

## ABSTRACT

Title of Dissertation: THE MALLEABILITY OF COGNITIVE CONTROL  
AND ITS EFFECTS ON LANGUAGE SKILLS

Erika Kristin Hussey, Doctor of Philosophy, 2013

Dissertation directed by: Professor Michael Dougherty  
Department of Psychology

Cognitive control, or executive function (EF), refers to the mental ability to regulate and adjust behavior across domains in the face of interference, conflict, or new rules. Evidence from psycholinguistics suggests a role for cognitive control in a range of language processing tasks including syntactic ambiguity resolution and verbal fluency. Separate work demonstrates that EF abilities are malleable with extensive practice, such that training improvements transfer across domains to novel tasks that rely on the same underlying EF mechanisms (an effect dubbed ‘process-specificity’). In uniting these two growing literatures, this dissertation investigated the (causal) role of cognitive control for language processing through two longitudinal training interventions.

In one study, I demonstrated that practicing a battery of cognitive tasks conferred selective benefits on untrained reading tasks requiring syntactic ambiguity resolution. Compared to controls, individuals who responded most to an EF training task exhibited (1) higher accuracy to comprehension questions indexing offline reinterpretation, and (2) faster real-time recovery efforts to resolve among conflicting interpretations. A second experiment extended these findings by addressing the degree to which training on a

single EF task was necessary and sufficient to confer transfer to untrained, related language measures. Participants were assigned to practice a single training task that was minimally different from other training groups' tasks in terms of EF demands. By and large, participants who practiced a high-EF training task were exclusive in demonstrating a cross-assessment improvement profile consistent with a process-specific account: Pre/post benefits across a range of ostensibly different linguistic (verbal fluency, syntactic ambiguity resolution) and non-linguistic (Stroop, recognition memory) tasks were observed selectively for conditions with high-EF demands; no benefits were seen for cases when the need for cognitive control was minimized. Together, these findings provide support for the malleability of EF skills and suggest a critical (and perhaps causal) role for domain-general cognitive control in language processing. Further, the present studies indicate that within the right framework, and having appropriate linking hypotheses, cognitive training may be a viable way to improve language use.

THE MALLEABILITY OF COGNITIVE CONTROL AND ITS EFFECTS ON  
LANGUAGE SKILLS

by

Erika Kristin Hussey

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  
Doctorate in Philosophy  
2013

Advisory Committee:

Professor Michael Dougherty, Chair  
Dr. J. Isaiah Harbison  
Professor Susanne Jaeggi  
Dr. Jared Novick  
Professor Colin Phillips, Dean's Representative  
Professor L. Robert Slevc

© Copyright by  
Erika Kristin Hussey  
2013

## **Dedication**

To my late uncle, Richard Rued, whose enthusiasm for science inspired me at a very young age.

## Acknowledgements

Although this section is often reserved for thanking those who directly contributed to the experimental work that follows, the dissertation itself is merely a final product made possible because of the unwavering support of many people who made my experiences during graduate school positive and rewarding.

First off, my trajectory would have been unimaginably different had it not been for my advisors, Jared Novick and Mike Dougherty. As my primary advisor from day one, Mike's willingness to take me on as a student opened the door to tremendous opportunity. Without Mike taking a chance on me and endorsing my application, I would not have begun the journey that culminated in this degree! All along, Mike treated me as an equal, helping me to develop independent research agendas that were never discounted because of my limited experience as a scientist. Although we may have disagreed on some points, Mike was always compassionate, emphasizing that his challenges were aimed at strengthening the quality of my research. I am thankful for his constant eagerness to push me to develop scientific rigor. Moreover, Mike's flexibility and willingness to let me explore new disciplines—even when they did not overlapped entirely with his area of expertise—allowed me to mature professionally and ultimately develop collaborations with others, including one with my second advisor, Jared!

Prior to meeting Jared in my second year of graduate school, I (like many other junior doctoral students) lacked focus on how I envisioned my research program unfolding. Being fascinated by every topic that landed on my radar made it difficult to rein-in my interests; however, collaborating with Jared allowed me to bring together these disparate pieces in a theoretically meaningful way. As such, Jared has been

instrumental to my growth as a scientist: His willingness to entertain my excitement over half-baked ideas led to the joint development of novel, yet carefully-guided research questions. Following Jared's lead, I have learned to effectively express my work in conversation, during formal presentations, and in writing. Finally, Jared's aegis during the job application process has been truly incredible; words cannot express how thankful I am for the opportunities that have presented because of his eagerness to market my work to colleagues. Generally, I am indebted to both of my advisors for their obvious care and consideration for my well-being over the years, evidenced by the countless hours we spent batting around experimental designs and zeroing in on the locus of effects.

In addition to Mike and Jared, I am grateful for my advisory committee members over the years for their valuable feedback. Thank you to Bob Slevc, Susanne Jaeggi, Isaiah Harbison, Tom Wallsten, Tom Carlson, Daphne Soares, and Bill Idsardi for challenging me to think critically and forcing me to clearly articulate my thoughts during important graduate milestones. I must separately thank Colin Phillips for making it simple to transition into the language science community at Maryland; his encouragement over the years has played a significant role in my growth as a researcher and thinker.

I must also thank my coauthors, whose commitment to the projects presented here made completing each a cinch! Material creation, experimental design, task scripting, and data collection would not have been possible without the remarkable attention from and helpful assistance of Susan Teubner-Rhodes, Alan Mishler, Isaiah Harbison, and Kayla Velnoskey. Further, I am indebted to my colleagues within the Working Memory

Plasticity and Decision, Attention, and Memory labs, who appropriately challenged my ideas, with every effort to make my work bulletproof. Thanks especially to Nina Hsu, Jeff Chrabaszcz, and Joe Tidwell for being the incredible resources and friends throughout the final years of my graduate studies. Jeff and Joe particularly warrant special thanks for providing astute insight and knowledge regarding the sophistication of the statistics (and necessary R coding) used in this manuscript.

I thank Mike Shvartsman for his support and technical assistance during my initial foray with eye tracking. The intensive task of data collection and coding for the two training studies would not have been possible without the assistance of many reliable and skilled research assistants including Brad Zayac, Ruth Ludlum, David Alexander, Lauriane Stewart, Brogan Murphy, Remington Carey, Eric Atticks, Marie Hutton, Jennifer Sloane, Ryan Corbett, Carrie Clarady, and Alex Smaliy. Furthermore, the project ran smoothly due to extensive planning among members of the CASL TTO 3501 team.

In addition to my advisors, coauthors, and dissertation committee members, I am thankful for the guidance and support of Sharona Atkins, Mike Bunting, Barbara Forsyth, and Scott Weems who all sacrificed many hours to help design the first training experiment presented here. A special thank you to Amy Giardina and Michelle Falk for their invaluable help as CASL program managers. Finally, without the enthusiastic support of the current director of CASL, Amy Weinberg, this work would not have been so well advertised and received by the university community.

I am grateful for the comments from many journal reviewers and colleagues on the first two chapters of this dissertation, each of which shaped and improved the present work. Thank you to John Trueswell, Sylvia Gennari, Andriy Myachykov, Marina Bedny,



and Gerry Altmann for their thoughtful suggestions on earlier versions of Chapters 1 and 2. I am also indebted to the following individuals for their insightful and well-taken comments on Experiment 2: DJ Bolger, Sarah Brown-Schmidt, David Caplan, Kiel Christianson, Nelson Cowan, Gary Dell, Randy Engle, Susan Garnsey, Al Kim, Jared Linck, Dan Mirman, Akira Miyake, Yuko Munakata, Polly O'Rourke, Akira Omaki, Myrna Schwartz, and Duane Watson. Finally, I thank the attendees of the 2010 Conference on Architectures and Mechanisms for Language Processing (AMLaP), the 2011 and 2012 CUNY Conferences on Human Sentence Processing, and the 2010 and 2012 Annual Meetings of the Psychonomics Society for their comments on this work as it was presented in its earliest iterations.

Funding from various groups allowed me to pursue such large-scale training studies including that from the Center for Advanced Study of Language (CASL), the Graduate School (via the Ann Wylie Dissertation Award), and the Language Science NSF IGERT program. Further, I am indebted to Pam Komerak of the NACS program for nominating me for additional support, which granted me additional time to complete my dissertation. Pam and many others in the Psychology and Language Science groups made day-to-day affairs run incredibly smoothly, including Carol Gorham, Tony Chan, Lori Kader, Enamul Haque, Joanne Leffson, Trish Bell, Merle Henry, and Csilla Kajtar.

Getting this far would not have been possible without the support of my wonderful friends within the Neuroscience and Cognitive Science (NACS), Psychology, and Language Science programs, especially Anna Chrabaszcz, Melissa Pangelinan, Kevin Donaldson, Matt Miller, Greg Cogan, Sarah Helfinstein, Amanda Chicoli, Julian Jenkins, Jen Merickel, Tracy Tomlinson, and Walky Goode.

Going back to my scientific roots, applying to graduate school would not have been a thought without the support of my undergraduate advisors at Rutgers University. I am thankful for my initial positive experiences with Arnold Glass and Carolyn Rovee-Collier, who guided me as I first wet my feet in the area of cognitive psychology. Importantly, much of my decision to pursue scientific research was due to the encouragement of my first mentor and psychology professor, Gary Brill, whose enthusiasm for cognitive science inspired me to pursue a similar career path. Thanks to all of these important figures for supporting me as I zeroed in on the decision to pursue a doctoral degree.

Lastly and most importantly, I am thankful for the love and encouragement of my family. I couldn't ask for better supporters in my parents, sister, brother, and grandparents; I'm delighted that achieving this degree has made you so proud! Thanks for always standing behind me and providing timely pick-me-ups. Finally, I'm entirely indebted to my partner in crime, Jared Finch, who stuck by my side through the years, making me laugh during the most stressful moments and smile in the meantime. I am lucky to have Jared and his loving family as an extension of my own. Overall, my family's support over the years has made my life whole, and this journey would not have been possible without them.

## Table of Contents

<b>List of Figures .....</b>	<b>x</b>
<b>List of Tables .....</b>	<b>xii</b>
<b>List of Abbreviations .....</b>	<b>xiv</b>
<b>Chapter 1: A Review of the Extant Literature .....</b>	<b>1</b>
<b>1.1 Executive Function Training and its Transfer Across Cognitive Domains.....</b>	<b>9</b>
<i>1.1.1 Near-transfer of Training.....</i>	<i>11</i>
<i>1.1.2 Far-transfer of Training .....</i>	<i>13</i>
<b>1.2 The Role of Executive Function in Language Use and the Implications for Training .....</b>	<b>16</b>
<i>1.2.1 Syntactic Ambiguity Resolution.....</i>	<i>17</i>
<i>1.2.2 Lexical Ambiguity Resolution .....</i>	<i>19</i>
<i>1.2.3 Reference Resolution .....</i>	<i>20</i>
<i>1.2.4 Verbal Fluency.....</i>	<i>24</i>
<b>1.3 General Discussion and Dissertation Aims.....</b>	<b>27</b>
<b>Chapter 2: Experiment 1 - Training Executive Functions for Sentence Processing .</b>	<b>31</b>
<b>2.1 Experimental Preliminaries .....</b>	<b>33</b>
<b>2.2 Hypotheses .....</b>	<b>35</b>
<b>2.3 Method .....</b>	<b>39</b>
<i>2.3.1 Subjects.....</i>	<i>39</i>
<i>2.3.2 Design .....</i>	<i>39</i>
<i>2.3.3 Training Tasks.....</i>	<i>41</i>
<i>2.3.4 Transfer Task: Syntactic Ambiguity Resolution.....</i>	<i>44</i>
<b>2.4 Analyses and Results.....</b>	<b>48</b>
<i>2.4.1 Sentence Comprehension Accuracy.....</i>	<i>48</i>
<i>2.4.2 Real-time Reanalysis (Eye Movements) .....</i>	<i>65</i>
<b>2.5 Discussion of Experiment 1 .....</b>	<b>75</b>
<i>2.5.1 Summary .....</i>	<i>75</i>
<i>2.5.2 Using a Process-Specific Training Approach .....</i>	<i>79</i>
<i>2.5.3 Limitations and Caveats .....</i>	<i>85</i>
<b>Chapter 3: Experiment 2 – Process-Specific Training for Parsing and Non-Parsing Skills.....</b>	<b>95</b>
<b>3.1 Experimental Preliminaries .....</b>	<b>100</b>
<b>3.2 Hypotheses .....</b>	<b>107</b>
<b>3.2 Method .....</b>	<b>110</b>
<i>3.2.1 Subjects.....</i>	<i>110</i>
<i>3.2.2 Design .....</i>	<i>111</i>
<i>3.2.3 Training Tasks.....</i>	<i>112</i>

3.2.4 Transfer Tasks .....	117
<b>3.3 Analyses and Results.....</b>	<b>127</b>
3.3.1 Training Task Performance.....	127
3.3.2 Index of Training Effects: Posttest N-back-with-Lures .....	130
3.3.3 General Analyses for Pre/Post Measures .....	134
3.3.4 Stroop Task .....	135
3.3.5 Recognition Memory Task .....	143
3.3.6 Verb Generation Task.....	148
3.3.7 Lingering Garden-Path Recovery (Comprehension Accuracy) .....	154
3.3.8 Real-time Reanalysis of Garden-Path Sentences .....	158
3.3.9 Real-time Reanalysis of Relative-Clause Sentences.....	173
<b>3.4 Discussion of Experiment 2 .....</b>	<b>179</b>
3.4.1 Summary .....	179
3.4.2 N-Back Strategies.....	186
<b>Chapter 4: General Discussion .....</b>	<b>190</b>
<b>4.1 Tying Together Experiments 1 and 2.....</b>	<b>190</b>
4.1.1. Comprehension Accuracy .....	190
4.1.2. Real-time Recovery Efforts .....	193
4.1.3 Contrasting Experiments 1 and 2 .....	195
4.1.4 Necessary and Sufficient Features for Cognitive Control Training .....	198
<b>4.2 Understanding Processes with Signal Detection Models.....</b>	<b>200</b>
4.2.1 Analysis and Results of Posttest N-Back Task.....	200
4.2.2 Analysis and Results of Recognition Memory Task.....	204
4.2.3 Considering the Contribution of Separate Processes for Transfer .....	206
4.2.4 A Summary of Possible Trained Mechanisms .....	208
4.2.5 Training as a Method to Improve or Capture Baseline Abilities? .....	210
<b>4.3 Future Directions .....</b>	<b>212</b>
4.3.1 Considerations for Future Process-Specific Training Studies .....	212
4.3.2 Applications for Other Populations .....	216
4.3.3 Applications for Bilingualism .....	218
4.3.4 Caveats .....	222
<b>4.4 Closing Remarks .....</b>	<b>224</b>
<b>Appendix A.....</b>	<b>226</b>
<b>Appendix B.....</b>	<b>232</b>
<b>Appendix C.....</b>	<b>233</b>
<b>Appendix D.....</b>	<b>235</b>
<b>References.....</b>	<b>239</b>

## List of Figures

Figure 1. Longitudinal design of Experiment 1 .....	40
Figure 2. <i>N</i> -back performance curves across training sessions for responders and non-responders .....	56
Figure 3. Change from pretest to posttest in comprehension accuracy rates split by trainign group (untrained controls, <i>n</i> -back non-responders, and <i>n</i> -back responders) .....	57
Figure 4. Responsiveness on the three in-house ‘control’ training tasks for <i>n</i> -back responders .....	63
Figure 5. Trainees' and untrained controls' regression-path times across assessments launched from each sentence region for ambiguous and unambiguous items.....	69
Figure 6. Longitudinal design of Experiment 2.....	112
Figure 7. Sample trials of the global and local blocks of the recognition task.....	121
Figure 8. Training performance (raw and normalized) over the course of 16 training sessions for each training group (Lures, No-Lures, and 3-Back).....	128
Figure 9. Accuracy by each item type (targets, lures, and fillers) on the posttest <i>n</i> -back task for training group .....	132
Figure 10. Cross-assessment response times on the Stroop task, split by high- (Stroop Cost) and low-conflict (Stroop Benefit) conditions for each training group.....	138
Figure 11. Cross-assessment response times on the recognition memory task, split by probe-type (targets, fillers, and lures) on high- (local block) and low-conflict (global block) conditions for each training group.....	145
Figure 12. Cross-assessment verb generation latencies for nouns in terms of competition (high- and low-conflict) and retrieval demands (high- and low-association) for each training group .....	152
Figure 13. Cross-assessment accuracy to comprehension questions probing for lingering effects of misinterpretation of garden-path sentences split by high- (ambiguous sentences) and low-conflict (unambiguous sentences) for each training group.....	158
Figure 14. Cross-assessment regression-path time for each of four regions of garden-path sentences split by high- (ambiguous sentences) and low-conflict (unambiguous sentences) for each training group.....	165
Figure 15. Cross-assessment residual second-pass time for each of the four regions of garden-path sentences split by high- (ambiguous sentences) and low-conflict (unambiguous sentences) for each training group .....	172

Figure 16. Cross-assessment residual second-pass time for the four regions of relative-clause sentences split by high- (object-extracted) and low-complexity (subject-extracted) conditions for each training group.....	178
Figure 17. Signal detection measures indexing discriminability ( $d'$ ) and response criterion ( $\beta$ ) on each block—3-back and 6-back—of the posttest $n$ -back task for each training group .....	202
Figure 18. Cross-assessment signal detection measures ( $d'$ and $\beta$ ) for performance on the high-conflict local block of the recognition memory task for each training group .....	205

## List of Tables

Table 1. Explanations of the 8 training tasks used in Experiment 1.....	45
Table 2. Performance measures of responders and non-responders across the four training tasks .....	54
Table 3. Significant fixed effects from the best fitting mixed-effects models of comprehension accuracy data, testing for an Assessment (pretest vs. posttest) by Group (responders vs. non-responders vs. untrained controls) interaction separately for ambiguous and unambiguous items on each of the four training tasks .....	59
Table 4. Four sentence regions of reflexive absolute transitive (garden-path) sentences specified for fine-grain analysis .....	66
Table 5. Significant fixed effects from the best fitting mixed-effects models of regression-path time following entry into the final region of each sentence testing for an Assessment (pretest vs. posttest) by Group (task responders vs. non-responders vs. untrained controls) interaction separately for ambiguous and unambiguous items for each of the four in-house training tasks .....	72
Table 6. Significant fixed effects from the best fitting mixed-effects models of each regression-path time component (first-pass time, re-reading earlier regions, and second-pass time) testing for an Assessment (pretest vs. posttest) by Group (task responders vs. non-responders vs. untrained controls) interaction separately for ambiguous and unambiguous items .....	74
Table 7. Explanations of the 3 versions of <i>n</i> -back used in Experiment 2 .....	114
Table 8. Summary of the analyses of covariance (ANCOVAs) testing for the effects of training group (Lures vs. No-Lures vs. 3-Back) while controlling for pretest performance on posttest performance for the high- and low-conflict conditions of the four assessment tasks (Stroop, recognition memory, verb generation, and garden-path recovery) of Experiment 2.....	140
Table 9. Significant fixed effects from the best fitting mixed-effects models of comprehension accuracy data, testing for an Assessment (pretest vs. posttest) by Group (Lures vs. No-Lures vs. 3-Back) interaction separately for ambiguous and unambiguous items on each of the four training tasks .....	156
Table 10. Significant fixed effects from the best fitting mixed-effects models of regression-path time following entry into the final region only for garden-path materials testing for an Assessment (pretest vs. posttest) by Group (Lures vs. No-Lures vs. 3-Back) interaction separately for ambiguous and unambiguous items on each of the four training tasks .....	161
Table 11. Significant fixed effects from the best fitting mixed-effects models of residual re-reading time for garden-path materials testing for an Assessment (pretest vs. posttest)	

by Group (Lures vs. No-Lures vs. 3-Back) interaction separately for ambiguous and unambiguous items on each of the four training tasks .....	170
Table 12. Four sentence regions of relative-clause sentences specified for fine-grain analysis .....	176
Table 13. Significant fixed effects from the best fitting mixed-effects models of residual re-reading time for relative-clause materials testing for an Assessment (pretest vs. posttest) by Group (Lures vs. No-Lures vs. 3-Back) interaction separately for subject-extracted and object-extracted items on each of the four training tasks .....	177
Table 14. Summary of the training groups demonstrating reliable cross-assessment comparisons for the various conditions (high- and low-conflict; high- and low-difficulty) of the assessment tasks (Stroop, recognition memory, verb generation, parsing garden-path sentences, parsing relative-clause sentences) of Experiment 2 .....	182



## **List of Abbreviations**

AIC<sub>C</sub>: Corrected Akaike information criteria

ACC: Anterior Cingulate Cortex

ANOVA: Analysis of Variance

ANCOVA: Analysis of Covariance

BF (JZS BF): Bayes Factor (Jeffrey-Zellner-Siow)

EF: Executive Function

LIFG: Left Inferior Frontal Gyrus

LNS: Letter- Number Sequencing

OE: Object-Extracted (relative clause)

MCMC: Marcov chain Monte Carlo

SE: Subject-Extracted (relative clause)

SEM: Standard Error of the Mean

VLPFC: Ventrolateral Prefrontal Cortex

WM: Working Memory

## Chapter 1: A Review of the Extant Literature<sup>1</sup>

Cognitive control, also called executive function (EF), refers to a cluster of mental processes that permit the flexible adjustment of thoughts and actions across domains, allowing individuals to adapt to new rules and guide the selection of task-relevant over task-irrelevant information in an environment that varies continuously (Miller & Cohen, 2001). As we navigate our surroundings, we can frequently rely on a set of highly regularized functions that render certain tasks like driving a car or skimming a magazine article relatively automatic. Sometimes, however, new instructions or conflicting information compels us to override these reflexive actions and instead consider what might otherwise be a disfavored (or atypical) response. For instance, a resident of Chicago may be in the habit of making a legal right turn on red when driving at home, but this routine behavior could result in a costly ticket when she visits New York City, where turning on red is strictly prohibited! Likewise, imagine reading the following sentence upon skimming the magazine: *At the restaurant, the interns discussed the bill before suggesting edits to the senator.* One might initially interpret the word “bill” to mean the list of charges incurred for the meal, rather than its intended (though less common) interpretation, namely a draft piece of legislation. On the surface, both examples are quite different, but conceivably induce a similar experience: the detection of an incompatibility and the ensuing need to rein-in a highly familiar, yet currently inappropriate cognitive reaction (e.g., refrain from turning; revise the more frequent

---

<sup>1</sup> This section is a modified version of: Hussey, E.K., & Novick, J.M. (2012). The benefits of executive control training and the implications for language processing. *Frontiers in Cognition*, 3(158). doi: 10.3389/fpsyg.2012.00158

meaning, but current misanalysis, of “bill”). Such ‘interference resolution’ functions are an essential part of cognitive control (Botvinick, Braver, Barch, Carter, & Cohen, 2001) and help adapt information-processing strategies so individuals can regulate behavior in view of ever-changing goals, new contexts, or situation-specific demands.

