

In this design, it is crucially the experimental manipulation of cognitive control that allows the causal inference to go through. Instead of relying on participants to have naturally-varying cognitive control ability (as correlational tasks do), this design allows for the minimal change across conditions to be limited to changes in how much conflict a previous trial generates. It's worth noting here that

both types of tasks – correlational studies and conflict adaptation paradigms – are not mechanism-neutral. They each implicitly assume a particular notion for how the human system of cognitive control works. Correlational studies by their nature represent an attempt to measure cognitive control as a stable trait that is expressed to different (but internally consistent) extents in different populations or individuals when they perform tasks that require it. Conflict adaptation tasks, on the other hand, attempt to measure cognitive control as a processing state, the expression of which may change dynamically over time or with sufficient prompting from interference-heavy input. So while neither method is a-priori neutral without knowing more about the nature of our cognitive control mechanism, it's notable that the "state" framing potentially poses more of an issue for the "trait" framing than vice versa. That is: if an individual's level of activation of their cognitive control system varies greatly on a moment-to-moment basis, this will make it difficult to correlate an individual's cognitive control "level" on one task to another, since each task might reflect a snapshot at an activation peak or trough, leading to weak or inconsistent correlational data. If, however, cognitive control is relatively stable within an individual, this should be readily apparent in conflict adaptation data. That is, adaptation to conflict should be immensely difficult to observe across the board. Put another way, in correlational studies, being incorrect about the stable "trait" assumption will lead to underestimation of the extent to which two tasks correlate, whereas for conflict adaptation studies, being incorrect about the dynamic "state" assumption would be readily apparent, as it will lead to a lack of conflict adaptation effects, even within one domain.

By now, a series of studies has used the conflict adaptation paradigm with adults to investigate the effect of cognitive control engagement on real-time sentence processing. Hsu & Novick, (2016) interleaved garden-path sentence comprehension "put" trials with a classic color-word Stroop task, and Hsu, Kuchinsky & Novick (2021) did the same with Eriksen Flanker trials. Temporarily ambiguous garden path sentence comprehension is improved (as indexed by correct actions, looks to the correct goal and looks away from the incorrect goal) following incongruent Stroop trials, and similar results were found following incongruent Flanker trials. Importantly, the conflict adaptation effect was not found following similarly difficult Flanker-style trials that required sustained attention but were not designed to engage participants' cognitive-control system. Participants were presented with a trial that required them only to pay attention and press a button when arrows on the screen changed direction, but did not require them to down-weight an irrelevant stimulus. In this case, it's assumed that attentional control was up-regulated, but the difficulty of this task did not lead to a difference in performance on the subsequent sentence-processing task. In these studies, then, it is specifically the upregulation of cognitive control and not sustained attention that leads to improved performance (or adaptation) when encountering subsequent conflict during sentence processing.

These results also show carry-over from the non-linguistic tasks that require execution of cognitive control to a sentence processing task, when the sentences require re-parsing, but not for minimally different sentences that do not. This indicates that the system that resolves conflict during sentence processing is domain-general, since it is the same one that is implicated in non-linguistic tasks. Importantly,

up-regulating this system by completing an incongruent trial of a non-linguistic task has a direct impact on subsequent sentence processing performance, a result that allows us to draw a causal link between the two processes.

1.3: Incremental sentence processing and cognitive control in children

1.3.1: Incremental sentence processing in children

Over the last few decades, a series of visual-world eye-tracking studies have shown that children also process sentences incrementally. In a seminal set of experiments, Trueswell et al. (1999) presented children with temporarily ambiguous garden-path sentences, in a visual-world eye-tracking task similar to Tanenhaus et al. (1995). Children were presented with sentences such as "Put the frog (that's) on the napkin into the box" while their actions and eye-movements were measured. Children exhibited a so-called "kindergarten-path" effect: they were misled by their initial parse of the sentences and had pronounced difficulty reaching a final, adult-like interpretation. Children had significantly more difficulty interpreting temporarily ambiguous sentences without the overt complementizer "that's" than the unambiguous version, as seen in both their online eye-movements and offline act-out actions. Specifically, on approximately 50% of trials, children made incorrect "hopping" actions, moving a frog to an empty napkin before moving it to an empty box, indicating that they established an initial VP-attachment parse of the target sentence, and did not fully re-analyze the sentence so that the PP "on the box" attached to the NP, modifying the frog. Similarly, children looked significantly longer to the empty napkin even after the sentence was disambiguated, compared to the

unambiguous overt-complementizer condition (and also when compared to adults), indicating that they had not successfully abandoned their initial parse.

This work provided evidence that the child parser, like its adult counterpart, eagerly builds structure from pieces of the whole. With each new word, children's sentence processing mechanism, like adults', makes guesses about overall syntactic structure. Subsequently, various studies have both corroborated and expanded upon these results.

Hurewitz et al. (2000) conceptually replicated these findings, and further extended them to show that young children will happily use restrictive modifiers in production, given appropriate context. They presented five-year-old children with an identical set-up, but prior to the "put" task asked them questions that encouraged them to use restricted modification. Children readily did so in the production task but continued to avoid a modification reading in the "put" task. For example, they were asked questions like "Which frog went to Mrs. Squid's house?" and, after watching a short vignette with two frogs, readily answered "the one on the book." However, these same children still tended to continue looking at the empty napkin when told "put the frog on the napkin into the box." These results provide evidence that children's difficulty with "put" sentences is not attributable to difficulty with restricted modification or scene-referential knowledge in general, but rather to the type of mental revision procedures necessary when interpreting sentences that are temporarily syntactically ambiguous. In particular, the authors suggest that children's difficulty with these sentences stems from their reliance on the "highly reliable verb

preference" for *put* to take a locative PP, as well as their more general resistance to dropping their current syntactic analysis in favor of a new one.

Weighall, (2008) also directly replicated the findings of Trueswell et al., (1999), and extended them to show that they are robust to contextual interference: children continue to make errors even when contextual cues support an NP-attachment interpretation (e.g. "Put the frog on the red napkin into the box" where the frog in question is on a red napkin, and the open napkin is a different color). These results are consistent with children relying on the cue they receive from the verb "put" (that it must take a goal) at the expense of other, potentially less reliable contextual information, such as how speakers choose to modify noun phrases. These results indicate that, even in the presence of a highly suggestive context, children continue to focus exclusively on particular parsing cues at the expense of others. To the extent that these parsing cues conflict and that conflict has to be mediated to reach a final interpretation for the sentence, children's developing system of cognitive control is implicated.

1.3.2: Cognitive control development in children

Parallel to the literature on developmental parsing, a great deal of work has shown that in non-linguistic tasks, children have difficulty overcoming their initial mental commitments. In particular, school-age children have been shown to have great difficulty inhibiting a prepotent response in the face of evidence that it is not the correct one for a task at hand. In a large study of children and adults from age 4 to 45, Davidson et al. (2006) demonstrated this by measuring performance as a function of age on various tests of cognitive-control execution. Tasks included Simon spatial

incompatibility (on incongruent trials participants had to press a left button when a circle was on the right side of their screen), and used either arbitrary stimuli (e.g. circles) or iconic stimuli (arrows). These tasks were designed to vary memory load and inhibitory control demand separately. The authors found that, while even the youngest children had above-chance accuracy on incongruent trials, on all inhibition tasks accuracy increased with age and reaction times decreased. These results indicate that children have more difficulty than adults in executing cognitive control in non-linguistic domains, even when controlling for general slowness to push buttons or differences in attention.

Similar studies have shown these strong effects of age on a variety of other tasks that engage the cognitive-control system. Diamond et al. (2007) showed an increase in performance with age (and with training) on a Flanker task, where children had to focus on a central stimulus while ignoring surrounding competitors (Eriksen & Eriksen 1974). The same age trajectories have also been found on Flanker tasks using EEG measures with children and older adults (Friedman et al., 2009; Anderson, 2002). 8-12 year-olds also perform worse on both a classic Flanker task (which the authors claim to be a measure of "interference suppression") and a go/no-go variant said to measure of "response inhibition" (Bunge et al., 2002). These same children also exhibited non-adultlike neural activity in the pre-frontal cortex during these tasks. Taken together, these results demonstrate a stark developmental pattern where across various methodologies, young children have more difficulty than (young) adults in engaging their cognitive-control system, and this difficulty is lessened as children age.

From this work, it is clear that the system that governs cognitive-control execution improves with development, alongside children's ability to navigate syntactic ambiguity. However, it remains unclear how exactly these two systems interact. The following sections will outline evidence that both language development and cognitive control development play a role in helping children parse ambiguous sentences in an adult-like manner.

1.4: Language and cognitive control development as explanations for ambiguity processing in children

1.4.1: Language development and ambiguity processing

Following up on the work on the Kindergarten Path effect, Anderson et al. (2011) demonstrated that, on the whole, children tend to ignore visual cues to disambiguation. They presented five-year-old children with the same "put" task and sentences as Trueswell et al. (1999), in a mouse-tracking paradigm. They also gave the same children a vocabulary test (the Peabody Picture Vocabulary Test), and sought to correlate children's performance on the two measures. They found that vocabulary scores did correlate with children's ability to use these extra-linguistic cues: children who scored higher on the PPVT were more likely to correctly revise their parse for temporarily ambiguous sentences like "put the frog on the napkin into the box" when there were multiple frogs, which made the modification more likely for participants who made use of the referential context. This, they note, suggests a potential role for language experience in overcoming parsing ambiguity, though the precise nature of that role remains unclear.

Following this work, a few other studies have corroborated the finding that language experience plays a role in helping children overcome kindergarten-path errors. Huang, Leech & Rowe, (2017) presented five-year-old children with passive sentences that required syntactic revision (e.g. “The seal is quickly eaten by it”), and found that differences in language experience that related to children’s socioeconomic status facilitated syntactic retrieval during sentence comprehension. Huang & Hollister (2019) took this line of reasoning a step further: they presented children with a similar kindergarten-path interpretation task, and measured children’s performance on that as well as a test of language experience (the Diagnostic Evaluation of Language Variation–Screening Test) and a test of cognitive control development (a modified Stroop task). They found that when children’s socioeconomic status was better predicted by the DELV-S than the Stroop task, children were better able to revise their initial interpretations on the sentence task. This, the authors suggest, demonstrates a role for the development of linguistic knowledge in overcoming kindergarten-path errors, though the precise nature of what children may need to learn remains elusive.

1.4.2: Cognitive control and ambiguity processing: correlational evidence

In contrast to attempts to correlate children’s general language development with garden-path processing, several studies have now attempted to establish the link between children’s immature cognitive control recruitment and their sentence processing ability, by looking at the kindergarten path effect. In particular, recent attempts have been made to correlate children’s performance on cognitive-control tasks with their performance on “put” tasks. For example, Woodard, Pozzan, &

Trueswell (2016) had 5-year-old children complete three executive function tasks, and the "put" task described in Trueswell et al., (1999). The non-linguistic executive function task battery consisted of a Flexible Item Selection Task/Card Sort (Jacques & Zelazo, 2001) where children were asked to sort cards from the game SET along multiple dimensions (e.g. sort first by shape, then by color), a Day/Night task (Gerstadt, Hong, & Diamond, 1994) where children were shown a picture and were asked to say its opposite (e.g. saying "night" to a picture of a sun), and a Flanker/No-go task (Rueda et al., 2004; de Abreu et al., 2012), a version of the Flanker task that included trials where no response was required. The authors replicated the results of Trueswell et al. (1999), namely that children continued to look at the incorrect goal, failing to revise their initial commitments after reaching the point of disambiguation. They then looked for correlations between performance on the executive function tasks and the "put" sentence task. They used "Flanker switch cost" as a dependent measure, defined as "the combined z-score for the RT and error difference between switch and no-switch trials," where switch trials have a different congruency type to the previous trial and no-switch do not. They found a negative correlation between Flanker switch cost and looks to the correct goal in the sentence task, as well as a correlation between Flanker switch cost and act-out errors in the sentence task. That is, better flanker performance (on this particular switch cost measure) correlated with increased looks to the correct goal and correct act-out actions in the "put" task. However, none of the other executive function measures correlated with "put" task performance (including the simple subtraction of performance on congruent-incongruent Flanker trials), providing mixed results for the claim that children have

difficulty with the put task due to immature cognitive-control development. It should be noted that Woodard et al. do not themselves explicitly implicate cognitive control as the mechanism that underlies garden-path recovery, but rather "cognitive flexibility," as this term better describes the process necessary to have high performance on the switch cost measure in the Flanker task.

Following this study, Qi et al. (2020) investigated a potential competence vs. performance distinction within garden-path processing: they hypothesized that the Kindergarten Path effect may be caused by children needing more time than adults to process their input, but that 5-year-old children have the underlying competence to ignore late-arriving information. To support this hypothesis, they showed that the Kindergarten Path effect is effectively mitigated when children are presented with a slower speech rate. Relevant to the current work, Qi et al. also correlated Kindergarten Path performance with performance on other, ostensibly non-linguistic tasks: a Simon says task and a Flanker task. They found that children who performed more accurately on incongruent Simon says and Flanker trials were also more likely to have a smaller ambiguity effect in their act-out performance, though these correlations did not survive Bonferroni corrections for multiple comparisons. It should also be noted that these effects potentially conflict with those in Huang & Hollister (2019), who did not observe a correlation between cognitive control measures and recovery from ambiguity using mixed-effects analyses. This raises questions about the stability of correlating individual differences in executive function measures with real-time ambiguity resolution in general. If individual

children do not show stable performance on these measures, methods that take advantage of these within-individual fluctuations may provide cleaner data.

If, as Section 1.2.3 argues, domain-general cognitive control is needed for garden-path recovery, then why are correlations between these non-linguistic and linguistic tasks that seem to tap into the same, domain-general system apparently sporadic? The answer may lie in the use of individual differences as a method to manipulate the amount of participants' cognitive control engagement, which presents at least two interpretive challenges:

i) It is possible that different tasks require use of our cognitive control system to different extents. If the difference between the conflict-inducing and control trials in task A (e.g. Flanker) is significantly larger than this difference in task B (e.g. Stroop), correlations may be difficult to detect. Correlational studies often operationalize their cognitive control measure by assigning participants cognitive control "scores," that are reflective of the relative difference between their performance (e.g. reaction time or accuracy) on congruent versus incongruent trials of a task like Flanker or Stroop. Even assuming individuals have relatively stable scores over time, a participant may have wildly different scores across tasks. This might be either due to differences in task demands, or because task A and task B require use of the participant's cognitive control system to different extents (e.g. the difference between the congruent and incongruent trials differs across tasks). Because of this, task A may correlate with a subsequent task C (e.g. Garden-path sentence comprehension), while task B does not show enough variation across participants to correlate, even if both task tap into the same underlying mental system.

ii) Issues may arise from characterizing cognitive-control engagement as a stable trait of an individual, when evidence shows that this engagement status is a state that is transient, and operates to shift attention toward the most informative dimension of the input when multiple dimensions conflict (Gratton et al., 1992). The greater the variability within an individual at different testing timepoints, the more this noise will drown out the stable variable of interest (Nozari & Dell, 2011; Nozari & Novick, 2017; for a discussion of this distinction as it relates to bilingualism, see Salig et al., 2021). As the next section will show, a way to circumvent this methodological briar patch is to simply treat cognitive control as a mechanism that adjusts to meet information processing demands in the moment, and manipulate its relative engagement level in order to observe its impact on real-time sentence processing.

4.2.3: Cognitive control and ambiguity processing: causal evidence

Given that correlations between garden-path recovery and cognitive control measures appear only intermittently for children, this leaves open two possibilities: either children's failures to revise when processing garden-path sentences are not due to a relative deficit in their domain-general cognitive-control system at all, or correlational data is an unreliable measure. While Hsu & Novick (2016) demonstrate that cognitive-control engagement improves garden path sentence processing in adults, they leave open questions about how the upregulation of cognitive control affects young children's parsing and interpretation. One possibility is that, unlike for adults, for children, upregulating cognitive control does not put the system into a state

where conflict is more easily overcome. Evidence in support of this is that Stroop-to-sentence conflict adaptation results do not port over to child studies cleanly.

Huang et al. (2016) sought to extend the results of Hsu & Novick (2016) to 5-year-old children. They presented children with a child-friendly Stroop task, where children had to name cartoon dogs whose names were color words that conflicted with their fur color on incongruent trials. Interleaved with these, children performed a "put" task. They found that performance on the Stroop task did indeed influence performance on the sentence task (implicating domain-general cognitive control in sentence processing for children), but not in the same way it did for adults. Instead of adapting to conflict, incongruent Stroop trials made children worse at successfully revising their initial parse directly afterward. Specifically, children were more likely to look to the incorrect goal and less likely to look at the correct goal on ambiguous sentences following incongruent Stroop trials. More time spent considering the incorrect goal indicates that children committed more strongly to a goal-parse of "on the napkin" following cognitive-control engagement. On the surface, these results appear to support the Depletion view – children, having more limited cognitive control resources than adults, deplete some of this resource on Incongruent Stroop trials, leading to difficulty using their cognitive control system to ignore their initial parse while interpreting subsequent ambiguous sentences. This raises an important question: does children's domain-general cognitive control system interact with their language processing system in a way that is different from how these systems interact for adults?

Given evidence that the cognitive control system that is necessary for syntactic revision is indeed the same one used for non-linguistic tasks such as Stroop and Flanker, it is reasonable to conclude that the evidence cited in Section 1.3.2 can be brought to bear on developmental sentence processing. Namely, we can conclude that children's cognitive control system is immature compared to that of adults, and that this immature system also governs the choice of how to proceed in revising an incorrect parse of a sentence. However, a fundamental question remains: in what way is the child's cognitive control system immature? What underlying mechanism can explain the apparent adaptation effects for adults, but depletion effects for children?

The Depletion view would state that children have a pool of control resources, which is used up on a trial where cognitive control has to be executed. It is useful to note that this hypothesis predicts depletion effects across the board for children, even in non-linguistic tasks that involve back-to-back conflict-inducing stimuli. However, findings for conflict adaptation/depletion in children have been somewhat mixed. Iani et al. (2014) tested 6-8 year-old children on a Simon task, where they were instructed to respond to a red or blue square by pressing a corresponding button. On incongruent trials, the square appeared on the opposite side of the screen from its response button, and on congruent trials it appeared on the same side. They found that following incongruent trials, the Simon effect (incongruent minus congruent reaction times) was reduced for first graders (6-7-year-olds) but not for slightly older second graders (7-8 year-olds). Larson et al. (2012) found similar mixed effects with slightly older participants: They tested 9-year-old children on a color-word Stroop task, and found

conflict adaptation effects for both children and adults, with no significant differences between the groups.

In a similar study, Waxer & Morton (2011) tested children (9-11 years-old), adolescents (14-15 years-old), and adults (18-25 years-old) on a Dimensional Change Card Sort (DCCS) task (Zelazo, 2006), in which a cue (the letter "S" or "C") indicated on each trial whether to sort an item based on shape or color. On congruent trials, the item could be sorted only based on one dimension, and on incongruent trials it could be sorted based on both. They found that adults and adolescents adapted to conflict: their performance improved on incongruent trials following other incongruent ones. Children, however, showed the opposite effect: their performance on incongruent trials suffered following incongruent trials compared to when following congruent ones. In contrast, Ambrosi, Lemaire, and Blaye (2016) found explicit conflict adaptation results for 5-6 year-olds performing Flanker, Simon and Stroop tasks, mirroring performance for adults.

These results paint an inconsistent picture, but one which does not depict universal depletion effects when children encounter multiple stimulus conflict-inducing trials. Adaptation effects found by Iani et al., Larson et al., and Ambrosi et al. seem to undermine the Depletion account, and instead favor one where, similarly to adults, cognitive control engagement serves to upregulate a system that eases dealings with conflicting stimuli by biasing processing toward task-relevant cues (e.g. ink color in a color-word Stroop task). However, why do we see depletion-like effects when children interpret ambiguous sentences following incongruent Stroop trials (as demonstrated by Huang et al. 2016)? These results can be made parsimonious if we

assume that children and adults differ in what their cognitive-control system pushes them to focus on when it is engaged during sentence processing. Huang et al., (2016)'s results may result from Stroop conflict helping children to focus on strong, reliable processing cues and inhibit the signal they get from less reliable ones. This would still lead children astray if they're relying more on verbs and ignoring the subsequent corrective preposition. This "Cognitive biasing" hypothesis says that encountering information processing conflict engages children's cognitive control system and biases them to weight the importance of reliable cues more heavily than unreliable ones. The rest of this dissertation aims to provide evidence for and further specify this view.

Chapter 2: Conflict Adaptation with Active/Passive ambiguity

The findings outlined in Chapter 1 cannot distinguish between a biasing account and a depletion story, as they often predict similar outcomes, albeit for different reasons. For example, for the results of Huang et al. (2016), under the Depletion view, children should have more difficulty reaching an adult-like final interpretation for garden-path sentences following incongruent Stroop trials because their control resources are depleted. Under the Biasing view, the same result is predicted but now due to the fact that children's control system is more heavily biased to use reliable cues, and verbs are generally reliable. The goal of this chapter will be to distinguish between the Depletion and Cognitive Biasing accounts. To do so, we will need an experiment wherein the cue that is ordinarily reliable is no longer the cue that leads children to build an incorrect parse, but rather the cue that leads them to an adult-like interpretation. If children follow this cue more after cognitive-control engagement, this will provide evidence in support of the Cognitive Biasing view (as this account predicts that cognitive-control engagement should push children to rely on ordinarily reliable processing cues), and against the Depletion view (as this account predicts that children should have difficulty revising following cognitive control engagement, regardless of cue reliability).

2.1: Experiment 1: Conflict Adaptation with Passives – Early Novel Words

To distinguish between the Depletion and the Cognitive Biasing hypotheses, 5-year-old children were presented with temporarily ambiguous sentences where the cue to recovery is generally a reliable one. Since the Depletion hypothesis is that

children's cognitive control system is instantiated as a distinct quantity of mental resources, navigating ambiguity during sentence processing should tap into this pool and use a particular quantity of its contents. As it's defined here, this hypothesis already takes as given that the resource pool in question is domain general, and can be tapped into by non-linguistic tasks in addition to sentence processing ones. Non domain-general accounts will not be considered, following the conclusion of Section 1.2.3, that the existence of conflict adaptation effects provides strong evidence that the cognitive control system used for linguistic tasks is not wholly distinct from the system we use for non-linguistic operations.

In contrast, the Cognitive Biasing hypothesis is that children's cognitive control system is instantiated not as a finite pool of resources but as a biasing mechanism, which can be relatively up- or down-regulated. Under this hypothesis, once children encounter a task that requires the use of their cognitive-control system, the system boosts signal toward task-specific representations that are relevant for the overall goals of the task. As described in Section 1.2.2, this hypothesis borrows from classic models of cognitive-control engagement in adults, such as Botvinick et al. (2001). In this work, the authors set out to model the internal conflict monitoring system, and account for how it might respond to increases in control level. In other words, this work is an attempt to explain how conflict adaptation (or the "Gratton Effect") might be mentally instantiated. They model conflict adaptation in the Stroop task (see Figure 1.1) as involving a conflict monitoring node, that bears a weighted connection to a task demand system. The task demand system then contains separate sub-systems for display color and the word itself in the Stroop task. These sub-

systems then connect to more specific nodes that encode specific colors (e.g. red ink or the word "green"). These more specific nodes both feed into one response node, creating an information bottleneck. When conflict is encountered between the word and ink color in a Stroop task (e.g. the word green is presented in red ink), both colors are activated within the response node. The initial conflict monitoring node detects this information overload in the response node, and feeds information back down to the task-demand system, telling it to up-weight the information coming from the display color, and/or down-weight the irrelevant information coming from the word node. Thus, when conflict is encountered a second time, the information from the display color is more highly weighted than information from the display word, and the conflict in the response node is more easily overcome. This provides a mechanistic explanation for how conflict adaptation is possible during successive trials in a Stroop task.

The Engagement/Cognitive Biasing hypothesis is inspired by this model, and posits that children's cognitive control system interacts with their language processing system in much the same way as Botvinick et al. (2001) spell out for the Stroop task. As a general case, when children encounter conflict, a domain-general system that encodes task demands is activated, and within that system a node that encodes task-relevant representations is activated (this is the equivalent of the "display color" node for Botvinick et al., 2001, but meant to apply to the general case, not just Stroop). This node is then connected to task-relevant representations that encode specific details of the task at hand. For a Flanker task, for example, it would be the direction the relevant/middle arrow is facing. When processing an ambiguous

sentence, it would be the cue indicating the intended final parse (e.g. the start of the second prepositional phrase in "Put the frog on the napkin into the box" for an *adult* comprehender). Success occurs when these task-specific representations are relatively more active than other, prepotent task-specific representations (e.g. the direction flanking arrows are facing). When task-relevant representations are upregulated, conflict at the response node is alleviated as this node receives the signal from one task-specific node more strongly than the other. Note that this system does not encode encountering conflict as using up a finite store of a mental resource, but instead as a network that can be differentially more or less "engaged," depending on the weights between the nodes' connections.

While there may be ways to incorporate something like a pool of mental resources into the Cognitive Biasing hypothesis, it is fundamentally a different characterization of what the mental system that mediates representational conflict looks like. The two hypotheses described here also come apart in that they make different predictions when it comes to children's ability to adapt to successive trials that involve conflict. The Depletion account predicts that, following a high-conflict task, children will have used up some quantity of their cognitive-control ability and will therefore exhibit worse performance if the following trial also involves conflict. Or at least, performance should not improve. Cognitive biasing predicts the opposite: if cognitive-control engagement boosts activation of task-relevant representations, encountering conflict on one trial should lead to an alleviation of conflict and therefore better performance on a subsequent conflict trial, *as long as* participants do indeed treat task-relevant cues as task-relevant. The cognitive biasing account

predicts a decrement in performance when non-task relevant cues are treated as such, which may come about when these cues are ordinarily reliable.

To test these two hypotheses, children were presented with a paradigm that manipulates their state of cognitive control engagement: Stroop and Sentence trials were interleaved, and conflict is encountered both in the (non-linguistic child-appropriate) Stroop task and in the sentence-processing task. As the previous chapter outlined, the advantage to using a conflict adaptation paradigm over a correlative one is that it allows us to draw a causal inference between the two types of tasks: since cognitive-control engagement is directly experimentally manipulated (as opposed to measuring free variation within a particular population), if performance on the sentence task is influenced by the type of Stroop trial children receive immediately before it, this indicates that interpreting the sentence required use of the same mental process.

Depletion and Cognitive Biasing therefore make different predictions when verbs are cues to recovery from the garden path, instead of misleading cues. The Depletion hypothesis predicts that, after encountering conflict, children should continue to be misled by subsequent conflict, regardless of where verbs occur. The Cognitive Biasing hypothesis predicts that, after encountering conflict, children should adhere more to the interpretation provided by verbs as reliable cues. Children should therefore be misled by verbs when they are cues to an incorrect sentence interpretation (as in the "put" case).

Previous studies of the Kindergarten Path effect, including Huang et al. (2016), have conflated these two dimensions: "put" sentences use the

subcategorization information from the verb to lead children astray. If cognitive control engagement allows children to up-weight cues that are ordinarily reliable indicators of what the eventual parse should be, this explains why children have more difficulty recovering from misparsing following cognitive control engagement in the "put" task: they rely more heavily on the cue that is ordinarily a reliable one for sentence processing, instead of focusing on the information that is task-relevant in this particular task.

To test this, Experiment 1 will use temporarily ambiguous active/passive sentences, where, unlike "put" sentences, verbs are cues to recovery from misinterpretation. If encountering an incongruent Stroop trial immediately prior to a passive sentence improves recovery from misinterpretation on the sentence task, this will indicate that engaging children's cognitive control system pushes them to rely more on reliable cues (or less on unreliable ones), allowing them to successfully ignore their agent-first bias and focus on the disambiguating information provided by the verb, supporting the Cognitive Biasing hypothesis. If instead incongruent Stroop trials impair recovery from misinterpretation for passive sentences, this will indicate that engaging children's cognitive control system uses up a portion of their mental capacity devoted to mediating ambiguous input, and the Depletion account will be supported.

2.1.2: Participants

For Experiment 1, 32 children aged 4;0 to 6;6 (mean 5;1, 18 female, 14 male) were recruited from schools in the University of Maryland, College Park community. All children heard English as their primary language, and assented to participate in

the study. Children received a small donation to their schools for participating. All procedures were approved by the University of Maryland Institutional Review Board.

2.1.3: Materials

Children were presented with two interleaved trial types: Stroop and Garden-path sentences:

Child Stroop stimuli: Since the younger children in this study were not yet of reading age, a modified, child-appropriate Stroop task was used (Huang et al., 2016). Children were presented with cartoon dogs, whose fur color could be blue, brown, red, or green. The task was to say the name of the dog, also a color word, while ignoring the dog's fur color. On congruent trials, the dogs' names matched the fur color (e.g. a blue dog named "Blue"). On incongruent trials, the dogs' names mis-matched their fur color (e.g. a red dog named "Green", see Figure 2.1).

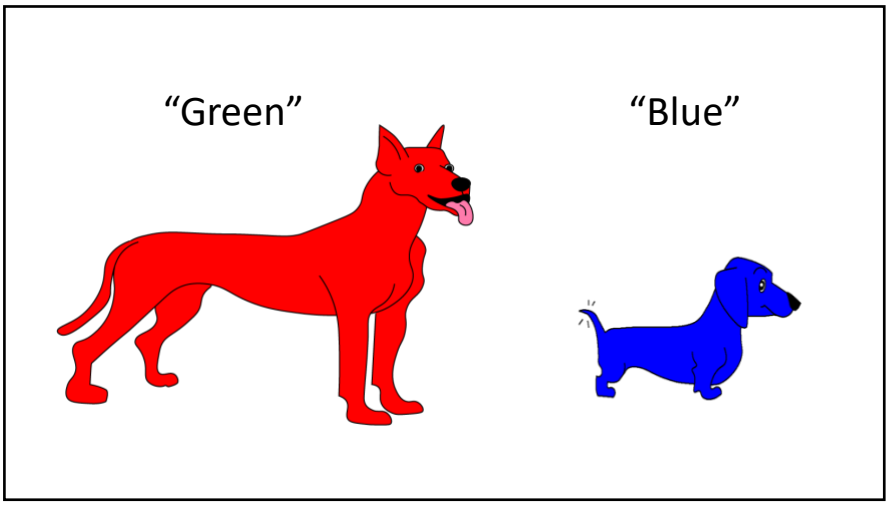

Figure 2.1: Incongruent and Congruent Dog Stroop stimuli

Children were trained on the names of the dogs before encountering target trials: they were shown at least 10 trials each of congruent and incongruent dogs, and

proceeded once they could reliably name them. To minimize conflict on congruent trials, the breed of dog also indicated the trial type: small, dachshund-like dogs were used for congruent trials, and larger, Labrador-like dogs were used for incongruent trials. This helped children identify the trial type, reducing the chance that they would experience interference from other dog names on congruent trials. Importantly though, measures of interest for this task require comparing performance following congruent vs. incongruent trials, so while children may have experienced some minimal interference effects on congruent trials, incongruent trials should still require the use of cognitive control to a much greater extent.

Garden-path sentences: Sentences consisted of simple active/passive pairs, with novel or unspecified creatures as one participant in the event. Novel creatures were subjects, as in "The Furpin will be quickly chasing/chased by the monkey." Novel creature names consisted of two-syllable nonce words. A follow-up sentence informed children that their task was to identify (point at or tap) the novel creature (e.g. "Click on the *Furpin*"). The experimenter then performed the clicking action for the child. [1]

While hearing these sentences, children viewed images of the scene they described, with multiple potential referents for the novel word. For example, children saw an image with a cartoon monkey in the center of their screen, with a small, patientive novel creature on one side of it, and a large, agentive creature on the other side (see Figure 2.2). While the known animal was always presented in the center of

[1] Though children were told to "click on" the creatures, this instruction was often interpreted either as pointing at or tapping, as on a touch screen. When this wasn't the case, "click on" commands were interpreted (as intended) as a joint task for the child and experimenter to perform together, with the child pointing and the experimenter making the click action.

the screen, the side of the large vs. small novel animal was counterbalanced across trials.

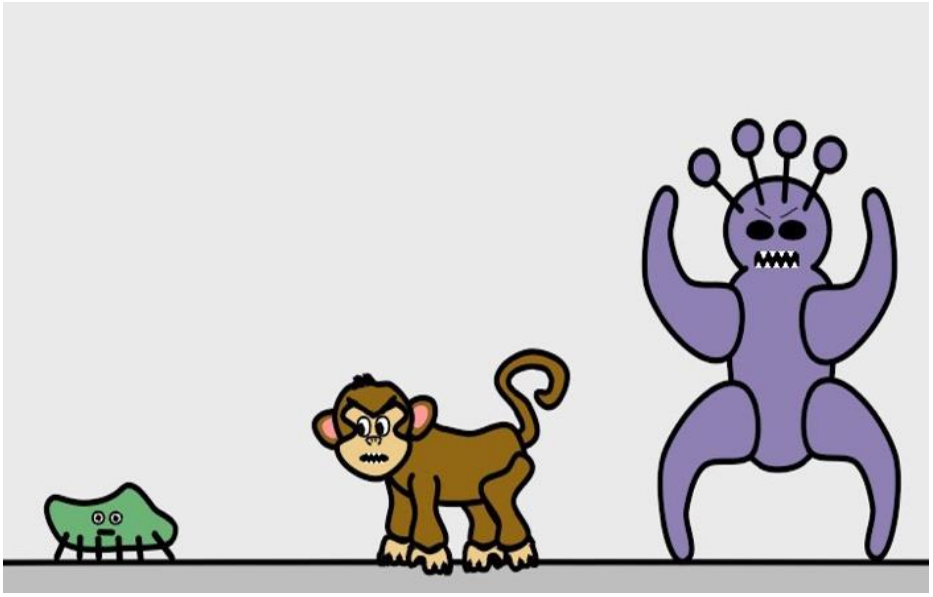

Figure 2.2: Visual scene for Experiments 1 &3. All scenes featured a known creature center-screen (e.g. a monkey), a novel creature as a likely agent (purple creature) and a novel creature as a likely patient (green creature).