As many researchers have argued, executive functions encompass a *collection* of cognitive processes that help guide goal-directed behavior; that is, cognitive control is not a unitary construct but comprises separable components (Botvinick et al., 2001; Miller & Cohen, 2001; Norman & Shallice, 1986). In addition to the interference-resolution processes outlined above, other EFs include task-switching, updating, and information monitoring, each of which can operate over visual, spatial, or verbal domains (Friedman & Miyake, 2004; Miyake et al., 2000; Smith & Jonides, 1999) and thus may be recruited across a variety of tasks including selective attention, decision-making, working memory (WM), error monitoring, and language processing (Badre & Wagner, 2007; Botvinick et al., 2001; Thompson-Schill, Bedny, & Goldberg, 2005; *inter alia*). With regard to interference-resolution functions in particular, converging data from neuropsychological patients and brain-imaging studies of healthy adults suggest that, across a range of WM, attention, and language tasks, posterior regions of left ventrolateral prefrontal cortex (VLPFC) commonly support the ability to resolve among competing sources of evidence, regardless of domain (Thompson-Schill et al., 2005).

In the first chapter of this dissertation, I discuss how a burgeoning literature demonstrates that EFs can be trained through ample practice—that such abilities are seemingly not fixed, but malleable—and that performance increases throughout the course of training generalize to novel tasks that were not part of the training protocol.

Some examples of transfer include benefits on unpracticed tasks tapping fluid intelligence (Jaeggi, Buschkuhl, Jonides, & Perrig, 2008), working-memory updating (Dahlin, Neely, Larsson, Bäckman, & Nyberg, 2008; Li et al., 2008), and task-switching (Korbach & Kray, 2009)—that is to say, transfer benefits have been observed across a *range* of executive functions.

I am especially interested in the implications that these training-transfer findings have for language processing under conditions of conflict, given that domain-general interference-resolution and cognitive-control functions have been associated with assorted linguistic abilities including the resolution of lexical (Bilenko, Grindrod, Myers, & Blumstein, 2009; Copland, Sefe, Ashley, Hudson, & Chenery, 2009; Khanna & Boland, 2009; Vuang & Martin, 2011) and syntactic ambiguities (Novick, Trueswell, & Thompson-Schill, 2005; Trueswell, Sekerina, Hill, & Logrip, 1999; Ye & Zhou, 2009), verbal fluency (Kan & Thompson-Schill, 2004; Novick, Kan, Trueswell, & Thompson-Schill, 2009; Robinson, Blair, & Cipolotti, 1998; Schnur et al., 2009), and perspective-taking during natural dialogue (Brown-Schmidt, 2009; Nilsen & Graham, 2009; for reviews, see Novick et al., 2005; Novick, Trueswell, & Thompson-Schill, 2010). For each language ability detailed below, I couch my hypotheses within a *process-specific* account (see Dahlin et al., 2008; Shipstead, Redick, & Engle, 2010; 2012), which in the training literature posits that post-intervention, performance increases on novel tasks largely depend on the extent of overlap between the training and transfer measures, both in terms of the shared cognitive processes and underlying neural systems needed to complete them. That is, if a certain component of EF (e.g., interference resolution) is targeted and improved through training, then transfer measures relying on common

processes should be influenced accordingly, irrespective of domain or modality. In view of this, I will focus my initial discussion on a few language comprehension and production tasks that fit within the VLPFC-mediated process-specific function typically referred to as ‘interference resolution’ (however, I acknowledge other brain systems involved in a wider array of EFs, and consider the implications of this for training and the effects on language in the final chapter). The training studies discussed in Chapters 2 and 3 implement a subset of these language tasks as untrained assessment measures.

As sketched in the driving and reading examples earlier, conflict (or interference) here will refer to conditions that contain the presence of mismatched or incongruent sources of evidence. Specifically, ‘conflict’ designates cases in which current situation-specific demands generate an incompatibility between how an input stimulus should be characterized (dubbed *representational conflict*), given how the input is normally considered. Such interference is often called ‘prepotent conflict,’ because individuals must override their dominant (prepotent) biases in support of atypical alternatives (Botvinick et al., 2001). For instance, the Stroop task is a canonical representational conflict task involving the need to countermand a prepotent bias that is generated by a lexical representation (which gives rise to an automatic reading response), in favor of a perceptual (color) representation. A comparable type of representational conflict occurs in the form of ‘underdetermined conflict,’ in which multiple candidate representations are equally reasonable and thus compete for selection (Botvinick et al., 2001). Importantly, brain-imaging findings suggest separable neuroanatomical involvement for representational conflict versus *response conflict* (or *response selection*; see Milham et al., 2001; Milham, Banich, & Barad, 2003; Nelson, Reuter-Lorenz, Sylvester, Jonides, &

Smith, 2003). My major focus here is on the implications of interference-resolution training at the representational level on particular language-performance measures such as lexical and syntactic ambiguity resolution (in comprehension) and verbal fluency (in production). Both prepotent and underdetermined representational conflicts recruit posterior regions of VLPFC (Brodmann areas 44 and 45) across language and memory domains, meeting the requirements for a test of process-specificity (see Novick et al., 2010; see also Milham et al., 2003 and Nelson et al., 2003, which demonstrate VLPFC recruitment for representational conflict resolution but anterior cingulate recruitment for response-level conflict resolution).

Generally, I believe that—considering the mounting evidence showing the effectiveness of various types of EF training in different populations (Jaeggi, Buschkuhl, Jonides, & Shah, 2011; Klingberg et al., 2005; Westerberg et al., 2007)—there is room to establish new research investigating if EF training protocols that focus on selective sub-processes (i.e., conflict resolution) could be used successfully as an intervention technique to mitigate problems in general language use that arise under high-EF (i.e., high-interference) demands.

Indeed, there is tantalizing evidence supporting process-specific transfer to conflict/interference-related language measures, drawn not from a long-term training paradigm per se, but rather from another type of intervention designed to *fatigue* selective cognitive processes common to WM and language processing tasks. These so-called ‘resource depletion models’ offer an interesting framework to understand *negative* transfer to tasks relying on temporarily exhausted EFs shared across ostensibly different domains (Persson, Welsh, Jonides, & Reuter-Lorenz, 2007; Van der Linden, Frese, &

Meijman, 2003). That is, rather than boosting general-purpose EFs through long-term practice, as is the case with training studies, resource depletion paradigms rely on short-term “overuse” of a particular cognitive process. For example, after performing a complex task that places high demands on EF capacities, these resources are rendered temporarily unavailable for continued use; therefore, performance *decreases* on transfer measures that rely on the common “worn out” EF (Persson et al., 2007; Van der Linden et al., 2003; see also Synder et al., 2011 for similar findings among anxious individuals).

In one study (Persson et al., 2007), interference-resolution abilities were fatigued through an intensive session of an item-recognition task with high interference-resolution demands. In this task, participants indicated whether a probe item (e.g., *C*) appeared in an immediately prior memory set (e.g., *r, f, c, l*) (see Monsell, 1978). Frequently, subjects could respond correctly due to familiarity alone: familiar probes required a ‘yes’ response and unfamiliar ones a ‘no’ response. However, relying on familiarity on some ‘no’ trials was prone to error, because they contained a probe (e.g., *G*) that was not among the current memory set (*j, p, v, m*) but *was* among the items in the *prior* trial (*g, k, v, p*). Thus, these trials required subjects to override a prepotent familiarity bias (and ‘yes’ response) and instead re-characterize the probe stimulus as ‘familiar-but-irrelevant,’ and respond ‘no.’ Such ‘recent-no’ trial types, when compared to ‘non-recent-no’ trials (when the probe did not appear in either the current or preceding sets) routinely recruit left posterior VLPFC (Jonides & Nee, 2006). Important for the current discussion, after subjects completed this task and “fatigued” the interference-resolution process, they subsequently demonstrated selective performance declines on VLPFC-mediated, high-EF conditions on a verbal fluency task, in which they had to generate an associated verb to a

given noun (e.g., *scissors* → *cut*; high-EF items had many possible associated verbs, like *ball* → *kick*, *throw*, *catch*, *bounce*, and thus contained underdetermined response conflict; see Thompson-Schill et al., 1998). This pattern of negative transfer was not observed for (1) subjects who received exposure to only low-EF trials during their intensive practice session (i.e., no recent-no trials); or (2) individuals who practiced a different task before the verb generation task, namely a stop-signal task that recruits mainly right-hemisphere networks and a different subcomponent of EF (response inhibition; see also Friedman & Miyake, 2004). Together, this suggests that the process-specificity observed across intervention and transfer tasks operates on a short time scale, such that as interference resolution is temporarily depleted, other tasks relying on shared cognitive and neural resources are affected accordingly.

Although these effects are transient, the selective transfer findings are nonetheless critical: they demonstrate that interference resolution abilities are at least temporarily malleable, and this malleability can subsequently affect language processing under similar conditions of high interference resolution demands. Consequently, throughout this dissertation, I ask: considering evidence for process-specific transfer, on a short time scale, across memory and language tasks that commonly rely on VLPFC-mediated interference-resolution functions, might one observe longer-term effects on language measures as well, when interference resolution is boosted via extensive practice? That is, is there *positive* transfer—namely, performance *increases*—when individuals consistently train interference-resolution functions over time? I hypothesize that the answer should be yes, given the evidence that other executive functions (e.g., task-switching, etc.) are both trainable and transferrable.



Although I outline below some potential benefits of interference resolution training on language use, I also discuss some caveats that should be considered, including individual differences in training success (not everyone responds similarly to training or reaps the same profile of benefits, cf. Chein & Morrison, 2010; Jaeggi et al., 2011), limitations that may be involved in training special populations, and the need for explicit linking hypotheses between training and any expected transfer: namely, there must be a theory that bridges the hypothesized underlying cognitive processes from one task to another (i.e., from an intervention task to a transfer task). Transfer from training to untrained assessment tasks cannot be expected, or explained, without a well-formulated process-specific theory (Shipstead et al., 2010; 2012). To this end, I also speculate that the magnitude of transfer effects is contingent upon the *degree* to which a targeted EF contributes to and shares critical features with an outcome measure. This is particularly important if, as some researchers suggest, EF is not a unitary construct but is comprised of separable, multi-component processes such as interference resolution, updating, and task-switching (Dahlin et al., 2008; Miyake et al., 2000; Persson et al., 2007).

Throughout this first chapter, I integrate the extant training and psycholinguistic literatures to develop testable hypotheses from an emerging picture within EF training research. The following section begins with a brief review of cognitive training studies demonstrating transfer to novel tasks that are ostensibly different from those practiced during the training regimens, but share specific processing demands. I then turn to research on the role of interference resolution in language use, sketching some hypotheses and implications the training findings have for new work aimed at improving language processing under high EF—particularly high interference-resolution—demands.

That is, if interference-resolution is malleable (which seems to be the case given the resource depletion work outlined above), I hypothesize that training such processes should also show transfer to untrained measures of interference resolution within the linguistic domain, patterning with other training-transfer findings. The theory bolstering this claim comes from work (drawn from patients, children, and brain imaging studies of adults) indicating that interference-resolution and cognitive-control measures play an important role in language tasks that I outline below.

### **1.1 Executive Function Training and its Transfer Across Cognitive Domains**

A recent flurry of research is devoted to testing if general-purpose cognitive abilities can be enhanced through consistent practice with WM tasks that recruit brain regions within the cortico-striatal network key to executive functioning. Although interventions geared toward improving psychological faculties, specifically intelligence, were pioneered decades ago (see Feuerstein, 1980; Sternberg, Ketron, & Powell, 1982), Klingberg and colleagues have recently reinstated the notion by training domain-general cognitive abilities as a means to remediate populations with diminished WM resources including stroke patients (Westerberg et al., 2007), children with attention deficit hyperactivity disorder (Klingberg et al., 2005), and older adults (Brehmer et al., 2011). Ever since, cognitive training programs have undergone significant study, particularly in healthy adults, to examine whether normally-functioning individuals' EF abilities can be improved, and what generalized outcomes consistent training might have on everyday performance on non-trained tasks. To this end, researchers have been investigating questions related to dosage-dependence (does more practice yield more transfer?; Jaeggi et al., 2008), the extent to which training transfers to untrained but related measures

(Chein & Morrison, 2010; Karbach & Kray, 2009; Li et al., 2008; Morrison & Chein, 2011), if training tasks must adapt to individuals' performance to be effective (Brehmer et al., 2011; Klingberg et al., 2005), and individual differences in training success (Jaeggi et al., 2011).

Here, I focus on the extent to which training generalizes to novel tasks. The typical training study is designed as a pre/post longitudinal experiment in which subjects are assessed on some cognitive capacity immediately before and again after an extensive intervention. In some cases, the intervention comprises practice with a single training task (Dahlin et al., 2008; Jaeggi et al., 2008; 2011; Li et al., 2008), whereas in others, a battery of training tasks is administered (Karbach & Kray, 2009; Klingberg et al., 2002; 2005). Regardless, the training tasks are different from those completed at the pre/post assessment sessions, with the intervention component typically lasting for several hours distributed over a few weeks. Upon conclusion of the regimen, trainees return to the lab and complete follow-up assessments, namely complementary versions of the tasks that were done just prior to training, to evaluate whether performance on assessments has reliably improved, thereby providing evidence for “transfer.”

Transfer has been documented for untrained tasks that share obvious features with well-practiced training tasks, an effect sometimes referred to as “near-transfer.” For instance, performance increases on WM training tasks generalize to structurally similar (but new) WM assessments (Karbach & Kray, 2009; Li et al., 2008; see below). However, “far transfer” can also be observed, namely to assessments that appear, on the surface, wildly different from the training tasks completed throughout the intervention regimen (Dahlin et al., 2008; Jaeggi et al., 2008; 2011; Kloo & Perner, 2003). This latter

form of transfer is possible provided that training and assessment tasks share certain essential underlying EFs (as well as overlapping neural resources; see Jonides, 2004; Sayala, Sala, & Courtney, 2006; Shipstead et al., 2010; 2012).

### *1.1.1 Near-transfer of Training*

Near-transfer effects emerge when the nature of the processed information—including stimulus type, task structure, and response type—is similar across training and assessment tasks (but see Morrison & Chein, 2011 for an alternative definition of near-transfer). For instance, in one report (Li et al., 2008), trainees practiced a spatial 2-back task, during which they had to monitor the locations of sequentially-presented squares on a 3x3 grid and respond whenever the current location matched the location seen two trials earlier. Compared to a no-contact control group, trained participants demonstrated post-intervention improvements on a spatial 3-back task, providing evidence for near-transfer to a more difficult, but otherwise identical task. Another type of near-transfer occurs when the *type* of information (i.e., the stimuli) being processed is changed across training and transfer tasks, while the response-level requirements remain constant, resulting in a structural continuity between both tasks. For example, in the same study by Li and colleagues (2008), trainees also improved on *numeric* 2- and 3-back tasks, where instead of remembering locations on a grid, subjects indicated when a serially-presented number (0-9) matched the identity of a number presented two (or three) trials previously. The authors argued that transfer to a numeric *n*-back task provided support for a task-specific response strategy shared across stimulus modalities: Although the spatial 3-back and numeric *n*-back tasks differ from the spatial 2-back training task, all require the same

basic strategy, namely, information must be monitored and updated in a predictable fashion.

In addition to the above findings, Karbach and Kray (2009) observed that increases in task-switching abilities—an EF based on mental shifting across different goals or rules—as a consequence of training generalizes to performance on novel tasks with similar switching demands. Specifically, their training regimen involved making two-alternative forced-choice judgments about pictures (trees/flowers), based on two separate characteristics (e.g., identity vs. color), such that the relevant characteristic (or rule) changed predictably across trials. Stimulus types (fish/birds, trees/flowers, sports/music, planes/cars) and response categories (identity, number, color, and rotation) varied across sessions within the training regimen. An assessment of near-transfer involved responding to a novel set of stimuli (fruits/vegetables) using number and identity as response categories; compared to a non-switching active-control group, the task-switching trainees showed greater posttest improvement in switching costs, i.e. the difference in response time on switch (color followed by identity judgment) versus non-switch trials.

These examples highlight two sources of near-transfer: training and outcome measures tap the same underlying EFs (e.g., monitoring and updating), and both tasks provoke similar processing demands through a shared task structure (task-specific aspects). Consequently, it is difficult to disentangle the source of near-transfer effects, as two possibilities may account for any observed pre/post changes: (1) the trained EF shared by both tasks may have been improved, or (2) a task-specific strategy may have been developed. Indeed, in cases of near-transfer, the training and transfer tasks need not

tap the same underlying EFs, since transfer could occur simply with improvements at task-specific aspects of the paradigm. Near-transfer effects might be unsurprising: practicing an  $n$ -back task improves  $n$ -back performance, and therefore transfers to other  $n$ -back tasks (perhaps regardless of domain); likewise, practicing a categorization task-switching task generalizes to a similar task with novel categories. But, the extent to which these near-transfer effects are driven by the shared EFs across training and assessment tasks, the surface-level features (stimulus or response characteristics) that are isomorphic between both sets of tasks, or through a combination of both factors is unknown.

### *1.1.2 Far-transfer of Training*

Training studies designed to show far-transfer effects help to elucidate the role of shared EFs; by design, the surface-level properties—stimuli or required responses—of the training and assessment tasks are quite different. Consequently, contrary to near-transfer findings, far-transfer effects are assumed *not* to rely heavily on the structural (task-specific) similarities across training and assessment tasks, and instead result mostly from improvements on underlying EFs important to both the training and assessment measures (Shipstead et al., 2010). In other words, the goal of far-transfer training is rooted in improvement of specific processes engaged during tasks with dissimilar structures, often spanning domains (again, sometimes referred to as *process-specific training*).

For instance, in one set of studies, subjects practiced a dual  $n$ -back memory task involving simultaneous updating of shape locations and the identity of heard letters, such that a target was defined as an item repeating  $n$ -trials previously in either modality (Jaeggi et al., 2008). Trainees showed subsequent improvements on Raven's Advanced

Progressive Matrices, a transfer task that requires participants to select a textured shape from a set of possible response items, which fits a sequence of other textured shapes to complete a particular pattern with one absent piece (Jaeggi et al., 2008; 2011). The response and surface-level properties of *n*-back and Raven's are distinct, as one task involves monitoring a continuous stream of letters or block locations for familiar instances, and the other requires reasoning to identify the missing element that completes a 4x4 matrix containing orderly patterns across rows and columns; thus, to observe transfer, there must be an underlying process common to both tasks that is enhanced through intensive *n*-back training. The authors reasoned that this shared process centered around a common need to employ attentional control, such that their training procedure—which forced trainees to practice constant shifting of attention to new stimuli—facilitated this ability, thereby enabling transfer to Raven's, which similarly involves updating and selection among multiple representations (via the control of attention). Importantly, because the training and transfer measures were characteristically so different, the authors argued that task-specific elements could not explain the observed generalization, effectively ruling out near-transfer as an explanation for their findings. Rather, training boosted a part of the EF system—here, multiple-task management and attentional control processes—important for a range of cognitive tasks, including Raven's performance. Indeed, separate work demonstrates that *n*-back and Raven's activate a similar network of neural regions, providing additional support for resources common to both tasks (Burgess, Gray, Conway, & Braver, 2011).

Additional evidence of process-specific training comes from demonstrations of selective far-transfer from an updating task (letter running-span) to a structurally different

assessment measure (number *n*-back) that requires a similar updating EF; critically though, such transfer was not demonstrated on the Stroop task, which relies on a separable EF—interference resolution (Dahlin et al., 2008). During the letter running-span task, participants must recall the last four items of a study list that terminates unexpectedly, forcing them to continuously update the correct response set from a fleeting memory store; similarly, their version of *n*-back required subjects to monitor and refresh representations as new information is processed and deemed relevant. Running-span and a standard number *n*-back task recruit similar striatal regions, corroborating their underlying reliance on a common EF. Contrastingly, tasks requiring interference resolution, like Stroop, require subjects to re-characterize an automatized response (reading) in order to promote atypical, but task-relevant information (color name); such tasks rely on a separable neural profile (compared to that required for updating tasks) including a network of frontal and parietal regions. Dahlin and colleagues (2008) demonstrated that training on running-span confers benefits to assessment measures that share updating demands and corresponding neurological profiles (*n*-back), while those with little or no such overlap (Stroop) show negligible improvement. In sum, the amount of far-transfer to untrained tasks following intervention depends on the degree of overlap among cognitive and neural resources shared by the training and the transfer tasks.

Given these training and far-transfer effects for a range of EFs (e.g., attention control, memory updating), one might also hypothesize that transfer from general-purpose EF training to certain tasks of language processing might occur as well. That is, the language tasks are not trained per se, but tap particular cognitive functions (interference resolution) that may be trainable through an extensive regimen targeting



common processes (or neural resources). As hypothesized below, the result could be an alleviation of language processing difficulty under conditions that place heavy demands on the EF system in healthy, and perhaps even in special populations. I focus on a select few of these language conditions in the following section, concentrating specifically on a functional-anatomical association between interference-resolution processes of EF, and regions within left VLPFC that support them (for an extensive review, see Novick et al., 2010). I sketch how this association is important for production and comprehension abilities in healthy adults, young children, and patients with circumscribed VLPFC damage.

## **1.2 The Role of Executive Function in Language Use and the Implications for Training**

One priority in psycholinguistics has been to study how non-linguistic cognitive abilities contribute to language production and comprehension. EF abilities have emerged as a candidate characteristic, defining in part those individuals who can better coordinate rapidly among multiple sources of linguistic (syntactic, semantic) and extra-linguistic (pragmatic, contextual) evidence across a range of communicative tasks. Given the breadth of work on various EFs for language, I focus only on the role of interference-resolution training for a handful of language tasks. As sketched above, interference resolution refers to the re-characterization of information in the face of competing sources of evidence. Regarding language processing, good interference-resolution skills enable readers and listeners to avoid comprehension errors in the face of ambiguity (e.g., by consulting top-down evidence to override misinterpretations), produce the right word among competing options, and take an interlocutor's perspective when assessing

common-ground information during natural, unscripted dialogue (see Novick et al., 2005; 2010). Indeed, patients with circumscribed damage to left posterior VLPFC consistently underperform on high-interference conditions on non-linguistic tasks such as Stroop and the ‘recent-no’ task described above (Hamilton & Martin, 2005). Moreover, this general interference-resolution disorder in patients has been tied to their concomitant deficits on language tasks that generate similar EF demands, for example, when dominant meanings of lexical ambiguities must be countermanded (Bedny, Hulbert, & Thompson-Schill, 2007), when initial interpretations of syntactic ambiguities must be reprocessed (Novick et al., 2005; 2009), or when object names must be selected among categorical competitors (Schnur et al., 2009). As such, by training general-purpose interference-resolution abilities—supported by regions within VLPFC—in healthy adults, I hypothesize that there should be systematic improvements in high-EF conditions on language tasks requiring shared demands for interference resolution. Below, I provide examples of when interference-resolution abilities appear to interact with particular language processing skills and outline the implications these associations have for process-specific training in extreme populations as well.

### *1.2.1 Syntactic Ambiguity Resolution*

Readers and listeners process sentences in real-time, committing to an interpretation incrementally as words and phrases are encountered moment-by-moment (Altmann & Kamide, 1999; Tanenhaus, 2007). One consequence of incremental processing is temporary ambiguity: the first analysis individuals assign sometimes turns out wrong. Cognitive control has been tied to individuals’ ability to adjust interpretations when late-arriving evidence signals that their initial analysis was incorrect (Novick et al.,

2005). Such cases of interference/conflict (the so-called “garden-path effect”) elicit temporary processing difficulty in reading (Frazier & Rayner, 1982; Staub & Rayner, 2007; *inter alia*) and confusion during spoken comprehension (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Individuals must then engage in a process that permits them to revise and capture the intended interpretation.

Evidence for the role of interference-resolution in this recovery process comes from populations with underdeveloped or impaired cognitive control such as young children (whose PFC development is protracted; see Huttenlocher & Dabholkar, 1997) and patients with focal damage to left posterior VLPFC. Both populations fail to initiate cognitive-control functions across assorted non-syntactic measures (e.g., Stroop, the recent-no, and other analogous tasks; e.g., Khanna & Boland, 2009; Hamilton & Martin, 2005), and both groups similarly fail to revise sentence interpretations following early misanalysis (Novick et al., 2009; Trueswell et al., 1999; Weighall, 2008; see also Christianson, Williams, Zacks, & Ferreira, 2006 for similar patterns in older adults). The linking assumption is that the discovery of a misinterpretation deploys interference-resolution to resolve the incompatibility between representations of sentence meaning: the one initially assigned and the one in need of recovery, similar to the controlled processes required to resolve interference during incongruent Stroop trials, or interference from familiar but currently irrelevant items in the ‘recent-no’ task (Hamilton & Martin, 2005; Novick et al., 2005; 2010). Interestingly, healthy adults undergoing functional neuroimaging demonstrate co-localized neural activity within left posterior VLPFC when performing both syntactic and non-syntactic tasks requiring interference

resolution, corroborating the necessary involvement of shared, domain-general processes presumed from special populations (January et al., 2009; Ye & Zhou, 2009).

This convergence of findings suggests an opportunity to alleviate the processing difficulty associated with temporary ambiguities that arise during sentence processing by targeting the EFs (through training) that appear to be domain-general, i.e. common across certain syntactic and non-syntactic tasks. Indeed, I tested this hypothesis in the studies reported in Chapters 2 and 3 of this dissertation.

### *1.2.2 Lexical Ambiguity Resolution*

Research examining comprehension at the single-word level suggests a role for interference resolution when the dominant meaning of an ambiguous word (e.g., *bill*, as the tab issued by a restaurant) must be overridden to retrieve its subordinate meaning (an outline of a prospective law) (Bedny et al., 2007). Questions posed in this literature examine whether good conflict-resolution skills enable context-dependent meaning selection, and conversely, whether poor abilities impair it. Researchers have found that better conflict resolution is related to young children's contextual sensitivity: context *can* be used by kids to countermand dominant, but inappropriate meanings of an ambiguous word; however, the use of top-down information is largely dependent on the maturity of their EF abilities, as indexed by a separate task of conflict resolution and inhibitory control (Khanna & Boland, 2010). Correspondingly, neuropsychological patients with poor conflict resolution show inadequate lexical ambiguity resolution when the subordinate meaning is activated by local contextual information (Balota & Faust, 2001; Bedny et al., 2007), suggesting that such patients have difficulty suppressing context-inappropriate meanings of ambiguous words (Copland et al., 2009; Grindrod & Baum,

2003; Vuong & Martin, 2011). Finally, across several studies, regions within VLPFC—the same areas involved in lesion-deficit analyses of patients showing interference-resolution impairments—are active in healthy adults during lexical-decision tasks necessitating resolution of meaning competition, suggesting that VLPFC-mediated EFs trigger to resolve increased competition associated with accessing the less frequent meaning of an ambiguous word (Bilenko et al., 2009).

Lexical ambiguity resolution abilities may also be enhanced, hypothetically, through interference-resolution training tasks designed to target EFs central to overriding dominant biases and implementing cognitive control (provided the effects are large enough to observe improvement; this may be particularly true in clinical patients). Future research might test whether EF training, with the right tasks, could garner improvements in integration among top-down contextual and lexical sources of evidence, particularly when these latter sources give rise to multiple conflicting meanings. There are obvious implications for clinical patients with word-comprehension deficits stemming from poor interference-resolution abilities.

### *1.2.3 Reference Resolution*

When conversational participants interact, they establish what is known as ‘common ground,’ or shared beliefs. Brown-Schmidt (2009) has demonstrated that variations in cognitive control abilities can explain healthy individuals’ occasional inattentiveness to common ground information; that is, objects visually accessible only to the listener are occasionally (incorrectly) favored as a referential interpretation over objects accessible to both partners. Specifically, individual differences in interference resolution may determine if a listener can successfully override perspective-inappropriate

interpretations of referential ambiguities uttered by their partner. As such, interference resolution may predict how easily semantic and pragmatic information is integrated in order to rule out incorrect interpretations during natural dialogue.

Indeed, a study testing young children corroborates this account by showing that although 5-year-olds can distinguish common versus privileged knowledge during conversation, the preference for their own perspectives—assessed by gaze duration to inappropriate privileged-ground alternatives—is predicted by measures of interference resolution and inhibitory control (Stroop, a tapping task, and the *bear/dragon* puppet task), all of which require resolving among conflicting representations by overriding a dominant rule/bias (Nilsen & Graham, 2009). That is, children with poorer cognitive control demonstrated exaggerated looking times to high-EF referential alternatives inaccessible to the speaker but hidden (or “privileged”) so that only the listener (the child) can see them (e.g., a small duck when “*Look at the duck*” is uttered and competes with the target that is common knowledge, i.e., a large duck). Namely, children with better task performance on high-EF conditions were more likely to override their egocentric view and modify their behavior to be consistent with information shared by both communicative parties, and did so *selectively* for high-EF items evidenced by spending less time gazing at inappropriate privileged-ground alternatives.

Adults occasionally show similar consideration of perspective-inappropriate interpretations when a speaker utters a referential ambiguity, failing to be sensitive to common-ground information immediately. This behavior is also related to individual variation in interference-resolution abilities. For instance, during one ‘visual-world’ task (Brown-Schmidt, 2009), participants assisted the experimenter in revealing the identity of

subject-privileged pictures on a display by answering the experimenter's questions. Generally, addressees consulted common-ground information to resolve temporarily ambiguous request, like, *What's above the horse with the glasses?*, when two horses might be referenced, one wearing glasses and another wearing shoes. If the item above one of the horses (the horse with shoes) was previously grounded, then subjects directed their gaze toward the unmentioned target and the horse (with glasses) located below it, as the ambiguity unfolded. Crucially, however, the better an addressee was able to use perspective information to avoid considering inappropriate interpretations (i.e., understanding the question to mean the already-revealed object) was determined by his Stroop performance. That is, subjects with better cognitive control were quicker to resolve referential conflict by directing their attention away from grounded items and toward previously unmentioned items.

Although interference-resolution measures account for the individual differences in perspective-taking ability in children and adults, common-ground assessment likely requires multiple different kinds of EF (e.g., memory for perspective). However, it is important to note that the only experimental conditions predicted by Stroop performance are those that impose high interference-resolution demands. This raises the question: If relevant EF skills can be targeted and enhanced via interference-resolution training (for instance, using a training-appropriate version of the Stroop task as in Brown-Schmidt, 2009), would individuals (particularly children) subsequently be less likely to consider unintended interpretations in cases of referential ambiguity? That is, one might hypothesize that EF training, within a process-specific interference-resolution framework, will result in a generally sharper ability to promote relevant sources of information like

context and pragmatics, and suppress currently irrelevant ones (e.g., one's privileged perspective) through top-down control.

Indeed, there is indirect support for this. Work by Kloo and Perner (2003) provides evidence for far-transfer across structurally dissimilar tasks of information re-characterization within a theory of mind context in young children, who were either assigned to card-sorting training or false-belief (perspective taking) training. The card-sorting task involved categorizing cards with two distinct features (e.g., two yellow apples, one green apple), with the relevant dimension changing (from number to color) after each set of cards were fully sorted. The false-belief task required children to answer questions about a conflicting situation in which a puppet acted on a certain puppet, but said he acted on another puppet. To assess the training-mediated effects of card-sorting and theory of mind, two novel assessments were implemented: The card-sorting transfer task included incorporating multiple rules for new cards (sort by number then color) and sorting an entirely different set of cards on novel dimensions. The false-belief-transfer measure was a traditional Sally-Ann task using the same puppets from training. Reciprocal far-transfer was observed for both types of training—individuals receiving false-belief training improved on card sorting, and those trained on card-sorting showed benefits on the Sally-Ann task—suggesting the presence of a shared object re-description process. Note that a similar card-sorting task resulted in transfer to “task-switching” measures in a report of near-transfer highlighted earlier (Kray & Karbach, 2009). Both sets of results point to the malleability of EFs important for perspective taking, namely, object re-description (given by the Kloo and Perner findings) and task-switching (consistent with Kray and Karbach's work). To this end, task-switching ability is apt to



overlap with interference resolution (object re-description), as switching between multiple rules involves overriding old features and rules in favor of newly relevant ones, a type of information recharacterization that is a hallmark of interference resolution. A carefully designed training regimen—for example, by comparing task-switching training with interference resolution training—may illuminate the overlapping contributions of each EF for each false-belief and perspective-taking tasks similar to those outlined above.

#### *1.2.4 Verbal Fluency*

During language production, the ease with which a lexical item is generated depends partly on the degree of competition from other candidate words. Competition demands are particularly high when multiple semantically related words are equally plausible contenders for selection (a classic case of underdetermined representational conflict; see above discussion). Items with high versus low name-agreement, for instance, present different levels of conflict during naming tasks, such that low name-agreement items associated with many alternative labels (e.g., couch/sofa/loveseat) elicit more competition, reflected by longer naming latencies, thus requiring the use of VLPFC-mediated interference/conflict resolution to select among the competing alternatives (Kan & Thompson-Schill, 2004; Novick et al., 2009). High name-agreement items (e.g., images that invoke a single label, like apple), by contrast, have fewer alternative labels to choose from, rendering them less dependent on interference-resolution processes, and thus easier to access and produce. Furthermore, selection costs are compounded when cases of high-competition (low name-agreement) are crossed with increased retrieval demands (e.g., low association-strength between a cue and its most accessible response),

such that items with multiple weak associates are most difficult to output (Snyder et al., 2011).

This high- versus low-name-agreement asymmetry has been examined in nonfluent aphasic patients with VLPFC damage—the same patients mentioned above who exhibit generally poor interference resolution and cognitive control on a variety of nonlinguistic interference resolution tasks like Stroop and the recent-no task. This population demonstrates exaggerated effects of production difficulty for high-competition conditions that require the recruitment of interference-resolution resources, such that they take significantly longer or even fail to produce these items altogether relative to low-competition items (Novick et al., 2009). Patients with this neuroanatomical profile have difficulty with other verbal fluency tasks, including completing sentences when the options are open-ended (and therefore ambiguous), versus when the to-be-completed fragments provide a highly constrained context, yielding little competition from possible alternative continuations (Robinson et al., 1998; Robinson, Shallice, & Cipolotti, 2005). Similarly, healthy speakers take longer to produce the names of pictured objects when they are presented in semantically homogeneous (e.g., snake, cow, dog, ant) versus mixed contexts (e.g., snake, bus, axe, chair) due to the increase in lexical-semantic competition among semantically related competitors (Belke, Meyer, & Damian, 2005; Hodgson, Schwartz, Brecher, & Rossi, 2003). In one study, nonfluent aphasics with circumscribed VLPFC damage generated more errors when naming objects in homogeneous contexts; a companion neuroimaging experiment further showed that even healthy adults with a greater VLPFC response to naming under homogeneous conditions are prone to more naming errors compared to individuals with less VLPFC activation (Schnur et al., 2009).

Careful consideration of the literature suggests that language production under conditions of conflict appears to be modulated by general EF abilities, like those governing interference resolution on Stroop-like tasks. Consequently, training tasks tapping these same underlying neural resources may, hypothetically, be drawn on as tools to boost word selection abilities under elevated interference-resolution demands. The idea is that better interference-resolution skills acquired through training might generalize to an increased ability to resolve among semantically related lexical items that compete for selection, carrying important implications for clinical interventions in populations with deficits in verbal fluency that accompany a more general deficit in interference resolution. Indeed, the EF training intervention study presented in Chapter 3 includes measure of verbal fluency to test this hypothesis.

Furthermore, training may also have consequences for selecting among competing alternative names during states of elevated anxiety. One study reveals that more anxious individuals (evaluated by a composite score of anxious apprehension) are impaired relative to less anxious subjects when they must generate an associated verb (in response to a given noun) under high retrieval demands, an effect mediated by VLPFC (Snyder et al., 2011). This suggests that EF resources are depleted in cases of anxiety (Gray, Braver, & Raichle, 2002), which can negatively affect word selection processes under elevated EF demands (e.g., high competition items). Future research on interference-resolution training, therefore, might also address whether the right interventions can be used to offset such effects of anxiety and other deleterious affective states in both production and comprehension (but see Beilock & Carr, 2005).

### **1.3 General Discussion and Dissertation Aims**

Overall, I reviewed a sample of language tasks that depend heavily on posterior regions of left VLPFC, which support interference-resolution abilities in a variety of populations. Among these measures there is great overlap in the EF processes involved to carry them out successfully, whether it means employing interference-resolution to produce the right word, resolve lexical ambiguities, take a speaker's perspective to avoid errors in interpretation despite referential ambiguity, or recover from temporary misanalysis during sentence parsing. I believe that in view of these convergent findings, the theory that interference resolution and cognitive control contributes to language use may lead to the hypothesis that these domain-general cognitive control processes could be the target of extensive training regimens, the result of which could be attenuated processing difficulty during language use across a range of tasks, as indexed through measures of far-transfer. Such hypotheses are motivated also by the demonstration of positive transfer effects in non-linguistic cognitive domains following regimens targeting other EFs.

Given prior evidence for far-transfer from WM training tasks to other measures such as task-switching, updating, and false-beliefs, the major goal outlined here, based on a theory of the role of left VLPFC and cognitive control in language processing, would be to design training studies in search of generalized effects to language measures, in hopes of mitigating difficulties under certain production and comprehension conditions during everyday language use. Considering the patterns just reviewed suggesting a shared role for domain-general interference-resolution processes across a variety of language processing tasks, a common training regimen targeting this EF could, hypothetically, be

successful in correcting problems observed in each of these tasks.

The aim of this dissertation is to test for the causal relationship of general-purpose interference resolution for language use. I accomplished this through two intervention studies, hypothesizing that practice with an interference resolution task would confer selective benefits to high-EF conditions across a wide range of measures, but not to those conditions where the need to deploy EF is removed. To elaborate, transfer was only expected where interference resolution is targeted through training so as to affect shared processes that facilitate performance on particular linguistic and non-linguistic tasks; WM training tasks not involving interference-resolution are not expected to confer transfer to untrained conditions of language and non-language tasks relying on EF. The complementary effect was also expected: Conditions of assessment tasks that do *not* rely on interference resolution skills were not expected to improve following extensive exposure to interference resolution training.

In Experiment 1, I tested for the effect of training on a battery of general-purpose cognitive tasks (including canonical WM tasks and one interference resolution task) on syntactic ambiguity resolution. I hypothesized that improved EF following training should generalize to real-time sentence processing and comprehension. Notably, transfer effects should be restricted to parsing conditions under high EF demands, namely when readers must revise an early parsing commitment after encountering new evidence that conflicts with their developing interpretation. Because EF training is hypothesized to transfer only to tasks requiring common underlying EF mechanisms, no training-related changes are expected under low EF demands, when reinterpretation is unnecessary. Such selectivity would be evidence for a successful process-specific training-transfer effect.

Moreover, trainees demonstrating improvement on WM training tasks that do not tap interference resolution were not expected to show this selective performance boost; these subjects were, instead, hypothesized to show no improvement from pretest to posttest on measures indexing high or low interference resolution. Finally, I hypothesized that the participants assigned to the (untrained) no-contact control group would show no cross-assessment improvement on ambiguous or unambiguous items, owed to their lack of experience practicing tasks that require EF.