Filler sentences were included that involved 3 known animals and an unambiguous statement such as "The lion and the tiger will be loudly roaring." These were followed by imperatives that targeted one of the forementioned animals (e.g. "Click on the tiger"). These trials served to obscure the target manipulation, balance trials so that the target was sometimes the middle object, and also provided children with some reprieve from the harder tasks, since these filler items were generally easy for them.

Audio recording: Test sentences, questions, and fillers were recorded by a female native speaker of American English. Recording was done in a sound-proof booth to minimize environmental noise, and child-directed prosody was used, but the audio was otherwise naturalistic.

2.1.4: Procedure

Children were tested in a quiet room, seated approximately 24 inches away from a computer monitor. Eye-movements were measured using an Eyelink 1000 table-mounted (remote) eye-tracker and stimuli were presented using ExperimentBuilder software (SR Research, Ontario, Canada). After being trained on the dog Stroop task, children underwent calibration, during which they looked at 5 dots on the screen, in turn. Next, children saw counter-balanced trials of the dog Stroop and garden-path sentence tasks. The order of trials was such that both halves of the experiment contained an equal number of congruent and incongruent Stroop trials, as well as active and passive sentences. While the order of Stroop and sentence trials appeared random to the child, tasks were psuedorandomly intermixed to form an equal number of pairs: 4 sets each of congruent-active, incongruent-active, congruent-passive, and incongruent-passive. Children also heard an additional 20 filler sentences, as well as 20 Stroop trials over and above the 16 paired with target sentences (for a total of 36 sentences and 36 Stroop trials). Children were tested on 2 lists, each containing either passive or active versions of each sentence, so that no child heard both the passive and active version of a particular item. Children were instructed to remain relatively still but were otherwise free to move as they liked. For this reason, mid-session recalibration was sometimes necessary. A drift-correct was done between every trial, during which children had to stare at a circle in the center of the screen to continue. Occasionally, when calibration was off by more than two degrees, the session was briefly paused and children underwent re-calibration.

Sessions lasted approximately 30 minutes, and children generally reported enjoying participating.

2.1.5: Analysis

Time-course data was averaged by word region. For each, the average proportion of fixations to the likely agent and likely patient creatures were calculated. Proportion of fixations was determined by interest area bounding boxes of equal size for the large, likely agent, the medium-sized known animal and the small, likely patient, though the boxes for the smaller creatures contained a larger proportion of blank screen around the creature. This was done to ensure that slight miss-calibration would not result in under-counting looks to the smaller creatures. While this sometimes resulted in looks to the empty space above the likely patients being counted as looks to the likely patient, this was seen as preferable to underestimating patient looks, and it was reasoned that looks launched into the general vicinity of the smaller creatures were still likely to be driven by the intention to look at the creatures themselves.

Region	Average length in milliseconds (sd)
The	167(45)
Novel noun	812(115)
Will be	357(50)
Adv	638(111)
Verb stem	315(139)
Verb morphology	420(112)
the	140(45)
Known noun	442(89)

Table 2.1: Average duration of each target utterance region (in ms)

Since target sentences were disambiguated following the verb stem, the critical region for this experiment consisted of the period of time directly following the onset of verb morphology, until the onset of the following sentence. Audio recordings were naturalistic and word lengths varied slightly across items, so in order to create consistency in the analysis sentences were aligned to the onset of verb morphology. In order to account for the time to plan and execute a saccade, looks were shifted 200ms (Altmann & Kamide, 2004; Matin, Shao, & Boff, 1993), so that the region of interest reflected the first point at which looks may have been driven by an interpretation of the sentence as active or a passive reanalysis. As Table 2.1 shows, the region in question lasted 1,002ms on average.

Prior to analysis, fixations were cleaned in DataViewer (SR Research, Ontario, Canada). Blinks and other artifacts were removed. In the critical window, looks outside the areas of interest were excluded, and the remaining proportion of

looks to the likely agent and likely theme areas were analyzed. Logistic mixed effects regression models were fit to these data with random intercepts and slopes for participants and items, and with sentence type and prior Stroop condition as fixed effects.

2.1.6: Results

Children who exhibited trackloss on more than 33% of trials were excluded from analysis (2 children). Additionally, children who did not complete the entire experiment were excluded (2 children). Finally, while participants generally understood the game, two additional children were excluded for failure to accurately respond on over 50% of the filler items. This resulted in 6 total children being excluded, and data from the remaining 32 children was analyzed.

Stroop results: Children were fairly accurate in responding during congruent Stroop trials: average accuracy for congruent trials was 75%. For incongruent Stroop trials, children had more difficulty, per design. Average accuracy for incongruent trials was 53%. Since children had only two seconds to respond on these trials, failures were generally failures to respond in time, not errors in response color word (85% were failures to respond and 15% were incorrect color word responses). Of these incorrect color word responses, 67% were errors on incongruent trials wherein children responded with the lure color (e.g. saying "Red" to a red dog named "Green," where "Green" was the correct response). Other incorrect word errors were a combination of incorrect responses on congruent trials and audible but unintelligible responses.

Previous Item Type	Current Item Congruency	Accuracy
Stroop-congruent	Congruent	0.85
Stroop-congruent	Incongruent	0.64
Stroop-incongruent	Congruent	0.82
Stroop-incongruent	Incongruent	0.69
Target sentence-congruent	Congruent	0.89
Target sentence-congruent	Incongruent	0.47
Target sentence-incongruent	Congruent	0.85
Target sentence-incongruent	Incongruent	0.42

Table 2.2: Stroop accuracy by previous item type (Stroop or sentence, congruent or incongruent)

Additionally, Stroop accuracy varied slightly as a result of previous trial (see Table 2.2). When the prior trial was a congruent Stroop trial, children were slightly more accurate at responding to a successive congruent Stroop trial. Notably, a numeric conflict adaptation effect was found when analyzing only instances of two successive Stroop trials: when trial N-1 was an Incongruent Stroop trial, children were slightly more accurate to respond to a successive incongruent Stroop trial (see Figure 2.3). A mixed effects regression model with random slopes and intercepts for subjects and items confirmed that the interaction between Stroop type and prior Stroop item type was not significant (β=0.53, SE=0.85, t=.624, *p*=.53). However there was a significant main effect of current Stroop trial type such that children were more accurate on congruent Stroop trials than on incongruent ones (β=1.44, SE=0.47,

t=3.05, *p*=0.002). As the answers children gave to Stroop trials were verbal, reaction time was not measured.

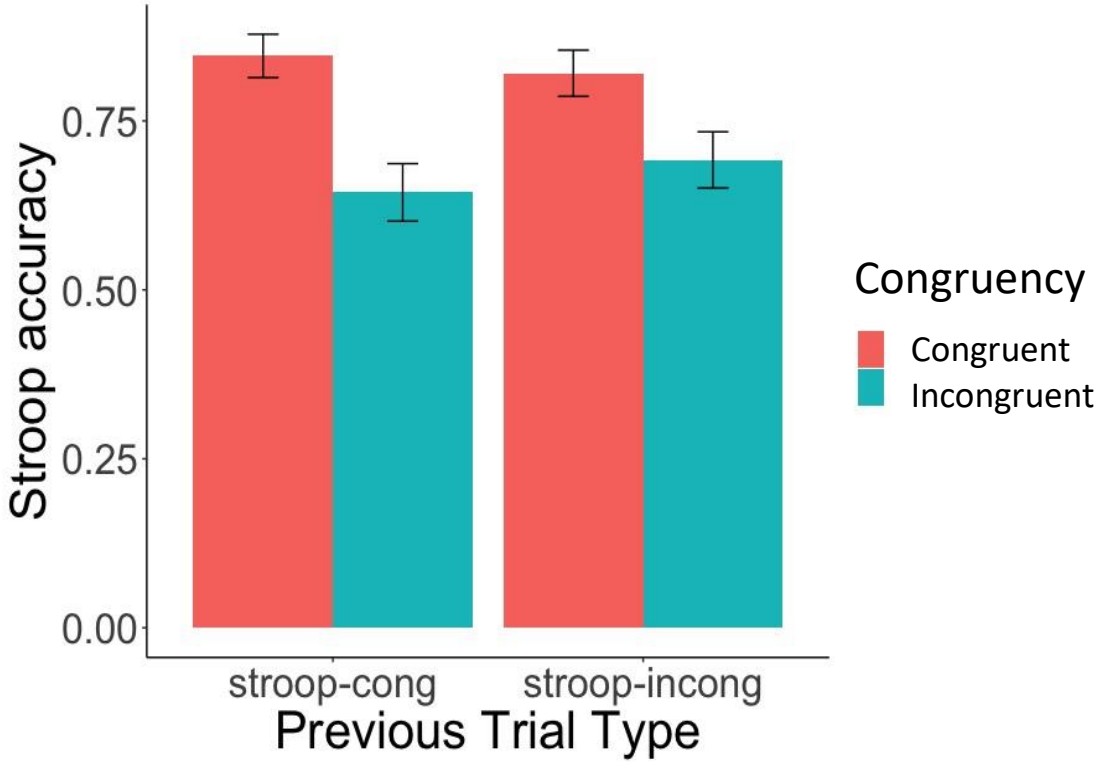

Figure 2.3: Stroop accuracy by prior Stroop trial

Act-out results: Following the target sentences, children were asked to click on the novel animal, and their responses were recorded. Their overall accuracy was 63%. In general, children were fairly accurate at choosing the intended target of the novel word for actives, but less so for passives (see Table 2.3 for by-condition break-down). This can be compared to their overall accuracy for looks during the critical region for target sentences, described below, which was similar for passives at 44%, but lower for actives at 59%. Act-out accuracy did not vary as a result of prior Stroop trial type, as confirmed by a mixed effects regression model with random intercepts and slopes for participants and items (β=0.36, SE=0.65, t=.55, *p*=.58). There was, however, a

main effect of sentence type such that children were more accurate for active sentences than for passive ones (β=2.86, SE=0.47, t=6.07, p<.001). Reaction-time data for the act-out task were not analyzed, as children pointed at their choice for novel word referent, and the experimenter performed the actual click. This measure was therefore deemed to introduce too much room for subtle bias on the part of the experimenter to be a meaningful representation of children's mental state.

Condition	**Act-out Accuracy**
Congruent-active	86.7%
Congruent-passive	39.1%
Incongruent-active	89.1%
Incongruent-passive	37.5%

Table 2.3: Act-out accuracy by condition for Experiment 1

Sentence results: Children showed an overall preference toward looking at larger creatures/animals. During the period of time between the point of disambiguation and end of the sentence, children looked toward the known animal 38% of the time, to likely agents 26% of the time, and likely themes 20% of the time. The remaining 16% of the time children were looking elsewhere on the screen or offscreen.

Stroop x Sentence interaction: As Figure 2.4 shows, children's looking patterns after disambiguation were less accurate on actives following incongruent Stroop trials, but more accurate on passives following incongruent Stroops. This interaction was

significant (β=.74, SE=0.04, t=18.9, *p*<.001). There was also a main effect of sentence type such that active sentences resulted in more looks to target (β=0.69, SE=0.01, t=65.2, *p*<.001) and a main effect of prior Stroop trial type such that incongruent Stroop trials led to more correct looks to target (β=0.04, SE=0.01, t=3.15, *p*=0.002).

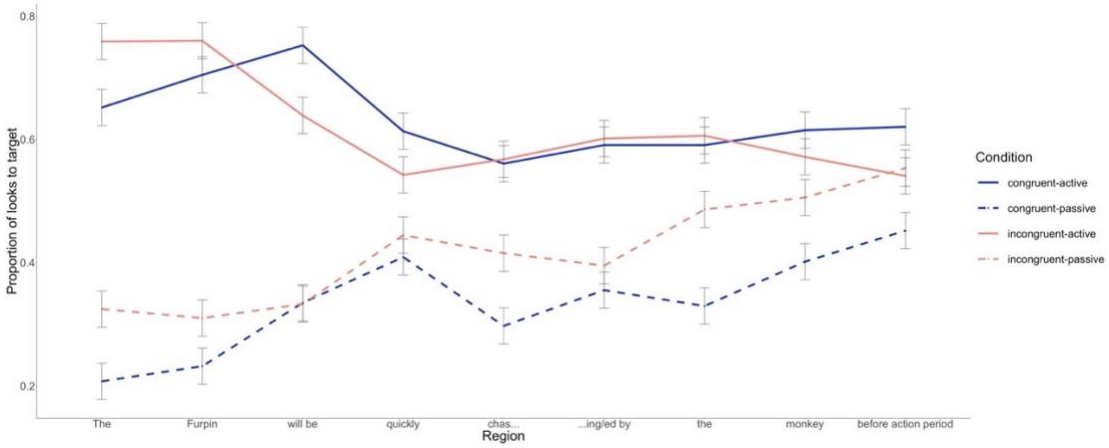

Figure 2.4: Time-course analysis by region for Experiment 1. Y-axis represents proportion of looks to the target character (the likely agent for active sentences, and the likely patient for passives). The region before the action period refers to the pause after final noun offset but prior to the onset of the instruction sentence (e.g. "Click on the Furpin"). The critical region of analysis for this study consisted of the last 4 bins from verb morphology until the onset of the instruction sentence.

As is clear from Figure 2.4, in several conditions children exhibited "psychic" effects, where they seemed to be already looking to the likely agent or theme reliably before the point of disambiguation. Since these looks cannot be due to the experimental manipulation, they were excluded in a subsequent "Switch" analysis. In this, all trials where children were already looking to the likely agent or theme and in which they continued to look at that item during the critical region were excluded from analysis, resulting in the removal of 31% of trials in total. The StroopxSentence interaction was then reexamined for these data. Again, it was found that there was an interaction such that while incongruent Stroop trials did not cause children to look

more toward the target animal for active sentences, they did push children to look more accurately in passive sentences (β=.26 SE=0.05 t=4.86 p<.001). See Figure 2.5 for a graph of these Switch analysis results.

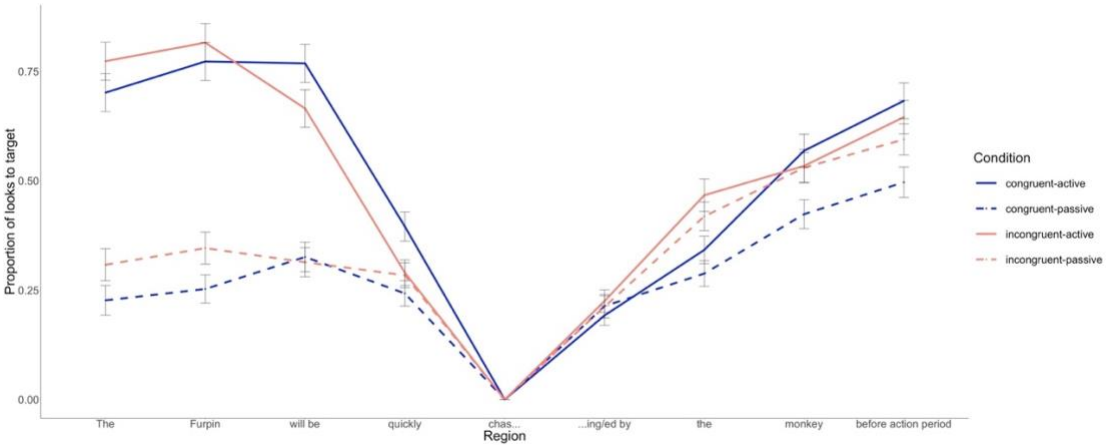

Figure 2.5: Switch analysis for Experiment 1. Trials in which children were looking toward the target during the verb stem were removed, to ensure that effects after the point of disambiguation were due to the manipulation, instead of a continuation of prior gaze.

2.1.7: Discussion

Children in Experiment 1 showed the expected pattern of results on Stroop trials – they were more accurate on congruent Stroop trials than incongruent ones, suggesting that the dog Stroop task did effectively engage their cognitive control system. Additionally, at least numerically, they showed a classic conflict adaptation effect: children were marginally more accurate on incongruent Stroop trials that followed other incongruent trials, and less accurate on congruent Stroop trials that followed incongruent trials. This interaction, however, was not significant. Since speed of response was not measured, it is additionally possible that children exhibited a speed-accuracy tradeoff in the Stroop task that was not observable using the current method.

It should be noted that it is unclear in this case whether performance following inaccurate Stroop trials should be excluded. Children attempting yet failing to successfully complete an Incongruent Stroop trial may engage their cognitive control system just as they do then they succeed. In these cases, it is unclear whether the signal carried forward to the sentence trials is different than the one carried forward when they succeed. Additionally, most children simply timed out of the Stroop task but did not make an incorrect naming decision. In these cases, they may have been undergoing the same process as they underwent on successful trials, but at a protracted rate. While it is likely true that on some proportion of these trials, children were simply ignoring the Stroop task, and in these instances excluding incorrect trials ought to reduce experimental noise, it is also possible that an analysis excluding these trials would eliminate exactly the trials that children found challenging.

Of course, difficulty with incongruent Stroop trials could originate from a variety of places. For one, in this task, incongruent Stroops may have required a higher working memory burden than congruent trials, since children had to remember the names of the dogs on incongruent trials but could simply respond on the basis of observed fur color on congruent trials. Children were also aware of the increased difficulty of incongruent trials, and therefore lower accuracy on these trials could reflect a difference in metalinguistic task demands (e.g. children may have felt more anxious about their answer on incongruent trials because of the higher working memory and/or cognitive control burden). Neither of these possibilities, though, predict that incongruent trials would boost children's performance at interpreting subsequent sentence trials. If anything, the predictions ought to flip: children who

were anxious about more difficult trials should perform worse at subsequent testing points. In addition, there is little evidence that working memory training is effective in either adults or children (Shipstead et al., 2012; Melby-Lerbåg & Hulme, 2013). And where training is shown to be effective, it does not usually generalize to novel contexts. It would therefore be unlikely that children's greater use of their working memory system on incongruent Stroop trials would lead them to more effectively parse passive sentences a second later.

Overall, children also performed more accurately on active sentences than passives. They consistently looked more to the large, agentive creature throughout the trials. After the verb, looks to the likely agent increased for actives and decreased for passives overall, suggesting that children used verb morphology to interpret the sentences in an adult-like manner, and were not always hopelessly garden-pathed by the assumption that the first-mentioned noun would be the agent of the sentence. However, there is also evidence that children *were* garden-pathed to some extent: they looked more to the likely agent following the initial novel noun, and continued to look at the likely agent after the verb in passive sentences more than half of the time.

Importantly, performance on Stroop and sentence trials interacted: children were even more likely to look to the likely patient for passive sentences after disambiguation if the prior Stroop trial was incongruent, compared to when it was congruent. Children's interpretation of active sentences was also modulated by Stroop trial, in the opposite direction. After incongruent Stroop trials, children were slightly less likely to look to the likely agent on active trials. While this might be an initially

surprising effect, it may help explain prior discrepancies between the results of conflict adaptation studies for adults and children. In the adult literature, the effect of a decrease in performance on control (or ostensibly non-conflict) trials following high-conflict trials has been called post-conflict slowing. This has been observed in typical Simon and Stroop tasks, and has been previously pointed to as a counter-effect that may obscure conflict adaptation effects (Verguts et al., 2011). Children's decrement in active performance following incongruent Stroop trials is likely a similar effect – for any number of reasons, they may find it difficult to stay on task following a high-conflict trial in general. It is likely that post-conflict slowing is at work in the passive conditions here as well, and that children's increased performance on passives following incongruent Stroop trials occurs in spite of this effect.

Note that the existence of a post-conflict slowing effect in general is not itself evidence for the Depletion account. Since under this account encountering conflict uses up resources dedicated to the mediation of representational conflict, it predicts that performance should be diminished only when multiple instances of conflict are encountered in a row. Or at least, performance should drop more on the second of two high-conflict trials in a row (e.g. passives following incongruent Stroop trials) than on a low-conflict trial following a high-conflict trial (e.g. actives following incongruent Stroop trials). It is unclear how such an account would explain the larger decrement in performance on active trials following incongruent Stroop trials, or even an equal decrement in performance across all trial types, if that is indeed what is happening under the hood.

In contrast, the Engagement hypothesis readily explains the present results. If children's cognitive-control system is put into a more highly engaged state following the completion of an incongruent Stroop trial, and if the subsequent sentence processing task requires the use of the same system, it follows that performance on the sentence processing task should increase following the successful navigation of an incongruent Stroop task.

While seemingly explanatory, this conclusion remains on tentative grounds without further support. One potential cause for concern is that the results are consistent with children always looking more toward the likely theme following incongruent Stroop trials. On passives following incongruent Stroops, they look more to the likely theme leading to higher accuracy. But the same is true for active trials – children look more toward the likely theme for actives following incongruent Stroop trials as well, leading to lower accuracy for active sentences following conflict. While there is reason to believe, as stated above, that this is due to a generic post-conflict slowing effect, it leaves open the possibility that children's performance is due to a tendency to look toward less-obvious visual referents following high-conflict trials, or indeed just to look around the visual scene more. If this is the case, then children may be ignoring the sentences entirely, and any conclusion about cognitive-control engagement facilitating conflict mediation during sentence processing is unfounded based on these results.

A second potential concern for Experiment 1 is that children may not have completely understood the intended relationships between characters in the visual scene. This concern arose from the fact that children often asked things like "which

one is the Furpin?" when instructed to click on it. Of course, images were designed to convey agent- or patient-hood, for example by giving the larger, agentive character sharp teeth and the smaller, patientive character a scared look. It's also true that this may be a potential concern for any task involving novel words. Still, since the characters were static to allow or quick trial transitions, it was not clear that children always understood the intended relationships between the characters. If children misunderstood this, it might have created undue noise in the dataset, and obscured the intended effect. By extension, a version of this experiment with less ambiguity in the intended relationships between the visual referents was predicted to yield a larger effect.

2.2: Experiment 2: Conflict Adaptation with Passives – Pronoun Version

The aim of Experiment 2 was to alleviate the two above concerns, that the results of Experiment 1 arise for some other reason than cognitive control engagement facilitating children's sentence processing, and also that children may have been confused about the relationship between the static figures in the visual scene. While the conflict adaptation structure and Stroop trials were preserved from Experiment 1, several updates were made to the visual scene and sentences.

To address the second concern, images were replaced by a trio of known animals, all in relatively canonical positions relative to each other (e.g. a dog chasing a cat chasing a mouse, as in Figure 2.6). This was done in an effort to reduce potential confusion about the intended likely agent and likely patient. Since the animals were known, instead of novel words, children heard sentences like "The cat will be quickly

chasing/chased by it" had to guess which animal was being referred to with the pronoun.

To address the first concern, children were presented with sentences in which successful interpretation did not straightforwardly equate to ignoring a likely agent in a visual scene, or attending more to a likely theme. In Experiment 2, the order of the known and "novel" nouns was reversed. Since sentences were of the form "The cat will be quickly chasing/chased by…", with the middle animal mentioned first, looks to the smallest animal (e.g. the mouse) indicated an active interpretation and looks to the largest animal (e.g. the dog) indicated a passive interpretation. Here, if children were ignoring the sentence and were instead inclined to look more toward smaller, less assuming visual objects following cognitive control engagement, they should still look more toward the likely patient following incongruent Stroop trials. Now, however, these looks are divorced from looks indicated by potential garden-path recovery, as looks toward the likely patient are consistent with an active interpretation.

Alternatively, if children now look more toward the likely agent following incongruent Stroop trials, this will provide converging evidence for Experiment 1, and indicate that children are more likely to successfully re-parse the sentence as a passive when their cognitive-control system is in a more highly engaged state.

2.2.1: Participants

For Experiment 2, 32 five-year-old children age 4;0 to 6;6 (mean 5;1, 20 female, 12 male) were recruited from schools in the University of Maryland, College Park community. Four additional children were tested but were excluded from the final sample due to inability or desire to complete the experiment. All children heard English as their primary language, and assented to participate in the study. Children received a small donation to their schools for participating. All procedures were approved by the University of Maryland Institutional Review Board.

2.2.2: Materials

In Experiment 2, no novel creatures were used. One participant in the visual scene was referred to by a pronoun, as in "The cat will be quickly chasing/chased by it." Children saw known animals in canonical positions (e.g. a large dog chasing a medium-sized cat chasing a small mouse). Though the animals were known, children were still required to attend to the verb in order to figure out the referent of "it." Images were standard clipart or cartoon representations of the characters in question.

Sentences always mentioned the animal displayed in the center of the screen, which was both the agent and patient of the verb, as the first NP. Likely agent and patient creatures were presented to the left and right of this animal, with the order counterbalanced across trials but consistent for a particular item. Children saw 16 target sentence trials, each preceded by a congruent or incongruent Stroop trial.

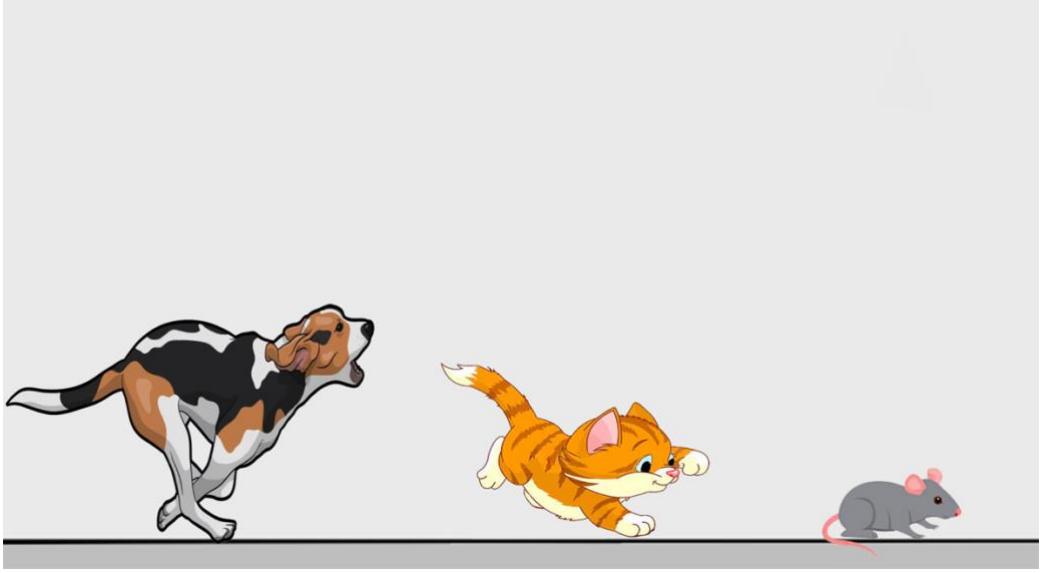

Figure 2.6: Image of a sample trial for Experiment 2. Visual scenes always consisted of 3 known creatures in relatively canonical positions, with orders counterbalanced between items (but consistent within an item).

2.2.3: Procedure

The experiment proceeded similarly to Experiment 1: Children were tested in quiet room, seated approximately 24 inches away from a computer monitor. Eye-movements were measured using an Eyelink 1000 table-mounted (remote) eye-tracker and stimuli were presented using ExperimentBuilder software (SR Research, Ontario, Canada). After being trained on the dog Stroop task, children underwent calibration, during which they looked at 5 dots on the screen, in turn. Next, children saw counter-balanced trials of the dog Stroop and garden-path sentence tasks. The order of trials was such that both halves of the experiment contained an equal number of congruent and incongruent Stroop trials, as well as active and passive sentences. The relative order of dog Stroop trials was identical to Experiment 1. Additionally, the psuedorandomization of Stroop and sentence trials was kept the same. The 20 filler sentences and 20 filler Stroop trials were identical to Experiment 1 and presented in the same order. Children were again tested on 2 lists, each containing either passive or active versions of each sentence, so that no child heard both the passive and active version of a particular item. Sessions lasted approximately 30 minutes, and children generally reported enjoying participating.

2.2.4: Analysis

Experiment 2 used a similar method of analysis to Experiment 1. Children for whom there was trackloss for more than 33% of their total data were excluded (1 child). Since this game involved only animals children knew and was subsequently

quite easy for the older children, 2 children opted to make it more "fun" by systematically responding incorrectly for every trial. While it was clear that these children understood the parameters of the game, it was judged that their inclusion would either add undue noise if left in, or result in an unfaithful characterization of their looks if it was systematically reversed back during data analysis. For these reasons, these children were excluded as well. This resulted in 3 total additional children being excluded, and data from the remaining 32 children was analyzed.

Time-course data was again averaged for each word in target sentences, and the average proportion of fixations to the likely agent and likely patient creatures plotted (see Figure 2.8). Interest area bounding boxes were again of equal size for the likely agent, the medium-sized animal and the small likely patient, applying the reasoning from Experiment 1.

The critical region for Experiment 2 was again directly following the onset of verb morphology, until the onset of the following sentence, with sentences aligned to the onset of verb morphology, and looks were shifted 200ms.

Region	**Average length in milliseconds (sd)**
The	192(66)
Known noun	710(109)
Will be	384(83)
Adv	755(107)
Verb stem	370(124)
Verb morphology	285(118)
It	250(74)

Table 2.4: Average duration of each analysis region for Experiment 2

Fixation exclusion and cleaning criteria were identical to Experiment 1 (blinks and artifacts were removed prior to analysis, and looks outside of the regions of interest were excluded). Logistic mixed effects regression models were fit to these data with random intercepts and slopes for participants and items, and with sentence type and prior Stroop condition as fixed effects.

2.2.5: Results

For Experiment 2, correct looks on passive sentences were again consistent with ignoring both the agent-first bias and a visual preference to look at larger animals, and instead looking to the small, patientive animal.

Stroop results: Children's average accuracy at congruent Stroop trials was 69%, as compared to 75% for Experiment 1. Their average accuracy for incongruent Stroop trials was 46% (again relatively comparable to 53% for Experiment 1). For Experiment 2, 86% of Stroop inaccuracies were due to a lack of any response, while

only 14% of errors were the result of naming the dogs an incorrect color word. Of these inaccurate word responses, 83% were the result of children naming the dogs by the lure color and the remaining 17% of response errors were due to children responding but unintelligibly, or responding with a non-lure color.

Previous Item Type	Current Item Congruency	Accuracy
Stroop-congruent	Congruent	.81
Stroop-congruent	Incongruent	.52
Stroop-incongruent	Congruent	.74
Stroop-incongruent	Incongruent	.70
Target sentence-congruent	Congruent	.56
Target sentence-congruent	Incongruent	.47
Target sentence-incongruent	Congruent	.74
Target sentence-incongruent	Incongruent	.56

Table 2.5: Stroop accuracy by previous item type for Experiment 2

Stroop accuracy also varied as a result of previous trial (see Table 2.5 for full details). When the prior trial was an incongruent Stroop, children were more accurate at responding to a successive incongruent Stroop trial (see Figure 2.7). A mixed effects regression model with random slopes and intercepts for items confirmed that the interaction between Stroop type and prior Stroop item type was significant such that children were more accurate on incongruent Stroop trials following other incongruent Stroop trials, but less accurate on congruent Stroop trials following incongruent Stroop trials (β=1.53, SE=.66, t=2.32, p=.02). There was also a

significant main effect of current Stroop trial type such that children were more accurate on congruent Stroop trials than on incongruent ones (β=1.78, SE=.37, t=4.87, p<.001). The verbal answers children gave to Stroop trials were again not measured for reaction time, leaving open the possibility of a speed-accuracy tradeoff not observable from accuracy data alone.

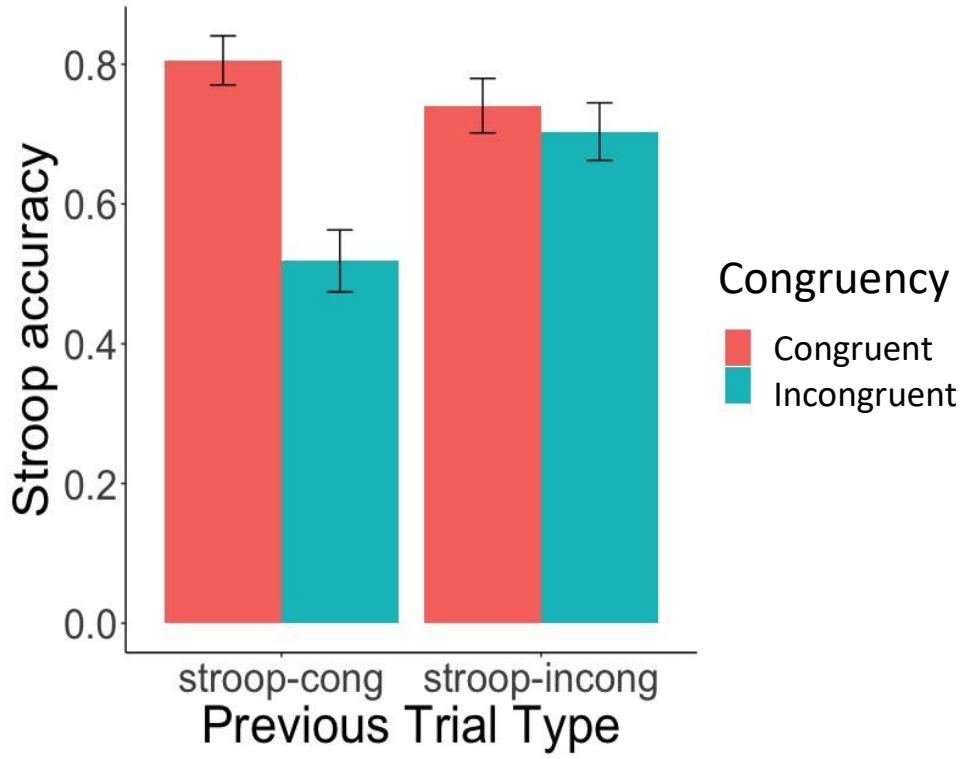

Figure 2.7: Stroop accuracy by prior Stroop trial for Experiment 2

Act-out results: Children's act-out accuracy for Experiment 2 varied less across trial types than their accuracy for Experiment 1. Overall, they chose the intended referent of the pronoun on 53.7% of trials (see Table 2.6 for by-condition break-down). This was relatively comparable to their overall accuracy for looks during the critical region for target sentences. As in Experiment 1, act-out accuracy did not vary significantly as a result of prior Stroop trial type. Since characters in this experiment were always

known animals, some children opted to name the animal as their response instead of making a "clicking" action. While unexpected, this was encouraged as it reduced the likelihood that children would become uncalibrated as a result of the clicking movement.

Condition	Act-out Accuracy
Congruent-active	53.2%
Congruent-passive	52.3%
Incongruent-active	59.1%
Incongruent-passive	50.5%

Table 2.6: Act-out accuracy for Experiment 2

Sentence results: Despite the switch to all known animals, children showed less of an overall preference toward looking at larger creatures. Children looked to the center named creature a bit less than in Experiment 1, overall 27% of the time between the point of disambiguation and the end of the sentence. They looked to likely agents during this period 14% of the time and likely themes 13% of the time. The remainder of the time children looked elsewhere on the screen or offscreen. Here, correct looks on passive sentences were consistent with ignoring the agent-first bias, and following the visual preference, and looking to the large, agentive creature. Correct looks on active sentences were the result of ignoring the visual preference but following the adult-like interpretation of the sentence to look at the small, patientive creature.

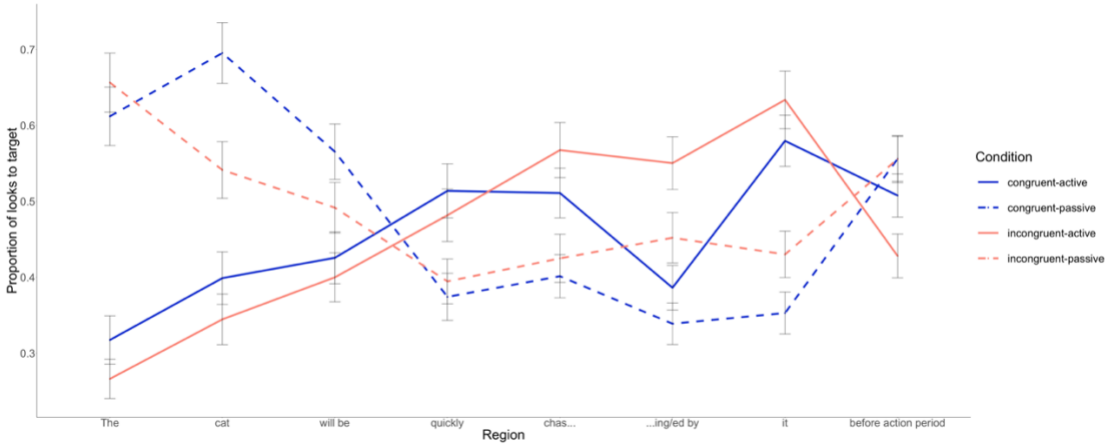

Figure 2.8: Proportion of looks to target across trial types for Experiment 2

Stroop x Sentence interaction: As in Experiment 1, there was an interaction between prior Stroop trials and children's performance on active vs. passive sentence trials (Figure 2.8). For active sentences, children were even less accurate at looking to the target animal following incongruent Stroop trials (vs. Congruent Stroop trials) than in Experiment 1. For passive sentences, however, children's looks were even more accurate following incongruent Stroop trials (vs. Congruent Stroop trials). This interaction was significant (β=0.35, SE=0.13, t=2.67, *p*=0.007). There was also a main effect of sentence type such that children looked more at targets during the critical region of active sentences than for passive sentences (β=0.89, SE=0.37, t=238, *p*=0.01). While numerically children's looks were also more accurate following incongruent Stroop trials than following congruent ones, this main effect was not significant (β=0.63, SE=0.39, t=1.64, *p*=0.10).

As can be seen in Figure 2.8, children once again sometimes fixated the target animal prior to the point of disambiguation. To ensure that looks were due to the experimental manipulation, switch analyses were performed, as for Experiment 1 (Figure 2.9). Switch analyses resulted in the removal of 30% of trials. As can be seen

in Figure 2.10, the StroopxSentence interaction is still present despite the removal of these trials (β=1.06, SE=0.28, t=3.79, p<0.001). There was also still a main effect of sentence type such that children's looks were more accurate for active sentences (β=1.4,1 SE=0.64, t=2.20, p=0.02).

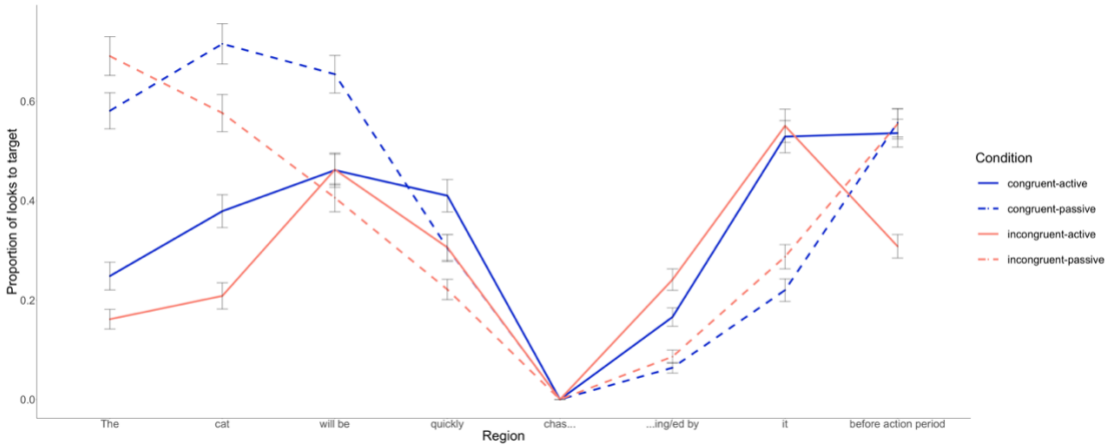

Figure 2.9: Switch analysis results for Experiment 2

2.2.6: Discussion

The aim of Experiment 2 was to rule out potential confounds of Experiment 1, and demonstrate that when children's cognitive-control system is engaged, this system allows them to navigate sentential ambiguity more easily. Experiment 2 served to replicate the results of Experiment 1 in the absence of novel words and novel creatures. Children showed a similar interaction effect: on active sentences following incongruent Stroop trials performance decreased, but this effect was ameliorated for passive trials. This once again indicates a causal relationship between the non-linguistic cognitive control task and the parsing task, although the two tasks may appear to make use of different cognitive systems on the surface. The shared

aspect between the two, the need to override a prepotent mental representation to focus on a different, task-relevant one, is therefore implicated.

A secondary aim of Experiment 2 was to rule out the possibility that children were simply ignoring cues from the sentence in Experiment 1 and looking more to less-assuming items in a visual scene following cognitive control engagement. If that had been the case, children should have looked more toward likely patients on both active and passive sentences. Instead, here, children were more likely to look toward likely agents for passive sentences following incongruent Stroop trials.

Here, Stroop data reveal a similar story to Experiment 1. Children performed as expected - they were more accurate on congruent Stroop trials than incongruent ones. Here also, children demonstrated a typical Stroop effect such that incongruent trials preceding other incongruent trials lead to an increase in accuracy. These results are numerically consistent with those of Experiment 1, and indeed now the interaction is found to be significant.

Additionally, Act-out data once again accorded with eye-movement data numerically, but failed to show a significant effect. This result is not unexpected. While act-out actions are presumed to reflect the outcome of online processing and therefore are expected to follow looking-time patterns, act-out actions are a coarser-grained measure of children's sentence interpretations. They are therefore significantly removed from revealing subtle, online distinctions in children's sentence interpretations.

One relatively unexpected finding of Experiment 2 was that, compared to Experiment 1, children performed even worse at active sentences following

incongruent Stroop trials. In other words, the post-conflict slowing effect observed in Experiment 1 was greater in Experiment 2, even in the low-conflict active trials. One possible explanation is that the target answer for actives in Experiment 2 was the patient in the visual scene, and incongruent Stroop trials made children even less likely to look toward this unassuming character. If so, this would indicate that fears about children ignoring the sentence cues in Experiment 1 and looking more to the likely patient regardless of active or passive morphosyntax were unfounded.

Regardless of the cause of the increased post-conflict slowing effect, this potentially explains another relatively surprising finding – the smaller "boost" children got for passives after incongruent Stroop trials, compared to Experiment 1. While it was predicted that increasing the obviousness of the relationship between the referents in the visual scene would lead to a larger effect, the opposite was observed. This can be explained by a larger post-conflict slowing effect, essentially working more to cancel out the benefit provided by children's cognitive control system being differentially more engaged after incongruent Stroop trials.

A next step in demonstrating that it was indeed the conflict engendered by having to ignore the agent-first bias that lead to children's overcoming of post-conflict slowing in Experiments 1 and 2 is to observe the effect of parsing a minimally different sentence that does not provide children with a relatively weak parsing cue to overcome. In essence, if the "boost" children are given on passives is indeed due to cognitive control engagement allowing them to ignore the unreliable parsing cue, then by taking out this cue we ought to see that this effect can be "knocked out." This was the goal of Experiment 3.

2.3: Experiment 3: Conflict Adaptation with Passives – Late Novel Words

For Experiment 3, the known and novel nouns were reversed from Experiment 1. Correct looks on passive sentences were now consistent with following the visual preference to look at larger creatures. Importantly, these sentences have been shown to not lead children to commit to an agent-first bias, and as such, the Stroop task was not expected to improve passive performance. Prior work has shown that presenting children with ambiguous Passive sentences of this type with known NP1s indicates that the NP will be previously-established in the discourse context, and therefore reduces children's reliance on the agent-first bias, while novel NP1s indicate new, unfamiliar entities and increase reliance on it. This is particularly true when known NP1s signal given entities relative to novel NP2s, where children appear to largely withhold the agent-first bias (Huang & Arnold, 2016; Huang & Ovans, 2022).

2.3.1: Participants

For Experiment 3, 32 additional five-year-olds (4;0-6;6, mean 5;1, 18 female, 14 male) were recruited. Five additional children were tested but were excluded from the final sample due to inability to complete the experiment (3) or unexpected equipment failure (2). As in the first two experiments, children were recruited from schools in the University of Maryland, College Park community. All children heard English as their primary language, and assented to participate in the study. Children received a small donation to their schools for participating. All procedures were approved by the University of Maryland Institutional Review Board.

2.3.2: Materials

In Experiment 3, novel creatures were again used, as in Experiment 1. Visual scenes, Stroop trials, and fillers were all identical to Experiment 1, as was the pseudorandomized trial order. Unlike Experiment 1, sentences always mentioned the animal displayed in the center of the screen, which was either the agent or patient of the verb, as the first NP. Likely agent and patient creatures were again presented to the left and right of this animal, and mentioned with a novel word in the sentence this time as the second NP, as in "The monkey will be quickly chasing the Furpin… Click on the Furpin."

Audio recording: Audio files for filler trials were the same as the ones used in Experiment 1. For target sentences, audio files were re-recorded from Experiment 1 (instead of re-spliced) to sound naturalistic, but were recorded under similar conditions and with the same speaker using child-directed prosody in order to keep changes to a minimum.

2.3.3: Procedure

The experiment proceeded similarly to Experiments 1 and 2. The set-up, eye-tracking equipment, calibration, trial order, Stroop trials, fillers, and session duration were all identical to previous iterations.

2.3.4: Analysis

As in the prior two experiments, children for whom there was trackloss for more than 33% of their total data were excluded (2 children). Three children did not

complete the experiment and were subsequently excluded. For two children, sudden equipment failure meant they were not able to complete the experiment and they were excluded as well. This resulted in 7 additional children being excluded. Data from the remaining 32 children was analyzed.

Time-course data was again averaged for each word in target sentences, with the average proportion of fixations to the likely agent and likely patient creatures plotted (see Figure 2.11). Interest area bounding boxes matched those from Experiment 1.

The critical region for Experiment 3 was again directly following the onset of verb morphology, until the onset of the following sentence, with sentences aligned to the onset of verb morphology, and looks were shifted 200ms. The region in question lasted 1174ms on average, see Table 2.7 for full details.

Region	Average length in milliseconds (sd)
The	119(62)
Known noun	761(98)
Will be	340(54)
Adv	738(106)
Verb stem	408(155)
Verb morphology	403(87)
The	124(52)
Novel noun	647(122)

Table 2.7: Average duration of regions in Experiment 3

Again, fixation exclusion and cleaning criteria matched Experiment 1. Blinks, artefacts, and looks outside the areas of interest were excluded, and the remaining proportion of looks to the likely agent and likely theme areas were analyzed. Logistic mixed effects regression models with random intercepts and slopes for participants and items were fit to these data, with sentence type and prior Stroop condition as fixed effects.

2.3.5: Results

Here once more, correct looks on actives were indexed by looking toward the likely theme while correct looks on passives meant looking toward the likely agent.

Stroop results: Children were slightly more accurate than in Experiment 1 at responding during congruent Stroop trials: average accuracy for congruent trials was 70%. For incongruent Stroop trials, children again had more difficulty. Average accuracy for incongruent trials was 49%. 81% of times when they did not succeed were failures to respond in time, and the other 19% were errors in response color word. Of these incorrect word errors, 53% were the result of children naming the lure color word, while the remaining 47% were the result of children making unclear responses or unrelated color words.

Previous Item Type	Current Item Congruency	Accuracy
Stroop-congruent	Congruent	.78
Stroop-congruent	Incongruent	.58
Stroop-incongruent	Congruent	.80
Stroop-incongruent	Incongruent	.60
Target sentence-congruent	Congruent	.60
Target sentence-congruent	Incongruent	.53
Target sentence-incongruent	Congruent	.73
Target sentence-incongruent	Incongruent	.43

Table 2.8: Stroop accuracy by previous trial type for Experiment 2

As in Experiment 1, Stroop accuracy varied only very slightly as a result of previous trial (see Table 2.8). When the prior trial was an incongruent Stroop trial, children were slightly more accurate at responding to a successive incongruent Stroop trial (see Figure 2.10). A mixed effects regression model with random slopes and intercepts for items confirmed that the interaction between Stroop type and prior Stroop item type was not significant (β=.04, SE=.49, t=.08, *p*=.94). However, there was once more a significant main effect of current Stroop trial type such that children were more accurate on congruent Stroop trials than on incongruent ones (β=.94, SE=.27, t=3.54, *p*<.001). The verbal answers children gave to Stroop trials were again not measured for reaction time, leaving open the possibility of a speed-accuracy tradeoff not observable from accuracy data alone.

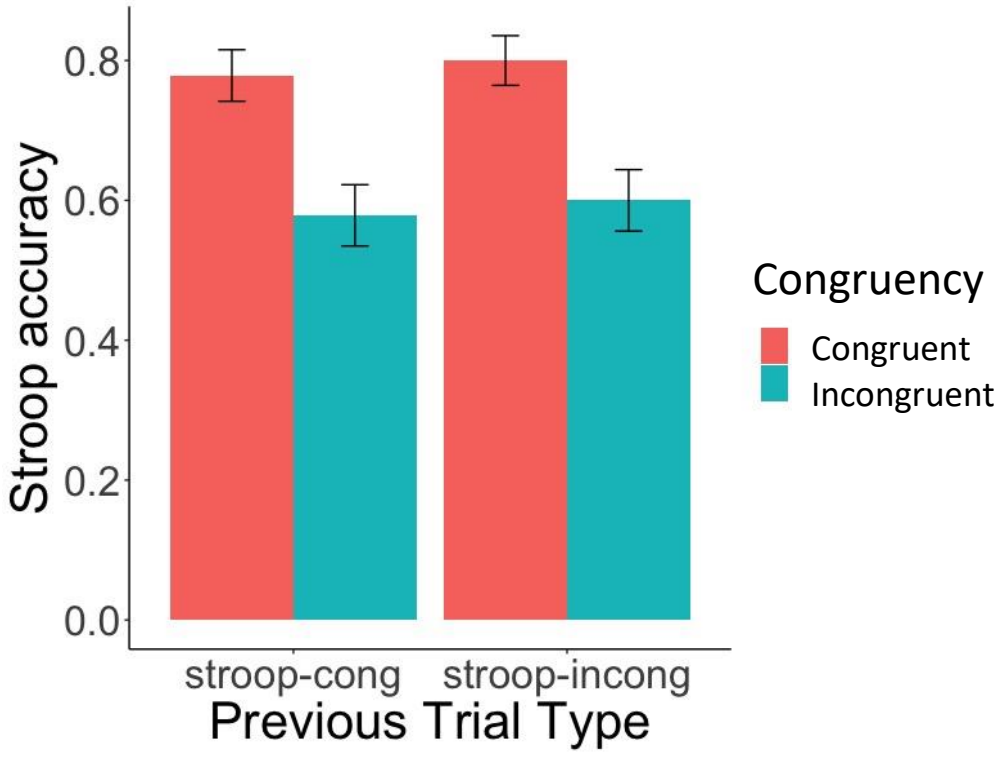

Figure 2.10: Stroop accuracy by previous Stroop trial type

Act-out results: Children were once again fairly accurate in their act-out actions for Experiment 3. They selected the target animal on 77.6% of trials, compared to 63.1% in Experiment 1 where children were presented with the same visual scene (see Table 2.9 for by-condition break-down). This was a bit higher than their overall accuracy for looks during the critical region for target sentences. As with the previous two experiments, act-out accuracy did not vary as a result of prior Stroop trial type, but was numerically similar to the looking-time interaction.

Condition	**Act-out Accuracy**
Congruent-active	75.0%
Congruent-passive	89.6%
Incongruent-active	64.6%
Incongruent-passive	81.3%

Table 2.9: Act-out accuracy for Experiment 3

Stroop x Sentence interaction: Unlike Experiment 1, there was now no interaction between prior Stroop trials and children's performance on active vs. passive sentence trials (Figure 2.11). For active sentences, children exhibited the typical post-conflict slowing effect. They were less accurate at looking to the target animal following incongruent Stroop trials than when following Congruent Stroop trials. For passive sentences, children's looks were also less accurate following incongruent Stroop trials (vs. Congruent Stroop trials). There was a main effect of sentence type such that children were more accurate for active sentences than for passives (β=1.41, SE=.64, t=2.20, *p*=.02). There was no main effect of prior Stroop trial type (β=.78, SE=.65,

t=1.19, *p*=.23). There was also a marginal interaction such that children were more accurate for passive sentences following congruent Stroop trials, but not for actives (β=2.13, SE=1.13, t=1.88, *p*=.059).

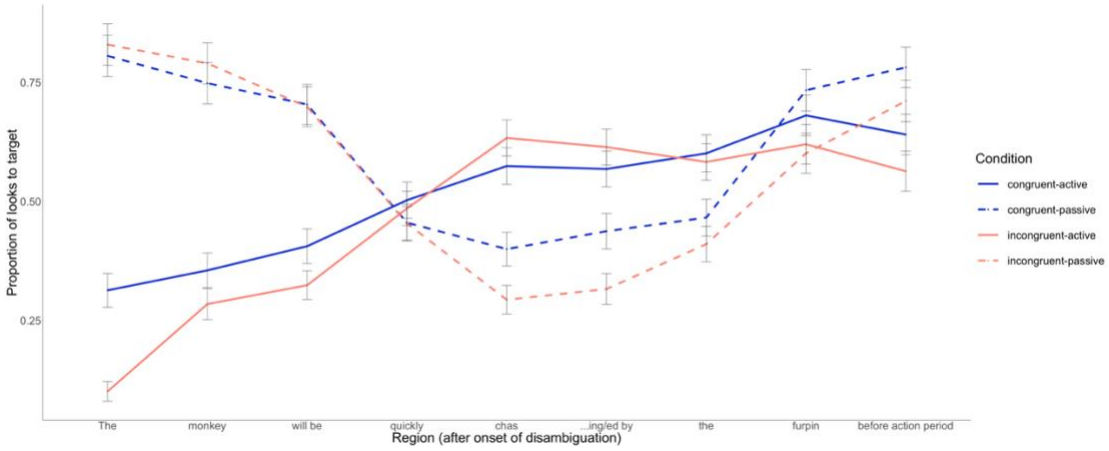

Figure 2.11: Proportion of looks to target creatures for Experiment 3

As can be seen in Figure 2.11, children again sometimes fixated the target animal prior to the point of disambiguation, and therefore switch analyses were performed. This resulted in the removal of 34% of trials (see Figure 2.12 for switch analysis results). The StroopxSentence interaction was significant after the removal of these trials, such that children looked more toward the target for passive sentences following congruent Stroop trials, but not for active sentences (β=3.92, SE=.15, t=27.0, *p*<.001). There was also still a main effect of sentence type such that children

looked more accurately for active sentences (β=.83, SE=.03, t=25.9, p<.001).

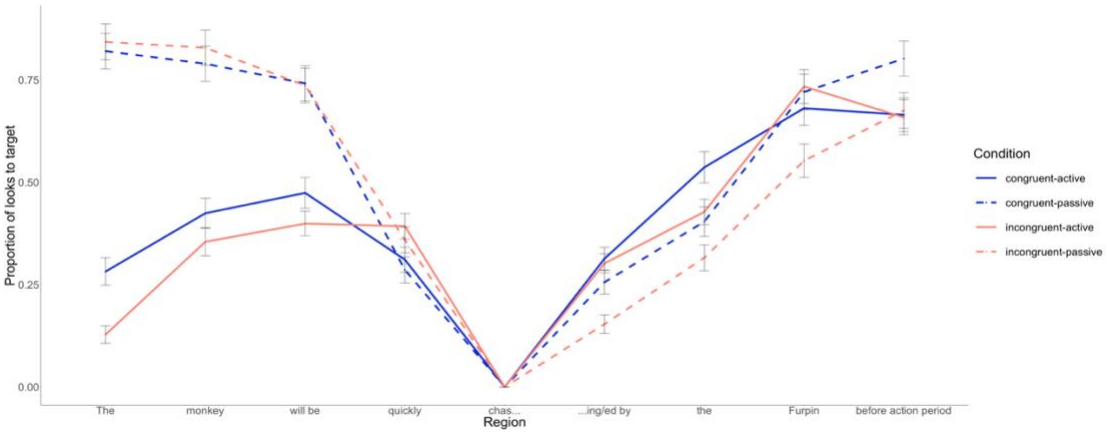

Figure 2.12: Switch analysis for Experiment 3

2.3.6: Discussion

In Experiment 3, children were presented with sentences that varied minimally from the ones they were presented with in Experiment 1, with only the order of the known and novel nouns reversed. Yet this change had a large impact on children's subsequent experience in navigating syntactic ambiguity.

To begin with, Stroop data reveal a relatively similar story to Experiments 1 and 2. Children performed as expected - they were more accurate on congruent Stroop trials than incongruent ones. Additionally, Act-out data once again accorded with eye-movement data numerically, but failed to show a significant effect. This result is again not unexpected, for the reasons outlined above. As expected, children were also more likely to look to the likely agent than the likely patient on all trials.

The post-conflict slowing effect was again observed in active sentences: children were less accurate at looking to the likely patient after disambiguation for

actives following incongruent Stroop trials than they were following congruent Stroop trials. Interestingly, though, this effect was no longer reduced for passives.

This result seemingly accords with the conclusions of Huang & Arnold (2016), who showed that while novel words tend to generate a (weak) agent-first bias, when known nouns are used in their place children appear not to actually commit to the agent-first bias at all. This they argue, is because there is reason to fall back on the agent-first bias for novel nouns. In the absence of semantic information, the bias is a reasonable scaffold to use to quickly understand who did what to whom when parsing incrementally. On the other hand, known nouns being used to refer to something that has two clear potential roles in the visual scene gives children pause. When encountering "The cat…" while seeing a cat engaged as the agent of one action and the patient of another, it makes sense to hold off judgement until more information is received.

The goal of Experiment 3 was to take advantage of this distinction. Given a temporarily ambiguous sentence with known nouns preceding novel nouns, in which children have been previously shown not to commit strongly to the agent-first bias, do we still see that cognitive control engagement helps children recover? The results of Experiment 3 indicate that this is not the case: when children do not initially commit to a bias to begin with, engaging cognitive control does not appear to help children recover from that bias.

Instead, Experiment 3 appears to reveal the effect of post-conflict slowing on the passive trials as well. In the absence of an agent-first commitment, this effect is revealed to apply across sentence types. This revelation makes it perhaps more

reasonable to assume that the results of Experiments 1 and 2 happen in spite of this post-conflict slowing effect. In effect, engaging children's cognitive control system appears to be helping them fight the effects of experiencing difficult Stroop trials to begin with.

2.4: General Discussion

The results of Experiment 1 demonstrate that following cognitive-control engagement, children are aided in their ability to recover from their misinterpretation of passive sentences, compared to an active/congruent baseline. These results argue in favor of the Cognitive biasing hypothesis over the Depletion account, since the Depletion account predicts worse performance on back-to-back trials that use up some amount of cognitive control. The Cognitive Biasing account, on the other hand, predicts the observed results: better performance on passive sentences following cognitive-control engagement, since children's upregulated cognitive-control system causes them to attend more to cues that are generally reliable, in this case verbs, which for passive sentences are a revision cue, not a misleading one. The Cognitive Biasing hypothesis similarly explains the discrepancy between the current results and those of Huang et al. (2016) – since in the "put" task verbs are misleading-but-reliable cues, attending to them more following cognitive-control engagement will lead children astray, as they find.

Experiment 2 adds to these findings, serving to rule out the possibility that children may simply be attending to the visual scene without parsing the sentences accurately. Not only do the results of Experiment 2 serve to conceptually replicate the central finding of Experiment 1, the confounding link between sentence type and

referent is broken here. Even when correct looks for passive sentences mean looking to the likely agent and correct looks for actives mean looking to the likely patient, cognitive control engagement appears to help children overcome their agent-first bias and reach the final interpretation for passive sentences.

Further evidence that cognitive control engagement specifically helped children overcome the agent-first bias comes from Experiment 3. Here, children were given sentences previously shown to not lead them to follow this bias at all. Now, cognitive-control engagement failed to help children overcome parsing ambiguity and instead lead to a decrement in performance, as with the relatively unambiguous active sentences. This helps to delineate the bounds of how children's cognitive-control system interacts with their parser. Namely, these results fit with the Botvinick et al. (2001) conflict monitoring account in that cognitive control takes effect when there are two competing alternatives, each with a comparable level of activation. When this is not the case, as in Experiment 3 where children are not lead to assume an agent-first bias, cognitive-control engagement does not lead to superior conflict mediation.

It should be noted that it remains unclear whether children's specific non-adult like performance on the Put task, in the absence of a conflict-adaptation paradigm, is due to the relative engagement status of their cognitive control system (see Choi & Trueswell, 2010, who demonstrate that in Korean where verbs are reliable cues but occur late, children still have difficulty with syntactic revision). That is, the claim here is not that children always rely on reliable cues when presented with representational conflict. The claim is that what it means for children to be in a more "engaged" state of cognitive control is that this system, which mediates

representational conflict, up-weights processing cues that are judged to be task-relevant. What this means in the context of sentence processing is that children seem to up-weight information that comes from ordinarily reliable parsing cues, such as verbs.

These results raise several further questions. For one, if children are attending more to "reliable" cues when their cognitive-control system is relatively up-regulated, what measure of reliability are they using? This will be addressed in the studies outlined in the next chapter.

Chapter 3: Evidence for reliability

The results of Experiments 1-3 provide support for several intriguing conclusions. For one, they are consistent with a growing literature showing that an individual child does not have a set amount of cognitive control ability, but rather has a mental system that could be in a more or less engaged state (Botvinick et al., 2001; Luna et al., 2010; Braver 2012; Hsu & Novick, 2016; Huang et al., 2016). As the previous chapter discussed, evidence for this comes from the finding that providing children with a task that requires relatively more or less conflict mediation influences their ability to mediate subsequent, unrelated conflict on a sentence processing task, even in a within-subjects experimental design.

These results now raise further questions about the precise nature of this control system and the dimensions of children's input it acts upon. Experiments 1-3 suggest that cue reliability is a criterion for parse re-weighting when children's cognitive control system is engaged. However, it remains unclear what aspect of children's input their cognitive-control system homes in on as it interacts with the developing parser. Do children track a metric of reliability, and attend more to cues that are more reliable when their cognitive-control system is highly engaged? Perhaps instead the cognitive control system operates over frequency, biasing children to allocate more attention to common words more generally? The results of the studies in Chapter 2 are consistent with at least three possible hypotheses in this vein, which I'll attempt to disentangle in this chapter:

A) The Reliability Hypothesis. Under this hypothesis, verbs, as a grammatical category, are reliable cues. In other words, the probability that the subcategorization

information gleaned from a verb will lead to an accurate guess about upcoming role assignment or an upcoming parse is high. For example, given the sentence fragment "The dog chased…" children have enough information to reason that there are two likely roles that need to be filled. Once the verb has been encountered, children may infer that "chased" requires an agent and patient (in this case chaser and chasee). Using an agent-first bias (Abbot-smith et al., 2017), they may infer that the dog is the chaser, and that there will be an upcoming noun has a high probability of being the chasee. There is a high probability that these actors will be mentioned in the sentence, and this information can be gleaned from the verb alone.

Under this account, the process of interpreting ambiguous sentences involves, in part, identifying parsing cues in your input and assessing the likelihood that each cue will be an accurate indicator of the eventual intended sentence parse. For example, a child hearing the fragment "Put the frog…" might reason that there are three roles that their interlocutor will fill: agent, patient, and the location of the putting event, since "put" is ditransitive and must take a PP as one of its objects. This gives the child some expectations about upcoming sentence structure: namely, that a prepositional phrase is expected and it will likely specify the location of the putting event. The higher the likelihood that each time the child hears "put" these roles are filled, the more reliable an indicator of upcoming structure "put" can be said to be. This hypothesis can be formalized as computing $P(S|V)$ where S is a particular sentence structure being observed (e.g. a PP attaching to the VP), and V is a particular verb token. $P(S|V)$ is therefore the probability that a particular sentence structure is observed, given that a particular verb V is encountered. For example, $P(VP->PP|Put)$

is likely quite high, while *P(VP->PP|Jump)* is likely quite low (I return to these claims and attempt to provide empirical support for their presence in child-directed speech in Chapter 4).

Under this hypothesis, children rely on verbs in Experiments 1-3 precisely because verbs are reliable cues in this way, and attending to them is an efficient strategy when multiple cues in a sentence are in conflict. This is akin to the role that cognitive control is thought to play in the Stroop task (Botvinick et al., 2001). There, conflict adaptation occurs when word and color cues are in conflict, and when participants are in a heightened control state, they are then more likely to attend to the cue that is reliable for the task at hand. In the case of the Stroop task, this is the cue that is in line with their goals in the task, namely the ink color. In the case of sentence processing, cues that reliably help you predict sentence structure serve the same purpose, and may therefore be what children rely on when their cognitive control system is differentially more engaged in a sentence processing task.

B) The Frequency Hypothesis. Verbs as a category are particularly frequent parts of speech (e.g. compared to prepositions). Under this hypothesis, children rely on verbs not necessarily because they are reliable but because they are both indicators of upcoming structure and occur frequently. The relative reliability of verbs as parsing cues, under this hypothesis, does not explain why children appear to use them more following cognitive-control engagement in Experiments 1-3. In particular, when sentential cues are in conflict (as in ambiguous sentences) and when children's cognitive-control system is relatively up-regulated, children may rely on verbs because their frequency makes them particularly familiar or recognizable. This

hypothesis can be formalized as children simply computing *P(V)*. For example, *P(Put)*, or the frequency of the word "Put" is likely quite high, while *P(Situate)* is likely much lower. This account explains children's reliance on verbs in the "Put" task and in Experiments 1-3, just as the Reliability Hypothesis does.