Experiment 2 extends Experiment 1 by implementing more carefully controlled training regimens, as well as additional assessment tasks with high-EF demands. In particular, participants in Experiment 2 were assigned to one of three training groups, each of which practiced a single version of the *n*-back task (mentioned briefly above) over the course of the training period: The Lures Group trained on *n*-back with lures, a task expected to boost interference resolution. An identical version of the *n*-back task absent lure items was given to the No-Lures Group, rendering this group a direct control for the Lures Group. A final 3-Back Group performed a non-adaptive version of *n*-back to remove a critical design component thought to drive training-transfer effects. Briefly, adaptivity forces participants to stay at the precipice of their best performance throughout training, providing an appropriate level of difficulty to keep them engaged and improving over the course of their study involvement. All participants in Experiment 2 performed a battery of tasks at pretest and posttest, including syntactic ambiguity resolution, verbal fluency, recognition memory, and Stroop. Built into these assessment tasks were conditions with high-EF demands (e.g., ambiguous sentences, nouns with many verb associates, blocks with interfering memoranda, and incongruent color words) and little-

to-no-EF requirement (e.g., unambiguous sentences, nouns with few verb associates, blocks with no interfering to-be-remembered items, and congruent or neutral color words). Given that only subjects assigned to the Lures Group practiced a training task compelling them to routinely resolve interfering representations, I hypothesized that this group (and no other) would improve selectively on the high-EF conditions—and show little to no performance boosts on the low-EF trials—of the assessment tasks. Likewise, no cross-assessment performance advantage should exist among subjects assigned to the remaining non-Lures Groups on the high-EF conditions. Moreover, Experiment 2 included difficult syntactic materials that have not been directly tied to non-mnemonic capacities (object-extracted relative clauses), thus presenting the opportunity to test whether interference-resolution training confers a general advantage for all cases where effortful cognition is required. An improvement by the Lures Group on these items would indicate that high-EF training confers a general advantage for all difficult items, rather than just those items with interference resolution demands. Taken together, both studies are expected to provide convergent evidence for the malleability of interference resolution skills and its effects on language processing.

## **Chapter 2: Experiment 1 - Training Executive Functions for Sentence Processing<sup>2</sup>**

In the first experiment of this dissertation, I investigated whether enhancing regulatory functions through EF training improves garden-path recovery in healthy adults. Given claims that EF and flexible cognition also play an important role in a range of other language processing tasks (see above), a positive result from the present research could open the door to exploring whether EF training may be an effective intervention tool for improving reading comprehension, particularly in rare instances when multiple evidential sources do not conspire to guide or facilitate processing (i.e., when various representations compete, resulting in small but reliable consequences for interpretation; see Chapter 1 and Hussey & Novick, 2012).

The training tasks implemented in Experiment 1 were developed based on benchmark working memory tasks, some of which are known to elevate demands for cognitive control (e.g., by increasing the need to resolve among conflicting representations in memory). It is important to note that a battery of working memory training tasks were administered to test the possibility that performance improvements on specific ones might contribute differentially to improvements in syntactic ambiguity resolution (see Method and General Discussion). Put differently, the training tasks that rely on resources common to garden path recovery should render correlated measures

---

<sup>2</sup> This section is a modified version of: Novick, J.M., Hussey, E.K., Teubner-Rhodes, S.E., Harbison, J.I., & Bunting, M.R. (2013). Clearing the garden-path: Improving sentence processing through cognitive control training. *Language and Cognitive Processes*. doi: 10.1080/01690965.2012.758297



across both tasks. As highlighted in Novick, Trueswell, and Thompson-Schill (2005; see also D’Esposito & Postle, 1999; Novick et al., 2009; 2010), some working memory tasks rely on a non-mnemonic capacity, which involves the EF ability to resolve conflicting (or interfering) representations—a general skill necessary for some linguistic tasks, including recovering from misinterpretation. Such cases require EF—in particular interference resolution—to rein-in initial mischaracterizations of the input. Not all working memory tasks necessarily share this interference-resolution property to the same extent; thus, it is possible that only those training tasks that tax the need to resolve among competing alternatives and re-characterize information will predict or correlation with performance improvements in syntactic ambiguity resolution abilities across assessments.

Therefore, I will examine whether performance increases on some working memory training tasks (*n*-back, Letter-Number sequencing, Block Span, and/or Running Span; see Method) contribute to changes in ambiguity resolution performance more than others. According to a process-specific training account, the amount of transfer to untrained tasks following intervention depends on the extent of overlap between the cognitive and neural resources shared by the training and the transfer tasks. As outlined in Chapter 1, garden-path recovery engages interference resolution processes supported by regions within left VLPFC. One of the EF training tasks, namely *n*-back with lures, has been shown previously to recruit regions within VLPFC, owing to the interference generated by lure trials (see Method below; see also Gray et al., 2003; Jaeggi et al., 2003; Owen, McMillan, Laird, & Bullmore, 2005). Thus, a process-specific account predicts that only those EF tasks that recruit common areas in VLPFC for interference resolution, such as the *n*-back task, should transfer to syntactic ambiguity resolution (Novick et al.,

2005); EF training tasks involving exclusively other functions like maintenance and manipulation of information in verbal or spatial working memory absent interfering representations (e.g., Letter-Number sequencing; Block Span) should demonstrate little or no transfer (or transfer that is sensitive to the degree of overlap between the training task and transfer measure). By including multiple training tasks that employ different components of EF to varying degrees in the training regimen, I was able to test whether EF training at the broadest level sufficiently improves garden-path recovery, or whether interference resolution training specifically is necessary to increase syntactic ambiguity resolution abilities, thus informing a deeper understanding of the domain-general cognitive control mechanisms that contribute to sentence reinterpretation.

## **2.1 Experimental Preliminaries**

Pre- and post-training assessments included a reading task using sentences containing a temporary syntactic ambiguity. Consider (1) and (2):

1. While the thief hid the jewelry that was elegant and expensive sparkled brightly. (Temporarily Ambiguous)
2. The jewelry that was elegant and expensive sparkled brightly while the thief hid. (Unambiguous)

In (1), the ambiguity springs from the verb “hid,” which can be used either reflexively (individuals can hide themselves), or transitively (individuals can hide objects). Here, the transitive interpretation is strongly supported due to the absence of a comma following “hid,” which would impose the reflexive analysis (Ferreira, Christianson, & Hollingworth, 2001). The presence of a plausible object (“the jewelry”) further supports the transitive interpretation (see Garnsey, Myers, Pearlmutter, & Lotocky, 1997). Hence,

readers rapidly interpret the sentence to mean the thief is hiding the jewelry. This analysis, however, is ultimately unviable because “the jewelry” turns out to be the subject of a new clause (“the jewelry sparkled...”), not a direct object. Upon encountering late-arriving disambiguating evidence that conflicts with the developing interpretation (“...sparkled brightly”), readers must initiate cognitive control processes in order to re-characterize their initial representation of sentence meaning, i.e., to resolve the conflict and revise their misinterpretation (Novick et al., 2005; 2009). In (2), the reversed clause order unambiguously signals the reflexive analysis; consequently, reinterpretation is unnecessary and interference resolution and cognitive control processes need not deploy. Another way to disambiguate (1) would be to simply add a comma following the verb (“While the thief hid, the jewelry...”). However, in order to provide sufficient room for accuracy improvement across assessments, I adopted the reversed clause order disambiguation in (2) to maximize the ambiguity effect between conditions, following prior work indicating nominally higher error rates when comparing (1) to this unambiguous construction, versus the comma-disambiguation construction (see Exp. 3 in Christianson, Hollingworth, Halliwell, & Ferreira, 2001).<sup>3</sup>

---

<sup>3</sup> Although I subscribe to constraint-based lexicalist perspectives of ambiguity resolution (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994; Novick et al., 2003; Novick, Trueswell, & Thompson-Schill, 2008; Trueswell & Tanenhaus, 1994), testing this theory against serial models (in which individuals start with a syntactically-driven interpretation and revise when needed; Frazier & Fodor, 1978) was not the focus of the current experimental efforts, as various constraints were not manipulated to differentiate these models. Moreover, under both accounts, the ambiguous sentences in the present experiment should initially lead to an incorrect transitive interpretation, which must be reconciled with conflicting input later in the sentence regardless of how it was developed. I, therefore, omit discussion of parsing-theory contrasts and describe my materials, as well as readers’ processing decisions, in simple terms that do not rely on a particular parsing framework (for a review and theoretical discussion of constraint-based theories with respect to cognitive control, see Novick et al., 2005).

Participants answered questions that probed for lingering effects of misinterpretation, for example, “Did the thief hide himself?” Full reanalysis does not always occur in ambiguous cases, resulting in erroneous ‘no’ responses (Christianson et al., 2006; Christianson & Luke, 2011). Importantly, all comprehension questions queried the correct (reflexive) interpretation, and *not* the initially conceived and consequently incorrect analysis (the transitive interpretation (e.g., “Did the thief hide the jewelry?”). This was designed as such to avoid vulnerability to memory effects, where error commissions (i.e., a ‘yes’ response to “Did the thief hide the jewelry?”) could be influenced by familiarity of the memory trace of the initial misinterpretation. In other words, even if a reader did correctly revise the sentence, they might respond erroneously because the question itself restated the incorrect transitive analysis, thereby reactivating the initial misinterpretation. Error commissions to such sentences could be furthermore shaped by plausible inferences or general world knowledge, since thieves are likely to hide jewelry. Instead, in order to correctly respond ‘yes’ to these questions—for instance, verifying that the thief was hiding himself—readers actually had to override the initial, incorrect transitive interpretation (that the thief was hiding the jewelry) to recover the alternative reflexive interpretation, which is a more straightforward indicator of garden-path recovery. Similarly, an incorrect ‘no’ response to these questions signifies a lingering commitment to the early direct-object analysis and, thus, recovery failure.

## **2.2 Hypotheses**

The present domain-general EF training regimen may support controlled revision. Hence, I hypothesize that interpretation recovery—reflected by comprehension accuracy for ambiguous sentences—should improve following training. Such performance

increases might be especially related to training tasks aimed at enhancing interference resolution abilities. No changes are expected in unambiguous cases where the need for cognitive control is removed.

To investigate the effects of EF training on *real-time* sentence processing and reanalysis, I recorded participants' eye-movements. Leftward saccades (regressions) to previously encountered material signal changes in moment-by-moment revision, and mark the launch of recovery functions (Frazier & Rayner, 1982; Sturt, 2007). I hypothesized that recovery efforts should improve following training, reflected by less processing difficulty upon encountering disambiguating (i.e., conflicting) evidence. Note that changes in eye-movement patterns should be associated only with reading behavior following entry into disambiguating sentence regions, where interference resolution and cognitive control processes are hypothesized to engage. Changes are not expected in other regions of ambiguous sentences, or anywhere in unambiguous sentences, if improvements are related specifically to enhancements in domain-general interference resolution abilities.

Furthermore, training-related improvements in garden-path recovery processes—indexed by both online and offline measures sketched above—may depend to a greater extent on performance increases on some training tasks than others. Theoretically, the extent of improvement on a training task targeting interference-resolution mechanisms might be especially likely to predict gains in garden-path recovery, because interference-resolution processes are thought to help recovery of alternative parsing options when other sources of evidence have guided the parser toward an incorrect syntactic characterization of the input (see Novick et al., 2005; 2009; 2010). In other words,

enhanced interference-resolution abilities may help readers better avoid misinterpretations by more rapidly countermanding early parsing decisions in real-time, reflected by more efficient changes in controlled revision processes (i.e., regressions) once a misanalysis has been discovered.

As highlighted by previous findings in the cognitive training literature, the ability to observe reliable effects of training across assessments hinges on whether individuals in the treatment group *actually improve* on the task(s) completed throughout the training regimen (see Chein & Morrison, 2010; see also Jaeggi et al., 2011). That is to say, there may be decisive differences within the group of trainees regarding the extent to which individuals respond positively to the regimen. Here, only those who successfully improve performance during interference-resolution training are expected to transfer these benefits to untrained measures of garden-path recovery (see the Discussion for caveats to this method). An important approach to analyzing this kind of study, therefore, is to differentiate training “responders” from “non-responders” (Chein & Morrison, 2010; Jaeggi et al., 2011). This can be done in two ways: (i) by using training responsiveness to the various tasks as a continuous variable to test the relation between the amount of improvement during the regimen and the amount of pre/post improvement in ambiguity resolution (i.e., transfer)—a multiple regression analysis that is consistent with prior training work; and (ii) by treating responsiveness to the various training tasks as a discrete variable, separating responders from non-responders via well-established statistical clustering methods (see Fraley & Raftery, 2002; 2011) and comparing these groups’ performance on the sentence processing task (using accuracy and reading time data as dependent variables) to the untrained group, which had no inter-assessment

training data to evaluate. I expected that only the responders would exhibit cross-Assessment garden-path recovery gains that are significantly greater than the other two groups. This latter analytic approach is novel for training studies. Generally speaking, the responders are the people of most theoretical and practical interest here.

To summarize, I hypothesize that:

1. Individuals' level of improvement on a training task targeting interference-resolution processes (*n*-back with lures; see Method) should predict gains in garden-path recovery, whereas performance increases on the three other working-memory training tasks, which do not involve interference resolution functions, should *not* predict test-retest changes in ambiguity resolution.
2. Those who show steady and significant improvement (“responders”) on a training task targeting interference resolution processes (*n*-back with lures) will differ reliably from the untrained control group—as well as from subjects in the training group who do not respond well to this task—regarding their cross-Assessment change in garden-path recovery.
3. By contrast, responders on the three other training tasks, which do not target this important cognitive control function by design, should behave similarly to untrained controls and the non-responders on those tasks in terms of cross-Assessment performance in syntactic ambiguity resolution. This should be the case if and only if the other training tasks do not tap (or, at least, tap less of) the proposed underlying EF shared by the *n*-back-with-lures task and syntactic ambiguity resolution (i.e., interference resolution).

## 2.3 Method

### 2.3.1 Subjects

Healthy native-English-speaking subjects were randomly assigned to a training or no-contact control group. Thirty-three participants were excluded from analyses (16 from the training group) for failing to complete all study phases. The final participant group comprised 43 individuals (training group:  $N=21$ , 15 women,  $M_{\text{age}} = 21.1$  years, age range = 18-39 years,  $M_{\text{education}}: 14$  years; control group:  $N=22$ , 15 women,  $M_{\text{age}} = 21.8$  years, age range = 18-36 years,  $M_{\text{education}}: 14.29$  years). None of the subjects had a history of neurological disorders, stroke, or learning disabilities, and no one reported taking medications to correct problems related to neuropsychological or neuropsychiatric impairment. All subjects had normal or corrected-to-normal vision and hearing.

### 2.3.2 Design

A double-blind pretest/posttest design was used; accordingly, neither subjects nor experimenters knew subjects' condition assignments. Different moderators held training and assessment sessions in separate labs, so that the experimenter who collected the assessment data was blind to the condition to which each subject had been assigned. Additionally, because subjects in the experimental and control conditions never interacted, they were in principle blind to each other's condition and unaware of the differences between them. The experimental group visited the training lab for 20 one-hour sessions in the three-to-six weeks ( $M=4.9$  weeks) intervening pretest (Assessment 1) and posttest (Assessment 2) (see Figure 1). Importantly, training did not involve practicing syntactic ambiguity resolution or reading of any kind. Thus, any demonstrated effects of transfer might reasonably be attributed to improvements in domain-general



processes, rather than to extra experience practicing linguistic- or syntactic-specific processes. Control participants received no contact during this interval (see Chein & Morrison, 2010; Jaeggi et al., 2008), but the interim between their assessments was also three to six weeks ( $M=5.1$  weeks), matched to the training group.

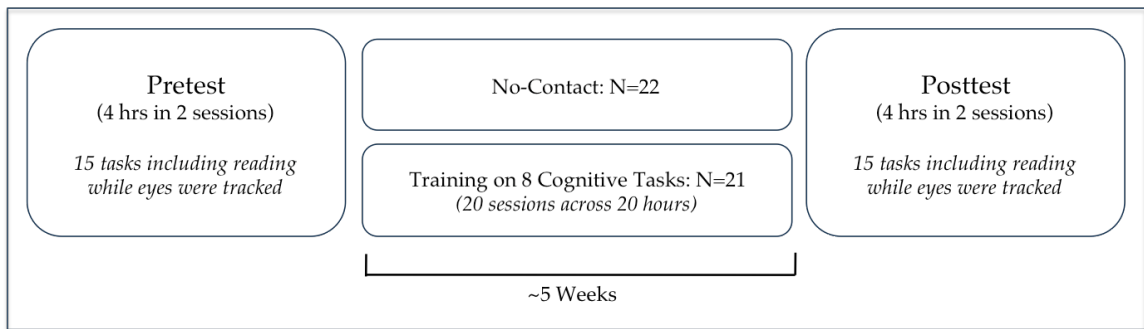


Figure 1. Longitudinal design of Experiment 1.

During each assessment, participants completed 14 short cognitive tasks and a reading task testing syntactic ambiguity resolution. I consider here data from only the syntactic ambiguity resolution task, as the additional cognitive assessments addressed independent research questions related to crystallized and fluid intelligence (and will therefore be reported elsewhere). Moreover, these other assessments were conducted and led by other researchers, and were largely completed during a separate task-administration session. None of the other cognitive assessments involved psycholinguistic tasks of any type. Each assessment battery was administered across two 2-hour sessions that were completed on different days within a two-week period.<sup>4</sup>

---

<sup>4</sup> All subjects also completed a third assessment, which occurred three months following Assessment 2 without additional training for the experimental group. Assessments 1 and 2 were of primary interest, as performance at Assessment 2 measured the immediate effects of training versus Assessment 1 (Assessment 2 was completed approximately one week after trainees finished the regimen). Assessment 3 was included to evaluate maintenance of training effects primarily for the non-syntactic measures of cognitive function (i.e., the assessments of fluid and

### 2.3.3 Training Tasks

In the interval between assessments, subjects in the training group completed 20 hours of practice on eight tasks, four of which were working memory tasks with EF characteristics designed to tax and improve their ability to regulate attention. A battery of four EF tasks was used in order to tap a broad array of executive-control functions (see below and Table 1), to test if gains on any particular training task(s) with emphasis on specific EF properties (e.g., interference resolution) could significantly predict ambiguity resolution improvements versus others. These four EF tasks were programmed in-house, and were developed based on paradigms commonly used in the neurocognitive literature. These included a letter *n*-back task with lures in non-*n* positions (targeting conflict/interference resolution processes); an auditory letter running-span task (targeting the capacity of attentional focus; see Bunting, Cowan, & Saults, 2006); a letter-number sequencing task (LNS, a complex span task targeting the manipulation of verbal stimuli in working memory); and a block span task (a complex span task targeting visual-spatial working memory). Previous research has implicated the recruitment of regions within left VLPFC during non-training versions of *n*-back with lures (Gray et al., 2003; Owen et al., 2005) and some versions of running span (Postle, Berger, Goldstein, Curtis, & D’Esposito, 2001; see Discussion). Posit Science contributed the remaining four training tasks from their brain-fitness software packages (Brain Fitness Program, Version 2.1; Insight, Version 1.1). These included “jewel-diver” (targeting divided attention through visual-tracking of multiple objects), “match-it” (targeting the ability to match auditory and visual representations of a phoneme), “sound-replay” (targeting phoneme

---

crystallized intelligence). We do not include Assessment 3 data, as they do not bear on my central hypotheses.

categorization and discrimination), and “listen-and-do” (targeting the ability to follow auditory instructions).<sup>5</sup>

Four tasks were administered per training session for approximately 15 minutes each. Over the 20 sessions, each task was repeated 10 times, and task difficulty adapted dynamically to individual levels to keep participants continually on the threshold of their best performance. Task order was the same for all participants. I describe the four in-house EF tasks briefly below, each of which is also detailed in the top 4 panels of Table 1.

**N-Back.** Sets of twenty-five single letters were displayed serially and participants indicated by button press whether the current letter had appeared  $n$  items previously (see first panel of Table 1). For example, if given the sequence *H-B-K-H* in a 3-back condition, the second *H* would be a ‘target’; in the sequence *H-B-K-T*, the *T* would be a ‘non-target’ because it does not match the 3-back stimulus, *H*. This version of  $n$ -back was intended to train conflict/interference resolution mechanisms by including ‘lure’ trials—recently presented letters that occurred either immediately before ( $n-1$ ) or after ( $n+1$ ) the  $n^{\text{th}}$ -back item (Kane, Conway, Miura, & Colflesh, 2007; see also Gray et al., 2003; Burgess et al., 2011). For example, if given the sequence *H-B-H-D-K* in a 3-back condition, the second *H* would be a lure (an  $n-1$  lure) because it was a 2-back, not a 3-back, stimulus. Thus, subjects would have to respond ‘non-target’ to this item. Because

---

<sup>5</sup> The four commercial Posit tasks primarily targeted low-level perceptual functions, and were included not because of any expected relation to syntactic ambiguity resolution, but because of theoretical overlap with the other pre/post cognitive assessments that subjects completed. The link between the Posit tasks and the other assessments addresses entirely separate research questions beyond the scope of—and unrelated to—the work presented in this paper. We mention them because subjects in the training group completed them during the interval between assessments, but hereafter we limit further discussion and analysis of these tasks because they will be reported in full elsewhere.

lures did not appear in the specified  $n$ -back location, participants had to override a tendency to respond based on familiarity alone and resolve the conflict between the correct representation and a familiar, but incorrect one (see General Discussion). Participants encountered three lure levels before  $n$  increased: no lures,  $n+1$  lures only, and both  $n+1$  and  $n-1$  lures. Task difficulty increased when participants achieved at least 85% accuracy by first increasing lure level incrementally and then by increasing  $n$ . Task difficulty decreased if participants fell below 65% accuracy, again by first decreasing the lure level and then by decreasing  $n$ . Difficulty values reflected both the value of  $n$  and the lure level.

**Running Span.** Anywhere from 12 to 20 letters were presented auditorily in a continuous stream (see second panel of Table 1). Each string ended unpredictably, after which participants immediately had to recall the last  $n$  items from a fleeting auditory memory store. Initially,  $n=2$  and  $n$  increased after participants successfully satisfied the criteria for progression at each of three presentation rates: 1,000 ms, 750 ms, and 500 ms. If a participant achieved 100% accuracy on four successive trials, then presentation rate increased in increments of 250 ms to a maximum rate of 500 ms. If the presentation rate was already 500 ms, then  $n$  increased by 1 and the presentation rate slowed to 1,000 ms. If mean accuracy dropped to 25% or below, then task difficulty decreased by slowing the presentation rate by 250 ms; if the presentation rate was already 1,000 ms, then  $n$  decreased by 1. Difficulty values reflected both the value of  $n$  and the presentation rate.

**Letter-Number Sequencing (LNS).** Pseudo-randomized sets of one or more sequences of interleaved letters and digits were presented visually (see third panel of Table 1). Participants were instructed to recall the numbers in ascending order first,

followed by the letters in alphabetical order, separately for each set sequence. The number of items within a sequence and the number of successive sequences presented before recall adapted to participant performance. If participants performed perfectly on four consecutive sets, the task increased in difficulty first by incrementally increasing the number of characters per sequence from two to six, and then by increasing the number of sequences per set from one to six. If participants completed less than two consecutive sets correctly, then the task decreased in difficulty by first reducing the number of characters per sequence followed by the number of sequences per set. Difficulty values reflected both the number of sequences per set and the number of blocks per sequence.

**Block Span.** Sets of one or more sequences of shaded blocks were presented in a 4x4 grid (see fourth panel of Table 1). After each set was presented, participants were instructed to recall the block locations for each sequence in the order of presentation. Initially, a set consisted of only one sequence of two blocks. If participants had perfect recall for four consecutive sets, then task difficulty increased first by incrementally boosting the number of blocks per sequence from two to five, followed by the number of sequences per set from one to six. Task difficulty decreased if participants completed less than two of four sets correctly by first decreasing the number of blocks per sequence and then the number of sequences per set. Difficulty values reflected both the number of sequences per set and the number of blocks per sequence.

#### *2.3.4 Transfer Task: Syntactic Ambiguity Resolution*

Separate but complementary versions of the ambiguity resolution task were developed so that participants never saw the same materials across assessments. Twenty-four verbs that could be used both transitively or reflexively (e.g., “hid”) were borrowed

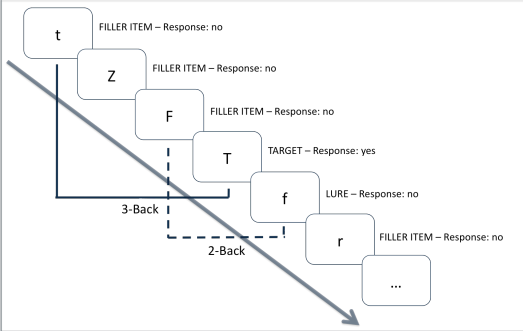
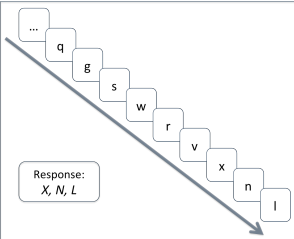
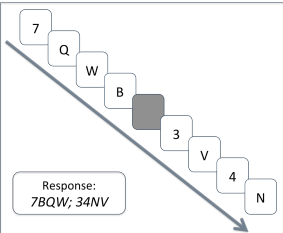
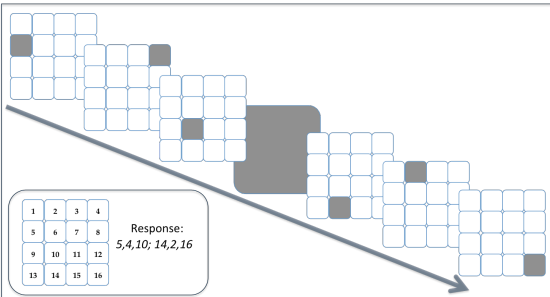
Task	Targeted Cognitive Ability	Example Trial
Visual Letter N-Back Task with Lures	interference resolution in working-memory	<p data-bbox="805 394 1333 424"><i>For 3-back condition with lure level of 1:</i></p> 
Auditory Letter Running Span Task	capacity of attentional focus	<p data-bbox="805 785 1105 814"><i>For condition n=3:</i></p> 
Letter-number Sequencing Task (LNS)	manipulation of verbal stimuli in working memory	<p data-bbox="805 1079 1105 1108"><i>For 2 4-item sequences:</i></p> 
Block Span Task	visuo-spatial working memory	<p data-bbox="805 1373 1360 1402"><i>For 2 sequences of 3 blocks:</i></p> 

Table 1. Explanations of the 4 in-house training tasks used in Experiment 1.

from Christianson et al. (2006) and were used to create 12 ambiguous and 12 unambiguous sentences per assessment (see examples 1 and 2 above). At each assessment, these 24 items were embedded within 90 filler sentences (borrowed directly from Christianson et al., 2006; personal communication), which did not contain syntactic ambiguities and sampled a variety of constructions to draw attention away from the ambiguity manipulation. This variety included transitive structures that resembled the experimental items but removed any critical temporary indeterminacy (e.g., “While the father prepared the burgers he covered them with pepper;” “The exterminator entered the school while the cockroaches scurried”). For each assessment, two lists were created: if an item in one list was ambiguous, it was unambiguous in its counterpart list. List administration was pseudorandom and counterbalanced across participants and assessments. Thus, each reflexive/transitive verb appeared only once per assessment. If a participant saw a particular verb in an ambiguous construction at Assessment 1, that verb appeared in an unambiguous construction at Assessment 2 in a different sentence. Hence, while verbs repeated, they did so only across assessments and appeared in new contexts and ambiguous/unambiguous frames.

A comprehension question about the correct reflexive interpretation was presented following every sentence. For ambiguous items, this meant that the questions probed for lingering effects of ambiguity and thus failure to revise and arrive at the correct interpretation (Christianson et al., 2006). For instance, in order to correctly answer “Did the thief hide himself?,” readers were forced to override the initially favored transitive analysis. The same question was presented for the unambiguous versions of an item. Thus, for all ambiguous and unambiguous items, the correct response was ‘yes.’

However, correct ‘yes’ and ‘no’ responses were balanced across the 114 total items (ambiguous, unambiguous, filler) at a given assessment. All sentences can be found in Appendix A.

**Apparatus.** Eye-movements were recorded using an EyeLink 1000 eye-tracker (SR Research), with vertical and horizontal eye position sampled every millisecond. Stimuli were presented via the UMass Amherst EyeTrack 0.7.10 Software (<http://www.psych.umass.edu/eyelab/software/>).

Participants were situated in the Eyelink’s forehead and chin rests. Viewing was binocular but the system was set to monocular recording. The eye-tracker was calibrated to an average spatial-resolution error of  $0.50^\circ$  or less and recalibrated as needed. Eye-movement data were excluded from one participant who could not be calibrated (a subject in the untrained control condition).

Each trial began with a fixation box in the position of where the leftmost character of the sentence would appear. Once a subject fixated this box, the sentence appeared automatically, replacing the fixation box; this procedure served as a trial-by-trial calibration check. Each sentence was presented in its entirety on a single line. Participants were instructed to read each sentence at a comfortable pace and press a button when finished to advance to the comprehension question, to which they responded ‘yes’ or ‘no’ via button press. Before the experiment, participants completed ten practice trials to ensure that they understood the procedure. Total task time averaged 40 minutes (range=25 to 50 minutes), including recalibration and a scheduled break.



## 2.4 Analyses and Results

Analysis of the training data revealed that participants showed the expected improvement on the four in-house training tasks (average effect size, Cohen's  $d=1.7$ ). However, did training gains transfer to syntactic ambiguity resolution? I focused on two measures of garden path recovery to address this question: sentence comprehension accuracy and real-time reanalysis using eye movements.

### 2.4.1 Sentence Comprehension Accuracy

**Analysis.** Using multiple regression, I examined the relation between individual training task performance and cross-assessment improvement in syntactic ambiguity resolution with Training Task as a factor, to understand the nature of the *continuous* relation between training improvement and transfer on a subject-by-subject basis. Crucially, this analysis allowed insight into whether trainees' gains on certain intervention tasks significantly predicted performance gains in garden-path recovery. Following the multiple regression analysis, I report cluster analyses that identified responders and non-responders on each of the four training tasks; I then entered these discrete responsiveness variables into multilevel mixed-effects models to test for Group-by-Assessment interactions, to determine if the responders' ambiguity resolution improvements differed reliably from both the non-responders *and* the untrained controls, who provided no training data between reading assessments. I conducted multilevel mixed-effects models using R's lmer function (lme4 library, Bates & Sarkar, 2007) due to their appropriateness for handling categorical data (see Jaeger, 2008). All accuracy data were first transformed using an empirical (e)logit function to correct potential problems related to heterogeneity of variance (see equation 5 in Barr, 2008). For clarity,

untransformed data are reported and illustrated in the figures. (Transformations did not result in any change in data patterns or significance values.) Such mixed-effects models were used both to statistically evaluate any test-retest improvement in the conditions of interest and to examine whether any reliable differences emerged among the groups (responders, non-responders, controls; see below) in test-retest changes. For all statistical models, Subjects and Items were crossed as random intercepts (Baayen, Davidson, & Bates, 2008; Jaeger, 2008; Quené & van den Bergh, 2008). In each analysis reported, I evaluated whether both random slopes and intercepts improved the fit of the models. Corrected Akaike information criteria ( $AIC_C$ ; see Burnham & Anderson, 2004) were used to determine whether the best-fitting model included random slopes. In every case, only random intercepts improved model fit (see  $AIC_C$ s in Tables 3 and 5); therefore, all models that I report in the main text exclude random slope terms.

I also conducted Jeffreys-Zellner-Siow (JZS) Bayes-factor (BF) tests to verify the results of each t-test reported below using R's `ttest.Quad` function (BayesFactorPCL library, Morey & Rouder, 2010; see Rouder, Speckman, Sun, Morey, & Iverson, 2009 for a detailed explanation of BFs of t-tests). JZS BF tests include a parameter,  $r$ , used to index expected effect sizes; because I have no hypotheses with respect to effect size,  $r$  was set a priori to a default value of 1.0. Cauchy priors were assumed for all BF tests implemented for each reported balanced one-way ANOVA model below using R's `onewayAOV.Quad` function (BayesFactorPCL library, Morey & Rouder, 2010; see Masson, 2011; Rouder, Morey, Speckman, & Province, in press). Note that where an ANOVA model is unbalanced or requires more than one factor of interest, BFs are not reported; thus, where mixed-effects models (including random effects of Subjects crossed

with Items) are reported, BFs are not conducted. Some comparisons are expected to support the null hypothesis, and JZS BFs provide a means to assess the degree to which this is indeed the case. Bayes-factor tests reflect the likelihood of support for the alternative hypothesis over support for the null hypothesis, such that for t-tests, coefficients less than 0.1 index strong support for the null hypothesis and those less than 0.3 index substantial support for the null hypothesis, while those greater than 3 index substantial support for the alternative hypothesis, and those greater than 10 strongly support the alternative hypothesis.

**Baseline Ambiguity Results.** To determine first, as a manipulation check, whether the ambiguous materials imposed the hypothesized difficulty compared to unambiguous items, I fit the accuracy data for Assessment 1 only, crossing Subjects and Items as random effects and including Sentence-Type (Ambiguous, Unambiguous) as the critical fixed factor. The best-fitting mixed-effects model included a reliable effect of Sentence-Type, revealing significantly more errors in ambiguous (41%) than unambiguous (10%) conditions ( $z=11.85, p<.001, BF=25.90$ ). This suggests that the ambiguous materials provoked the expected difficulty in interpretation-recovery at Assessment 1. Accordingly, I tested if training gains predicted improvements in garden-path recovery from Assessment 1 to Assessment 2 using multiple regression.

**Relating Garden-path Recovery Improvement to Training Responsiveness: Multiple Regression Results.** Because participants in a training group typically achieve different levels of training performance (cf. Chein & Morrison, 2010; Jaeggi et al., 2011), I investigated whether training gains on the four in-house intervention tasks—computed by subtracting subjects' first session performance from their final session performance—

were related to individual levels of garden-path recovery improvement, an approach consistent with prior training studies. Given that the regimen targeted a range of EFs (see Method), this analysis also permitted scrutiny of the specific training tasks that reliably forecasted gains in ambiguity resolution, thereby providing insight into whether practicing particular EFs contributed to increased sentence-reinterpretation abilities versus others.

Entering Training Task as an independent factor while controlling for Training Gains, I conducted a multiple regression analysis testing for the continuous relationship between performance increases on the four intervention tasks and post-intervention improvement in accuracy to comprehension questions following syntactically ambiguous sentences (the dependent variable). I ran separate models for ambiguous and unambiguous data because I maintained the a priori hypothesis that training-mediated differences should occur only in the high-interference, ambiguous condition, whereas no such effects were expected in the unambiguous condition. Interestingly, as hypothesized, the *n*-back task was the only training task to result in a main effect of comprehension accuracy improvement on ambiguous items ( $t(71)=-2.11, p<0.05, BF=0.57$ ; all other main effects:  $ps>0.16, BFs<0.17$ ). No training tasks accounted for such an effect in unambiguous items ( $ps>0.38, BFs<0.09$ ). Crucially, the interaction of Training Gains-by-Training Task nearly reached significance for the *n*-back task only ( $b=0.19, t(71)=1.92, p=0.05, BF=0.39$ ; the analysis of covariance interaction terms for the remaining training tasks:  $ps>0.11, BFs<0.10$ ), indicating that accuracy improvement on ambiguous sentences depends on performance increases on this task in particular. (Notably, there was sufficient variance in how responsive the trained group was for the three non *n*-back

tasks; see cluster analyses below for LNS, Block Span, and Running Span. Nevertheless, Gains-by-Task interactions still were not observed). An interactive relationship did not emerge for unambiguous items for any training task ( $p_s > 0.38$ ,  $BFs < 0.04$ ). Taken together, these results suggest that the greater improvement achieved through consistent practice with the  $n$ -back task, the more improvement achieved on a far-transfer task of syntactic ambiguity resolution. As hypothesized, I believe that this selective correspondence is due to shared processing attributes (i.e., interference resolution) across  $n$ -back-with-lures and garden-path recovery.

The multiple regression analysis allowed the use of responsiveness as a continuous variable to understand test-retest changes in garden-path recovery as they relate to performance increases on the four training tasks; I observed that only  $n$ -back gains reliably predicted garden-path recovery improvements. However, I cannot perform a similar analysis including untrained controls because this group had no inter-assessment data on which to base such a responsiveness variable. On the other hand, multilevel mixed-effects models allow for a direct comparison of controls and trainees.

Because the multiple regression analysis revealed important individual differences in responsiveness on the  $n$ -back training task, I employed hierarchical cluster analyses to statistically separate individual trainees who showed gains on this task from those who did not. This served to confirm the relation between  $n$ -back training gains and improvements in garden-path recovery and ultimately allowed me to compare both subgroups to the untrained subjects (thus further probing the interaction observed in the regression analysis for this task separately). Cluster analyses are a novel approach to examining individual differences in training responsiveness, as prior studies have either

tested for correlations between training gains and transfer effects (Chein & Morrison, 2010), or used a median split to define training responders and non-responders (Jaeggi et al., 2011) when evaluating how training variability relates to transfer success. In previous research, responders and non-responders have usually been analyzed separately in terms of transfer effects, rather than being statistically evaluated against each other. Here, I entered the cluster-defined responsiveness variable into a larger mixed-effects model to allow the model to determine whether a Group-by-Assessment interaction provided the best fit of the garden-path accuracy data, where the Group factor contained three levels: responders, non-responders, and untrained controls. In other words, the mixed-effects comparisons allowed me to ascertain whether the responders' accuracy reliably improved and if this improvement was significantly different from the other two groups.

**N-back Cluster Analysis Results: Identifying Responsive Trainees and Non-Responsive Trainees.** I identified individuals who responded well to the *n*-back task with a model-based cluster analysis using R's *mclust* function (*mclust* library, Fraley & Raftery, 2011), which implements maximum likelihood estimation and Bayes criteria to identify the number of naturally occurring clusters of subjects given the distribution of an outcome measure (see Fraley & Raftery, 2002). This analysis was conducted for *n*-back performance using training gains as the primary index of training responsiveness, where gains were computed by subtracting each participant's initial training-session performance from his or her final training-session performance (see also Jaeggi et al., 2011). This analysis identified two clusters of subjects: 13 "responders" (6 women,  $M_{\text{age}} = 22.4$  years; age range = 19-39 years,  $M_{\text{education}} = 14.6$  years) and 7 "non-responders" (6 women,  $M_{\text{age}} = 20.5$  years, age range = 18-34 years,  $M_{\text{education}} = 14.2$  years). As depicted in

Group	N	Mean	Slope	Gains	First Score	Final Score
Task: N-Back						
All Trainees	20	4.442	0.404	3.689	2.246	5.934
Responders	13	5.277	0.569	5.132	2.323	7.455
Non-responders	7	2.891	0.096	1.009	2.101	3.110
Responders vs. Non-responders (F-value):		63.07***	61.23***	85.66***	2.31	113.50***
Task: Letter-Number Sequencing						
All Trainees	21	5.344	0.160	1.585	4.112	5.697
Responders	4	6.329	0.337	3.437	4.209	7.646
Non-responders	17	5.112	0.118	1.149	4.090	5.238
Responders vs. Non-responders (F-value):		3.91 <sup>†</sup>	17.63***	30.87***	0.10	15.24**
Task: Running Span						
All Trainees	21	3.165	0.079	0.994	2.463	3.457
Responders	9	3.400	0.103	1.520	2.416	3.936
Non-responders	12	2.989	0.062	0.599	2.498	3.098
Responders vs. Non-responders (F-value):		4.92*	7.86*	39.22***	0.32	21.72***
Task: Block Span						
All Trainees	19	5.163	0.093	1.321	4.047	5.368
Responders	9	5.731	0.135	1.890	4.293	6.183
Non-responders	10	4.653	0.055	0.809	3.825	4.634
Responders vs. Non-responders (F-value):		30.12***	13.46**	65.13***	8.74**	57.64***

Table 2. Performance measures of responders and non-responders across the four training tasks. Groups were defined by a two-component cluster analysis (see text). Note that Block Span responders and non-responders differ at training-session 1, and LNS responders and non-responders show only a marginal difference in average training performance. <sup>†</sup>p<0.06, \*p<0.05, \*\*p<0.01, \*\*\*p<0.001

Figure 2, this particular responder/non-responder distinction demonstrates wide variability in terms of subjects' performance curves throughout the course of *n*-back training, an illustration that was confirmed by an analysis of variance (ANOVA). One-sample ANOVAs were conducted to compare the two clusters in terms of improvement;

in general, the clusters reliably diverged on a range of dependent measures, including (i) mean  $n$ -back training score over all 10 training sessions, (ii)  $n$ -back gains from session-to-session as indexed by slope, (iii)  $n$ -back gains from session 1 to session 10 (i.e., the measure by which the clusters were defined), and (iv) final  $n$ -back session score.

Importantly, responders did not differ from non-responders at the onset of training as reflected by session 1  $n$ -back score (see upper panel of Table 2). Taken together, this suggests that the cluster analysis segregated subjects into two meaningfully, systematically, and significantly different groups.

**Garden-path Recovery Improvement in Responders vs. Non-Responders vs. Untrained Controls: Mixed-Effects Model Results.** I compared test-retest performance across Assessments 1 and 2 to evaluate if the “responders” showed selective improvement in pre/post garden-path recovery performance that was statistically different from that of untrained subjects and “non-responsive” trainees who did not demonstrate gains on  $n$ -back. To do this, I fit the data for ambiguous and unambiguous materials in separate mixed-effects models with Subjects and Items as crossed random effects and both Assessment (1 vs. 2) and Group ( $n$ -back responders vs. non-responders vs. untrained controls) as fixed categorical factors, using the results of the above cluster analysis to determine the levels of each fixed Group factor. Similar to the multiple regression analysis earlier, I ran separate models for ambiguous and unambiguous data because I hypothesized a priori that training-mediated differences should occur only in the high-interference, ambiguous condition, whereas no such effects were expected in the unambiguous condition. Again, a categorical independent variable was used in lieu of a continuous measure of training responsiveness because untrained controls had no



analogous measure of inter-assessment gains (i.e., this group received no contact between assessments and therefore had no training data to contribute).