Support for this hypothesis comes from the finding that children who have recently heard a verb used in a particular syntactic frame will assume it will be used in that frame again (Peter et al., 2015). In other words, children experience structural priming when the stimuli used include biased verbs, and the priming effect is greater if the verbs are biased toward the primed structure to begin with. This indicates that when children are unsure of how to interpret their input, they default to parsing according to cues they are more familiar with, not necessarily ones that are more reliable. A limitation of this conclusion for purposes of the present work, though, is that it is unclear whether and how structural priming might interact with children's cognitive control system. It is possible that children parse according to a recently-used structure when their cognitive control system is up-regulated, but it is also possible that they do so in the face of general uncertainty, or all the time, as a general default.

C) The Privileged Role Hypothesis. Under this hypothesis, verbs as a category inhabit a privileged role among parts of speech, for a reason other than their frequency or reliability. For example, it has been argued that argument structure is projected from the verb, and that the relations between arguments and the verb are entailed by the verb itself (e.g. Dowty, 1979/2012; Chomsky, 1981; Stowell, 1981, see also Williams, 2015 for discussion). If this projectionist approach is correct and

additionally children are privy to knowledge of the verbs' privileged status, it may underlie their reliance on verbs. Put another way, verbs' importance is not borne out of their reliability or frequency, but from their preeminence in the sentence. While it may not be precisely clear how to characterize this "other" option, the studies in this chapter are designed to test its general ability to account for children's reliance on verbs following cognitive-control engagement.

The purpose of this chapter is to disentangle these three hypotheses. Support for the Reliability Hypothesis comes from studies of children's use of verb bias statistics. Snedeker & Trueswell, (2004) presented children with globally ambiguous sentences such as "Poke the bear with the stick," where even upon reaching the end of the utterance it is unclear whether in the intended parse the PP "with the stick" attaches to the NP, modifying the bear, or attaches to the VP, and specifies the instrument of poking. They manipulated how biased the initial verbs were toward these two analyses, and found that children relied on verb bias, but not referential information to make their decision. From this, the authors conclude that children are more likely to use highly reliable cues (like lexical constraints) to guide their parsing decisions when a choice must be made. Concurrent with these results, Trueswell & Gleitman (2004) also argued that children tend to "fall-back" on parsing cues that have proved reliable in the past when conflict arises between multiple potential parses of a sentence.

Further support for the Reliability Hypothesis comes from studies that have corroborated and extended these findings. Kidd, Stewart & Serratrice (2011) presented children with globally ambiguous (but biased) sentences like the ones used

in Snedeker & Trueswell (2004). Children relied on verb bias even in cases where referential information pushed adults to revise their initial analysis. For example, when presented with a sentence like "Cut the cake with the candle" and a display that contained a candle-less cake, a cake that had a candle in it, a lone knife, and a lone candle, children generally opted for a VP-attachment reading, using the candle to cut the cake, even despite its implausibility. Narrowing down the consideration set to action verbs vs. stative verbs increases children's chances of a successful parse still further. This, the authors argue, demonstrates that children are following a reliable parsing cue in the face of ambiguity: according to a corpus study conducted by Kidd & Bavin, (2007), choosing VP-attachment would lead to a correct parse more often than not, and children are taking advantage of this regularity.

Taken together, these results indicate that children tend to parse according to cues that are reliable when they are faced with ambiguous input. This is particularly evident when using stimuli that differ with respect to verbs' structural biases, where ambiguity is easily created and the reliability with which a particular verb saves the child from the ambiguous string can be tightly controlled. However, these results concern ambiguity processing in the general sense and do not perfectly speak to the question of whether children's developing cognitive control system biases them to parse using a reliability heuristic. To determine this, we must combine the methods used in these studies with a design that creates changes in children's cognitive control state in real time.

In the present study, participants were presented with a task that engaged children's cognitive control system in a conflict adaptation paradigm, as in

Experiments 1-3. Following this task, children saw sentences that systematically varied both the relationship between the verb and its ability to predict an upcoming parse (*P(S|V)*) while controlling for the frequency of the verbs themselves (*P(V)*). In order to rule out hypothesis C (that children rely on verbs as parsing cues because they inhabit a privileged role unrelated to their reliability or frequency), the comparison of interest will be between sentences that are minimally different such that the only variation is the particular verb used. It should be noted that this design does not explicitly test for the possibility that verbs inhabit a privileged role among parts of speech, but is instead designed to rule this out as the only dimension along which cognitive control engagement acts, regardless of reliability and frequency.[2]

3.1: Experiment 4: Imperative task, Instrument vs. Equi-biased verbs

3.1.1: Participants

60 Children aged 4;0 to 6;6 were recruited from the Infant and Child Studies Consortium Database at the University of Maryland, College Park. Children and their guardians participated virtually, communicating with researchers for set-up and troubleshooting via Zoom or another video conferencing application.

3.1.2: Procedure

For the Flanker task, children were instructed to help a cartoon protagonist follow the middle fish in a set of 5 (see Figure 3.1). They received two "first-round" practice Flanker trials that only contained one fish, and instructed them to press the

[2] Future work might consider a more direct test of this hypothesis, wherein verbs are not cues to recovery from misinterpretation at all.

"F" key on their keyboard when the fish pointed left, and the "J" key on their keyboard when the fish pointed right.[3] These first-round fish were included in order to orient children to the correct response buttons. Next, children saw 10 "second-round" practice Flanker trials. These trials contained five fish in a row, mimicking the eventual experimental trials. These trials were included so that children could practice responding to congruent and incongruent trials prior to the main interleaved portion of the experiment.

Before beginning the test phase, children were also presented with two practice sentence trials, that were set-up just as an experimental sentence trial but not included in data-analysis as they were used to explain the nature of the "pretend" task to children. On these trials, children were told they were engaging in a pretend task and were free to respond however they liked.

During the test phase of the experiment, children were presented with more Flanker trials, interleaved with sentence trials. On sentence trials, children heard globally ambiguous sentences instructing them to perform a task such as "Pretend to poke the elephant with the carrot", while viewing corresponding images on their screen. During the practice trials, children were told to press any button whenever they were "done pretending." Children's eye-gaze toward the corresponding images on the screen was then measured in order to assess real-time attachment preferences following cognitive control engagement.

[3] In the event that a child could not reliably distinguish their left and right, they were told to press the button that showed where the fish was facing.

Children completed a total of 8 Flanker-sentence pairs, along with 16 filler fish trials in order to disguise the manipulation. Sessions lasted approximately 20 minutes and children generally enjoyed the study design and found it easy to complete.[4] Parents were asked to complete the study in a quiet room, but due to the nature of online testing, the household environment naturally has more distractions than the lab. Despite this, children were generally highly engaged throughout the experiment.

As children performed the sentence-processing task, experimenters watched over their shoulder to record their act-out actions (via a separate device, often held up by a very patient parent). Since children were instructed to "Pretend to VERB the NOUN with the NOUN," their physical movements in response to this imperative can be used to reveal their attachment preferences as well, albeit not in as fine-grained a way as their eye-movement data. For the eight target trials, experimenters recorded whether children made a movement that seemed to be consistent with an instrument-like interpretation (e.g. for "Poke the elephant with the carrot" this might mean miming dragging the lone carrot to one of the elephants and poking at it with the imaginary carrot), or a modifier interpretation (e.g. using their finger to poke the elephant that had a carrot).

While children performed the entirety of the experiment on their webcam-enabled home computers, experimenters were also present for the entirety of the session on the secondary device (usually a phone, tablet, or second computer).

[4] When informally asked "was this boring or fun" and "was this easy or hard" (with positive and negative terms relatively counterbalanced), almost all children reported the session to be both fun and easy. While informal, these poll results are encouraging given the online nature of the experiment.

Children's guardians were asked to position the experimenters off to the side so that they could see both the screen on which the experiment was presented and child, in order to both ensure the experiment was presented correctly, help troubleshoot, and code the child's act-out actions.

3.1.3: Materials

Flanker task: Children were presented with a Flanker task containing congruent and incongruent trials. Because of the virtual format of the experiment, the dog Stroop task was not used to engage children's cognitive control system. This task requires real-time feedback during training that can be difficult to administer virtually, and relatively high-fidelity audio recording is necessary to capture children's verbal responses. Instead, a Fish Flanker task (Erikson & Erikson, 1974) was used in its place. This task has the additional benefit of having children directly interact with the computer, keeping their attention, and provides easily-analyzable reaction time data so that speed-accuracy tradeoffs can be assessed.

For this particular Fish Flanker task, children saw images such as the one in Figure 3.1, and were asked to press the "F" or "J" keys on their keyboard to indicate the direction of a middle fish, ignoring the ones flanking it. On 50% of Flanker trials, children saw a congruent image (all 5 fish facing the same direction), and on the remaining 50% they saw an incongruent image (with the central fish facing the opposite direction of the flankers). Performance on the flanker task was evaluated by both accuracy and reaction time in pushing the appropriate keys.

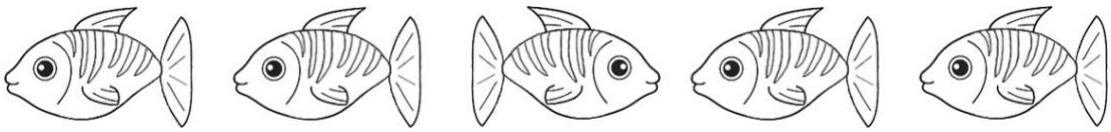

Figure 3.1: Example incongruent stimuli for the Flanker task used in Experiments 4 & 5

Sentence task: Interleaved with Flanker trials were globally ambiguous sentence trials. Sentences were verb-initial imperatives that instructed children to play a pretend game with images on their screen. For example, children heard "Pretend to poke the elephant with the carrot" while viewing an image of an elephant that has a carrot, a separate image of a lone carrot, and an image of an elephant that has a bowtie (e.g. Figure 3.2). The independent variable of interest was the attachment preference of the embedded verb (manipulated between subjects). In the Instrument condition, children heard verbs such as "poke" that reliably (according to adult norming data, discussed below) predicted an upcoming PP to attach to the VP, giving an instrumental reading (e.g. "Poke the elephant with the carrot" is more likely to mean "poke the elephant using the carrot" than "poke the elephant that has the carrot"). The goodness of fit of the prepositional object nouns were normed so as to not influence interpretations of verb bias (see norming section below). In the Equi condition, verbs were equally likely to predict VP or NP attachment, making them relatively unreliable predictors of a particular upcoming structure.

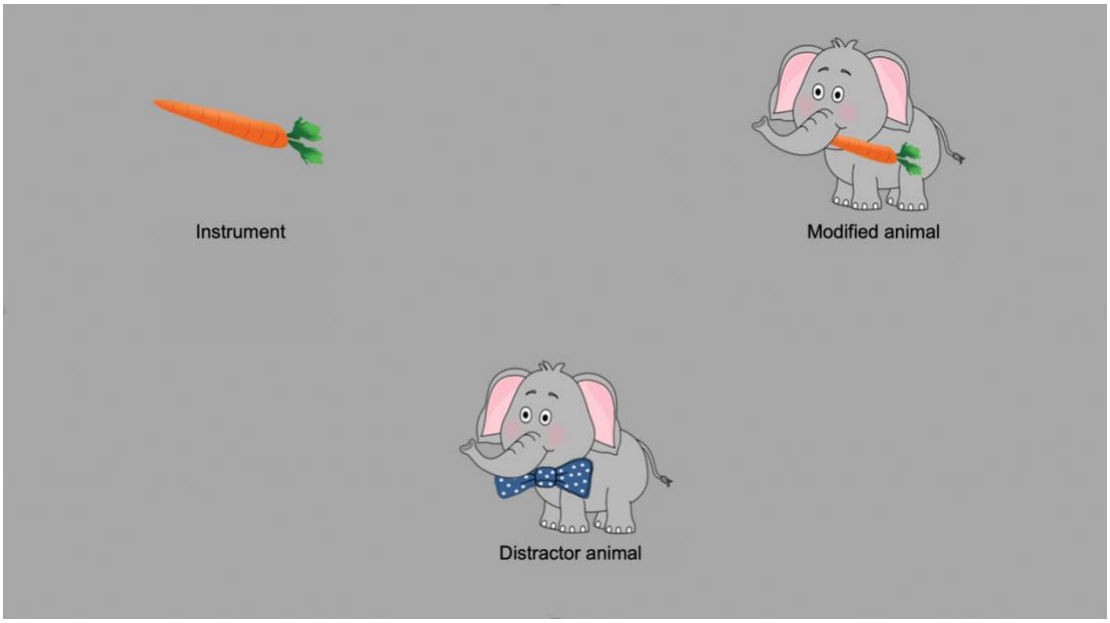

Figure 3.2: Sample visual stimuli for the sentence trials in Experiments 4 & 5. The text labeling the images is for illustrative purposes and was not included in the experiment itself.

3.1.4: Norming

Following Snedeker & Trueswell (2004), several rounds of norming were carried out to identify a particular set of stimulus items that would contain appropriately biased verbs, but would not bias participants toward modifier- or instrument-like responses for unintended reasons.

First, a set of norming experiments was carried out to determine the bias of each of a set of 24 verbs that children could be expected to know. The goal of this norming was to identify 8 modifier-biased, 8 instrument-biased, and 8 equi-biased verbs. Adult participants from Amazon Mechanical Turk were asked to complete a Cloze-task (Taylor, 1953) for each of several sentences. They were told the following instructions: *Please provide a short completion to the following sentence fragments.*

Fill in the first word or words you can think of that seems to complete the sentence in a normal way. The sentences should be grammatical English sentences. Participants then saw a series of sentence fragments for each verb like "She will push the person with..." and were asked to fill in the final word or phrase. Participants' responses were then coded as "Modifier" (e.g. *…the black hair)* "Instrument" (e.g. *…his hands)* "Ambiguous" (e.g. *…a box)* or "Irrelevant" (e.g. *…conviction).* Verbs that engendered irrelevant scores more than 50% of the time were not used. Verbs were considered modifier-biased if they produced modifier-like responses more than 60% of the time (though it should be noted that several modifier-biased verbs such as "pick" and "choose" were given modifier-like responses 100% of the time). Verbs were considered instrument-biased if they were given instrument-like responses more than 80% of the time, and equi-biased if they were not given modifier-like responses more than 60% of the time or instrument-like responses more than 80% of the time. While it should be noted that this scale skews the equi-biased responses toward the instrument-like end of the scale, the average responses for equi-biased verbs were squarely between the average responses for verbs that were considered instrument- and modifier biased. Namely, after taking out irrelevant scores, responses for modifier-biased verbs were modifier-responses 88% of the time (and instrument responses 11% of the time), the responses for instrument-biased verbs were modifier-responses 11% of the time (and instrument-responses 89% of the time) and the responses for equi-biased verbs were modifier-responses 38% of the time (and instrument-responses 62% of the time).

Since a pilot of the norming experiment revealed overwhelming instrument-like responses for this sort of Cloze task, several measures were taken to ensure that participants knew modifier-like responses were welcome. First, the following example was provided after the instructions: *For example, if the sentence fragment was "She will pat the dog that is...", you might continue it as: "She will pat the dog that is BROWN". Or if the sentence fragment was "He will eat the…", you might continue it as: "He will eat the DELICIOUS CHOCOLATE-CHIP COOKIE".* These instructions ensured that participants were aware that modifier-like interpretations were possible, but not necessary. A second measure to increase the overall plausibility of modifier responses was to interleave "that has" filler sentences into the target sentences. For example, participants also performed the cloze task on sentences like "He will toss the book that has the.." containing non-target verbs. A final measure was to always include "person" as the final noun that verb bias was tested with. Since the modifier reading becomes more plausible when it's generally more likely that the particular noun will be modified, "person" was chosen because people are a relatively good candidate for differentiation. (i.e. Some nouns like "ant" are prima facie unimportant to differentiate. It usually isn't important to specify which ant you're referring to. It's more likely that when referring to "the person" you'd feel the need to use a modifier like "with the hat").

While these measures increased the overall likelihood of modifier responses in this particular task, it seems unlikely that they would have manipulated the relative amount of bias of any particular verb *relative* to the others. For this reason, they were

employed as a way to normalize the instrument-skewed results, to more clearly see finer-grained distinctions between verbs.

Finally, **i**n order to ensure that the particular choice of noun-verb combinations didn't overshadow the verb bias manipulation, another set of norming experiments was carried out. Participants from Amazon Mechanical Turk were asked to judge the plausibility of the particular verb-noun combinations used in the study. Participants were asked to judge the fit of the final nouns used in the test sentences as instruments of the verbs. They were told the following instructions: *You will be asked to decide how reasonable it is to use certain objects as instruments for particular actions. Imagine that you had to do these actions USING an instrument of some kind. How good are the given instruments for this action? For example, it's plausible to use a fork to eat, but not very plausible to use a camera to eat. Don't judge by whether you personally could do these actions, but rather by how plausible it is that the scenario described might happen.* Participants were then presented with sentences that contained an unambiguous instrument reading, e.g. "Can you clean a fox using a brush?" and were asked to respond on a 1-7 Likert scale. End-points were labeled for participants, with a response of 7 being labeled as "Plausible" and 1 being "Not plausible."

Several rounds of this noun norming experiment were carried out to ensure that noun-verb pairs were neither too plausible nor too implausible. In the first round of plausibility norming, 24 participants were asked to judge the plausibility of eight nouns for each of the 24 verbs. Nouns that were given average ratings close to 4.0 on the scale of 1 to 7 were chosen. For verbs where no nouns were given average ratings

between 3.0 and 5.0 across the initial 24 participants' judgments, subsequent rounds of norming were carried out with new nouns until appropriate pairs were found.

Following these norming procedures, three lists of verbs were generated: 8 Modifier-biased, 8 Equi-biased, and 8 Instrument-biased. Across lists, the irrelevance scores (percentage of responses that were not consistent with either modifier or instrument interpretations) were below 25%. Results of an un-paired t-test revealed a non-significant difference between irrelevance scores for instrument (M= .08 SD = .12) and modifier (M= .18 SD = .14) lists (t(14) = 1.4, p=.185). Average instrument bias scores did significantly differ across lists, both for the comparison between modifier (M= .19 SD = .16) and instrument-biased (M= .89 SD = .07) verb lists (t(10) = -11.5, p= 6.87e-07). and instrument and equi-biased (M= .62 SD = .08) verb lists (t(14) = -7.2, p=5.11e-06).)

Modifier (M= .010 SD = .009) vs. Instrument (M= .002 SD = .002), and Instrument vs. Equi-biased (M= .005 SD = .009) verb lists did not differ in frequency (Modifier vs. Instrument t(8) = 2.3, p=.05, Equi vs. Instrument t(7) = .92, p=.38) as measured by Google N-gram (Michel et al., 2011) frequency scores from 2018 (the most recent year frequency data was available on the platform when the data were normed). Finally, Modifier (M= 3.77 SD = .59), Instrument (M= 3.93 SD = .70), and Equi-biased (M= 4.41 SD = .52) verb lists also did not differ in extent to which the final nouns were plausible fits with the verbs in the final round of norming, (Modifier vs. Instrument t(14) = 1.48, p=.64, Equi vs. Instrument t(13) = 1.56, p=.14).

While these sentence norming experiments are time-consuming, there is good reason to believe that they are necessary when conducting experiments that hinge on

these particular distinctions in the attachment biases of verbs. Qi, Yuan & Fisher (2011) demonstrated that non-linguistic general knowledge about the plausibility of events in the world can contribute greatly to comprehenders' interpretations of sentences with with-phrase attachment ambiguity. They presented 5-year-old participants with equi-biased verbs but used training dialogues and nouns designed to bias children toward particular attachment preferences. For example, children heard dialogues that included sentences like "What did Tim use to point at the Tiger? He pointed at the tiger with the red pencil" to bias a VP-attachment interpretation when children listened to subsequent sentences containing the same equi-biased verbs. Children's eye-movements then revealed a significant effect of training – children who heard a particular equi-biased verb in an instrument- or modifier-biased context were more likely to look at an image consistent with that bias. These results, the authors claim, demonstrate that children's interpretations of verb bias are relatively easy to manipulate with distributional information, even for verbs they already know well. Importantly, they point to the importance of controlling for information unrelated to the verb that may contribute to children's interpretation of with-phrase attachment, as the particular choice of dialogue and final noun proved capable of influencing their real-time interpretation of the bias of particular verbs.

3.1.5: Coding

Actions: Experimenter(s) assessed children's act-out actions in real time, as the experiment progressed. They categorized children's actions into six categories: Instrument reaches, wherein children dragged the lone instrument to one of the animals (ideally in a manner consistent with the verb, e.g. they mimed a "poking"

motion with the carrot for *poke*); Modifier reaches, wherein children interacted only with the modified animal (also ideally in a manner consistent with the verb, e.g. using their finger to poke the elephant); Distractor reaches, wherein children interacted only with the distractor; Both reaches, wherein the child performed both a modifier and an instrument reach; Neither reaches, wherein the child did not perform any action; and Unclear reaches, wherein the child performed an action but too far from the screen to determine which items they were pointing at. Both and Unclear reaches were ultimately collapsed, as it was determined that a "both" reach was similarly uninterpretable to an unclear reach. Experimenters also included in this category any additional anomalous actions, e.g. dragging the distractor animal to the modified animal. Since children's eye-movements were considered the main dependent variable of interest and in order not to bias children into interacting with the images in a particular manner, children were never corrected in their actions. After every experimental session with two or more experimenters present, coders compared action codes and resolved any disagreements.

Eye-movements: Children's eye-movements were captured via the webcams on their home computers. Trained coders analyzed these videos, and categorized looks into 7 different categories: looks to the (coder's) top left of the screen, looks to the (coder's) top right of the screen, looks to the bottom middle of the screen, looks to the center of the screen, looks to the child's keyboard, looks off screen (e.g. looking back at a parent), and trackloss (when looks could not be determined, usually due to motion blur, child's position, or because their eyes were temporarily obfuscated). While

coders analyzed the video files frame-by frame, they only marked time-points when the child's gaze changed from the previous code. Looks were coded using VCode (Hagedorn et al., 2008), and left/right codes were matched on to instrument/modifier looks using R (R Core Team, 2021) after coding was complete. Looks were coded from the first time-point at which the trial began (when images appeared on screen, which coincided with the start of the audio files), until children pressed a button to advance to the next trial. While most trials lasted fewer than 10 seconds, occasionally children were distracted and took more time to complete a trial. Looks after 30 seconds were excluded from further analysis, as it was deemed unlikely that these were still meaningly influenced by the sentence presentation, much less the Flanker condition from the prior trial. For further coding details, as well as validation of this analysis method and the virtual-world eye-tracking method in general, see Experiments 6 and 7.

3.1.6: Results

Data from 88 children (ages 4;0 to 6;6) were collected. 30 were run on the equi-biased verb condition, and 30 on the instrument-biased condition. The results are divided into three sections below: Flanker results, Act-out results, and Eye-movement results. Flanker accuracy and reaction time was collected to ensure that children were able to complete the task, and that children's cognitive control system was indeed engaged by the this version of the Flanker task. If it was, children should be slower and/or less accurate on incongruent Flanker trials than on congruent ones. Act-out data was collected as a measure of offline sentence processing, though since children were asked to do a "pretend" task many of their actions were ambiguous or

completely internal. Eye-movements, therefore, were the main dependent measure assessed, and looks to the Instrument vs. looks to the Modified animal were taken as the main assessment of how children interpreted the ambiguous sentences.

Flanker results: Even the youngest children understood and performed well at the online Flanker task (see Figure 3.3). This measure was introduced as a manipulation check, to ensure that the Flanker task conditions induced differing levels of cognitive control demands. Children's average accuracy was 72%. Trials on which children's reaction time was more than 2.5 standard deviations longer than the mean were excluded, which resulted in removing all trials longer than 15.7 seconds. Reaction time was fairly quick and was modulated as expected: children's average reaction

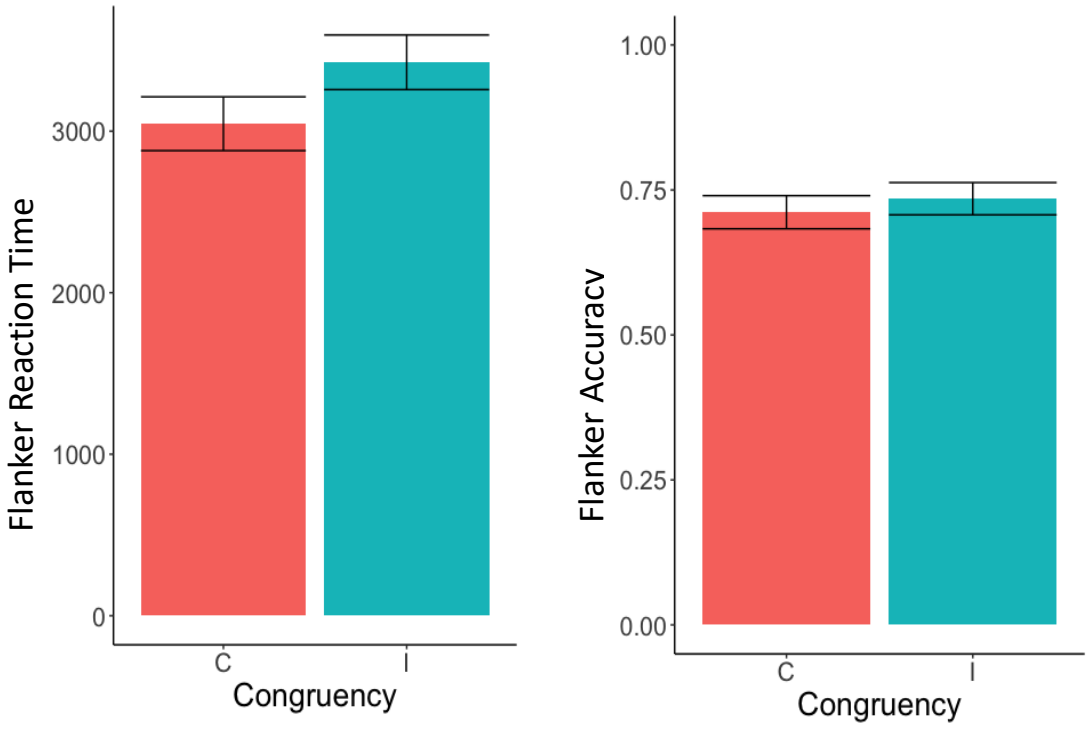

Figure 3.3: Flanker Reaction Time and Accuracy by Flanker trial type for Experiment 4

time to congruent fish trials was 3046ms, and for incongruent trials was 3426ms. A mixed effects regression analysis with random effects of subjects confirmed that this difference was significant ($t(57) = 3.9$, $p<.001$). A similar model confirmed that accuracy did not significantly vary with condition ($t(57) = 1.1$, $p=.282$).

Act-out results: As expected, children were not all consistent in their interpretation of the "pretend" command, limiting the interpretability of these act-out data. Still, some patterns emerged. Frequently, children made clear dragging motions with their hands or attempted to use their computer's mouse or trackpad to drag the actual image of the instrument object to one of the animals, which was recorded as an instrument response. This was the most common response and occurred on 298 trials, accounting for 43.8% of the total responses. Carrying out the action (e.g. poking) on the modified animal was also a relatively common response, occurring on 290 trials (30.7% of total responses). Some children opted for a strong interpretation for the "pretend" command, carrying out the entire process in their heads without making an observable action beyond moving their eyes. This response occurred on 48 trials (7% of total responses). In some cases, children's actions were unobservable because the device set-up did not allow experimenters to see their actions. This was the case on 94 trials (13.8% of total responses). Finally, in some cases, children performed an action inconsistent with the sentence (e.g. performing an action on the distractor item). This occurred on 31 (or 4.5% of) trials.

Overall, verb bias did not have a large effect on children's eventual actions. While children were slightly more likely to make modifier-like actions for equi-

biased verbs than for instrument-biased verbs, children were also more likely to make instrument-like actions after equi-biased verbs (see Table 3.1). As Figure 3.4 shows, children were more likely to make no response or an ambiguous response following instrument-biased verbs.

Children's actions were also broken down by whether the prior fish trial they had just completed was congruent or incongruent (see Table 3.2). As Figure 3.5 shows, children were more likely to make an instrument-like action following an incongruent fish trial than a congruent one. Consistent with this, children were less likely to make a modifier-like action following an incongruent fish trial than a congruent one. While children were slightly more likely to perform no action following an incongruent fish trial, they were also slightly more likely to perform an ambiguous action following a congruent fish trial.

Code	Verb Type	Count
D	E	20
D	I	11
I	E	159
I	I	139
M	E	114
M	I	95
N	E	9
N	I	39
X	E	28
X	I	66

Table 3.1: Children's act-out actions in Experiment 4, broken down by verb type, from a total of 680 trials. In the Code column, "I" refers to an instrument-like action while "M" refers to a modifier-like action. "D" indicates that children did a modifier-like action on the distractor object, "N" indicates the child did no action, and "X" indicates that the child did an ambiguous or non-codable action. In the Verb Type column, "E" indicates that the verb was equi-biased and "I" indicates that it was instrument biased. The Count column contains the total number of trials on which a particular code occurred for that verb type.

Code	Prior Fish	Count
D	C	13
D	I	12
I	C	118
I	I	130
M	C	86
M	I	76
N	C	16
N	I	20
X	C	39
X	I	34

Table 3.2: Children's act-out actions in Experiment 4, broken down by prior fish trial type, from a total of 680 trials. In the Code column, "I" refers to an instrument-like action while "M" refers to a modifier-like action. "D" indicates that children did a modifier-like action on the distractor object, "N" indicates the child did no action, and "X" indicates that the child did an ambiguous or non-codable action. In the Prior Fish column, "C" indicates that the flanker trial directly prior was congruent and "I" indicates that the flanker trial directly prior was incongruent. The Count column contains the total number of trials on which a particular code occurred for that prior flanker trial type.

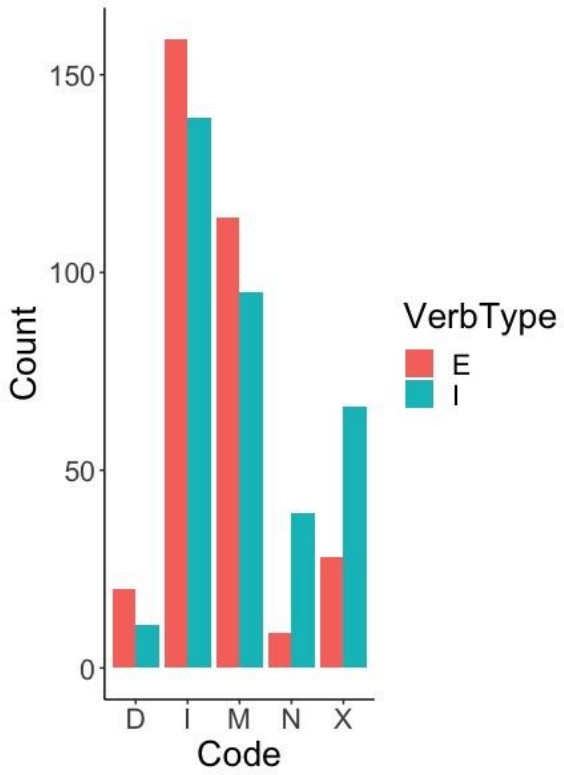

Figure 3.4: Count of the number of trials on which children performed particular act-out actions in Experiment 4, separated by verb type (see Table 3.1 for exact counts and coding key)

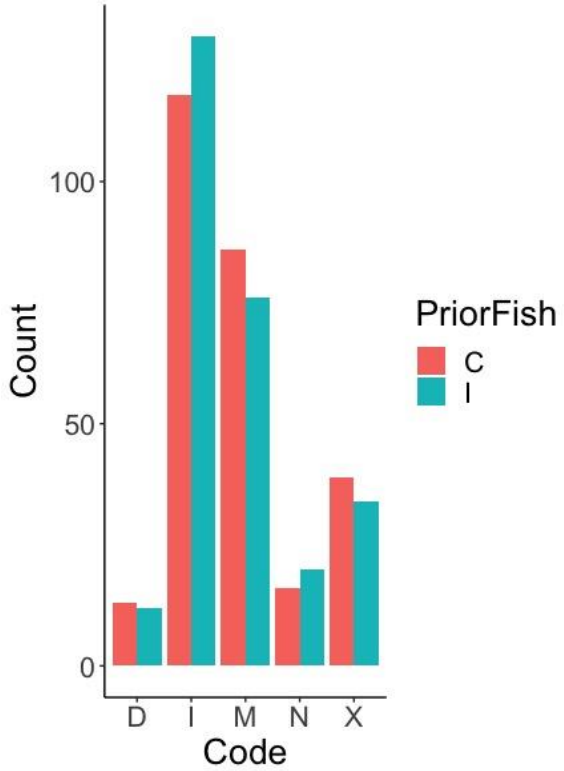

Figure 3.5: Count of the number of trials on which children performed particular act-out actions in Experiment 4, separated by prior Flanker trial type (see Table 3.2 for exact counts and coding key)

As mentioned above, these act-out actions sometimes proved difficult to code, either due to the vantage point of the researchers or children's position relative to the screen. For example, many children performed the "act-out" action in the air in front of the screen, making it difficult to determine which objects they were pretending to interact with. Additionally, children were told they were doing a pretend task and many children opted to do the "pretending" in their head, without making an explicit action. For this reason, these act-out results may be less reliable than the eye-tracking results presented below.

Eye-movement data: Two windows of analysis were used to assess children's looks. Following Snedeker & Trueswell, (2004), looks were analyzed following the onset of the prepositional object, as this time window is indicative of how children are using the PP to restrict reference. Snedeker & Trueswell broke this window into two parts: an early-PP window, 200-667ms after PPobject onset, and a late-PP window, 700-1167ms after PPobject onset. They found effects of verb-bias type on both the early and late PP windows, so this window was therfore collapsed in the present analyses into a general PP window, 200-1200ms after the onset of the final noun.

There is, however, a crucial difference between the present study and Snedeker & Trueswell (2004) – distractor instruments were not used, in order to make hand-coding from webcam videos easier by only requiring coding looks to three on-screen locations instead of four. Also for this reason, target animals and instruments were always presented on the top row. This means that children who were paying attention might realize that they don't necessarily need to wait until they hear the

instrument to perform the action. If that's the case, the instruction may be followed as soon as they hear the verb. For this reason, results were also analyzed from the verb onwards (spanning the window from 200ms after verb onset (first dotted line in Figure 3.6) to 1200ms after POnoun onset (final dotted line in Figure 3.6). As shown below, the choice of time-window did not affect results.

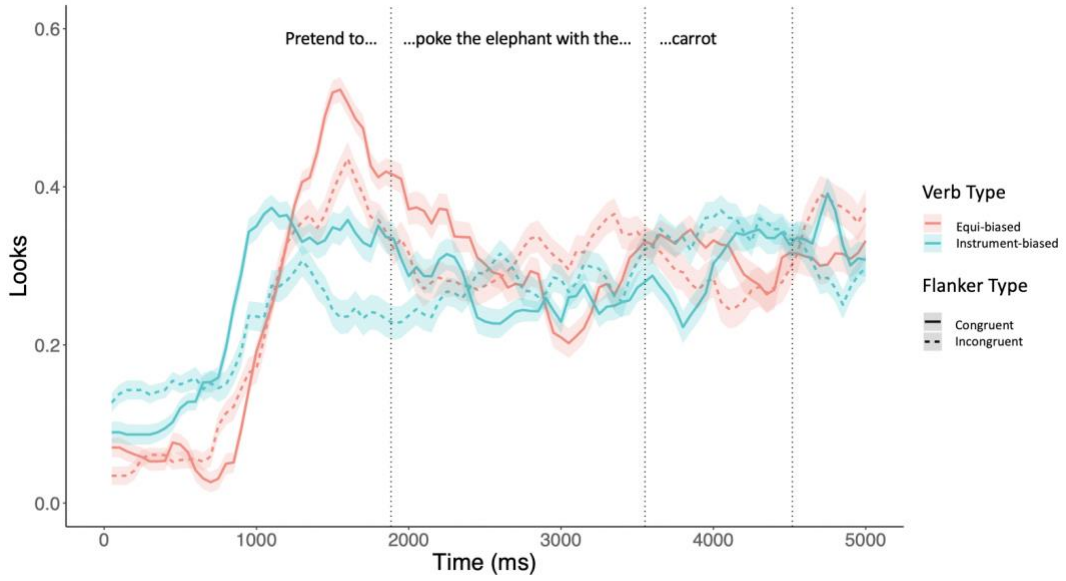

Figure 3.6: Looks to the modified animal for Experiment 4, separated by verb type and prior Flanker trial type. Dotted lines indicate verb onset, PO noun onset, and 1000ms after PO noun onset, respectively. Lines are adjusted 200ms to account for saccade planning.

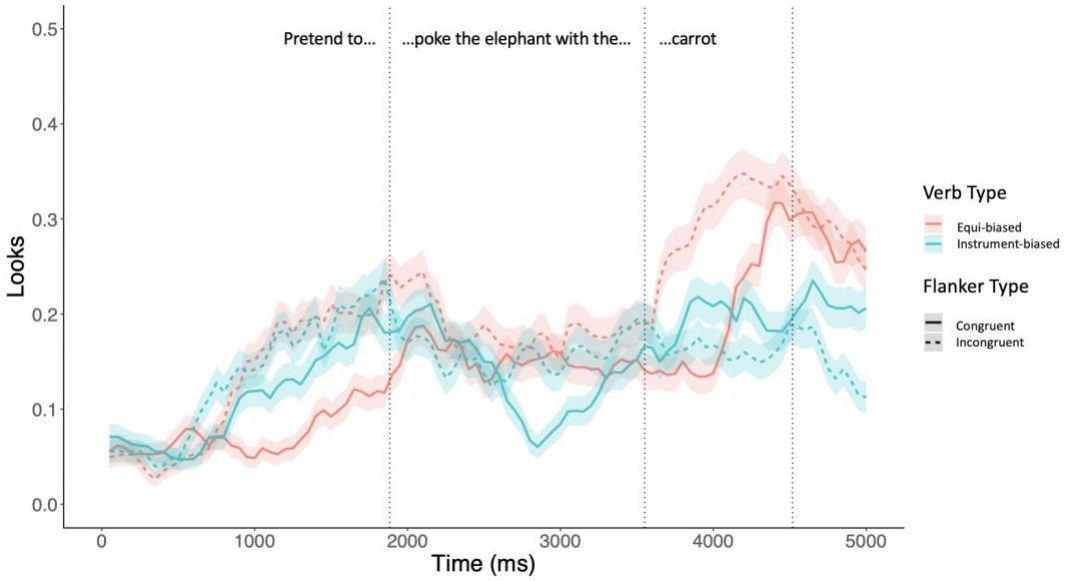

Figure 3.7: Looks to the lone instrument for Experiment 4, separated by verb type and prior Flanker trial type. Dotted lines indicate verb onset, PO noun onset, and 1000ms after PO noun onset, respectively. Lines are adjusted 200ms to account for saccade planning.

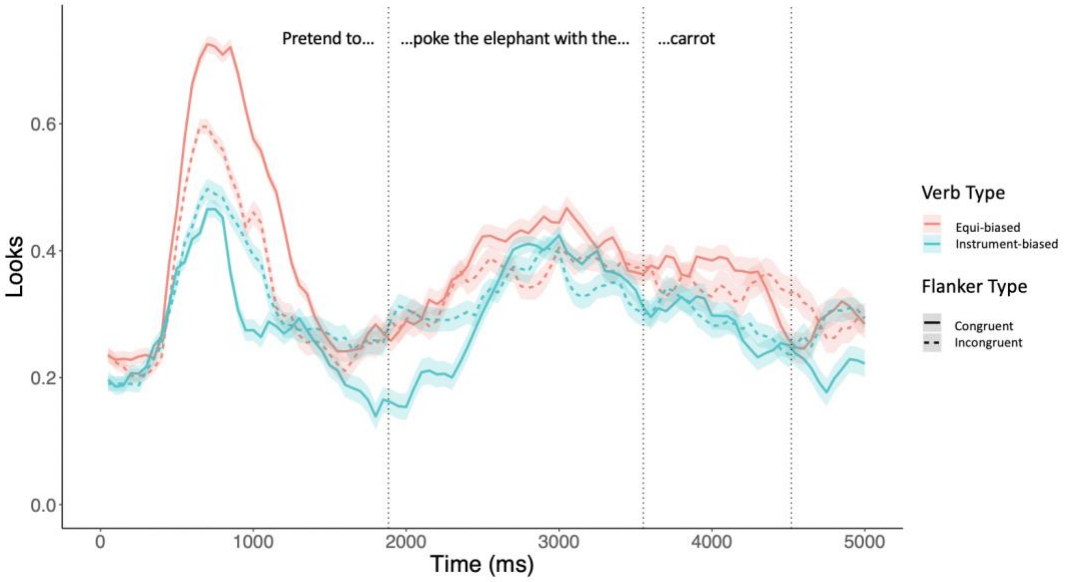

Figure 3.8: Looks to the distractor animal for Experiment 4, separated by verb type and prior Flanker trial type. Dotted lines indicate verb onset, PO noun onset, and 1000ms after PO noun onset, respectively. Lines are adjusted 200ms to account for saccade planning.

Figure 3.6 shows looks to the modified animal (e.g. Elephant with a carrot), while Figure 3.7 shows looks to the long instrument (e.g. carrot), and Figure 3.8 shows looks to the distractor animal (e.g. Elephant with a bow tie). In general, children were more likely to look to the modified animal and distractor than the lone instrument throughout the trial. This result is sensible, as the animals were likely more eye-catching overall than the lone instruments.

Figure 3.7 displays the primary measure of interest: looks to the lone instrument. A main effect of Flanker type was observed such that participants were more likely to look to the lone instrument following incongruent Flanker trials (Verb-onward window: $t(7) = 2.39$, $p=.05$; PO window: $t(7) = 3.27$, $p=.01$). There was no main effect of verb type in either window (Verb-onward window: $t(62) = .52$, $p=.61$; PO window: $t(63) = .29$, $p=.77$). Importantly (and unexpectedly), as Figure 3.7 shows, verb type interacted with prior Flanker trial type: for instrument-biased verbs, on trials directly following an incongruent Flanker trial, children were less likely to look at the lone instrument (Verb-onward window: $z=38.5$, $p<.001$; PO window: $z=51.2$, $p<.001$).

3.1.7: Discussion

The goal of this experiment was to determine whether children rely more on cues that are reliable predictors of sentence structure once their cognitive-control system is engaged. In this study, the equi-biased verb condition ostensibly presents weaker cues – encountering a particular verb will not provide strong evidence as to the attachment preference of the upcoming "with" phrase. If children rely more on

reliable cues following cognitive-control engagement, they should rely more on the instrument-biased verbs following incongruent Flanker trials.

On the surface, the results of Experiment 4 seem inconsistent with this Reliability hypothesis. When verbs were *less* reliable (more Equi-biased), the Flanker manipulation had a larger effect. Namely, incongruent Flanker trials that engaged children's cognitive control system led children to look to lone instruments more when the verbs were (seemingly) less strongly biased. It was predicted that cognitive-control engagement would lead children to rely on sentence processing cues that present reliable predictors of upcoming structure (e.g. strongly biased verbs), but not necessarily on ones that aren't (equi-biased verbs). Since verb lists did not vary in frequency, this would indicate that children can make use of their cognitive control system in a sophisticated way: They can use this system to better take *reliability* into account when navigating ambiguity, and are not simply relying on a metric that considers frequency or part-of-speech.

But all is not lost: There is good reason to believe that the particular verbs used in this study were not biased in the way suggested by the norming data presented above. As Experiment 8 will show, when analyzing child-directed speech, the equi-biased verbs used in this study were in fact more likely to be said with VP-attachment (and should perhaps be considered *more* instrument-biased than the verbs currently labeled as such). These claims will be returned to in Chapter 4, and will provide evidence in support of the Reliability hypothesis after all.

One additional thing that remains unclear from the previous study is the level of generalization children are drawing. Specifically, are children attending to the bias

of particular verbs, or following the VP-attachment preference of verbs in English more generally? To adjudicate these two possibilities, Experiment 5 introduces NP-attachment-biased verbs. These verbs are highly predictive of NP-attachment for an upcoming with-phrase, but are inconsistent with the bias of most verbs in English. If children interpret sentences in a with-phrase-as-modifier way for these verbs, this will indicate that children are indeed parsing according to the bias (and therefore reliability) of particular verbs.

3.2: Experiment 5: Imperative task, Instrument vs. Modifier-biased verbs

Experiment 5 compares the Instrument and Equi-biased verb conditions to a Modifier-biased verb condition. The goal of this manipulation is to determine the scope of reliability. Since most verbs in English are Instrument-biased, increased reliance on (what will turn out to be) Instrument-biased verbs in Experiment 5 could be compatible with two alternative explanations. Either children are following the language-general statistical information telling them to rely on VP-attachment when cues are in conflict, or they have successfully tracked the bias status of the particular verb. These two explanations will be distinguished in Experiment 5 – if cognitive-control engagement (still) causes children to assume VP-attachment, this will indicate that children are relying on a bias toward VP attachment that is generally reliable, regardless of the individual verb they encounter.

3.2.1: Participants

28 Children ages 4;0 to 6;6 were recruited from the Infant and Child Studies Consortium Database at the University of Maryland, College Park. As in Experiment 4, children and their guardians participated virtually, communicating with researchers for set-up and troubleshooting via Zoom or another video conferencing application.

3.2.2: Procedure

Children were given the same practice as in Experiment 4, and the same set of flanker trials, interleaved with sentence trials in an identical manner. Other details of the experimental set-up remained the same.

3.2.3: Materials

Sentence trials in Experiment 5 contained only Modifier-biased verbs. Children heard verbs such as "choose" that reliably predicted an upcoming PP to attach to the NP, giving a modified-noun reading (e.g. "Choose the elephant with the carrot" is more likely to mean "choose the elephant that has the carrot" than "choose the elephant using the carrot"). Verbs were normed as described in Experiment 4, and coding of the subsequent videos was done in a similar manner.

3.2.4: Results

Data from 28 children (ages 4;0 to 6;6) were collected. As before, results here are reported in three sections. Flanker results are reported to demonstrate that children's cognitive control system was engaged on incongruent trials, to a similar extent as it was in Experiment 4. Act-out results are reported as a measure of offline

performance. Eye-tracking results are again reported as the primary measure of children's online sentence interpretation.

Flanker results: As can be seen in Figure 3.9, flanker results for Experiment 5 very closely mirrored the results of Experiment 4. Children's average Flanker accuracy was 71% (compared to 72% for Experiment 4). Trials on which children's RT was more than 2.5 standard deviations longer than the mean were excluded, which resulted in removing all trials longer than 29.6 seconds. Reaction time was again modulated as expected: children's average reaction time to congruent fish trials was 3371ms, and for incongruent trials was 3976ms. A mixed effects regression analysis with random effects of subjects confirmed that this difference was significant ($t(37) = 2.87$, $p=.006$). A similar model confirmed that accuracy did not significantly vary with condition ($t(37) = .49$, $p=.627$).

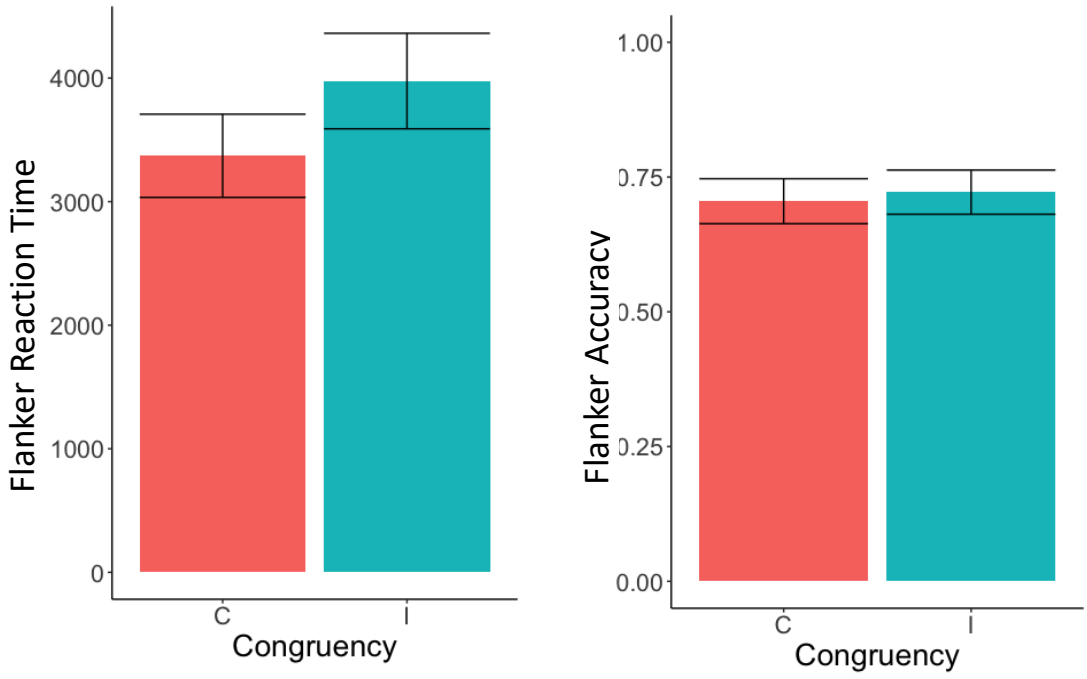

Figure 3.9: Flanker Reaction Time and Accuracy by Flanker trial type for Experiment 5

Act-out results: As in Experiment 4, children varied in their choice of how to act out the "pretend" commands. On 74 (18.5%) of trials, children did a clear dragging action of the lone instrument to one of the animals, compared to 43.8% for Experiment 4. On 136 (34%) of trials, children conducted the action on the modified animal (compared to 30.7% for Experiment 4). Many children in this experiment chose to carry out the action mentally, without making observable actions – this occurred on 120 trials, or 30% of the time (compared to only 7% for Experiment 4). On a further 55 trials (13.8%), children made unobservable actions (this result precisely matched this percentage of trials for Experiment 4, on which children also made unobservable or unclear responses 13.8% of the time). Finally, children performed an incoherent action (e.g. performing an action on the distractor item) on 15 trials, or 3.8% of the time (compared to 4.5% for Experiment 4).

In general, children made more modifier-like responses than in Experiment 4, and were also more likely to perform no explicit action. When the results are broken down by prior Flanker trial type (see Table 3.4), relatively few differences can be seen. As can be seen in Figure 3.10, children were relatively equally likely to make an instrument or modifier-like action if the sentence trial had been preceded by a congruent fish trial as when it had been preceded by an incongruent one. Children were, as in Experiment 4, also slightly more likely to perform no action following an incongruent Flanker trial than a congruent one.

Code	Prior Fish	Count
D	C	8
D	I	3
I	C	29
I	I	30
M	C	58
M	I	57
N	C	44
N	I	48
X	C	21
X	I	22

Table 3.3: Children's act-out actions in Experiment 5, broken down by prior fish trial type, from a total of 400 trials. In the Code column, "I" refers to an instrument-like action while "M" refers to a modifier-like action. “D” indicates that children did a modifier-like action on the distractor object, “N” indicates the child did no action, and “X” indicates that the child did an ambiguous or non-codable action. In the Prior Fish column, “C” indicates that the flanker trial directly prior was congruent and “I” indicates that the flanker trial directly prior was incongruent. The Count column contains the total number of trials on which a particular code occurred for that prior flanker trial type.

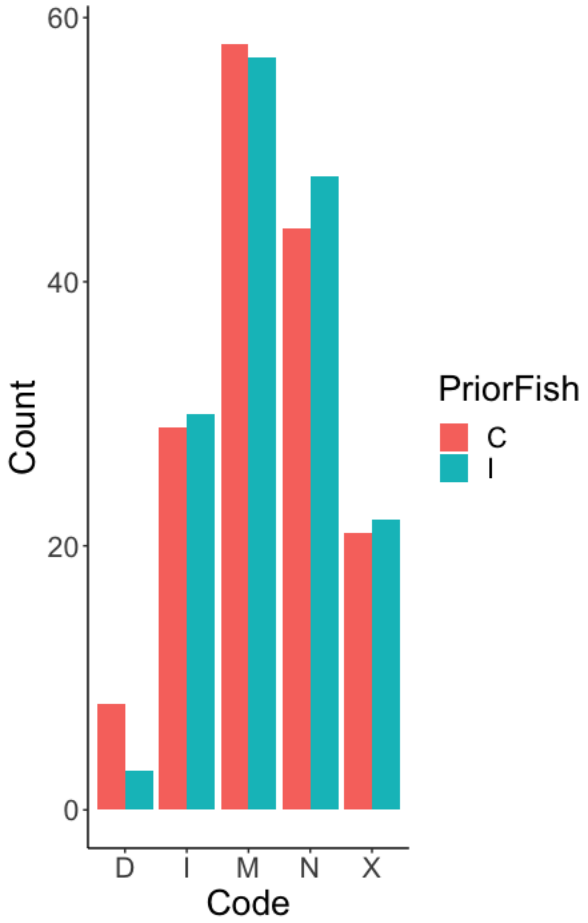

Figure 3.10: Count of the number of trials on which children performed particular act-out actions in Experiment 5, separated by prior Flanker trial type (see Table 3.3 for exact counts and coding key)

For the same reasons as those mentioned for Experiment 4, these act-out actions may be a faulty window into children's true thought process as they parsed the sentences in this study. For this reason, the eye-tracking results presented below were taken as a sounder record of how children interpreted the ambiguous sentences presented in this task.

Eye-tracking results: Using the same time window as Experiment 4, Figure 3.11 shows looks to the modified animal (e.g. Elephant with a carrot). Looks from

Experiment 4 are collapsed into one category and graphed as “instrument-biased” verbs (in red). Results from the 28 participants tested on modifier-biased verbs are presented in blue. As Figure 3.11 shows, participants who were shown modifier-biased verbs looked more to the modified animal than participants who were shown more instrument-biased verbs, indicating that participants did indeed parse according to verb bias. Figure 3.13 demonstrates a parallel finding: participants who heard modifier-biased verbs were comparatively less likely to look at the distractor animal (e.g. elephant with a bow-tie) than participants who heard instrument or equi-biased verbs. To some extent, looks to the distractor animal can be taken as indicative of VP-attachment, since participants interpreting the sentence as an instruction to act on an animal (e.g. poke an elephant) with an instrument might choose the distractor animal, whereas participants interpreting the sentence as an instruction to just act on a particular animal would have little reason to look at the distractor. Together, these results serve as a manipulation check, indicating that children were sensitive to verb bias, and this sensitivity is reflected in their eye-movement data.

Figure 3.12 displays the primary measure of interest: looks to the lone instrument. Participants who heard modifier-biased verbs were significantly less likely to look to the lone instrument than participants who heard instrument/equi-biased verbs, as confirmed by a mixed effects logistic regression model that included random effects of participants and items (Verb-onward window: z= 2.81 p=.004; PO window: z=28.36, p<.001).

Importantly, as Figure 3.12 also shows, looks to the instrument interacted with the prior Flanker trial type: for instrument/equi-biased verbs, on trials directly

following an incongruent Flanker trial, children were more likely to look at the lone instrument than on sentence trials following a congruent Flanker trial. Conversely, for modifier-biased verbs, on trials directly following an incongruent Flanker trial children were *less* likely to look at the lone instrument than on sentence trials following congruent Flanker trials. This effect was true in both the PO window (z= 19.39, p<.001) and the entire verb-onward window(z=30.9, p<.001). Figures 3.14 and 3.15 show this interaction in average looks to the instrument during these time-windows across Flanker condition and verb types.

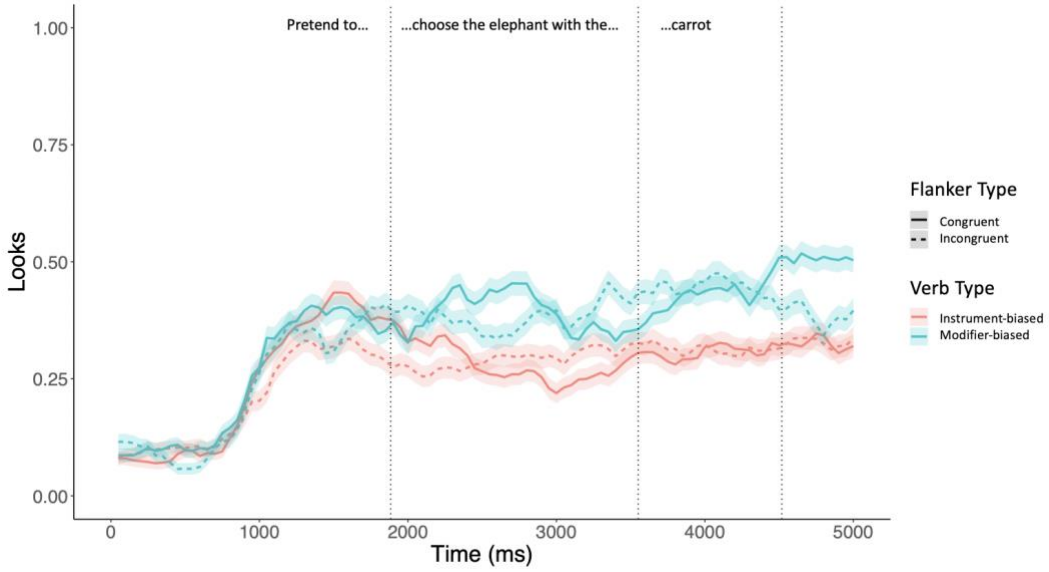

Figure 3.11: Looks to the modified animal for Experiment 5, separated by verb type and prior Flanker trial type. Dotted lines indicate verb onset, PO noun onset, and 1000ms after PO noun onset, respectively. Lines are adjusted 200ms to account for saccade planning.

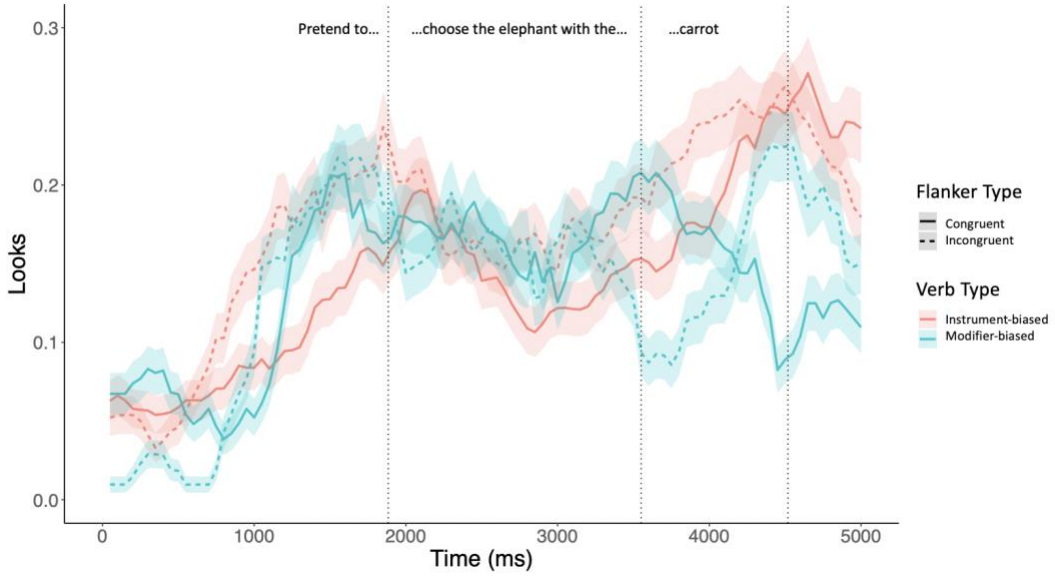

Figure 3.12: Looks to the lone instrument for Experiment 5, separated by verb type and prior Flanker trial type. Dotted lines indicate verb onset, PO noun onset, and 1000ms after PO noun onset, respectively. Lines are adjusted 200ms to account for saccade planning.

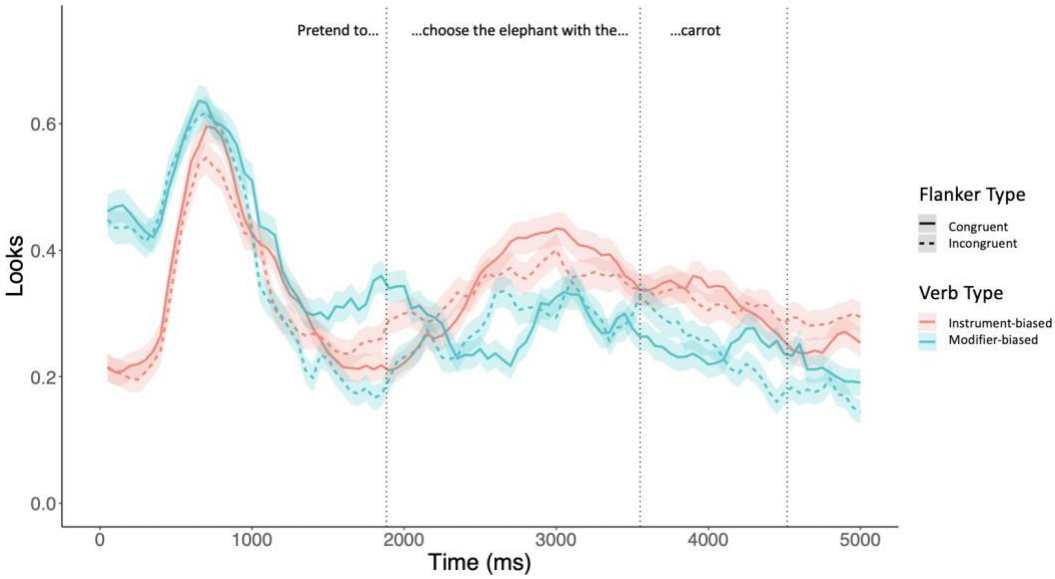

Figure 3.13: Looks to the distractor animal for Experiment 5, separated by verb type and prior Flanker trial type. Dotted lines indicate verb onset, PO noun onset, and 1000ms after PO noun onset, respectively. Lines are adjusted 200ms to account for saccade planning.

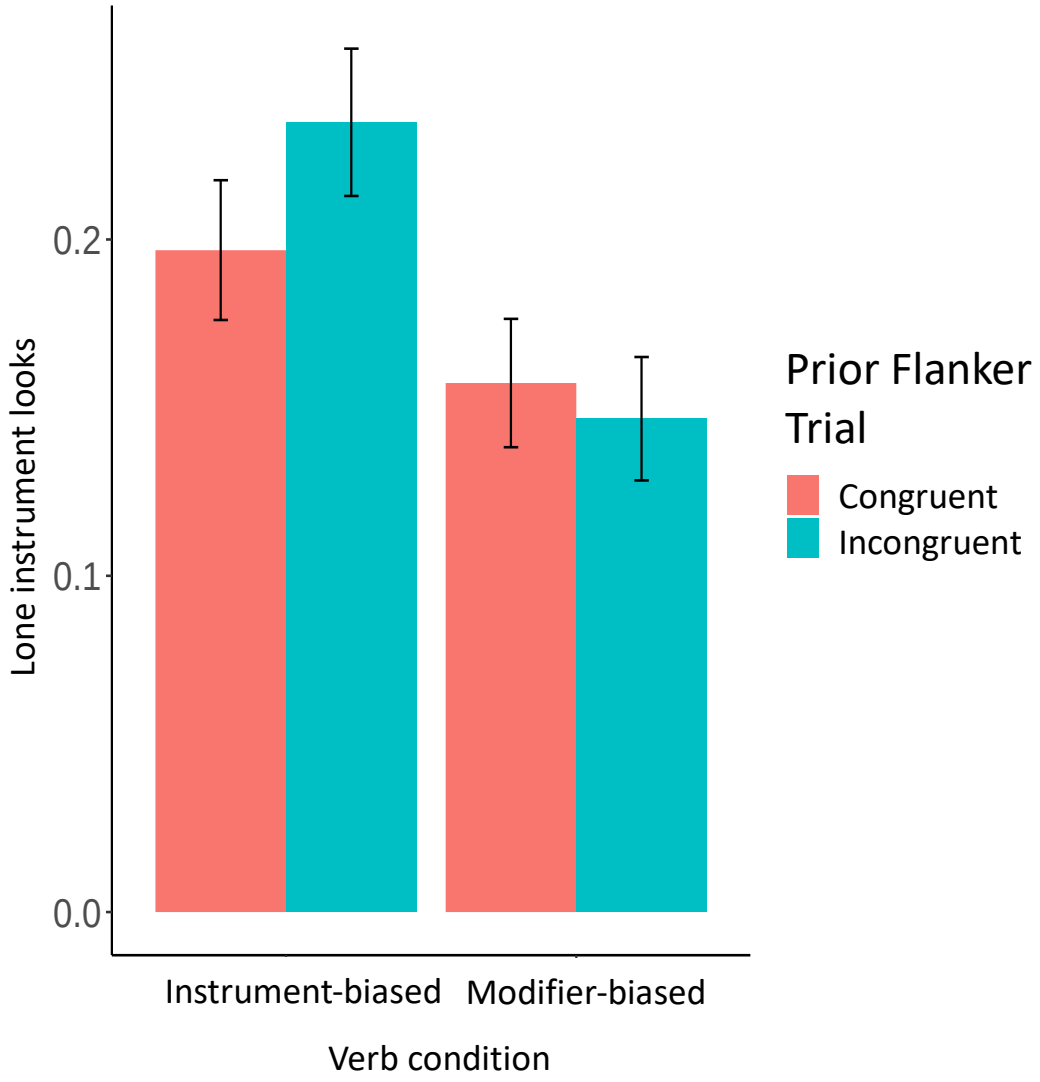

Figure 3.14: Bar graph of looks to the instrument in the PO region for Experiment 5

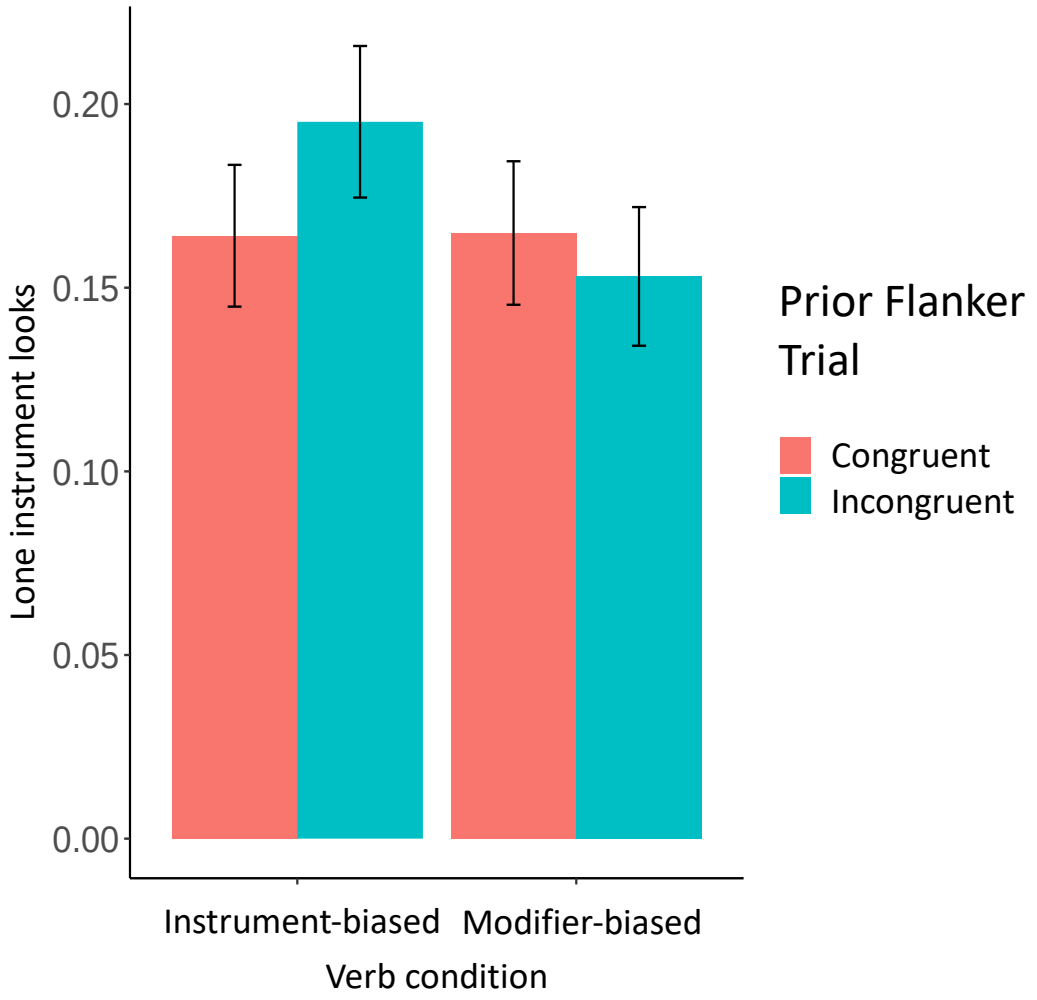

Figure 3.15: Bar graph of looks to the instrument in the Verb-onward region for Experiment 5

3.2.5: Discussion

The results of Experiments 4 and 5 together provide evidence for the Reliability Hypothesis. Following cognitive control engagement children parsed ambiguous sentences according to the particular bias of the verb and regardless of verb frequency.