I first tested if there were any unexpected differences across the three groups in terms of syntactic ambiguity resolution performance at Assessment 1. The best fitting mixed-effects model of Assessment 1 accuracy performance when Group and Sentence-Type were input as fixed factors included only Sentence-Type as a reliable fixed factor ( $z$ -value=8.89,  $p < .001$ ). That Group as a fixed effect did not improve the model fit indicates, importantly, equivalent performance among responders, non-responders, and untrained controls prior to intervention.

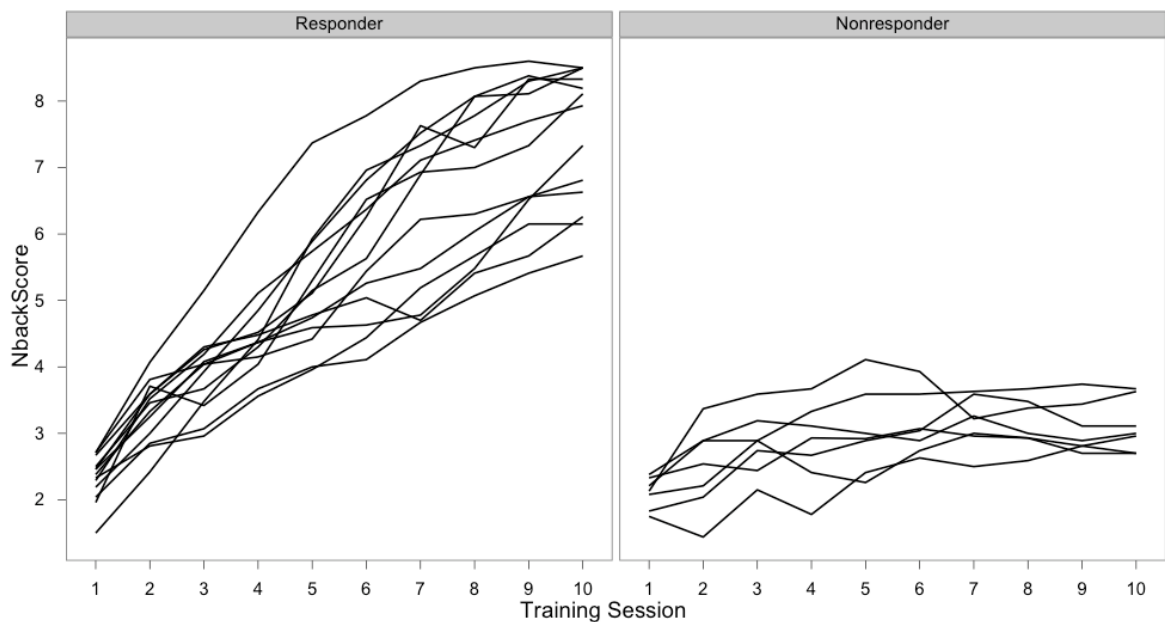


Figure 2. N-back performance curves by training session for responders and non-responders.

When analyzing cross-Assessment changes in accuracy, there were main effects of both Assessment and Group and a significant Group-by-Assessment interaction for the ambiguous ( $ps < 0.05$ ) but not the unambiguous sentences (see shaded panel of Table 3).

To investigate this interaction further, I fit the data for each group separately, crossing Subjects and Items as random effects and including Assessment (1 vs. 2) as a fixed factor. This revealed a significant main effect of Assessment for successfully trained subjects (i.e., *n*-back responders;  $z=-3.68$ ,  $p<0.001$ ), but not for subjects in the untrained control condition ( $z=-1.51$ ,  $p>0.13$ ) or the non-responder subgroup ( $z=0.83$ ,  $p>0.40$ ), such that only the intercept was reliable in the models for these two latter groups.

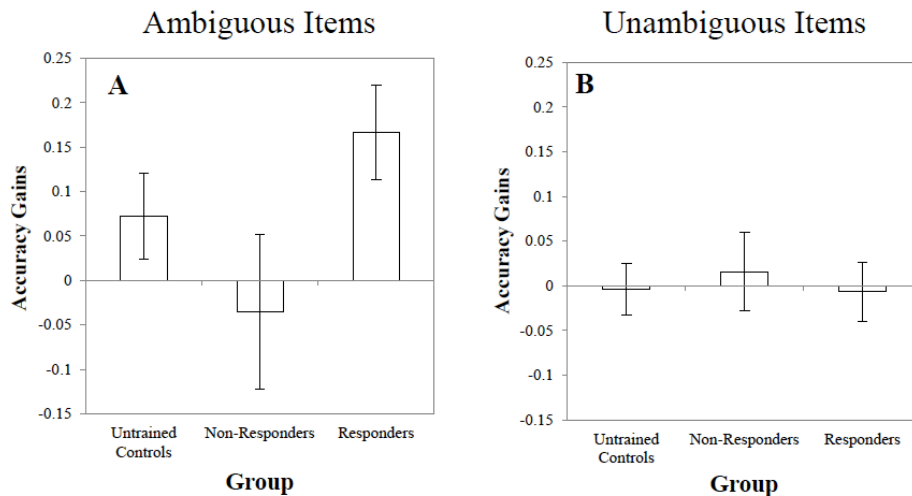


Figure 3. Change from Assessment 1 to Assessment 2 in comprehension accuracy rates split by Group (untrained controls, *n*-back non-responders, and *n*-back responders). The large positive difference score for *n*-back responders (see text) reflects that this subgroup had significantly better accuracy at Assessment 2 than at Assessment 1 for ambiguous items only, an increase that was reliably different from untrained subjects' and non-responders' performance changes (i.e., a Group-by-Assessment interaction). Error bars reflect  $\pm 1$  SEM.

Crucially, there was no such interaction for unambiguous sentences across sessions; indeed, the best-fitting model included only the intercept suggesting that the fixed factors did not account for accuracy patterns in unambiguous items (see shaded panel of Table 3). Figure 3 illustrates the magnitude of accuracy change across assessments for each group on ambiguous and unambiguous sentences; as can be seen,

responders' accuracy increases most on comprehension questions following ambiguous sentences ( $M=16.67\%$ ), compared to non-responders ( $M=-3.57\%$ ) and untrained controls ( $M=7.20\%$ ), who do not differ reliably in performance to ambiguous items across assessments. As expected, none of the groups demonstrate a cross-Assessment change in accuracy for unambiguous items; however, it is important to note that there was little room for change in this condition given that accuracy performance was near ceiling at Assessment 1.

Alongside the multiple regression patterns, these findings suggest that there may be important individual differences concerning who may benefit most from training—particularly from the interference resolution functions practiced through the present version of *n*-back (see General Discussion)—and, therefore, who should be expected to demonstrate reliable transfer to untrained measures of syntactic ambiguity resolution (see Jaeggi et al., 2011).

The relation between *n*-back training improvement and garden-path recovery gains is likely due to performance gains on the shared underlying interference resolution process. However, one possible interpretation of the results thus far is that successful trainees are not responding to the *n*-back task in particular, but rather that this subgroup merely enjoys a better capacity to learn generally from experience. Such a sharper ability to learn could, in principle, underlie both *n*-back improvement and greater test-retest improvement on syntactic ambiguity resolution. Although the results of the multiple regression analysis are suggestive against the “better learner” interpretation (because garden-path accuracy improvements depended selectively on individual training gains on the *n*-back task), it is possible that *n*-back responders were also the responders on the

Significant Model Parameters		Beta Estimate	SE	z-value	AIC <sub>C</sub> with / without slopes
<b>N-Back</b>					
<i>Ambiguous</i>	Intercept	0.79	0.31	2.59*	1170.38 / 1148.78
	Assessment	-0.44	0.20	-2.16*	
	Group	1.05	0.51	2.07*	
	Assessment x Group (Responders)	-0.73	0.37	-1.98*	
<i>Unambiguous</i>	Intercept	2.82	0.32	8.92***	609.19 / 589.42
<b>LNS</b>					
<i>Ambiguous</i>	Intercept	0.78	0.30	2.58**	8459.38/ 8496.18
	Assessment	-0.42	0.20	-2.06*	
<i>Unambiguous</i>	Intercept	2.82	0.31	8.96***	13358.1 / 13414.1
<b>Running Span</b>					
<i>Ambiguous</i>	Intercept	0.78	0.30	2.58**	8459.08 / 8492.98
	Assessment	-0.42	0.20	-2.05*	
<i>Unambiguous</i>	Intercept	2.81	0.31	9.047***	13363.1 / 13430.1
	Group	1.27	0.57	2.23*	
<b>Block Span</b>					
<i>Ambiguous</i>	Intercept	0.78	0.31	2.56*	7987.59/ 8005.69
	Assessment	-0.42	0.20	-2.08*	
<i>Unambiguous</i>	Intercept	2.81	0.32	8.89***	12787.1 / 12796.1

Table 3. Significant fixed effects from the best fitting mixed-effects models of comprehension accuracy data, testing for an Assessment (1 vs. 2) by Group (responders vs. non-responders vs. untrained controls) interaction separately for ambiguous and unambiguous items on each of the four training tasks. When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and Items were input into the model as crossed random effects. Excluding random slopes yielded better fits of every model, as indexed by lower AIC<sub>C</sub> values for models without random slopes as compared to those with random slopes. Thus the best-fitting models *without* random slopes are reported here. \*p<.05, \*\*p<.01, \*\*\*p<.001

three other training task but that those tasks did not permit sufficient variation in performance increases to observe any transfer effects.

To evaluate this possibility, I conducted additional cluster analyses to identify responders and non-responders on LNS, Running Span, and Block Span. The results showed that there was in fact significant performance variability on these three training

tasks, but that *n*-back responders did not necessarily also respond well to them, indicating that this subgroup was not selected for being better learners in general. Moreover, entering this responsiveness variable for the three other training tasks into mixed-effects models revealed that, as expected, increases on those tasks did not result in a Group-by-Assessment interaction regarding garden-path-recovery improvements, patterning with the non-significant contributions of each in the multiple regression model.

**Responsiveness to Other Training Tasks: Additional Cluster Analyses and Mixed-Effects Models.** I identified responders and non-responders to the three other training tasks (LNS, Block Span, and Running Span) using the same model-based cluster analysis outlined previously for *n*-back. The goal of this analysis was twofold: (i) to determine if the group of responders identified for the *n*-back task *necessarily* comprises the same individuals who responded well to the other training tasks; and (ii) to test whether responders on the other training tasks demonstrated a reliably greater improvement in sentence re-interpretation ability across assessments as compared to non-responders and untrained subjects. The second goal is particularly critical when entertaining a process-specific account of the present results, such that other tasks not designed to tap the EF of interest (interference resolution) should confer little pre/post benefit to syntactic reanalysis, or resolution of incompatible interpretations.

A model-based cluster analysis yielded only a single cluster for the LNS and Running Span tasks ( $N_s=21$ ), and three separate clusters for the Block Span task (non-responder group A:  $n=2$ ; non-responder group B:  $n=8$ ; responder group:  $n=9$ ). Because the model-based cluster analyses of LNS and Running Span gains did not reveal distinct groups of subjects, I employed an alternative cluster-analytic method whereby the model

must create a specific number of distinct clusters by maximizing the distance between them, such that the most similarly performing individuals coalesce within a cluster. Since I aimed to identify two broad clusters of individuals (responders and non-responders), I used a two-component clustering approach. The bottom three panels of Table 2 show that, by and large, the responder/non-responder groups did not differ in performance scores at the first training session (except for Block Span), but did reliably diverge in terms of average training performance, final session score, training gains, and performance slope. Together this indicates that the two-component clustering approach consistently defined two groups of subjects for each task that differed significantly on various measures of training responsiveness.<sup>6</sup>

To address whether *n*-back responders were also classified as “responders” on other tasks in the training regimen, I tallied the number of training tasks on which each participant was considered a responder. Figure 4 illustrates trainees’ propensity for general “responder” status given their *n*-back performance. Firstly, it is important to note in the figure that, indeed, the clusters of responders and non-responders identified across the four training tasks *do not systematically overlap*; that is, subjects showing improvements on the *n*-back training task may not have performed well on the other training tasks. In fact, most *n*-back responders (54%) were considered responders on only *one* of the other training tasks (out of three), and *none* of the *n*-back responders were responders on *all* training tasks (Figure 4). Moreover, three *n*-back responders (23%)

---

<sup>6</sup> Note that when applying this forced two-cluster approach to *n*-back performance data, the model identified the same individuals as responders—*n*=13—and non-responders—*n*=7—as previously categorized, replicating the model-based approach sketched in the main text above. This was also true for Block Span, except that the two subjects identified in the lowest-performing group were clustered here with the other non-responders into a single non-responder group, resulting in the following two clusters: responders—*n*=9—and non-responders—*n*=10.

were actually *non*-responders on LNS, Block Span, and Running Span. This pattern suggests that the *n*-back / garden-path recovery relation does not simply index a superior ability to learn, reflected commonly across improved performance on these two tasks. The reason is that one would expect those with a better or faster capacity to learn to demonstrate this capacity across *all* tasks. Instead, as hypothesized and demonstrated below, the relationship observed between *n*-back gains and a significantly improved ability to resolve temporary syntactic ambiguity was selective, which I believe reflects a positive response to practicing the EF functions common to *n*-back and garden-path recovery, rather than a sharper capacity to learn in general.

Secondly, I conducted multilevel mixed-effects models to test whether responders to the three other training tasks that were developed (which tapped different EFs than *n*-back by design; see Method and General Discussion) showed reliably better garden-path-recovery improvement across assessments than untrained controls and non-responders to those tasks. If not, then this would confirm the selectivity of—and suggest a special status for—*n*-back training in terms of its ability to tap and improve those domain-general EFs that are shared with sentence re-interpretation. As before, I fit the data for ambiguous and unambiguous materials separately with Subjects and Items as crossed random effects and entered both Assessment (1 vs. 2) and Group (responders vs. non-responders vs. untrained controls) as fixed categorical variables for each of the three other training tasks, as identified by the various task-specific two-component cluster analyses (Note: First, I tested if the three groups (responders, non-responders, controls) for each of the remaining three training tasks differed in syntactic ambiguity resolution performance at Assessment 1. Critically, in mixed-effect models that tested for a Group-by-Sentence-Type interaction

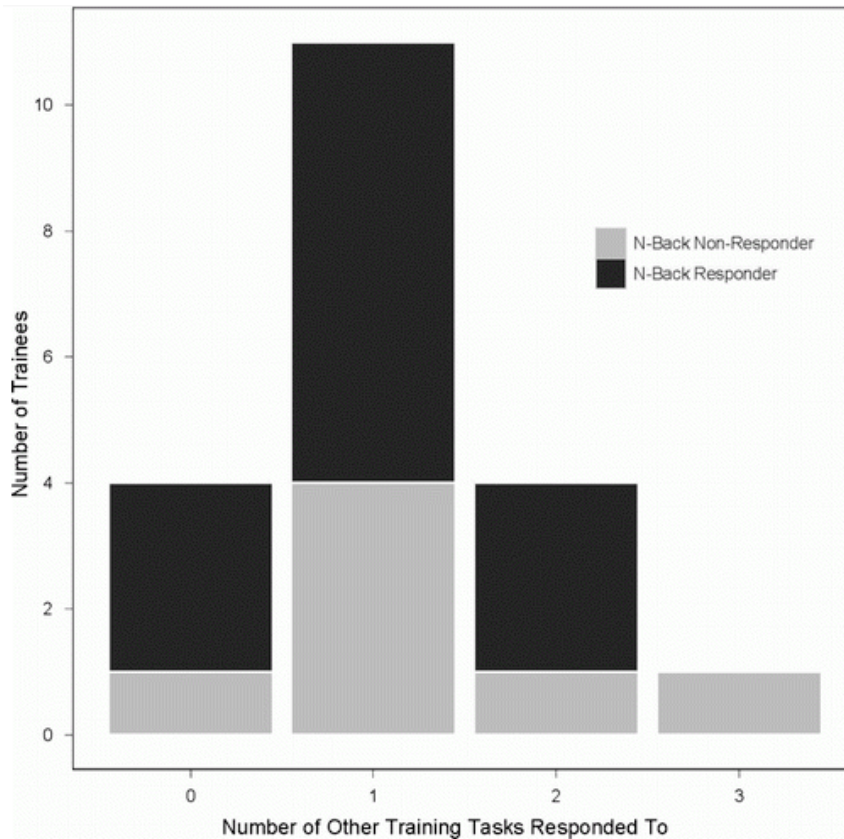


Figure 4. Responsiveness on the three other training tasks assuming n-back responsiveness. Responsiveness to each training task was defined by the output of a two-component cluster analysis (see text). Subjects varied in their ability to improve consistently across the other training tasks, irrespective of n-back gains, suggesting that n-back responders are not necessarily general-purpose learners. The x-axis depicts the number of other, non-n-back tasks that a trainee responded to based upon n-back responsiveness. It does not depict the total number of tasks to which a trainee responded. The “0” column, for example, shows that three n-back responders responded to 0 other tasks, and (only) 1 n-back non-responder did not respond to anything else.

at Assessment 1 (similar to what is reported above for the *n*-back task), only Sentence-Type improved the fit of each model ( $ps < 0.001$ ). That the Group factor was absent from each best-fitting model ( $ps > 0.26$ ) indicates no differences in garden-path recovery between responders, non-responders, and controls for LNS, Running Span, and Block Span prior to intervention).



As can be seen in the three lower panels of Table 3, there was a reliable fixed effect of Assessment for ambiguous items for the three other training tasks (LNS, Block Span, and Running Span), but no Group-by-Assessment interactions. Moreover, the best fitting models for unambiguous sentences included only the intercept for all tasks with the exception of Running Span (wherein non-responders differed from untrained controls and responders on accuracy to unambiguous sentences). Furthermore, as mentioned earlier, treating gains on these tasks as continuous variables in a multiple regression analysis—instead of forcing two categorical clusters of subjects—revealed that no task apart from *n*-back reliably predicted a relationship between garden-path gains and training gains. Taken together, the mixed-effects models and regression analysis may further reveal the importance of *n*-back training with lures, due to the interference resolution processes that are common to resolving temporary syntactic ambiguities.

Although I find this selectivity for *n*-back with lures, it is likely that this training task includes other working memory and executive functions besides interference resolution that are also shared with garden-path recovery, such as attention maintenance and memory updating. I merely wish to highlight that, whereas the other training tasks also involve attention maintenance and memory updating, *n*-back is the only task designed to target interference resolution, which is why it is the task of interest. This is not intended to imply that *n*-back with lures recruits no other cognitive processes of relevance. Moreover, despite the lack of transfer from the three other training tasks, I cannot exclude the possibility that they did not contribute anything to the observed transfer effects. I return to these important issues in the General Discussion.

#### 2.4.2 Real-time Reanalysis (Eye Movements)

**Analysis.** Changes in eye-movement patterns were selective and demonstrated better real-time reanalysis of temporary ambiguities post-training, corroborating and extending the patterns observed for changes in accuracy with respect to  $n$ -back responders. My primary reading measure of interest was regression-path time, which reflects the total time individuals take to read past a particular region, beginning with the eyes' first entry into that region from the left, until exiting that region rightward (see, e.g., Sturt, Scheepers, & Pickering, 2002; Stewart, Pickering, & Sturt, 2004). This measure considers leftward eye movements after encountering a region, when readers regress to reread earlier information, before moving on. Regression-path time thus reveals reading behavior directly after a reader's first encounter with a particular region.

As Stewart and colleagues argue, regression-path reading time is a valuable gauge of processing difficulty, perhaps even better than first-pass times. The reason is that readers frequently fixate a region before instantly regressing leftward; this initial fixation may therefore be short, resulting in a measurable but necessarily small first-pass cost, despite readers' experience of uncertainty or confusion (Stewart et al., 2004). When this occurs, significant evidence of processing difficulty should materialize in regression-path times, as this measure is responsive to both the length and frequency of regressions, thereby indexing revision cost and reanalysis (Sturt et al., 2002). Overall, this measure allows one to account for how long it takes a reader to pass a region of conflict (see below), and to determine how the associated processing difficulty changes as a function of training responsiveness. In other words, does the time-course of responders' reading

behavior improve (i.e., reduce in duration) immediately following a first entry into a region of conflict?

Because regression-path time is susceptible to exaggeration from eye movements to the left side of the screen—for instance, in preparation for a subsequent comprehension question—I truncated analysis at the final word of the sentence for any trial during which participants launched a leftward eye movement from that point, but did not then launch any rightward eye movements to continue re-reading the sentence. As my disambiguating region was always sentence-final (see Table 4, which defines the sentence regions), this truncation method was implemented to exclude extraneous regressions that were not associated with returning to later regions to continue processing the sentence.

<b>Sentence Type</b>	<b>Region 1</b>	<b>Region 2</b>	<b>Region 3</b>	<b>Region 4</b>
<i>Ambiguous</i>	While the thief hid	the jewelry	that was elegant and expensive	sparkled brightly.
<i>Unambiguous</i>	The jewelry	that was elegant and expensive	sparkled brightly	while the thief hid.

Table 4. Reflexive absolute transitive (garden-path) sentences were divided into four regions for fine-grain analysis. Note that for ambiguous items, region 4 is the critical disambiguating region whereas region 3 is the critical comparison region in unambiguous items, as it contains the same content. However, region 4 of unambiguous items (the final region) is also discussed (see text) because it occurs in the same sentence position as the critical disambiguating region of ambiguous items.

Given the evidence thus far that the *n*-back training task was the only one to yield performance differences that resulted in the relevant Group-by-Assessment interaction for the accuracy data, I mirrored the mixed-effects analysis for the eye-movement data to determine if regression-path durations patterned similarly. In other words, I report an analysis that compares reading behavior from the performance subgroups on the *n*-back task, responders and non-responders, to the untrained controls. As with the accuracy data,

I expected that responders' reading latencies (following entry into a conflict region of ambiguous items) would shorten reliably, whereas the other groups' reading behavior would remain unchanged (the critical Group-by-Assessment interaction). Finally, responders to the three non-*n*-back tasks should demonstrate no significant change relative to non-responders and controls, concomitant with the accuracy findings.

I conducted analyses on *correct trials only* as a means of measuring eye-movement patterns during successful garden-path recovery, that is, when one would expect readers to make leftward saccades in search of information to help them revise. Similar to the analyses I conducted for accuracy data, a multilevel mixed-effects model was used to fit the data for ambiguous and unambiguous materials separately with Subjects and Items as crossed random effects. Both Assessment (1 vs. 2) and Group (responders vs. non-responders vs. untrained controls) were included as potential fixed factors, with Group levels being defined by the results of the separate cluster analyses reported earlier for the four training tasks.

Baayen and colleagues (2008) argue that Markov Chain Monte Carlo (MCMC) simulations are useful for understanding the effects of each fixed parameter within mixed-effects models of continuous data, like the current regression-path-time data, because they handle missing data points well and provide numerical estimates of parameters that can be compared to those of a standard linear model. I analyzed cross-Assessment changes associated with entering the disambiguating region (e.g., “sparkled brightly”) of ambiguous sentences first because this is the only region expected to trigger interference resolution functions, given the introduction of new evidence that is incompatible with a reader's prior interpretation (Novick et al., 2005).

## Relating Real-time Processing Changes to Changes in Training

**Responsiveness: Mixed-Effects Models.** Table 5 shows the results of MCMC simulations<sup>7</sup> for all mixed-effects models that fit the total regression-path data from Region 4, which is the disambiguating region in ambiguous sentences. In unambiguous sentences, Region 4 was examined as a comparison, to match the region of analysis to the position of the critical region in ambiguous sentences, which necessarily contains different semantic content.

For ambiguous items, the model that included *n*-back responsiveness and Assessment as fixed effects (shaded panel of Table 5) revealed that an Assessment-by-Group interaction emerged ( $t=2.55$ ,  $p<0.05$ ), such that only *n*-back responders spent reliably less time passing this region at Assessment 2 versus Assessment 1 (a difference of 640 ms), as compared to *n*-back non-responders (a 41-ms difference) and untrained subjects (a 57-ms difference) (see also Figure 5).

Additionally, in the comparable model of unambiguous items—i.e., for the final region (Region 4) of an unambiguous construction—no reliable fixed effects or interaction terms emerged ( $p>0.37$  for the Group-by-Assessment interaction). Moreover, only the intercept was included for the other models that examined regression-path data launched from *all* other regions of both ambiguous *and* unambiguous sentences, including Region 3 of unambiguous sentences ( $ps>0.17$ ), which contained the same semantic content as the critical region of ambiguous sentences (e.g., “sparkled brightly”).

---

<sup>7</sup> I used R’s `mcmc` and `pvals.func` functions to perform all MCMC simulations to assess the significance of fixed effects of the mixed-effects models (MCMCpack library, Martin, Quinn, & Park, 2009; languageR library, Baayen, 2010)

(We did not create a table for the results of the mixed-effects models for the data in all other regions, but see Figure 5). Together this pattern indicates that the Group-by-

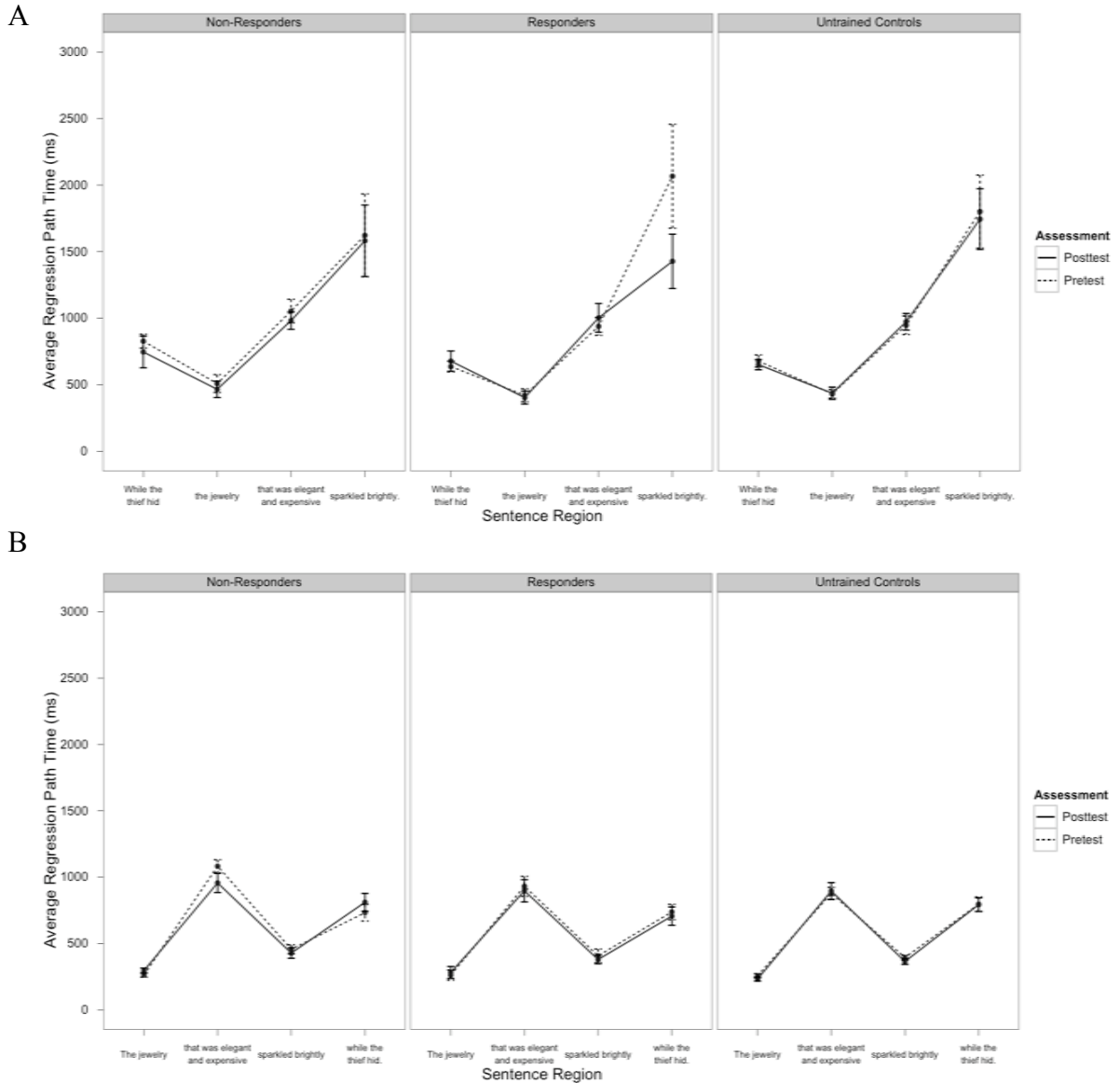


Figure 5. N-back responders', non-responders', and untrained controls' regression-path times across assessments launched from each sentence region for ambiguous (A) and unambiguous items (B). To remove the potential impact of outliers, I eliminated from analysis any per-region regression-path times that fell 2.5 standard deviations above or below a subject's mean across all conditions. (This Winsorization procedure affected less than 2.4% of the overall data across all regions and sentence types.) Only correct trials were analysed (see text). Error bars reflect  $\pm 1$  SEM.

Assessment interaction was selective for regression-path times stemming from the disambiguating region of ambiguous items, the region where interference resolution and controlled revision processes are hypothesized to engage. Importantly, at Assessment 1, mixed-effects models with Group as a fixed factor contained only the intercept as a reliable term for Region 4 of both ambiguous ( $t=7.08, p<0.001$ ) and unambiguous sentences ( $t=13.97, p<.001$ ). That Group as a fixed effect did not significantly improve model fit suggests equivalent regression-path times across groups prior to training in this critical region (as well as in all other regions of both sentence types:  $ps>.39$ ).

To further investigate the Group-by-Assessment interaction for ambiguous items, I fit the data for each group individually with Subjects and Items as crossed random effects and Assessment (1 vs. 2) as a fixed factor. This revealed a main effect of Assessment for responsive *n*-back trainees ( $t=4.27; p<0.001$ ), due to a drop in regression-path time associated with reading behavior post-entry into the disambiguating region (and only the disambiguating region) across assessments; in contrast, only the intercept was reliable in models testing for the effect of Assessment in the disambiguating region among the untrained controls ( $t=1.69; p>0.09$ ) and trainees who did not demonstrate improvement on the *n*-back task ( $t=0.31; p>0.75$ ), suggesting that neither group showed reduced times at Assessment 2.

#### **Responsiveness to Other Training Tasks: Additional Mixed-Effects Models.**

As can be seen in Table 5, only the intercept was included in the comparable models evaluating performance of responders, non-responders, and untrained controls on LNS, Running Span, and Block Span for both ambiguous and unambiguous data. Thus, the Group-by-Assessment interaction failed to emerge for responders on the other training

tasks, again indicating selective improvement for *n*-back responders and corroborating the interaction pattern I observed for the accuracy data (see three lower panels of Table 4). Said another way, the responders to each of the three other training tasks did not demonstrate any reading-time changes associated with the disambiguating region (or any other region) across assessments that were significantly different from the non-responders and untrained control subjects.

That the regression-path time of trainees who successfully improved on *n*-back was shorter following entry into the disambiguating region at Assessment 2 suggests that they had less difficulty recovering from confusion upon encountering late-arriving input that conflicts with their developing interpretation. Specifically, the time spent returning to earlier regions after encountering conflict—to obtain other evidence to facilitate revision—decreases after training as compared to before training. Again, no differences were found across assessments in any region of unambiguous items for any group, as expected.

Because regression-path analyses were conducted on correct trials only, the pattern of results suggests that when *n*-back responders arrive at the correct interpretation, they are, as a group, doing so in less time than they did before training, decreasing the duration associated with regressing out of the region of conflict and eventually reading past it. *N*-back responders' improved accuracy and shorter regression-path times after encountering disambiguating evidence may reflect better controlled revision following training: when confronted with new evidence that conflicts with developing interpretations, readers who undergo EF training (specifically, those who respond well to training on the *n*-back-with-lures task) spend less time regressing to



earlier material in order to recover successfully from their misanalysis, effectively gathering information more quickly to arrive at the correct sentence meaning. Although it

Significant Model Parameters		Beta Estimate	SE	t-value	AIC <sub>C</sub> with / without slopes
N-Back					
<i>Ambiguous</i>	Intercept	1675.24	219.12	7.645***	7398/7376.3
	Assessment x Group (Responders)	616.70	241.44	2.554*	
<i>Unambiguous</i>	Intercept	809.102	55.397	14.606***	12832.1/12812.1
LNS					
<i>Ambiguous</i>	Intercept	1656.7	230.7	7.182***	4451.93/4480.43
<i>Unambiguous</i>	Intercept	804.890	58.918	13.661***	7906.97/7932.67
Running Span					
<i>Ambiguous</i>	Intercept	1664.2	223.6	7.442***	5303.18/5328.88
<i>Unambiguous</i>	Intercept	802.136	55.406	14.478***	9386.74/9417.14
Block Span					
<i>Ambiguous</i>	Intercept	1650.91	218.38	7.56***	5544.56/5570.06
<i>Unambiguous</i>	Intercept	809.461	55.231	14.656***	9861.13/9884.73

Table 5. Significant fixed effects from the best fitting mixed-effects models of regression-path time following entry into the final region of each sentence testing for an Assessment (1 vs. 2) by Group (Task Responders vs. Non-responders vs. Untrained Controls) interaction separately for Ambiguous and Unambiguous items for each of the four in-house training tasks. Markov Chain Monte-Carlo (MCMC) simulations were conducted to test for the significance of each fixed effect, through which I generated 10,000 samples from the posterior distribution. When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and Items were input into the model as crossed random effects. Excluding random slopes yielded better fits of every model, as indexed by lower AIC<sub>C</sub> values for models without random slopes as compared to those with random slopes. Thus, the best-fitting models *without* slopes are reported here. \*p<.05, \*\*p<.01, \*\*\*p<.001

would be difficult, in my opinion, to attribute *n*-back responders' cross-Assessment changes launched from the disambiguating region to a better capacity to learn in general—because the changes occur following entry into the sentence region where information re-characterization is precisely hypothesized to deploy—this learning

interpretation is further discounted by the specificity of the interaction to *n*-back responders alone. This again suggests that selective improvement in reading behavior following entry into the disambiguating region is likely attributable to a positive response to consistent training on the *n*-back task, which shares EF processes with syntactic ambiguity resolution (see Discussion).

Finally, given that the disambiguating region is always the final region of the ambiguous experimental sentences, another potential explanation for the present regression-path-time findings is that training attenuates “wrap-up” effects. These are typically marked by longer sentence-final or clause-final reading times and have been attributed to clausal integration rather than reanalysis (Just & Carpenter, 1980; Rayner, Kambe, & Duffy, 2000; but see Warren, White, & Reichle, 2009 who argue that construction difficulty does not drive such effects). I acknowledge that the eye-movement patterns associated with the final region might be affected by wrap-up as well as reanalysis. Thus, to identify the locus of these reading time effects, I unpacked regression-path time into its three critical components for subsequent analysis: (1) time spent in the region before regressing out, (2) time spent outside of the region before returning, and (3) time spend in the region upon re-entry. If the regression-path time findings merely capture wrap-up effects (i.e., clausal integration is improved with training), one would expect an effect on just the measure indexing time spent upon reentering the region. However, if training mediates eye movements associated with reanalysis, one might hypothesize a change in the second component, when the reader is prompted to return to earlier regions to gather information after encountering a conflict. Table 6 provides the significant fixed factors associated with the best fitting linear mixed-

effects when Assessment and Group are entered into the model for each of the three components of regression-path time. Out-of-region regression-path time—the time spent

Significant Model Parameters		Beta Estimate	SE	t-value
<b>Initial Fixation Time</b>				
<i>Ambiguous</i>	Intercept	267.43	8.79	30.43***
<i>Unambiguous</i>	Intercept	239.16	6.96	34.39***
	Group (Responders)	-24.320	11.249	-2.16*
<b>Out-of Region Re-Reading Time</b>				
<i>Ambiguous</i>	Intercept	1576.2	183.2	8.6***
	Assessment x Group (Responders)	634.7	297.7	2.13*
<i>Unambiguous</i>	Intercept	694.23	111.56	6.22***
<b>In-Region Re-Reading Regression-Path Time</b>				
<i>Ambiguous</i>	Intercept	812.04	59.75	13.59***
<i>Unambiguous</i>	Intercept	709.78	46.41	15.29***

Table 6. Summary of the reliable fixed effects from the mixed-effects models of Group-by-Assessment for the three components of regression-path time in the disambiguating region (“sparkled brightly”) using responder/non-responder groups defined by a two-component clustering approach for each training task. Initial fixation time reflects the time spent in the disambiguating region prior to launching a leftward fixation outside of the region. Out-of-region regression path time captures the time spent reading earlier regions after leaving the disambiguating region prior to returning back to it. In-region re-reading regression path time measures the amount of time spent in the disambiguating region after visiting earlier regions.

re-reading earlier regions before revisiting the region of interest—bore a significant interaction such that responders show a reliable decrease from Assessment 1 to 2, while non-responders and untrained controls show no change. Initial fixation and in-region re-reading time showed no reliable interactions, indicating that the component driving the abovementioned regression-path time effect was out-of-region reading time. This finding is consistent with an interpretation favoring reanalysis (as opposed to clausal integration) as the process that improved following cognitive control training; therefore, even though the region containing conflict is sentence-final, wrap-up measures do not appear to be

mediated by training when regression-path time is decomposed into its critical components.

Moreover, even if the eye-movement patterns are affected by sentence-final wrap-up, the critical results are specifically related to ambiguity, as the Group-by-Assessment interaction was not found for the sentence-final region of the unambiguous items (see text above as well as Table 5 and Figure 5). In other words, if *n*-back training affected readers' ability to initiate sentence-final wrap-up, as opposed to their ability to deal with temporary ambiguity, then presumably such effects would have been found in unambiguous items as well.

## **2.5 Discussion of Experiment 1**

### *2.5.1 Summary*

I ascribe *n*-back responders' improved sentence reinterpretation to domain-general benefits of increased interference-resolution abilities through training. The reading task completed at both assessments is ostensibly different from the *n*-back task completed during intervention. Nevertheless, recovering correct interpretations following misanalysis relies on broad EFs that are shared across certain task types.

Overall, I provide further supporting evidence that syntactic ambiguity resolution depends on domain-general cognitive control mechanisms, even in non-clinical populations. These findings are a particularly strong addition to this line of research because most of the previous evidence has merely correlated neural activation patterns in response to cognitive control and garden-path tasks, or depended on pre-existing neuropathology to demonstrate a common deficit in syntactic and non-syntactic cognitive control abilities. Here, I directly manipulated conflict-resolution through training to test

its effect on syntactic ambiguity resolution processes in neurologically intact adults. For those subjects in whom I successfully increased interference resolution and cognitive control, I observed improvements in online and offline measures of garden-path recovery. For the first time, I show that garden-path recovery abilities are malleable in healthy young adults such that they can be improved by training the underlying EFs critical for revising misinterpretations, not just via practice with specific instances of ambiguous sentences. Notably, as hypothesized, performance increases on working memory tasks without the critical interference resolution feature are apparently insufficient to produce gains in syntactic ambiguity resolution (but may be necessary; see limitations and caveats below). The increases in syntactic ambiguity resolution performance are closely related to *n*-back gains, importantly, and not due to preexisting differences in either garden-path recovery ability (all three groups performed equivalently at Assessment 1) or *n*-back ability (responders and non-responders performed equivalently at the first session of *n*-back training). Finally, the present results speak to the role of regressive eye-movements in garden-path recovery, corroborating the notion that the regression-path measure in eye tracking may reflect the initiation of revision processes during real-time processing of syntactic ambiguity. Unique in this contribution is the result that readers may become less dependent on re-reading upon encountering conflicting evidence as their cognitive control becomes more efficient.

In particular, I demonstrated that subjects in the training group who achieved the greatest gains on the *n*-back training task, but not the other training tasks, subsequently showed better success at Assessment 2 (versus Assessment 1) in recovering the correct alternative interpretation of temporarily ambiguous sentences susceptible to misanalysis.

Furthermore, compared to the untrained group, and those in the training group who did not demonstrate consistent gains on *n*-back, the cross-Assessment performance change was reliably larger for the most successful trainees (responders), suggesting significantly increased accuracy to ambiguous items following the training regimen.

Equally compelling, this finding was accompanied by selectively shorter regression-path times launched from disambiguating regions where interference-resolution processes are hypothesized to deploy. Together these findings suggest that syntactic ambiguity resolution is a plastic cognitive skill that may be adaptable by training regulatory functions common to syntactic and non-syntactic measures. Notably, sentence-reinterpretation accuracy improved, and regression-path time decreased, for successful *n*-back trainees only under ambiguous conditions, when readers had to adjust processing to initiate recovery. Because I hypothesized that EF training would transfer only to tasks requiring common underlying EF mechanisms, no training-related changes were expected—and none were found—under low EF demands, namely when reinterpretation was unnecessary and thus did not prompt recovery processes to initiate (i.e., in unambiguous conditions).

I have noted several times the selective nature of the presented training results, specifically that transfer benefits are observed (1) only for ambiguous sentences when examining comprehension accuracy and (2) only where revision-processes are expected to be triggered when looking at regression-path time. Of course, with error rates of approximately 10% at Assessment 1 for unambiguous items, one could argue quite reasonably that such high performance was near ceiling (i.e., 90% accuracy), and therefore no change would be anticipated—or perhaps even possible—at Assessment 2.

Moreover, although changes in eye-movement patterns following training occurred selectively after entering the disambiguating region of ambiguous sentences, it could be argued that training mediated the efficiency of handling processing difficulty, which would be greater in ambiguous than unambiguous sentences, rather than reinterpretation abilities per se. I must therefore be cautious in concluding that selective improvement in interference resolution was the reason that the present training regimen benefitted comprehension and real-time processing of ambiguous but not unambiguous sentences.