This result has implications for the way in which children's cognitive control system interacts with their sentence processing system. Namely, this suggests that

when children are in a more highly engaged control state, they are more likely to parse according to cues that are ordinarily reliable predictors of sentence structure, and at a fairly fine-grained level. In particular, this suggests that being in a more highly engaged control state up-regulates domain-specific representations that are consistent with reliable cues for sentence prediction (of course, it is also possible that cognitive control engagement leads to down-regulation of domain-specific representations that are gleaned from unreliable cues, or both). The results of Experiment 5 begin to answer the question of "what makes a cue?". In this study, children appear to parse sentences according to the bias of particular verbs more following cognitive control engagement, suggesting that the grain size of the "cue" they're using to parse is at the word-level. These results even more strongly suggest that in Huang et al. (2016), when children were in a relatively up-regulated cognitive control state, they attended more to the particular biases of the verb "put," leading to depletion-like effects because they were more likely to expect an upcoming location.

Future work will replicate these experiments with adult controls. The goal of these control experiments will be to verify that a) the biases of the verbs are generally followed, and b) adults also rely on verbs that more reliably predict a particular upcoming structure, following cognitive control engagement. Since our hypothesis is that adults also rely more on reliable sentence processing cues when their cognitive-control system is relatively up-regulated, adults should perform similarly to children. To be sure, the Flanker task will need to be adjusted to be appropriately difficult for adult participants (e.g. by adding additional time pressure), but other aspects of the methods will remain largely unchanged. Prior work using a similar verb-bias

manipulation indicates that adults are at least sensitive to these distinctions. While it is not yet known whether cognitive-control engagement will affect the way adults use verb biases, it is at least clear that adults are more likely to parse "with" sentences according to a VP-attachment preferences when the verbs are lower in frequency (Ovans et al., 2019).

3.3: Experiments 6 & 7: Verification of "Virtual-World" eye-tracking procedures:

Experiments 4-5 make use of a relatively novel experimental paradigm: visual-world eye-tracking using online participants and hand-coding eye-movements (so-called "Virtual-world eye-tracking"). While the set-up for these experiments mirrors in-lab testing, there are several key ways in which testing visual-world eye-tracking experiments online may lead to differences from testing it in-lab (outlined in Section 3.3.1 below). For these reasons, Experiments 6 & 7 present attempts to replicate well-established visual-world eye-tracking results using online participants, in order to provide assurances for the experiments in this chapter that idiosyncrasies introduced by testing participants online should be minimal, and these results are comparable to what the results would have been had these been lab-based tasks.

3.3.1 Potential differences between lab and online visual-world eye-tracking

One potential source of variation in online visual-world eye-tracking is variation in participant monitor sizes. While during in-lab testing participants are generally all run on the same monitor and computer set-up, by necessity online participants are each participating on different machines, and the scene they're looking at may vary in width or placement. While this may introduce a challenge for

accurate eye-movement coding, it's also important to consider the effect that this sort of variability might have on the linking hypotheses inherent in visual-world eye-tracking experiments. Namely, it is assumed that participants' gaze reflects an underlying mental process that drives them to look at images that are more in line with their present interpretation of an utterance than images that are not (Huettig & Altmann, 2005; Huettig, Rommers & Meyer, 2011; Dussias, Kroff & Gerfen, 2013; Magnuson, 2019; Degen, Kursat, & Leigh, 2021). When eye-movements are found to be slower or less accurate in a particular condition, it can be inferred that this disruption reflects a different underlying mental process. However, when some participants have larger screens than others or are sitting closer such that the visual-world scene takes up a larger viewing angle, it may be more effortful to launch a saccade toward any particular image, as the distance the eye must travel is longer. Some participants may therefore make slower or fewer saccades in virtual-world eye-tracking for reasons unrelated to the task at hand, but because of the particular pressures introduced by their physical set-up. While it may be assumed that that these participants would be evenly split across conditions, it may still be imperative to check for outlying participants with particularly long gaze times on any one item.

Another potential issue with the virtual-world eye-tracking paradigm is that variation in where the camera is in relation to participants' screens may make coding difficult and result in a large amount of data loss. While many commercial built-in webcams are located centrally just above the screen, others are built-in below the screen and non-built-in cameras may of course be placed far away, at the participant's discretion. Additional coding/data loss concerns may arise if participants are not

particularly well-lit, or are too far away from their screens for coders to accurately measure small deflections in eye gaze.

On top of these concerns about codability, another source of noise may arise from the fact that subjects participating virtually are not in controlled environments. While traditional visual-world eye-tracking designs have been shown to be both valid and stable measures of real-time word-recognition (Farris-Trimble & McMurray, 2014) background noise has also been shown to slow the time-course of lexical processing (McMurray et al., 2017). Additionally, home or work environments may provide many eye-catching distractions (Bergefurt et al., 2021). These distractions may draw participants' eye gaze away from their screens. This, coupled with the lack of social pressure from not having an in-person experimenter nearby may and lead to a greater number of off-screen looks – essentially a greater degree of data loss in the experiment.

Finally, a major concern for virtual sentence processing experiments is that our inferences often rely on the precise timing between stimulus onset and eye-movements, often on the order of milliseconds. Online testing may introduce additional sources of lag that disrupt this timing. In particular, slow internet speeds may lead to delayed presentation of either audio or visual stimuli. Any lag in the resulting video files may also lead to eye-movements being recorded as slower than they were in reality.

3.4: Experiment 6: Word-recognition in the virtual world

In order to address the concerns enumerated above, it is imperative to directly compare in-lab visual-world and virtual-world eye-tracking results. Experiment 6 presents a conceptual replication of a well-known psycholinguistic finding using the virtual-world method, to determine whether any of the potential concerns about the method have merit. To this end, this experiment sets out to replicate Experiment 1 of Allopenna, Magnuson & Tannenhaus (1998), *"Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models."* This study was chosen both because it has been largely influential in the field (it has been cited over 1700 times), and because it has been replicated in laboratory settings (e.g. Farris-Trimble & McMurray, 2013). It is therefore unlikely that any failure to replicate the results of this experiment would be due to the effect itself being unstable, and failures can therefore be fairly safely blamed on the virtual implementation. This study also relies on fine-grained time-course data on the level of tens of milliseconds. Successful replication of such relatively fast effects would therefore bode well for effects with a protracted time-course, such as those usually measured in sentence-processing studies.

Recent attempts to use participants' webcams for visual-world data collection (e.g. Xu et al., 2015; Semmelmann & Weigelt, 2018) rely on automatic gaze-detection, but this requires participants to sit through lengthy calibration procedures and might result in significant data loss. While overall track loss is not reported, gaze detection accuracy can be off by over 200px.

Recently, some work has sought to systematically compare automatic gaze detection algorithms to hand-coded videos in an attempt to quantify the accuracy provided by automatic gaze detection algorithms, and has concluded that automatic gaze-detection can indeed result in significant data loss (Kandel et al., 2022). For these reasons, hand-coding videos remains a more reliable tool when visual-world data cannot be collected in-person.

A final drawback of prior validation work is that it focuses on quantifying overall attention to images on a screen, rather than fine-grained time-locking to spoken language input. Instead, replicating studies such as Allopenna et al. (1998) is an ideal way to validate remote testing, since looks to competitor objects in this study assess subtle mental processes that mediate between speech acoustics and word recognition over a distinct time-course.

3.4.1: Experimental Prospectus

In the original study, the authors demonstrated that listeners looked toward an image of a target (e.g.,"beaker") but also to cohort competitors (e.g., "beetle") immediately after word onset, as well as rhyme competitors (e.g., "speaker") toward word offset, over distractors (e.g., "carriage"). As predicted by TRACE models, this showed that listeners incrementally activate phonemic competitors during spoken-word recognition.

Three specific alterations to the design of the original study were made for Experiment 6, in order to increase the feasibility for online testing:

1. The current experiment makes use of a novel webcam eye-tracking paradigm, in which participants' faces are recorded through the cameras on their

computers (via PCIbex, Zehr & Schwarz, 2018). In this paradigm, looks are hand-coded by trained research assistants (e.g. Snedeker & Trueswell, 2004). While this method may introduce additional sources of variability, including screen size, speaker and webcam quality, internet bandwidth, background noise, and environmental distractions, it is the aim of this experiment to determine whether these changes will meaningfully affect the time-course of word processing.

2. The original study included partial-set trials (e.g., with two unrelated objects, target and rhyme competitor) as well as full sets. The present study used only partial-set trials in a Latin square design, reducing the trial number from 96 to 18. This ensured that cohort and rhyme competitor looks were independent, encouraged online participants to stay engaged for the duration, and reduced upload time of our video data.

3. In the original study, participants were instructed to put referents near other shapes on the screen. In the present study, they saw only the four stimuli (target, cohort & rhyme competitors, and distractor), spaced out to allow coders to detect fixation changes. Even when the relevant phonemes are mentally activated, participants may look less to the competitors due to their distance from each other. This change was made out of necessity, but is also directly serves as a way to validate the online testing format, rather than reducing the validity of the replication.

3.4.2: Participants

A total of 60 participants were recruited, 30 from Amazon Mechanical Turk (MTurk, https://www.mturk.com), and 30 from the University of Maryland SONA system (Sona Systems, https://www.sona-systems.com). The data from an additional

24 participants was unusable due to video upload failure (14 participants), poor positioning (5 participants), glasses glare (1 participant), or due to their having participated more than once (4 participants)[5]. Demographic information was collected for the Mechanical Turk participants: 7 were female 12 male and 11 chose not to report; 4 reported their race to be Asian, 3 Black, 11 White and 12 chose not to report; their mean age was 31.2 years, and participants hailed from at least 13 different states across the U.S. All demographics were self-reported through free-response fields. Participants were compensated with either $5 or class credit for participating.

3.4.3: Procedure

Once participants consented to use of their webcams, they were presented with 144 familiarization trials in which each of the images used in the experiment were presented and labeled (each trial lasted approximately 1-2 seconds). Participants were told "you'll see a series of images (twice each) and you'll hear their names to familiarize you with them" and did not have to provide a response to these trials.

Following this, participants were told "Now, you'll see the same images, and you'll be instructed to click on one." They then saw 18 trials with 4 images each (presented in the corners of the screen, as in Figure 3.16). On each trial, they heard a short sentence instructing them to click on one of the images (e.g. "Click on the carrot"). In addition to the target image, the three other images on each trial consisted

[5] It is unclear what possessed some participants to participate in the study more than once. For participants on Mechanical Turk, a few did the study multiple times under separate accounts (although this violates the platform's terms of service). Two additional participants participated twice on SONA several days apart, perhaps because they forgot having participated previously. In all cases only the first run was analyzed.

of 2 unrelated items and one item that was either a cohort competitor, rhyme competitor, or a third unrelated item.

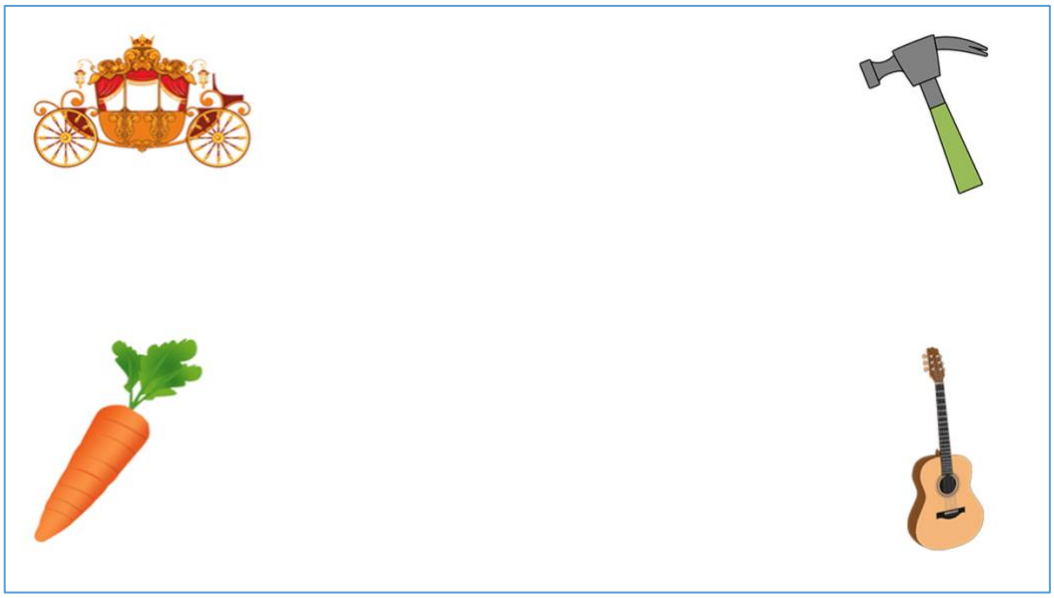

Figure 3.16: Sample trial scene from Experiment 6

3.4.4: Picture Norming

Prior to testing, images were normed for recognizability, frequency (using Google Ngram ratings for 2019, and neighborhood density, using the CLEARPOND database (Marian et al., 2012). An additional 24 participants recruited from Amazon Mechanical Turk were asked to label each image. From this, the percentage of participants who correctly labeled the item was calculated. The final lists of target words, cohort competitors, rhyme competitors, and unrelated items did not significantly differ in frequency, neighborhood density, or recognizability according to these metrics.

3.4.5: Avoiding coding concerns

To ensure that participants' gaze was optimally codable, they were initially instructed to sit such that they were directly in front of their computer and such that their eyes were clearly visible. They were asked to make sure to be lit from the front as much as possible, and to adjust so that there was little glare if they were wearing glasses. These instructions were given while participants were shown a video feed of themselves from their camera so that they could verify they were positioned appropriately.

An additional concern was that some participants might have camera settings that flip their image along a horizontal axis automatically. While most computer webcams automatically mirror the image they record, they often offer the option to reverse the image. Such inconsistencies would be catastrophic for visual-world data analysis, as participants with these settings would be looking in the opposite direction of where they were coded to be looking. To avoid this pitfall, participants were presented with two simple "catch" trials. On one, they were presented with an apple on the right side of their screen, and were asked to look at the apple while saying into camera whether it was on their right or left. Then, they were presented with a banana on the left side of their screen, and were instructed to stare at it while saying whether it was on their right or left. This allowed coders to check each participant to ensure both that the images were appearing on the correct side of their screen and that their camera was not systematically reversing their eye-movements.

One further measure was taken to address the other coding concerns noted at the beginning of Section 3.1. Participants were asked not to wear headphones, if

possible, so that the audio of the experiment was audible in the video recordings of participants' eye-movements. This allowed coders to analyze the audio from these video recordings and match it on to the audio files used in the experiment to determine whether there was any audio lag introduced by conducting the experiment online.

Another benefit of this request was that since some participants opted to wear headphones anyway, participants could be split by headphone use. This was taken to be a proxy for ambient noise level, with the assumption that participants wearing headphones were experiencing relatively little background noise, while participants without headphones were more likely to be exposed to the ambient noise of their environment. If such ambient noise had an effect on eye-tracking results, it is reasonable to assume this would lead to a difference in looking patterns between the headphone-wearing and non-headphone-wearing groups.

3.4.6: Results

Looks to the target were measured in the 1000ms time window after the onset of the target word (see Figure 3.17 below). As in the original study, in this time window participants looked significantly more to the target image than to the Unrelated distractor images ($t(72)= 6.04$, $p<.001$). Participants also looked significantly more to the cohort competitor than to the unrelated distractors ($t(43)= 3.91$, $p<.001$). Unlike the results of Allopenna et al., no significant differences were found in this window between participants' looks to the rhyme competitor and the distractor images ($t(43)= .95$, $p=.34$).

Additionally, participants' looks to the target image were slightly slower than they were in the original study. While for Allopenna et al., looks to the target began at approximately 200ms following target onset, here they began to diverge from distractor looks at approximately 400ms post target word onset.

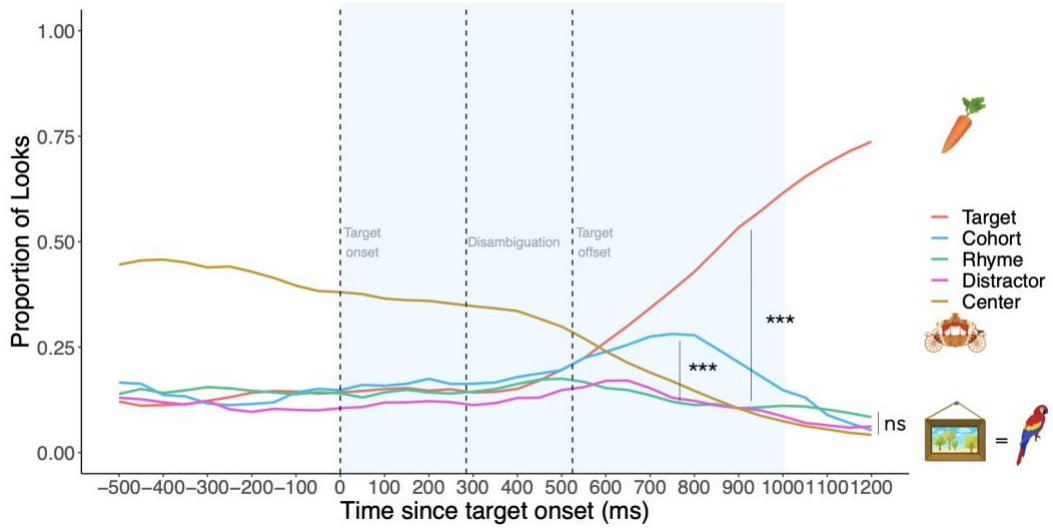

Figure 3.17: Looking-time results for Experiment 6

In an effort to determine why rhyme effects fail to show up using this method, the data were split in three different ways. First, results were split by target frequency, as measured by Google Ngram data. This was done because word frequency has previously been shown to affect the time-course of word recognition (Dahan, Magnuson, & Tanenhaus, 2001). As Figure 3.18 indicates, for both high and low frequency targets, no significant interactions were found between looks to target and frequency ($t(104)= .38$, $p=.70$), cohort competitor looks and frequency ($t(105)= .03$, $p=.97$) or rhyme competitor looks and frequency ($t(104)= 1.2$, $p=.25$).

Next, results were split by recruitment platform, with the thought that SONA participants more closely matched the demographics of participants in lab-based studies, so if participant demographics were to play a role in the lack of rhyme effects

they may emerge when the SONA participants are isolated. For MTurk and SONA participants, no significant interactions were found between looks to target and participant type ($t(254)= .72$, $p=.47$), cohort competitor looks and participant type ($t(981)= 1.04$, $p=.30$) or rhyme competitor looks and participant type ($t(981)= 2.1$, $p=.051$). Though Mechanical Turk participants came close to having significantly greater rhyme looks than SONA participants, these looks did not significantly differ from distractor looks ($t(982)= 1.67$, $p=.09$).

Finally, results were split by headphone use, as a proxy for background noise, as headphone users were likely to be less distracted by noise in their immediate environment. Once more for headphone users and non-users, no significant interactions were found between looks to target and headphone use ($t(267)= 1.05$, $p=.29$), cohort competitor looks and headphone use ($t(982)= .35$, $p=.72$) or rhyme competitor looks and headphone use ($t(982)= 1.04$, $p=.29$). Headphone users did look to target sooner however, after approximately 200ms from target onset, more closely matching the results of lab-based eye-tracking studies.

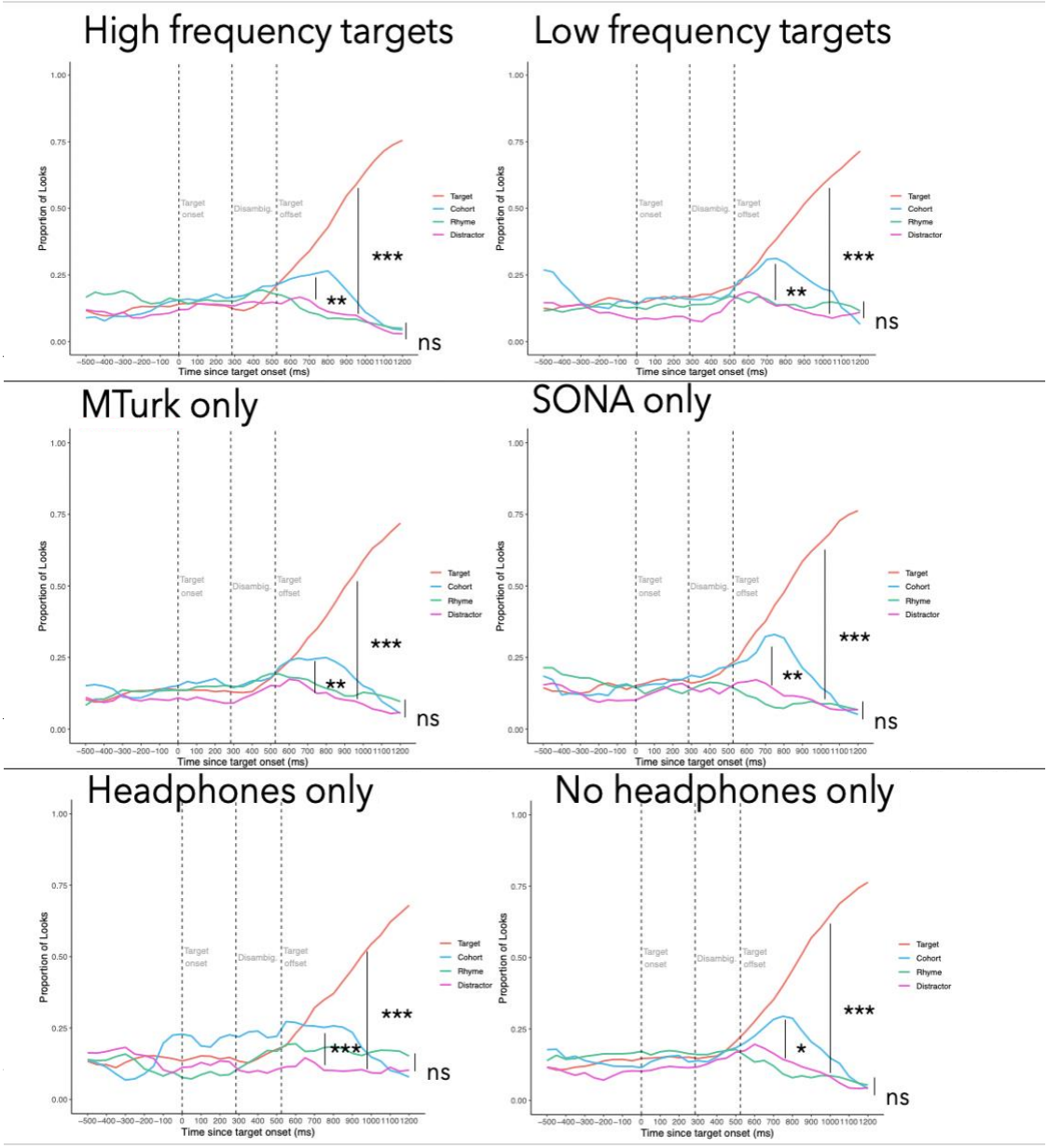

Figure 3.18:Post-hoc data splits for Experiment 6

3.4.7: Discussion

Both in the full set of results and in the various data splits outlined above, it was observed that participants readily looked to the target object and looked to cohort competitors following target word onset, but did not look to rhyme competitors more than to unrelated distractors. It was also observed that participants were relatively

slow to look to target objects compared to in-lab studies. This result was, however, seemingly nullified for participants who wore headphones, suggesting that background noise in participants' environments may have been to blame. This result is in line with word-recognition studies conducted with cochlear implant users, who also receive "noisy" input and are slightly slower to look at target words as a result (McMurray et al., 2017). Of course, participants were also instructed not to use headphones, so it is possible that the participants who opted to use the headphones were also more generally rushed, as indicated by their disinclination to carefully read the experiment instructions.

Overall, these results demonstrate that incremental word processing and some subtle frequency effects are observable in virtual testing. Webcam eye-tracking produces similar results to in-lab testing, but eye-movements are slower, and subtle effects like rhyme competition may be harder to detect. Even so, the presence of cohort competition provides evidence for this method's sensitivity to incremental processing, and provides validation for internet-based eye-tracking as a viable method for Experiments 4 & 5, as well as providing new, virtual avenue for visual-world sentence processing research for closely time-locked effects.

It remains unclear why this experiment did not reveal effects of rhyme competition. It seems unlikely to be due to differences in participant pool or background noise, and dividing the data by target word frequency similarly revealed no differences. In ongoing work, this experiment is being replicated (a further time), but with three key changes that may help to reveal these subtler effects: 1) More items will be used (increasing the trial number to 30 instead of 18) to increase the

power to detect these subtle effects, 2) items will be slightly closer together on the screen, so that it is less effortful to make a saccade and competition from parafoveal vision may increase, and 3) familiarization trials will be removed, as this process was lengthy and may have tired participants prior to the main portion of the experiment, making saccades less likely in general.

While these results are a relatively promising validation of the eye-tracking method used in the previous experiments, the lack of looks to rhyme competitors suggests that this method may not be as precise as automatic eye-gaze detection software used in the lab. One potentially reassuring factor is that the sentence-processing studies presented in the prior two experiments measure eye-gaze data at a longer time-scale. For this reason, Experiment 7 follows up on these results by attempting to replicate a well-known sentence-processing effect using virtual eye-tracking.

3.5: Experiment 7: Sentence-processing in the virtual world

The results of Experiment 6 indicate that while larger word processing effects are observable using a virtual-word paradigm, subtler effects such as rhyme interference may be washed out, and eye-movements may be delayed several hundred milliseconds relative to in-lab studies. Of course, the ambiguity processing effects discussed in this chapter occur over a more protracted time-scale, and may therefore be more readily observable in a virtual format. To test this, it's useful to establish whether sentence-level visual-world processing results are replicable in virtual testing.

To investigate this, this section presents a conceptual replication of Altmann & Kamide (1999), who demonstrated that during sentence processing, listeners use semantic information from verbs to constrain visual attention to objects that are most likely to be thereafter referenced given the context. Along with establishing an influential result (the original paper has been cited over 1800 times), the basic results of the paper have been replicated in-lab, again indicating that any failure to replicate the same findings can be blamed on the virtual format, with minimal doubt cast on the underlying processes themselves.

In the original study, the authors presented participants with visual scenes that contained a character with several objects in their vicinity. Meanwhile, participants heard sentences that contained more or less restrictive verbs that indicated how the character would interact with the objects around them. For example, they heard sentences like "The boy will eat/move the cake" while viewing a scene like the one in Figure 3.19. While only the target object (e.g. the cake) matched the selectional restrictions of the more restrictive verbs, the distractor objects also matched the restrictions of the less restrictive verb.

3.5.1: Participants

48 Participants were recruited from Amazon Mechanical Turk. Data from an additional 10 participants was excluded either because their videos failed to upload (6) or poor positioning (4). Of the remaining 48 participants, 16 were female, 23 male, 1 non-binary, and 8 chose not to report. 3 reported their race/ethnicity to be Asian, 8 Black, 25 White, 4 Hispanic or Latino, and 8 declined to respond. Their average age was 36.5 years, and represented 25 U.S. states. All demographics were

self-reported through free-response fields. Participants were compensated with $5 for participating.

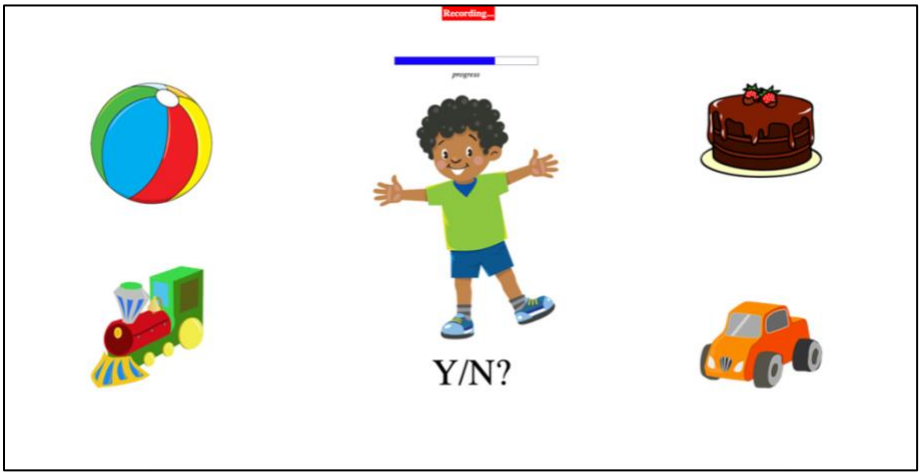

Figure 3.19: Sample trial scene in Experiment 7

3.5.2: Procedure

Participants were told "Your task will be to judge whether the sentences you hear apply to the pictures you see. For example, if you hear the sentence *The person will light the fire* and there is a picture of a fireplace, press Y for yes. If there is no fireplace, press N for no." They were then presented with 16 target trials during which a person (*The man/The woman/The boy/The girl/The baby*) was presented centrally, with four objects around them in the corners of the screen. After the sentence, a "Y/N?" prompt appeared in the lower center of the screen to prompt participants to respond. 16 additional filler items were created in which the mentioned object was not present in the scene, and participants' task was to judge whether the object was present or not. Items were presented in a Latin square design, so that while each participant saw each scene, they never saw the same scene with a filler and target trial.

While target items were kept as similar as possible to the original items used in Altmann & Kamide (1999), several items were updated for modern audiences (for example, it was determined that many participants would not know what a Filofax is – this was replaced with a folder). Since the original study did not report the images used in fillers, the additional filler trials were created to match the description of the ones in the original study, but with (presumably) different items. Images in these filer trials varied in whether the depicted a noun that matched the verb, in order to balance selectional restriction match across trials. Images were full-color clip-art style images deemed to be relatively uniform in art style.

3.5.3: Results

Looks were measured during the period of time between the onset of the verb and the onset of the noun, when participants hear restrictive verbs (see Figure 3.20). Looks to the referents that met those restrictions were compared for the restrictive vs. nonrestrictive verbs. For the non-target objects, a mixed effects analysis showed that participants looks did not significantly vary with verb type ($t(116)=.51$, $p=.61$). However, There were more looks to the target object for the restrictive than the non-restrictive verbs in the verb+determiner region before target noun onset ($t(116)=5.68$, $p<.001$).

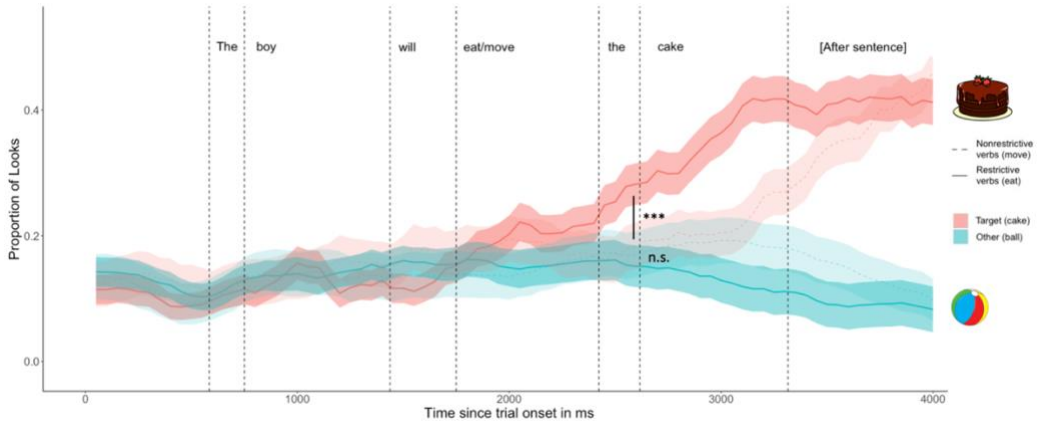

Figure 3.20: Looking-time results for Experiment 7

To determine whether background noise played a role in influenced the timecourse of participants looks to the target objects, data from headphone wearers (N=16) and non-headphone wearers (N=32) was analyzed separately, as in Experiment 6. For both groups, the results remained unchanged: a significant difference was found between target looks for restrictive vs. nonrestrictive verbs prior to final noun ("cake") onset such that participants looked more to the target for the restrictive verbs (With headphones: $t(37)=3.12$, $p=.002$; Without headphones: $t(78)=4.39$, $p<.001$). No such difference was found between looks to the other items on the screen for restrictive vs. non-restrictive verbs (With headphones: $t(37)=1.29$, $p=.19$; Without headphones: $t(78)=.87$, $p=.38$), (see Figure 3.21).

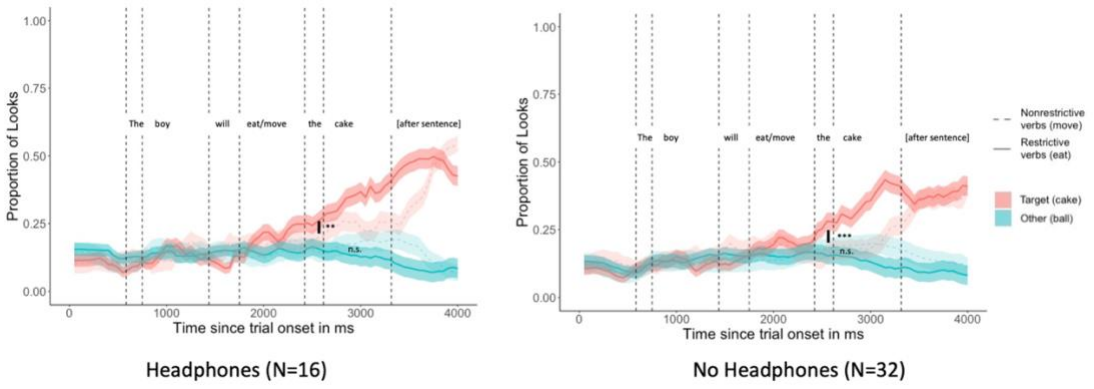

Figure 3.21: Looking time results for Experiment 7, split by headphone use

3.5.4: Discussion

These results conceptually replicate the findings of Altmann & Kamide (1999). It was found that participants launch quick, anticipatory eye-movements based on verb meaning. When participants heard more restrictive verbs such as "eat," they were more likely to look at referents that could be objects of that verb, even before they were named, as compared to when they heard less restrictive verbs like "move."

Overall, these results present a reassuring picture: sentence-level predictions are observable in virtual testing. In contrast to "wait-and-see" approaches (e.g. McMurray, Farris-Trimble, & Rigler, 2017; Van Petten & Luka, 2012), participants in these real-world settings seem to parse incrementally and launch anticipatory eye-movements to unnamed referents consistent with verb context, just as in lab-based studies. This also bodes well for the interpretation of Experiments 4 and 5: sentence-level visual-world processing results are indeed replicable in virtual testing, indicating that despite changes in format, concerns about whether the linking

assumptions of this new visual-world paradigm hold (due to noisy environments or variation in testing environments, etc.) ought to be minimal.

Chapter 4: Corpus analyses

An overall goal of this dissertation is to test the hypothesis that when children process sentential ambiguity they do not deplete a resource but instead engage a system that can re-weight the importance it places on certain types of input with repeated use. Further, when this system is engaged, children are better able to ignore unreliable cues. While the goal of Chapter 3 is to establish that children only rely on information from verbs specifically when they are reliable cues to structure-building, the notion of reliability outlined there requires further explanation and empirical support. The purpose of this chapter is to provide a detailed analysis of the input children receive, in order to determine whether the cues they rely on when their cognitive-control system is engaged are ones that generally do prove to be reliable indicators of who did what to whom.

Experiment 8 will seek to validate the corpus-based reliability of the particular verbs used in Experiments 4 & 5. Recall that the biases of these verbs have been determined based on adult cloze-task completion data, and may therefore differ somewhat in their reliability in child-directed speech. This experiment is a corpus analysis to determine the relative reliability of the verbs used in Experiments 4 & 5. If children's likelihood to parse those globally ambiguous sentences with VP or NP-attachment correlates with verbs' likelihood to predict upcoming withPP attachment in speech to children, this will provide further evidence that children are indeed calculating statistical reliability of particular verbs based on the sentences they hear.

A second question is whether the agent-first bias is really a less reliable predictor of sentence structure than verb morphosyntax, making it an unreliable

parsing cue for children. The conclusion from the prior studies is that children seem to ignore unreliable cues (not just their initial parse) during sentence processing when their cognitive-control system is comparatively more engaged. Experiment 9 is a second corpus analysis to determine whether the agent-first bias is indeed unreliable while information gleaned from verbs (e.g. the likelihood that "put" will precede a location) is reliably true in speech to children. The goal of this corpus analysis is to determine in what percentage of utterances children hear NP1s as agents (in which case the agent-first bias is reliable), and in what percentage as patients (where it's unreliable), and compare this to the percentages of when the subcategorization preferences of early verbs like "put" reliably predict structure. If NP1s as agents are less reliable predictors of argument structure than verbs like "put," this will indicate that children may be calculating fairly specific cue-reliability tradeoffs, down to the level of individual words across syntactic categories. If, instead, NP1s are not less reliable predictors of agent and patient-hood than verbs, this will indicate that children are calculating reliability at a much coarser level.

4.2: Experiment 8: Verb bias corpus analysis

The results of the previous experiments demonstrate that for children, engaging their cognitive control system has distinct, measurable effects on their sentence processing system. Experiments 1-3 demonstrated that when children's cognitive-control system is up-regulated by performing a task that engages the system, they are more likely to parse according to ordinarily reliable cues while discounting less reliable ones. Experiment 4-5 sought to confirm that this change in parsing strategy is to one that really does take reliability into account. If the

Reliability Hypothesis is correct, however, it predicts that on an item-level, children's reliance on verb bias following cognitive control engagement ought to increase with the increased bias strength of the verb. The more strongly biased the verb, the more children ought to follow that bias when they are in a more highly engaged control state.

As the previous chapter outlined, this was not quite the case: children appeared to follow an instrument bias more strongly for equi-biased verbs than for more instrument-biased verbs. One potential reason for this discrepancy may lie in the use of adult cloze-task norming data in establishing verb bias. The cloze task itself relies on adult productions, whereas the target phenomenon is children's stored representations of the bias of particular verbs. While the adult cloze results may approximate the bias that children experience, task characteristics may introduce discrepancies between norming data and children's interpretations of the biases of particular verbs.

The particular way cloze data were captured in this study potentially biased participants to produce more modifier-like responses. For example, sentences in the norming study asked participants to complete sentences like "She will [Verb] the person with…" The inclusion of "person" as the DO noun may have made it more likely that participants would respond with modifier responses overall (by design), as people are inherently differentiable and thus relatively likely to be modified (as noted in Chapter 3, compare to the sentence "she will [Verb] the ant with…" Since ants are less inherently differentiable, the choice to complete the sentence with an instrument-like response instead of a modifier-like response becomes more appealing). This may

mean that the verbs classified as “equi-biased” would really be classified as instrument-biased in a neutral context.

Finally, it’s possible that the delineation of three categories of verb-class is an artificial one, that doesn’t accurately match the mental categories that children or adults possess. After all, the likelihood that a particular verb will predict NP or VP attachment for an upcoming “with” PP can be measured on a continuous scale or a binary one (e.g. perhaps the slightest bit of bias in either direction serves to make a verb biased in that way). All of these reasons suggest that the adults’ cloze norming data leaves something to be desired when it comes to approximating children’s notion of verb bias, so it stands to reason that children’s performance in Experiments 4 & 5 may not perfectly correlate with verb biases measured by the cloze data – a lack of correlation between these two measures is not necessarily a knock against the Reliability Hypothesis.

The goal of Experiment 8 is to measure children’s biases for particular verbs in a way that is less removed from their own experiences with these verbs. Specifically, Experiment 8 aims to measure the reliability of each verb in Experiments 4 & 5, based on the input that children have heard. To that end, a corpus of child-directed speech was analyzed, and each verb was coded for the proportion of times it predicts an upcoming with-phrase to branch from the VP or NP.

To provide support for the Reliability Hypothesis, it is predicted that the more a verb is consistently used in a particular grammatical environment, based on this corpus measurement, the more likely children will be to rely on this verb’s bias following cognitive-control engagement.

For each verb, the proportion of utterances in which it was used in an instrument-biased way (out of the total number of instrument and modifier codes) was calculated to establish that verb's instrument score according to the corpus analysis. These scores were then compared to the adult cloze ratings from Experiment 4 & 5, and correlated with the likelihood for which children looked at the lone instrument during these experiments.

It was predicted that the more strongly biased the verb was found to be based on the corpus data, the more likely children would be to look at the lone instrument during the critical window in the prior experiments, regardless of verb category established by the adult cloze data. Further, it was predicted that this correlation would be stronger for the verbs that followed incongruent Flanker trials, indicating that cognitive control engagement from these trials led to stronger reliance on verb bias for the more biased verbs.

4.1.1: Corpus Selection

The corpora to be analyzed were drawn from the North American English corpora within the CHILDES library (MacWhinney, 2000). Corpora on children older than 8 years or who were not typically-developing were excluded, since these are less likely to directly reflect the experiences of the children who participated in Experiments 4 & 5. Indeed, the majority of corpora used, with the exclusion of MacWhinney (MacWhinney, 1991) & Gelman (Gelman et al., 1998) contained speech only to children younger than 6;6, which was the age cutoff for these experiments. These corpora comprise speech to 1,247 different children, and vary greatly in their breadth (number of children) and depth (number of utterances said to

each child). The utterances were measured in lab, home, and school environments, during a variety of activities (e.g. play time, meal times, book reading), and span several decades (many corpora date back to the 1970s and 1980s, while the most recent ones are from the 2010s).

To assess verb bias, utterances of interest were first limited to sentences containing the verbs used in Experiments 4 & 5, within 20 words of the word "with" using AntConc (Anthony, 2020) software. While not completely indicative of the target structure, this liberal criterion allowed for casting a wide net: the aim was to take a wider-then-necessary sample of verbs and pare it down, rather than risk excluding target utterances. This method also did not require corpora to be parsed (and parsed correctly) for inclusion in the analysis. This subset of the CHILDES corpora resulted in 7,825 utterances for subsequent analysis.

4.1.2: Coding Method

For each utterance, trained research assistants combed through and identified instances where these verbs were used in conjunction with "with" phrases and identified whether they were used with NP attachment, VP attachment, or in a non-target way. For each utterance, coders were given several utterances before and after the coding target, in the event that the situational context was necessary to determine the intended PP attachment. Coders assigned one of 4 different codes to each utterance, summarized in Table 4.1, below.

Code	Meaning	Example utterance
M	Modifier-biased usage (NP-attachment)	"I **found** this yellow canary bird with a black stripe on each wing"
I	Instrument-biased usage (VP-attachment)	"You going to **cover** the doll's feet with the blue blanket?"
O	Other usage (e.g. "along with")	"You can **look at** it later with me"
A	Ambiguous usage	"**Hit** the drum with the xylophone"

Table 4.1: Coding schema for Experiment 8

Coders also had the option of indicating that they believed an utterance was unambiguous, but that they were not personally sure what the correct code ought to be. In many instances, utterances were technically ambiguous (as in the "I" example in Table 4.1), but one interpretation was far more likely than others given the context (discussion of where to put a blanket).

25% of utterances were double-coded for reliability. Although inter-coder reliability was adequate (80%), to ensure accurate codes all instances that were deemed by at least one coder to be of the correct form (both useable and the with-phrase specified the PO of the verb in question) were double checked. This analysis resulted in 1332 total utterances, or an average of 55 instances per verb.

4.1.3: Results

Corpus analyses: Of the 24 verbs analyzed, 23 of them appeared in the corpus with a with-phrase specifying its prepositional object: "jab" did not, so this verb was subsequently excluded from further analyses. In order to establish the strength of each verb's bias, the percentage of times verbs were used in an instrument-like manner (of the total number of modifier- or instrument-like usages) was calculated as the verb's

instrument score. Figure 4.1 shows the instrument scores (in red) for each verb, sorted by score from lowest to highest (from the corpus data). This is shown in conjunction with the verb condition they were assigned based on the Cloze norming data in Experiments 4 & 5. While it is clear from this figure that the verbs considered to be modifier-biased based on adult data do indeed have the lowest instrument scores, the division between equi- and instrument-biased verbs is far less apparent. While there was a significant difference between modifier-biased and instrument-biased verb groups such that modifier-biased verbs had lower instrument scores (t=3.28, p=.005), there was not a significant difference between instrument-biased and equi-biased verbs' instrument scores, though equi-biased verbs were numerically higher than instrument-biased ones (t=.68, p=.51),

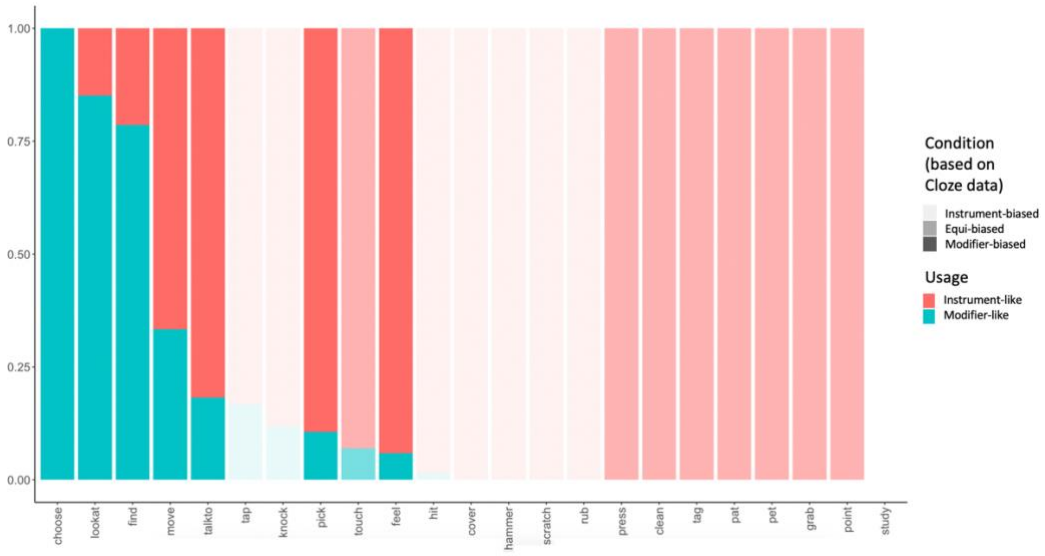

Figure 4.1: Proportion of Instrument-like usages in the corpus data (red) vs. Modifier-like usages. Shading represents cloze data category.

While this lack of differentiation between instrument- and equi-biased verbs shows a disconnect from the adult cloze task results, this homogeneity mirrors children's relatively similar looking-time data for equi- and instrument-biased verbs from Experiment 4. Of interest, then, is whether these corpus data correlate with children's propensity for VP-attachment interpretations.

Correlation with child looking-time results:

As Figure 4.2 below shows, children's looks to the lone instrument appear to predict verbs' corpus-derived instrument score. This was confirmed by a significant correlation between the two measures (R^2= 41.44. See Table 4.2 for linear regression model results).

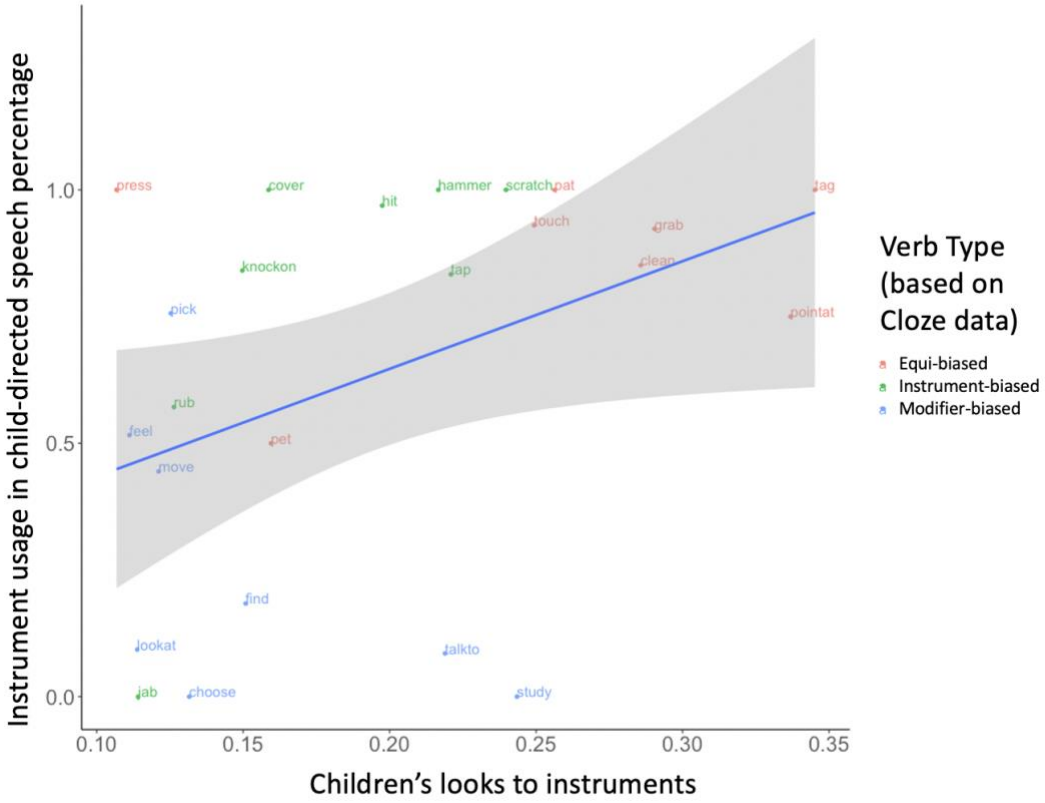

Figure 4.2: Children's likelihood of looking to the lone instrument in Experiments 4 & 5 is plotted on the X-axis, while for the same verbs, the likelihood that parents used them in an instrument-like way is plotted on the Y-axis. These measures are correlated (see Table 4.2 for full details)

All Verbs	*Estimate*	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	0.22	0.21	1.08	0.29
Instrument Looks	2.12	.99	2.14	0.04*
Verbs following incongruent Flankers	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.09	0.26	.378	0.72
Instrument Looks	2.56	1.20	1.14	0.05*
Verbs following congruent Flankers	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.37	0.36	1.01	0.34
Instrument Looks	1.48	1.91	.78	0.46

Table 4.2: Estimates of coefficients from Experiment 8 correlations between corpus Instrument scores and children's looks to the lone instrument in Experiments 4 & 5

This correlation was slightly stronger for verbs following Incongruent flanker trials (R^2=56.05) than for verbs following Congruent flanker trials (R^2=23.84), though these correlations did not significantly differ (t=.49, p=.63). See Figure 4.3 below for a visual representation.

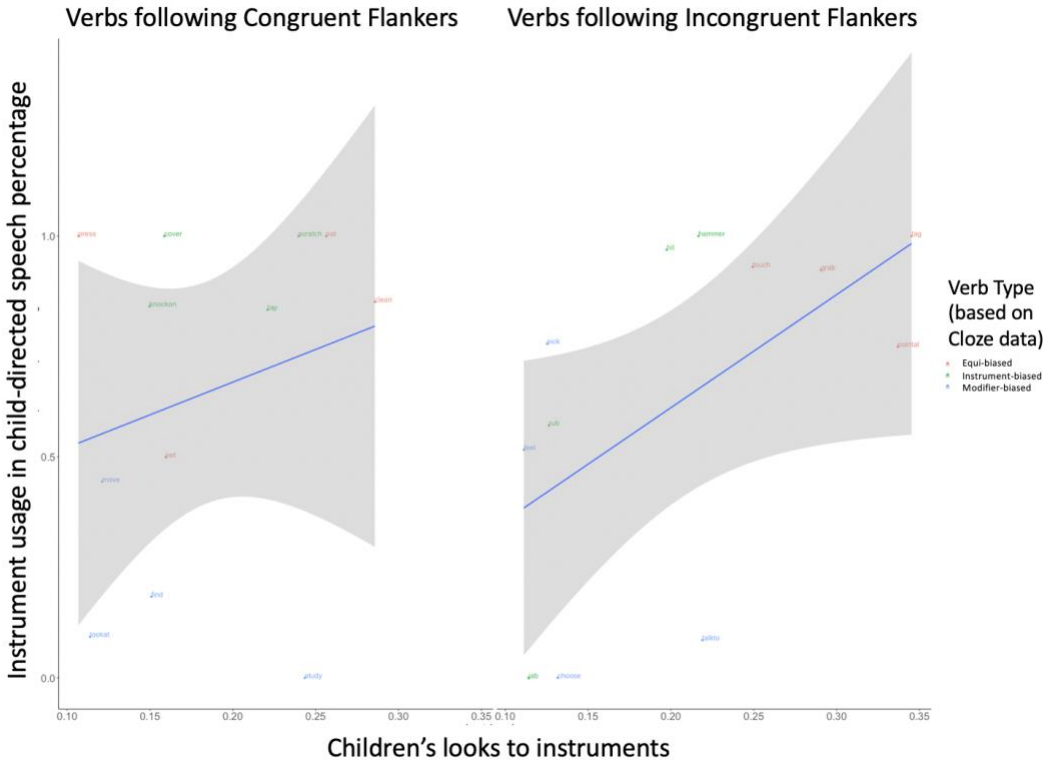

Figure 4.3: Children's likelihood of looking to the lone instrument in Experiments 4 & 5 is plotted on the X-axis, while for the same verbs, the likelihood that parents used them in an instrument-like way is plotted on the Y-axis. When verbs followed incongruent Flanker trials (right panel), the correlation is slightly stronger than when verbs followed congruent Flanker trials (left panel). See Table 4.2 for full model details

Visual world data reanalysis:

Finally, children's looking-time data from Experiments 4 & 5 were re-analyzed, with verb-type now categorized on the basis of instrument vs. modifier usage in the corpus data. These results are presented in Figure 4.4. As in Experiments 4 & 5, verbs we categorized as being instrument-biased if they were used with VP-attachment in the corpus in more than 80% of utterances, modifier-biased if they were used with VP-attachment in fewer than 40% of utterances, and equi-biased if they were used with VP-attachment in 40% to 80% of utterances. Under this re-analysis of verb types, it is first important to note that there was a main effect of verb type such

that children hearing instrument-based verbs looked significantly more to instruments than children did when hearing equi-biased verbs (β=0.06, SE=0.008, t=6.99 p<.001) and modifier-biased verbs (β=0.03, SE=0.009, t=3.05, p=0.002), indicating that children were indeed sensitive to the verb-bias divisions made on the basis of the corpus data from speech to children.

Additionally, these looks interacted with Flanker type. Children hearing instrument-biased verbs were significantly more likely to look to the lone instrument following Incongruent Flanker trials than when these verbs followed congruent Flanker trials (β=0.04, SE=0.009, t= 4.51, p<.001). Children hearing modifier-biased verbs, on the other hand, were significantly *less* likely to look to the lone instrument following incongruent Flanker trials than when these verbs followed congruent Flanker trials (β=0.03, SE=0.02, t=2.07, p=0.03). For equi-biased verbs, preceding Flanker trial type had no effect (β=0.02, SE=0.01, t=1.25, p=0.21).

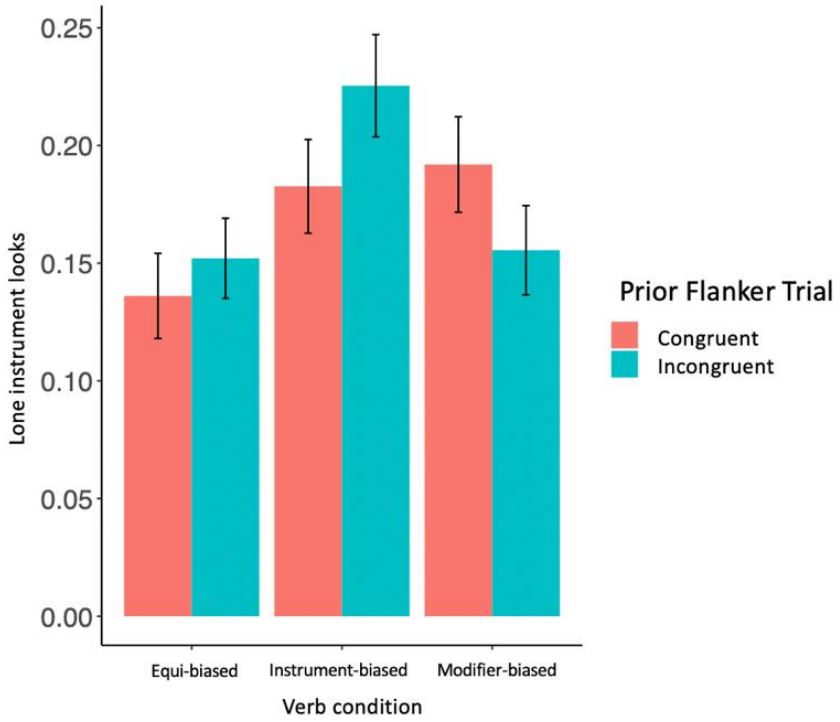

Figure 4.4: Children's looks to the lone instrument, split by verb condition (as measured by corpus usage data) and prior Flanker trial type.

4.1.4: Discussion

The corpus analysis results in Experiment 8 explain why children didn't seem to differentiate instrument- and equi-biased verbs in Experiment 4 – according to the distribution of these verbs in children's listening environments, this distinction doesn't match adults' norming data using a cloze task. In fact, the verbs previously classed as equi-biased were ones that were quite frequently used with an instrument specified. For example, when children heard the verb "pat," they heard sentences like "**Pat** it with your hand," and when children heard "touch" they heard sentences like "You may not **touch** anything with your dirty hands." These differences may stem from differences in the kinds of utterances parents are likely to say to children, and the kinds of utterances adults expect to hear in their daily lives. For an adult, patting or touching things in the abstract might lend itself to consideration of the object of the

verb. For a child, however, it may be much more important to specify the way in which the patting or touching is done (particularly if the child's primary touching instruments are dirty!).

The finding that children's looks correlate with the bias of the verbs, according to corpora of child-directed speech, suggests that their looks in Experiments 4 & 5 are modulated by the extent to which particular verbs predict upcoming structure in the corpus. When the verbs they hear are more likely to be used with instruments, children are more likely to arrive at a VP-attachment parse, and vice versa.

Moreover, the modulation of this correlation by prior Flanker type suggests that an upregulation of cognitive control increased children's reliance on verb bias. This finding, that children are more likely to be parsing according to the bias of the verb found in the corpus for the verbs that follow incongruent Flanker trials, lends support for the Reliability Hypothesis - the effect of those incongruent Flanker trials was to lead children to parse according to what they believe the bias of that particular verb to be: its reliability in predicting upcoming sentence structure.

Finally, and perhaps most importantly, when verbs are re-categorized into new verb types on the basis of how they are used in input to children, a clear pattern emerges. Incongruent Flanker trials lead children to parse according to verb type even more fervently. That is, children hearing instrument-biased verbs are more likely to interpret the with-phrase as VP-attached in these cases following cognitive control engagement. Likewise, children hearing modifier-biased verbs are more likely to interpret the with-phrase with NP-attachment following cognitive control

engagement. Whereas children hearing equi-biased verbs do not parse significantly differently on the basis of cognitive control engagement state. This provides some relatively clear evidence in favor of the Reliability Hypothesis: when children's cognitive control system is in a more up-regulated state, they are more likely to take their parsing cues from verbs when those verbs themselves more reliably predict a particular parse.

Turning back to Experiments 1-3, further support for the Reliability Hypothesis could be garnered by showing that in those studies as well, children were relying on the more reliable sentence processing cues following cognitive control engagement. In these studies, an underlying assumption was that for children, the likelihood that the verb "put" will specify a location is reliable, which lead to an increase in garden-path effects in Huang et al., (2016). On the other hand, it was assumed that the agent-first bias is a relatively unreliable parsing cue (at least compared to the verb cue in the passive sentences), leading children to stop relying on it quite as strongly following cognitive control engagement. Some initial evidence for this can be seen in the results of Experiment 3 – children happily abandon the agent-first bias given a discourse context that induces less uncertainty about the identity of NP1s. Further support for the Reliability Hypothesis would be established by empirically measuring the relative reliability of these cues in speech to children, and it is just this measurement that Experiment 9 seeks to establish.

4.2: Experiment 9: Put vs. agent-first corpus coding

Prior work has shown that for "put" imperatives with PP attachment ambiguity (e.g. "put the frog on the napkin into the box"), cognitive-control engagement increases children's online Goal interpretations of "on the napkin" (Huang et al., 2016). By hypothesis, this is because children regard the initial verb as a highly reliable cue that PP1 will attach to the VP and will be a location for the putting event. The goal of Experiment 9 is to compare, given children's input, the relative reliability of the verb "put" in predicting a location for the putting event in the proximate PP, to the reliability of children's agent-first bias. Since these cues exist at different grain sizes (the agent-first bias is a prediction about word order while subcategorization frames provide information about a particular verb), it is first necessary to quantify reliability in a way that can be measured on both of these levels. Here, reliability is measured as the relative likelihood that following one of these cues will lead to an adult-like interpretation for the sentence at hand. In other words, the likelihood that when the verb *put* occurs, a location will be (explicitly) specified will be compared to the likelihood that the first NP in a sentence will be the agent of that sentence. To that end, the number of instances of *put* and NP1 in selected corpora of speech to children will be measured, and the relative proportion of these instances in which they reliably predict goals and active structures, respectively, will be measured.

It is predicted that "put" will be a more reliable parsing cue than the expectation that initial NPs will be agents, lending credence to the claim that cognitive-control engagement does indeed lead children to rely more on cues that are more reliable, and less on cues that are less predictive of upcoming structure.

4.2.1: Corpus selection

To assess the likelihood that "put" would predict an upcoming goal and that an initial NP would be the agent of the main verb in an utterance, speech to children that was tagged for argument roles was needed. For this reason, the Adam, Eve, and Valian sub-corpora from the Pearl & Sprouse derived CHILDES corpus was used, as these corpora contain speech to children that is parsed and tagged for theta role assignment (Pearl & Sprouse, 2011).

For the analysis of "put" sentences, utterances were then limited to sentences that contained the verb *put* (1,303 utterances). For the agent-first analysis, utterances were instead limited to utterances that contained the tag "NP-1" (5,916 total utterances).

4.2.2: Results

Put results: Of the total number of sentences that contained *put* in the Pearl & Sprouse sub-corpora, 636 of these utterances were followed by a "Goal" argument role tag in the same utterance. In other words, for 48.81% of Put utterances a goal was explicitly mentioned.

Agent-first results: Of the total number of sentences that contained a noun phrase tagged as NP-1, 1087 were also labeled with an Agent tag and 1649 were labeled with a Theme tag (the remaining 3,180 utterances had other role assignments such as "experiencer," and were ignored in this analysis). In other words, of the initial NPs that were tagged as agent or theme, only 39.73% of them were indeed agents.

Further analysis of the particular sentence tags yielded some perhaps unsurprising results: The majority (61.6%) of the sentences in which the initial NP

was coded as a Theme were WH words coded as initial NPs. Excluding these, the analysis flips: of the initial NPs that were tagged as agent or theme, 61.3% of them were agents.

4.2.3: Discussion

The goal of this chapter is to provide further empirical support for the notion that children are calculating the reliability of various cues to sentence processing, and that this relative reliability heuristic is relied upon in cases of signal conflict, like structural ambiguity. The results of Experiment 9 are somewhat mixed – when children encounter the verb "put," it predicts a goal approximately half the time. On the other hand, the agent-first bias is less likely to lead children to the correct answer if wh-words are in their consideration set. If so, this potentially lends support for the notion that in Huang et al. (2016) children relied on the cue from "put" because it was relatively reliable, whereas in Experiments 1-2 children were better able to ignore the agent-first bias following cognitive control engagement because it was unreliable. If not, the results flip – the agent-first bias is a slightly more reliable cue than "put," though this is perhaps not proof that the agent-first bias provided a particularly reliable cue in Experiments 1-2, as even so it's still only accurate 2/3rds of the time.

It is perhaps worth considering why children rely on the agent-first bias at all considering that it's not the most robust heuristic. One reason may be that children encounter it frequently, if not reliably. Even in the present corpus analysis, sentences tagged with initial NPs were more than 4 times more common than sentences tagged with "put," despite it being a relatively common verb in child-directed speech. The

agent-first bias also allows children to make guesses about the nature of NP1s under conditions of uncertainty.

In general, there is fair reason to believe that the statistics of child-directed speech will differ in meaningful ways from those of adult speech, where sentential ambiguity is concerned. Prior work has found differences in child-directed and adult speech for "put" sentences, as measured by surprisal values at each word (Ovans et al., 2020). While surprisal at the point of disambiguation was high for adult corpora, it was relatively low for child-directed corpora, compared to other words in the sentence. This indicates that children may not be receiving as strong of an error signal as adults are when garden-path sentences are disambiguated, and highlights the need for evaluation of the input that children hear, instead of relying on adult norms.

Overall, the results of this chapter converge to lend credence to the notion that Experiments 1-5 do indeed support the Reliability Hypothesis. When children's cognitive control system is engaged, they are more likely to choose a parse that is suggested by the cue in the sentence that is ordinarily reliable for the task at hand. The results presented here suggest that while the presence of the verb "put" is a relatively reliable indicator that a goal will be explicitly mentioned, simply encountering a noun phrase is not necessarily a strong indicator that this noun phrase will be an agent. It may follow from this, when children's control system is up-regulated, they are more likely to expect an upcoming PP to act as the goal of *put*, but are less likely to follow the agent-first bias. Similarly, when speech to children is measured, more evidence is garnered for the idea that children are more likely to

parse according to sentence cues that are good predictors of upcoming structure following cognitive-control engagement.

Chapter 5: Conclusion

Across the nine studies presented here, the goal of this dissertation is to determine how children's still-developing cognitive-control system interacts with the developing parser. The experiments outlined in Chapter 2 indicate that when children's cognitive-control system is engaged, they rely more on parsing cues that they know to be reliable predictors of upcoming structure, and use these to re-rank potential parses of the sentence they're hearing accordingly. Building off these results, the experiments in Chapter 3 more precisely characterize the notion of reliability that children are calculating, and which cognitive-control engagement biases them to rely on. Finally, the corpus work in Chapter 4 extends these results, verifying that children are relying on cues that are independently shown to be reliable in the speech that they hear.

This dissertation began with an apparent puzzle: On the one hand, children have more difficulty than adults navigating ambiguity in non-linguistic tests of cognitive control such as Stroop and Flanker (Diamond et al.,2007; Bunge et al., 2002; Davidson et al., 2006; De Luca et al., 2010; Zelazo et al., 2014). This performance appears to mirror their difficulty in reaching a final, adult-like interpretation for garden-path sentences. Successful interpretation of temporarily ambiguous sentences seems to require the same control machinery as successful navigation of a Stroop task: Two mental representations suggested by the input are in conflict, and one must be ignored while the other is followed for the purpose of performing a task-specific action. For adults, conflict adaptation studies demonstrate fairly conclusively that there is a shared mechanism underlying garden-path and

Stroop processing (Hsu & Novick, 2016; Hsu, Kuchinsky & Novick, 2021). Given these findings, we might expect children's performance on garden path sentences to be correlated with their performance on non-linguistic cognitive control tasks, if their difficulties in interpreting garden paths stem underlyingly from their difficulties at executing cognitive control. Why, then, do these correlations often fail to appear?

This lack of correlation is puzzling under the assumption that adults succeed at navigating cognitive control tasks because they have more of a control resource that children are still accumulating. This interpretation of cognitive control is more or less in line with what it means to have more of other types of executive function, such as working memory. Those with longer working memory spans have more "slots" for information storage, and this leads to increased performance at non-linguistic measures of this, such as n-back tasks (e.g. Anguera et al., 2012).

However, an initial question that should be asked is what this control resource would look like, and what it would mean to have more or less of it. The ability to decide between two disparate mental representations and choose the one that is more relevant to the task at hand isn't the type of process that seems to require "slots" of this sort. Instead, as it's conceptualized in the adult cognitive control literature, having "more" cognitive control means more efficiently being able to filter mental activation so that the task-relevant representation is more highly active than the irrelevant one (e.g. Botvinick et al., 2001; Braver et al., 2007; Braver 2012). Cognitive control being a system that differentially boosts task-relevant domain specific representations can explain the conflict adaptation phenomenon in adults – one way to be "better" at executing cognitive control is to have a tighter connection

between the domain-general system that signals a need to attend to task-relevant information and the domain-specific system that identifies what that information is in the moment.

Back to the original puzzle - how could we conceptualize a system like this, that explains why children's difficulty with garden paths doesn't always relate to their performance on non-linguistic cognitive control tasks? One explanation, presented here, is that they're using a heuristic like reliability – when their cognitive control system is more engaged, they attend to the stimuli that are ordinarily reliable indicators for the task at hand. When these ordinarily reliable cues lead children astray, depletion-like effects are found, and when they are cues to the adult-like interpretation, children appear to adapt to conflict as adults do.

The data presented here are difficult to explain under a depletion account. If children's relatively non-adultlike performance on the "put" task is due to them having less of a resource, and each successive instance in which cognitive control is needed uses up some up this resource, it is difficult to explain how adaptation to conflict when encountering it a second time would occur. In particular, in Experiments 1 & 2, when children have just completed an incongruent trial of a cognitive control task and have therefore just had to ignore one tempting stimulus in favor of another, more task-relevant one, a depletion account predicts that children would be subsequently worse-equipped to ignore the initial parse they built when processing a garden-path sentence seconds later.

Instead, these data argue for something like the Reliability Hypothesis, wherein when children encounter conflict following cognitive control engagement,

they are more likely to activate representations that are suggested by input cues that are ordinarily reliable ones for the task at hand. All told, there are several major reasons to doubt an account that makes reference to resource depletion as an explanation for children's failure to revise during garden path sentence processing. For one, it's unclear how a system that instantiates control as an amount of a particular resource could explain differences between adult and child performance to begin with. Such an account would have to explain how, for adults, having more of a particular mental resource (e.g. storage space in memory) would lead to a greater ability to ignore irrelevant stimuli. Additionally, resource depletion is in conflict to how the cognitive control system is conceptualized in the adult literature, where cognitive control is described as a biasing system that can lead to increased activation of task-relevant cues when it is more up-regulated.

Overall, the studies presented here outline a fairly specific positive proposal for how, at a mechanistic level, a domain-general system that mediates conflict might interact with a domain-specific language processing system. Figure 5.1 provides a schematic of this positive proposal. Using as a starting point the model introduced by Botvinick et al., (2001) for how conflict monitoring leads to conflict adaptation in a Stroop task, this is meant to represent a general case of how cognitive-control may be engaged and lead to changes in real-time parsing decisions.

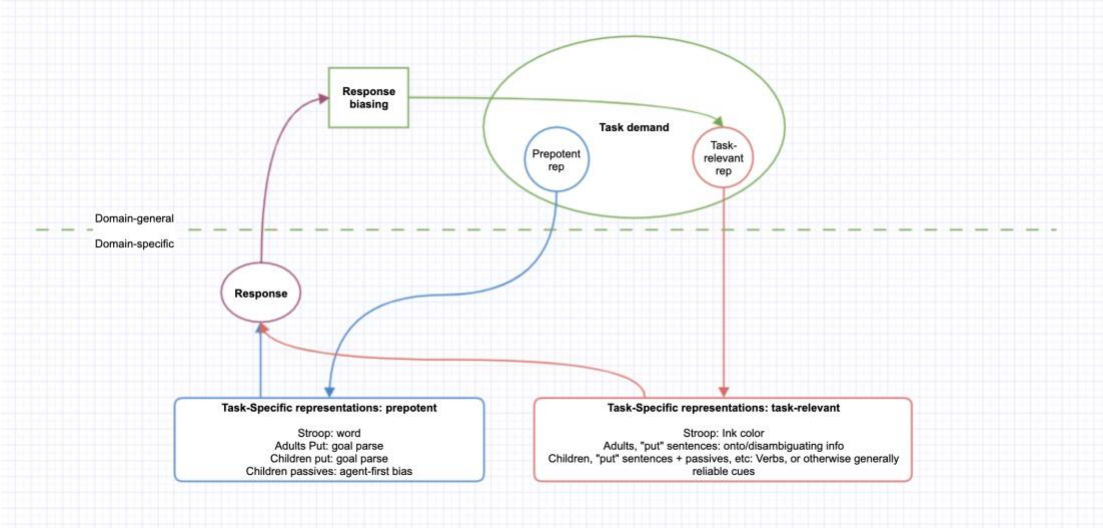

Figure 5.1: Proposed schematic of how domain-general cognitive control might affect domain-specific linguistic processes

Once conflict is encountered (e.g. on an incongruent Stroop or Flanker trial), a domain-general task-demand system (the green oval in Fig. 5.1) that encodes a distinction between prepotent and task-relevant representations is activated. In a classic color-word Stroop task, the task-relevant information (in red) would be the ink color while the prepotent, salient information (in blue) would be the color word. While the domain-general system does not contain these task-specific representations, it must be connected to other domain-specific systems that do.[6] In making a response (e.g. a button press, in purple), the prepotent and task-relevant information is brought into conflict, creating a situation that requires adjudication between two different ways of characterizing the input. This information is then relayed to a general system that biases responses toward task-relevant goals (green square), thus allowing for the

[6] Some nice neural evidence in showing this comes from Hsu, Jaeggi, & Novick (2017), who demonstrated with that the conflict-inducing trials in a Stroop task, N-back task and Recent probes task differentially recruit the left inferior frontal gyrus (LIFG) and coordinate with different posterior brain areas that are necessary for the particulars of the tasks (e.g. visual word form area during Stroop, and memory areas during the n-back task).

possibility of adaptation to meet task demands across disparate domains. When an ambiguous sentence is then encountered after an incongruent Stroop trial, connections to the task-relevant nodes are strengthened – this is an illustration of what it might mean for cognitive control to be relatively up-regulated. For adults interpreting an ambiguous sentence, they are better able to attend to task-relevant information in the task: the disambiguating words.

For children, the process is ostensibly the same, the major difference is found in the scope of what is considered "task-relevant" for a sentence processing task. If children treat information that is *generally* reliable (but perhaps not so for the particular sentence they're hearing) like verbs as particularly task-relevant, this will result in them relying more heavily on these cues when their cognitive-control system is more engaged. This strategy is a sensible one for the developing parser: in the face of conflicting cues, it is logical to have a system that's pre-biased to weight the one that has been reliable in the past more heavily. However, this strategy will lead to errors when those normally-reliable cues mislead. The role of language experience, then, is to narrow down the scope of task-relevance. With each subsequent brush with a particular structure (say, a "put" imperative with multiple prepositional phrases), it becomes clearer which information ought to guide parsing decisions when multiple cues are in conflict (here, the onset of the second PP). Heuristics like relying on verbs can be pruned away in favor of a more sophisticated representation of task-relevant cues to accurate parsing. As children age, widening the scope of what might be considered a reliable cue (e.g. beyond verbs) will help with recovery from misinterpretation in the long run.

5.1: Further questions and future work

One question that might arise is whether the main hypotheses presented here (Depletion vs. Cognitive biasing) are indeed mutually exclusive, or whether they could both be mentally instantiated, but perhaps at different levels of analysis. If the human cognitive control system approximates the model proposed in Figure 5.1 or ones proposed by Botvinick et al. (2001), it is not clear that a pool of "control resources" would be compatible with these models. Perhaps the speed or consistency with which the domain general task demand system signals to the part of the system that encodes domain-specific representations could be measured in a continuous manner, but this seems like a large deviation from the type of mental resources that are ordinarily invoked (e.g. Woodard et al., 2016; Qi et al., 2020; Powell & Carey, 2017; Wehbe et al., 2020; Ryskin, Levy, & Fedorenko, 2020). Alternatively, if the cognitive biasing account is wholly incorrect and the human cognitive control system does contain something resembling a resource that breaks down or is used up over time with each successive demand on the control system, the adaptation-to-conflict results seen here, in Ambrosi et al. (2016), and with adults in Hsu & Novick (2016) are difficult to explain.

A further question that may arise from these results is the nature of the conflict adaptation process. It stands to reason that if children's looks in the sentence trials of Experiments 1-5 are affected by similarity in a mental process initiated from preceding Stroop/Flanker trials, that the conflict adaptation effects discussed here ought to be reversible. That is, does conflict adaptation happen from sentence-to-Stroop/Flanker? While in theory the effect ought to be measurable in reverse, the

studies described here were not designed to test the effect in this direction, and therefore the number of C-C/C-I/I-C/I-I trials in this direction are inconsistent. Additionally, in these studies there was a longer delay between the time point at which a decision has to be made during sentence trials to Stroop/Flanker trials than Stroop/Flanker to sentence, as children were free to respond with as much time as they liked during the sentences. In Experiments 1-3, Stroop trials timed out after 2 seconds, and in Experiments 4-5 Flanker trials did not time out but children responded within a few seconds on average. In contrast, children often took up to 30 seconds to respond during sentence trials. This may also help explain the lack of interaction between Stroop/Flanker trial type and act-out results – after a few seconds the effect of prior trial conflict has diminished. Future work will seek to find a neurological signal of cognitive control engagement in an attempt to estimate the precise timing of this effect.

One might wonder whether the particular choice of cognitive control task used here (i.e. dog-Stroop and Flanker) might affect the results, given that it is common to use several different such tasks in studies assessing children's overall cognitive control abilities (e.g. Woodard et al., 2016; Diamond et al., 2007). While it has been claimed (e.g. Braver, 2012) that different tasks may tap into different sub-categories of cognitive control, it remains the case that all tasks that are ostensibly cognitive control tasks have in common the need to ignore one prepotent stimulus and instead focus on another, task-relevant one. It should therefore be the case that any one of the various tasks that require this ought to engage the cognitive control system, albeit they may do so to different extents. Relatedly, a recent meta-analysis of attempts to

find correlations between cognitive control tasks found that these tasks often fail to correlate, and have in fact been correlating less and less over time (Rouder et al., 2019). The authors conclude that high degrees of trial noise are responsible for this lack of correlation – even though they may all be tapping into the same underlying system, other mental systems must be used for particulars of each task (e.g. color and word processing for Stroop, direction processing for Flanker), and performance of many of these systems all factor into the final dependent measures on these tasks (e.g. accuracy or reaction time). In other words, while the particular choice of non-linguistic cognitive control task may matter because different tasks may introduce different amounts of noise, the basic fact that they each ought to make use of the same underlying system remains the same.

A further question may be how feasible the model presented in Figure 5.1 is as a description of how real-time parsing interacts with cognitive control. That is, in the model, conflict is encountered when two competing cues are simultaneously active yet a response must be made. In contrast, during parsing, even for garden-path sentences which are tailor-made to induce a cognitive control burden, information is presented to the listener in a linear manner. For this reason, competing cues may not necessarily simultaneously present in the input. The model presented here therefore assumes that even when competing parses are not generated simultaneously by the input listeners receive, initial, "incorrect" parses and revised, "correct" ones both exist in opposition to each other (e.g. it assumes that some aspect of the initially-built parse is held in working memory long enough to still need to be actively ignored when revised parses are generated). While a parse being initially-built may be a factor

that leads to that parse being a particularly tempting one to go with, this may be in contrast with later-built parses that are generated using ordinarily-reliable cues, like the information structure suggested by the main verb.

A lingering question from the results presented here is how the cognitive biasing and reliability hypotheses apply to adult conflict adaptation effects, and what changes they predict occur as children age. The Cognitive Biasing hypothesis is meant to apply to adults as well, whereas the Reliability Hypothesis is meant as an explanation of why children's performance often differs from adults'. While children may be led astray by using a reliability heuristic to determine which aspect of their input to attend to when multiple cues conflict, the inference is that over time adults learn which cues in their input are task relevant for the particular task at hand, instead of relying on cues that are ordinarily reliable.

This raises a related, and important question: what exactly makes something a "task"? The claim presented here is essentially that what it means for children's cognitive control system to "improve" over time and allow them to eventually succeed at cognitive control tasks is a greater understanding of the boundary conditions for the particular task at hand. For example, they must determine that when processing a sentence like "Put the frog on the napkin into the box," the part of their input that is particularly relevant to the current task is the disambiguating preposition, and not, in this case, the ordinarily reliable subcategorization information from the verb "put." What it means to rely more on the "task-relevant" cue when you begin a new task, then, is to up-weight the importance of the domain-specific mental representation that you've decided ought to be connected to your domain-general

representation of what's task-relevant, within the task-demand system. In this way, cognitive control needn't be constantly on the lookout for anything that could be relevant for any next task. Instead, the domain-general system that encodes task-relevant representations is pre-activated and acts to boost the domain-specific representation of that relevant cue the next time the cognitive control system is needed.

Together, the studies presented in this dissertation suggest that throughout development, our cognitive control systems remain consistent in the way they interact with domain-specific sub-systems such as sentence processing. We up- or down-weight cues on the basis of how likely they are to be task-relevant. When processing garden-path sentences, a task-relevant cue might be one that leads to a sensible understanding of who did what to whom. For comprehenders successfully comprehending "put" sentences, the final PP "on the box" provides such a cue by disambiguating the sentence. But some comprehenders (e.g. 5 year-olds) may be casting a wider net and treating the cue from the verb as more task-relevant, as that's often a good bet based on their prior experience with verbs. Part of what it means for children's cognitive control system to mature is to fine-tune their assessment of which cues count as relevant for the particular task at hand.

5.2: Conclusion

Overall, the aim of this dissertation has been to present a way of thinking about the influence of cognitive control on developmental sentence processing that is both coherent and makes use of the strides that have been made in our understanding of cognitive control engagement as a general process. The studies described here

attempt to do so by presenting evidence that cognitive control acts as a biasing mechanism that influences which cues in the input comprehenders are likely to attend to when multiple cues conflict. Children are led astray when ordinarily reliable cues lead to non-adultlike parses, but given friendlier input where reliable cues lead to adultlike interpretations as well, these cues can help children succeed.

Bibliography

Abbot-Smith, K., Chang, F., Rowland, C., Ferguson, H., & Pine, J. (2017). Do two and three year old children use an incremental first-NP-as-agent bias to process active transitive and passive sentences?: A permutation analysis. PloS one, 12(10), e0186129.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of memory and language, 38(4), 419-439.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73(3), 247-264.

Altmann, G. T., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. The interface of language, vision, and action: Eye movements and the visual world, 347-386.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. Cognition, 30(3), 191-238.

Ambrosi, S., Lemaire, P., & Blaye, A. (2016). Do young children modulate their cognitive control?. Experimental Psychology.

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. Child neuropsychology, 8(2), 71-82.

Anderson, S. E., Farmer, T. A., Goldstein, M., Schwade, J., & Spivey, M. (2011). Individual differences in measures of linguistic experience account for variability in the sentence processing skill of five-year-olds. Experience, variation, and generalization: Learning a first language, 7, 203-221.

Anguera, J. A., Bernard, J. A., Jaeggi, S. M., Buschkuehl, M., Benson, B. L., Jennett, S., ... & Seidler, R. D. (2012). The effects of working memory resource depletion and training on sensorimotor adaptation. Behavioural brain research, 228(1), 107-115.

Anthony, L. (2020). AntConc (Version 3.5. 9)[Software]. Waseda University.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In Psychology of learning and motivation (Vol. 8, pp. 47-89). Academic press.

Baddeley, A., & Della Salla, S. (1996). Executive and cognitive functions of the prefrontal cortex. Philosophical Transactions of the Royal Society: Biological Sciences, 351, 1397-1404.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. Trends in cognitive sciences, 12(5), 193-200.

Bergefurt, L., Weijs-Perrée, M., Maris, C., & Appel-Meulenbroek, R. (2021, January). Analyzing the Effects of Distractions While Working from Home on Burnout Complaints and Stress Levels among Office Workers during the COVID-19 Pandemic. In The 3rd International Electronic Conference on Environmental Research and Public Health (pp. 1-9).

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. Psychological review, 108(3), 624.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. Trends in cognitive sciences, 16(2), 106-113.

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. Variation in working memory, 75, 106.

Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J., & Gabrieli, J. D. (2002). Immature frontal lobe contributions to cognitive control in children: evidence from fMRI. Neuron, 33(2), 301-311.

Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J., & Gabrieli, J. D. (2002). Immature frontal lobe contributions to cognitive control in children: evidence from fMRI. Neuron, 33(2), 301-311.

Burgess, G. C., & Braver, T. S. (2010). Neural mechanisms of interference control in working memory: effects of interference expectancy and fluid intelligence. PloS one, 5(9), e12861.

Choi, Y., & Trueswell, J. C. (2010). Children's (in) ability to recover from garden paths in a verb-final language: Evidence for developing control in sentence processing. Journal of experimental child psychology, 106(1), 41-61.

Chomsky, N. (1981). 1981: Lectures on government and binding. Dordrecht: Foris.

Clayson, P. E., & Larson, M. J. (2011). Conflict adaptation and sequential trial effects: Support for the conflict monitoring theory. Neuropsychologia, 49(7), 1953-1961.

Cohen, J. D., & Huston, T. A. (1994). 18 Progress in the Use of Interactive Models for Understanding Attention and. Attention and performance XV: Conscious and nonconscious information processing, 15, 453.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. Cognitive psychology, 42(4), 317-367.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. Neuropsychologia, 44(11), 2037-2078.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. Neuropsychologia, 44(11), 2037-2078.

De Luca, C. R., & Leventer, R. J. (2010). Developmental trajectories of executive functions across the lifespan. In Executive functions and the frontal lobes (pp. 57-90). Psychology Press.

Degen, J., Kursat, L., & Leigh, D. D. (2021). Seeing is believing: testing an explicit linking assumption for visual world eye-tracking in psycholinguistics. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 43, No. 43).

Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, *318*(5855), 1387-1388.

Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. Science, 318(5855), 1387-1388.

Dowty, D. R. (1979). Word Meaning and Montague Grammar. Reidel, Dordrecht.

Dussias, P. E., Kroff, J. V., & Gerfen, C. (2013). Visual world eye-tracking. Research methods in second language psycholinguistics, 93-126.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception & psychophysics, 16(1), 143-149.

Farris-Trimble, A., McMurray, B., Cigrand, N., & Tomblin, J. B. (2014). The process of spoken word recognition in the face of signal degradation. Journal of Experimental Psychology: Human Perception and Performance, 40(1), 308.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. Cognitive psychology, 14(2), 178-210.

Friedman, D., Nessler, D., Cycowicz, Y. M., & Horton, C. (2009). Development of and change in cognitive control: A comparison of children, young adults, and older adults. Cognitive, Affective, & Behavioral Neuroscience, 9(1), 91-102.

Gelman, S. A., Coley, J. D., Rosengren, K. S., Hartman, E., & Pappas, A. (1998). Beyond labeling: The role of maternal input in the acquisition of richly structured categories. Monographs of the Society for Research in Child Development, 63(1), i-148.

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: performance of children 312–7 years old on a stroop-like day-night test. Cognition, 53(2), 129-153.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. Journal of Experimental Psychology: General, 121(4), 480.

Hammond, K. R., & Summers, D. A. (1972). Cognitive control. Psychological review, 79(1), 58.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behavior research methods, 50(3), 1166-1186.

Hsu, N. S., & Novick, J. M. (2016). Dynamic engagement of cognitive control modulates recovery from misinterpretation during real-time language processing. Psychological science, 27(4), 572-582.

Hsu, N. S., Kuchinsky, S. E., & Novick, J. M. (2021). Direct impact of cognitive control on sentence processing and comprehension. Language, Cognition and Neuroscience, 36(2), 211-239.

Huang, Y. T., & Arnold, A. R. (2016). Word learning in linguistic context: Processing and memory effects. Cognition, 156, 71-87.

Huang, Y. T., & Hollister, E. (2019). Developmental parsing and linguistic knowledge: Reexamining the role of cognitive control in the kindergarten path effect. Journal of experimental child psychology, 184, 210-219.

Huang, Y. T., Gerard, J., Hsu, N., Kowalski, A., Novick, J. (2016, March). *Cognitive-control effects on the kindergarten path: Separating correlation from causation* [Conference presentation]. 29th Annual CUNY Conference on Human Sentence Processing, Gainsville, FL, United States.

Huang, Y. T., Leech, K., & Rowe, M. L. (2017). Exploring socioeconomic differences in syntactic development through the lens of real-time processing. Cognition, 159, 61-75.

Huang, Y. T., & Ovans, Z. (2022). Who “it” is influences what “it” does: Discourse effects on children's syntactic parsing. Cognitive Science, 46(1), e13076.

Huang, Y. T., Zheng, X., Meng, X., & Snedeker, J. (2013). Children’s assignment of grammatical roles in the online processing of Mandarin passive sentences. Journal of memory and language, 69(4), 589-606.

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. Cognition, 96(1), B23-B32.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. Acta psychologica, 137(2), 151-171.

Hurewitz, F., Brown-Schmidt, S., Thorpe, K., Gleitman, L. R., & Trueswell, J. C. (2000). One frog, two frog, red frog, blue frog: Factors affecting children's syntactic choices in production and comprehension. Journal of psycholinguistic research, 29(6), 597-626.

Iani, C., Stella, G., & Rubichi, S. (2014). Response inhibition and adaptations to response conflict in 6-to 8-year-old children: Evidence from the Simon effect. Attention, Perception, & Psychophysics, 76(4), 1234-1241.

Jacques, S., & Zelazo, P. D. (2001). The Flexible Item Selection Task (FIST): A measure of executive function in preschoolers. Developmental neuropsychology, 20(3), 573-591.

Joey Hagedorn, Joshua Hailpern, and Karrie G. Karahalios. VCode and VData: Illustrating a new Framework for Supporting the Video Annotation Workï¬‚ow Extended Abstracts of AVI 2008.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and language, 49(1), 133-156.

Kandel, M., Yacovone, A., Slim, M., & Snedeker, J. (2022, March). *Webcams as windows to the mind: comparing web-based eye-tracking methods* [Conference presentation]. The 35th Annual Conference on Human Sentence Processing, Santa Cruz, CA, United States.

Kidd, E., & Bavin, E. L. (2007). Lexical and referential influences on on-line spoken language comprehension: A comparison of adults and primary-school-age children. First Language, 27(1), 29-52.

Kidd, E., Stewart, A. J., & Serratrice, L. (2011). Children do not overcome lexical biases where adults do: The role of the referential scene in garden-path recovery. Journal of Child Language, 38(1), 222-234.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. Science, 302(5648), 1181-1185.

Larson, M. J., Clawson, A., Clayson, P. E., & South, M. (2012). Cognitive control and conflict adaptation similarities in children and adults. Developmental Neuropsychology, 37(4), 343-357.

Lassotta, R., Omaki, A., & Franck, J. (2016). Developmental changes in misinterpretation of garden-path wh-questions in French. Quarterly Journal of Experimental Psychology, 69(5), 829-854.

Luna, B., Padmanabhan, A., & O'Hearn, K. (2010). What has fMRI told us about the development of cognitive control through adolescence?. Brain and cognition, 72(1), 101-113.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psychological review, 101(4), 676.

MacWhinney, B. (1991). The CHILDES project: Computational tools for analyzing talk. Erlbaum.

MacWhinney, B. (2000). The CHILDES project: The database (Vol. 2). Psychology Press.

Magnuson, James S. 2019. Fixations in the visual world paradigm: where, when, why? Journal of Cultural Cognitive Science 3(2). 113–139.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. Perception & psychophysics, 53(4), 372-380.

Marian, V., Bartolotti, J., Chabal, S., Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. PLoS ONE 7(8): e43230.

Mazuka, R., Jincho, N., & Oishi, H. (2009). Development of executive control and language processing. Language and Linguistics Compass, 3(1), 59-89.

McMurray, B., Farris-Trimble, A., & Rigler, H. (2017). Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally. Cognition, 169, 147-164.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. Developmental psychology, 49(2), 270.

Norman, D. A., & Shallice, T. (1986). Attention to action. In Consciousness and self-regulation (pp. 1-18). Springer, Boston, MA.

Nozari, N., & Novick, J. (2017). Monitoring and control in language production. Current Directions in Psychological Science, 26(5), 403-410.

Nozari, N., Dell, G. S., & Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. Cognitive psychology, 63(1), 1-33.

Omaki, A., Davidson White, I., Goro, T., Lidz, J., & Phillips, C. (2014). No fear of commitment: Children's incremental interpretation in English and Japanese wh-questions. Language Learning and Development, 10(3), 206-233.

Ovans, Z., Huang, Y., & Feldman, N. (2020, January). The (un) surprising kindergarten path. In Proceedings of the 42nd Annual Meeting of the Cognitive Science Society.

Ovans, Z., Oppenheimer, K., & Huang, Y. T. (2019, March). *Online parsing strategies are influenced by verb-specific and language-general biases* [Conference presentation]. 32nd Annual CUNY Conference on Human Sentence Processing, Boulder, CO, United States.

Pearl, L., & Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. Language Acquisition, 20(1), 23-68.

Peter, M., Chang, F., Pine, J. M., Blything, R., & Rowland, C. F. (2015). When and how do children develop knowledge of verb argument structure? Evidence from verb bias effects in a structural priming task. Journal of Memory and Language, 81, 1-15.

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(4), 940.

Powell, L. J., & Carey, S. (2017). Executive function depletion in children and its impact on theory of mind. Cognition, 164, 150-162.

Qi, Z., Love, J., Fisher, C., & Brown-Schmidt, S. (2020). Referential context and executive functioning influence children's resolution of syntactic ambiguity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 46(10).

Qi, Z., Yuan, S., & Fisher, C. (2011). Where does Verb Bias Come From? Experience with Par-ticular Verbs Affects Online Sentence Processing.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. Journal of experimental psychology, 81(2), 275.

Rouder, J., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. PsyArXiv Preprints.

Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., & Posner, M. I. (2004). Development of attentional networks in childhood. Neuropsychologia, 42(8), 1029-1040.

Ryskin, R., Levy, R. P., & Fedorenko, E. (2020). Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. Neuropsychologia, 136, 107258.

Salig, L. K., Valdés Kroff, J. R., Slevc, L. R., & Novick, J. M. (2021). Moving from bilingual traits to states: Understanding cognition and language processing through moment-to-moment variation. Neurobiology of Language, 2(4), 487-512.

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. Behavior Research Methods, 50(2), 451-465.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective?. Psychological bulletin, 138(4), 628.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. Cognitive psychology, 49(3), 238-299.

Stowell, T. A. (1981). *Origins of phrase structure* (Doctoral dissertation, Massachusetts Institute of Technology).

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of experimental psychology, 18(6), 643.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217), 1632-1634.

Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. Journalism quarterly, 30(4), 415-433.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. Cognition, 73(2), 89-134.

Trueswell, J., & Gleitman, L. (2004). Children’s eye movements during listening: Developmental evidence for a constraint-based theory of sentence processing. The interface of language, vision, and action: Eye movements and the visual world, 319346.

Ullsperger, M., Bylsma, L. M., & Botvinick, M. M. (2005). The conflict adaptation effect: It’s not just priming. Cognitive, Affective, & Behavioral Neuroscience, 5(4), 467-472.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. International Journal of Psychophysiology, 83(2), 176-190.

Veen, V. V., & Carter, C. S. (2006). Conflict and cognitive control in the brain. Current Directions in Psychological Science, 15(5), 237-240.

Verguts, T., Notebaert, W., Kunde, W., & Wühr, P. (2011). Post-conflict slowing: cognitive adaptation after conflict processing. Psychonomic Bulletin & Review, 18(1), 76-82.

Vuong, L. C., & Martin, R. C. (2014). Domain-specific executive control and the revision of misinterpretations in sentence comprehension. Language, Cognition and Neuroscience, 29(3), 312-325.

Waxer, M., & Morton, J. B. (2011). The development of future-oriented control: An electrophysiological investigation. NeuroImage, 56(3), 1648-1654.

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., ... & Fedorenko, E. (2021). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. Cerebral Cortex, 31(9), 4006-4023.

Weighall, A. R. (2008). The kindergarten path effect revisited: Children’s use of context in processing structural ambiguities. Journal of experimental child psychology, 99(2), 75-95.

Williams, A. (2015). Arguments in syntax and semantics. Cambridge University Press.

Woodard, K., Pozzan, L., & Trueswell, J. C. (2016). Taking your own path: Individual differences in executive function and language processing skills in child learners. Journal of experimental child psychology, 141, 187-209.

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv:1504.06755.

Zehr, J., & Schwarz, F. (2018). Penncontroller for internet based experiments (ibex). URL https://doi. org/10, 17605.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. Nature protocols, 1(1), 297-301.

Zelazo, P. D. (2015). Executive function: Reflection, iterative reprocessing, complexity, and the developing brain. Developmental Review, 38, 55-68.