However, there are several reasons leading me to believe that the selectivity of the results is not due to ceiling effects. Firstly, closer inspection of the error rates for unambiguous items at Assessment 2 shows some degree of individual variability. Average error proportions ranged from 0.059 to 0.11 across groups, with non-responders having the lowest error proportions after training, suggesting that individuals vary in their performance for these items at least nominally and may have some room to improve. Although one could potentially make the “ceiling” argument against the regression-path time results, it is nevertheless somewhat difficult to evaluate what ceiling performance might be. Certainly readers are bound to a lower reading-time limit by how fast their eyes can move. But, judging by the sometimes-long regression-path times in Figure 4 (in both ambiguous *and* unambiguous sentences, depending on the region), it is uncertain that readers could not have read anything in less time at Assessment 2. Indeed, the figure reveals that aside from “sparkled brightly” (the disambiguating region in ambiguous sentences), reading times were quite similar between ambiguous and unambiguous sentences, when considering regions with the same content (as opposed to regions with the same position in the sentence). Notably, average regression-path times in the

adjectival clause (e.g., “that was elegant and expensive”) were as long as one second in both sentence types, which seems to allow ample room for improvement, especially given that regression-path time is computed only when subjects regress out of the region, which need not occur at all. I believe, consequently, that it is rather informative that there were no test-retest differences in any region of unambiguous items regarding regression-path times for any group. It is equally informative that there were no test-retest differences in any region *except* following entry into the disambiguating region for *n*-back responders. Together, I believe these data patterns suggest an element of sensitivity and selectivity, such that performance improves only in regions associated with syntactic conflict.

### *2.5.2 Using a Process-Specific Training Approach*

Why was the *n*-back task particularly critical in capturing training and transfer success across both dependent measures of interest, namely (1) accuracy to questions probing for persistent effects of misanalysis and thus a failure to revise; and (2) regression-path times launched selectively from the disambiguating region? One explanation is that *n*-back gains, and only *n*-back gains, were related to garden-path-recovery improvement because of the controlled processing needed to resolve among the conflicting representations generated by interference lures. In a standard *n*-back task, participants can typically depend on familiarity to correctly identify which letter is a target because lures appear merely incidentally, and likely not often enough to warrant not relying on familiarity as a reliable cue. However, the introduction of lures after participants reached a certain performance criterion forced trainees to rein-in such a familiarity bias; when encountering a lure, they instead had to initiate conflict/interference-resolution processes to successfully override familiarity-based



evidence and re-characterize the stimulus as familiar but not in the relevant *n*-back location. Prior work has highlighted such an information re-characterization function as crucial for resolving syntactic conflict as well: during parsing, domain-general interference-resolution processes engage when individuals encounter input (e.g., “sparkled brightly”) that is incompatible with their developing analysis (see January et al., 2009; Novick et al., 2005; 2009; Novick et al., 2010). When a reader comes across such conflicting evidence and discovers the misinterpretation, he or she must “slam on the brakes” and deploy interference resolution processes that allow for a re-characterization of the current representation of sentence meaning, and for finding the correct, intended alternative.

Moreover, the *n*-back task has been shown to recruit posterior regions of the left inferior frontal gyrus (LIFG) within VLPFC during high-EF (lure) trials (Gray et al., 2003); this patch of cortex is routinely identified as the crucial neural underpinning of conflict/interference resolution in working memory (see Jonides & Nee, 2006) and has been implicated in cognitive control during sentence reinterpretation in both patient and neuroimaging studies (January et al., 2009; Novick et al., 2005; 2009; 2010; Ye & Zhou, 2009). In fact, the lure version of the *n*-back task is quite reminiscent of the working memory assessment—the ‘recent probes’ item recognition task—that was used to diagnose a interference-resolution impairment in the patient with VLPFC (indeed, LIFG) damage described in the Introduction (see also Hamilton & Martin, 2005). This patient’s deficit extended to a failure to override syntactic misanalysis and recovery from misinterpretation (Novick et al., 2009). In the recent-probes task, subjects responded to a probe (e.g., *D*) regarding whether it appeared in an immediately prior memory set (e.g., *s*

*f d m*) (see also Jonides et al., 1998; Monsell, 1978; Thompson-Schill et al., 2002).

Although subjects could frequently use stimulus familiarity to judge correctly—yes or no—whether the probe had appeared or not, a small subclass of ‘no’ trials introduced conflict and therefore susceptibility to error if one relied on a familiarity bias alone. On such ‘conflict’ trials, the probe (e.g., *H*) did not appear in the directly preceding memory set (e.g., *k p w n*), so the correct response was ‘no,’ but it had been seen one trial earlier (e.g., *h l w p*). As such, these so-called ‘recent-no’ trials, akin to lures in the current study, exploited lingering familiarity of the probe owing to its recent presentation; subjects therefore had to override a dominant familiarity bias because it might yield an incorrect ‘yes’ response, and instead re-characterize the probe representation as ‘familiar-but-irrelevant.’ The patient’s unusually high error rate under such conditions, compared to ‘non-recent-no’ trials (where there was no interference from the preceding memory set), identified a selective interference resolution impairment, which affected his parsing abilities under similarly circumscribed conditions. In particular, he demonstrated a failure to revise (or re-characterize) early parsing misanalyses and recover an alternative interpretation of sentence meaning when there was conflict between two incompatible syntactic representations (Novick et al., 2009; for convergent neuroimaging data see January et al., 2009; Ye & Zhou, 2009; for a review see Novick et al., 2010).

Thus, the linking assumption is that the need for interference resolution seems to be shared across a range of tasks, including garden-path recovery and working memory tasks that manipulate such demands, for instance the ‘recent-no’ trials of the item-recognition task and the *n*-back task with lures. Consequently, subjects in the study who showed consistent performance increases on the *n*-back task—reflecting an enhanced

ability to resolve among conflicting representations—demonstrated concomitant increases in garden-path recovery performance, likely because of the common interference-resolution process that was targeted through training. I reiterate that the other training tasks (except possibly Running Span, see below) were designed explicitly not to tap this function by excluding any manipulation of demands for information re-characterization. Finally, prior research exploring the utility of similar interference-resolution training tasks demonstrates far-transfer to other language measures that tap cognitive control resources supported by the posterior LIFG (e.g., in a transient manipulation of cognitive “fatigue;” see Persson et al., 2007). The illustration of transfer to syntactic ambiguity resolution in the current study may be considered an extension of that finding. Overall, these results are consistent with earlier studies that tie specific language processing abilities to domain-general cognitive control skills that putatively recruit regions within posterior LIFG (though see caveats below for further discussion).

One reason that improved performance on the other (non-*n*-back) training tasks, by contrast, failed to predict improvements in garden-path recovery may be because they did not, in their design, expressly involve conflict/interference resolution. The EFs tapped in these tasks included the manipulation and storage of visual-spatial information (Block Span) and the reorganization of alphanumeric stimuli in working memory (LNS), which may be theoretically harder to connect to syntactic processing and recovery from misinterpretation specifically. As well, demands for information re-characterization were intentionally not parametrically or dynamically manipulated in these tasks as they were in *n*-back. Of course, verbal working memory is apt to play a role in garden-path recovery during spoken language comprehension, and even in reading studies that use a moving

window paradigm where readers cannot review the input once it has past (see, for instance, Fedorenko, Gibson, & Rohde, 2006). Future work should test the relative impact of working memory training (including auditory working memory training)—with and without interference-resolution aspects—on garden-path recovery using alternative experimental paradigms, in both the reading and spoken domains. Next, although the LNS task involved reordering verbal information in working memory, which is likely involved in garden-path recovery to some extent, this training task required the repeated application of a specified rule in a predictable manner (always sorting numbers in ascending order and letters alphabetically), which may have been too superficial to involve any deeper re-characterization of representations that is necessary for revising misinterpretations.

It is also worth noting that complex working memory span tasks—a category into which LNS and Block Span both fall—typically correlate only weakly with lure-variants of *n*-back, demonstrating a divergence that may be linked to the cross-task asymmetry in interference-resolution demands (Kane et al., 2007; see also Jaeggi et al., 2008). In other words, not all working memory tasks necessarily share this feature, which might therefore be an important design component to consider in studies aimed at creating process-specific overlap between certain training and transfer measures. Also, to my knowledge, neither LNS nor versions of Block Span have been shown to recruit regions of VLPFC common to syntactic ambiguity resolution and other information-recharacterization tasks outside the parsing domain (e.g., the Stroop, *n*-back, and Recent-No tasks; see Haut, Kuwabara, Leach, & Arias, 2000, for a brain-imaging study of LNS,

which shows activation in orbitofrontal and dorsolateral prefrontal areas, with greater peak activations in the right hemisphere).

One result, however, that may be surprising is that responders on the Running Span training task did not demonstrate reliable garden-path recovery improvements, as this task *can*—depending on design—involve updating and incidental proactive interference from earlier items, processes that rely on regions within left VLPFC (cf. Postle, 2003; Postle et al., 2001). In Running Span (Pollack, Johnson, & Knaff, 1959), a sequence of an unpredictable number of items (e.g., letters) is presented, after which the last  $n$  items must be suddenly recalled. Hockey (1973) showed that presentation rate can dramatically alter the nature of the task (see also Bunting et al., 2006). When item-presentation rate is fast (e.g., 3 items/s), the task is conducive to a lower-effort strategy in which items are passively held until the list ends (i.e., when retrieval from a capacity-limited attentional store can occur). With a slower presentation rate (e.g., 1 item/s), participants can—when they are explicitly instructed—adopt a higher-effort strategy in which working memory is continually updated through rehearsal (Hockey, 1973). Thus, during retrieval in the slow-rate procedure, individuals must resolve interference from earlier stimuli that are concurrently being maintained. Questionable, however, is whether anyone would spontaneously adopt a higher-effort strategy when presentation rates change within an experiment, as in this task: based on Hockey (1973) and Bunting et al. (2006), active updating becomes increasingly difficult and almost impossible at fast presentation rates, such as the 500-ms condition. Although active updating is possible in the 1,000-ms condition, it is unlikely that participants would switch strategies in alternation (see Bunting et al., 2006). The faster rate used in the task may have therefore

eliminated the interference-resolution aspect of the task, explaining why Running Span did not predict garden-path recovery improvements. Moreover, there may be an important distinction here between the *attention control* processes needed for Running Span—in which listeners must rapidly *collect* information from a fleeting sensory memory store—and the *cognitive control* processes needed for *n*-back with lures—in which subjects must *re-characterize* information given conflicting internal representations.

Another explanation is that Running Span did not allow the same continuous improvement as *n*-back: participants, particularly the responders, demonstrated steady gains on *n*-back across all 10 sessions, but not on Running Span. For example, 51% of participants showed *more* improvement between the first and second training sessions of Running Span than between the second and any subsequent session, suggesting that training performance reached asymptote quickly. By contrast, less than 9% of participants (in fact, a subset of only non-responders) showed this pattern for *n*-back; most participants continued to improve throughout the regimen. Taken together, the clearer overlap with conflict/interference resolution and participants' consistent pattern of training gains suggest that *n*-back with lures may be a critical training task for eliciting and evaluating improvements in garden-path recovery.

### 2.5.3 Limitations and Caveats

One limitation of the current study is the absence of an “active” control group that also comes into the lab and practices tasks without the crucial EF elements (namely, interference resolution). Including such a group in future work would address the issue of whether demand characteristics or motivational factors alone are driving the effects. For

instance, it would be informative to test whether those who practice *n*-back without the interference-lure component do not show improvements in garden-path recovery, thus providing greater confidence in the claim that this particular EF is at the heart of the observed increases in trainees' reinterpretation abilities. Such a contrast—*n*-back with lures versus *n*-back without lures—would help isolate the locus of the current training results. Despite this drawback, I note the specificity of the transfer findings to (1) key individuals and (2) key features of the reading task that require EF, specifically interference resolution, which I believe together exclude purely placebo or motivational explanations. Firstly, transfer was observed specifically from *n*-back responders as opposed to non-responders, untrained controls, and other types of responders (i.e., LNS, Block Span, and Running Span responders). Secondly, *n*-back responders' overall accuracy to comprehension questions improved for ambiguous but not unambiguous items, suggesting that the offline effects were restricted to cases when readers had to revise interpretations (i.e., resolve among interfering representations; see Novick et al., 2005). The same pattern held for regression-path times associated with regressions from key conflict regions of ambiguous sentences. Given that the control group and *n*-back non-responders (as well as the responders to the three other training tasks) did not show these effects, I am confident that the observed findings cannot be attributed solely to practice. Indeed, I observed important individual differences in training achievement, which corresponded to successful transfer to syntactic ambiguity resolution: only the *n*-back responders demonstrated significantly greater improvements in both online and offline measures than the untrained subjects, despite non-responders having the same amount of training as those who responded well.

Additionally, the pre/post eye-movement comparison in both ambiguous and unambiguous conditions allowed me to examine whether any training-related changes in real-time reading patterns were restricted to regions requiring interference resolution, or whether *n*-back responders read sentences in less time across all sentence regions. This latter finding would have suggested broad increases in processing speed and reading efficiency, as might be predicted by a motivational account, irrespective of the need to initiate cognitive control when late-arriving evidence conflicts with one's initial interpretation. However, changes in reading-time patterns were much more precise: post-training, *n*-back responders' regression-path durations were shorter after entering the disambiguating region of ambiguous items, where new evidence signaled an incompatibility with the favored transitive analysis. Upon encountering such conflict, responders had an easier time recovering the reflexive interpretation, indexed by spending less time regressing to earlier material from the point of confusion. Importantly though, *n*-back responders' eye-movements, like those of untrained participants and *n*-back non-responders, did not change across assessments after entry into any regions of unambiguous sentences, or non-disambiguating regions of ambiguous sentences. If these improvements were due to other variables such as increased motivation, then trainees would have been expected to improve "across the board," rather than only after entering disambiguating regions of ambiguous sentences. Given the size and specificity of the observed improvement in garden-path recovery, I believe that cross-assessment gains are due to a positive response to EF training, especially conflict/interference resolution training. Because the overall effects of training emerge in parallel across two different measures (accuracy and regression-path time) and changes in reading-time stem from



exactly the expected region, it seems highly unlikely that the results are spurious.

Nevertheless, future studies should run the relevant control conditions sketched above to confirm this interpretation, particularly the role of conflict/interference resolution (see Experiment 2 below).

As noted earlier, the convergent findings that improvements were limited to *n*-back responders essentially rules out the possibility that trainees' enhanced ambiguity resolution was the result of a generally better capacity to learn, which could have, in theory, brought about the common improvements found for *n*-back and garden-path recovery. However, additional analyses confirmed that *n*-back responders did not necessarily respond equally well to the other training tasks (clusters of responders and non-responders did not overlap across training tasks; see Figure 4). Thus, these individuals were not selected on the basis of a generally greater ability to learn or motivation to perform well. In addition, the groups of responders identified for LNS, Block Span, and Running Span did not demonstrate improved interpretation-recovery abilities at Assessment 2 in accuracy or regression-path times compared to those tasks' non-responders and untrained subjects. Thus, I believe that the most parsimonious interpretation of these data is that *n*-back training improved interference-resolution functions, which resulted in improved sentence re-interpretation at Assessment 2. That responders to the three other tasks, moreover, did not outperform non-responders and untrained controls argues strongly against the possibility that mere practice effects are at the heart of the present findings.

Nevertheless, I cannot discount the possibility that the three other working memory training tasks were necessary (though insufficient) components of the training

regimen, as they were administered as part of a training battery (see Chapter 3 for a study specifically designed to isolate *n*-back). Indeed, the lack of transfer from LNS, Block Span, and Running Span might be termed a null effect. Although I have made theoretical arguments for why *n*-back with lures should be necessary (and perhaps sufficient) for contributing transfer, I am unable to say with absolute certainty why transfer from the other tasks may have failed. Future research might address this issue by demonstrating that these other working memory tasks can indeed increase performance on other cognitive and linguistic measures, such that there is a process-specific double dissociation (see, for instance, Dahlin et al., 2008). For example, it would be important to know if two experimental manipulations (e.g., *n*-back with lures training versus LNS or other complex-span training) affect language processing outcomes differentially; if the *n*-back manipulation affects cognitive control and syntactic ambiguity resolution and not, say, working memory span and the processing of other linguistic material, and the complex-span manipulation shows the opposite pattern, then one could make even stronger and more specific claims about attributing the observed effects to cognitive control training in particular. While interference resolution and cognitive control can be theoretically and empirically linked to the observed effects (alongside extant neurocognitive data), follow-up research must further tease apart process-specificity in additional experiments that do not rely on inferences from absent transfer effects (see Hussey & Novick, 2012, for similar arguments).

Having said this, I reiterate a similar point made earlier: I firmly believe that traditional “span” functions such as storage, processing, and maintenance factor into language interpretation irrespective of ambiguity or interference, for instance in spoken

comprehension tasks or in moving-window reading paradigms where the demands for mnemonic properties of working memory are high (see, e.g., Fedorenko et al., 2006). Thus, cognitive control—a non-mnemonic aspect of some working memory tasks—is just one explanation for what is shared across *n*-back with lures and syntactic ambiguity resolution; this does not necessarily preclude the likelihood that other working memory processes involved in *n*-back (e.g., updating) are also affecting performance. This perspective is correspondingly apt if the other non-interference training tasks are contributing *something* to the current results, which again, I cannot rule out entirely. Indeed, I do not claim that cognitive control is the only contributor to the observed findings, but rather that interference-resolution processes are a necessary aspect of the current training-transfer relation, given evidence that such functions are critical to syntactic ambiguity resolution (Novick et al., 2009). Again, I selected *n*-back with lures as the task of interest because, relative to the three other training tasks, it was the only one designed to target interference resolution processes. This is not meant to imply that *n*-back targets *only* interference resolution.

I have relied on parsimony to interpret the data in terms of interference resolution, consistent with neuroscience studies showing an important role for posterior areas within the LIFG in both syntactic and non-syntactic cognitive control (e.g., Novick et al., 2005; January et al., 2009; Jonides & Nee, 2006; Ye & Zhou, 2009). However, an important caveat is that these same interference-responsive VLPFC regions can be involved in a number of other cognitively demanding tasks (see, e.g., Duncan, 2010). In other words, patches of cortex within VLPFC that are recruited for interference resolution are unlikely to be involved *uniquely* in interference resolution aspects of cognitive control. Therefore,

other cognitive functions, as discussed above, cannot be excluded as contributing to training gains on the basis of neuroimaging data alone. In sum, although a preponderance of behavioral and neuroimaging data points to a role for cognitive control in garden-path recovery, this does not mean necessarily that there exists a neural dissociation between interference resolution and other executive functions. It does suggest, however, the importance of using multiple methods to offer converging data for a particular hypothesis. The data certainly fit with converging neurocognitive evidence from both clinical and healthy populations, providing another important approach that yields findings compatible with the cognitive control account. Nevertheless, the caveats outlined here suggest that, until both behavioral and neural dissociations are demonstrated, one can conclude only that the training-transfer results are consistent with the notion that cognitive control is the mediating ability across *n*-back-with-lures and syntactic ambiguity resolution. Other working memory and executive functions certainly remain as possible contributors. To what extent they are contributing is an open empirical question and must be addressed in follow-up research using designs as those sketched above.

Despite the present findings dovetailing with extant results involving developmental, neuropsychological, and healthy populations, altogether supporting a process-specific account of cognitive control for language processing, one alternate explanation for the present pattern involves a statistical artifact that arises as function of individual differences in correlated measures (Dimitrov & Rumrill, 2003; Hays & Hadorn, 1992; Linn & Slinde, 1977; Steiner & Norma, 2006; see also Tidwell, Chrabaszcz, Thomas, Mendoza, & Dougherty, 2013). All process-specific training approaches presume that untrained assessment measures—or conditions therein—that tap

resources common to those trained during the intervention enjoy selective benefits. One potential limitation to this explanation, however, arises when individual differences in trainability are considered, namely, when gains in one task (i.e., a training task) predict improvements on another (i.e., assessment task). There is much *theoretical support* for these analyses: Only participants improving on an underlying skill are expected to benefit on other tasks relying on this skill (see Chein & Morrison, 2010; Jaeggi et al., 2011). By this account, subjects who benefit most from training should demonstrate the greatest pre/post gains for tasks relying on interference-resolution abilities, like garden-path recovery (as I have shown and argued in this chapter).

Nevertheless, analytic approaches capitalizing on individual differences in training success to predict improvements on pre/post measures may be prone to a statistical artifact favoring such relationships. That is, training gains predicting assessment gains necessarily follows whenever there is a correlation between the trained and criterion variables: If variable A is linked to variable B, then as variable A improves, so will variable B (see Tidwell et al., 2013). Thus, individual differences in training provide irrefutable support for a *correlation* between a trained and untrained measure (if gains on training predict gains in performance on another task), but leaves open the question of the presence of a *causal relationship* between variables. Importantly, training designs (by virtue controlled manipulations for a subset of participants) strongly implicate a causal relationship between trained and boosted untrained measures when compared directly to control conditions. However, an analysis that considers individual-differences in trainability (e.g., median splits of trainees or responder/non-responder analyses) might undermine this causal relationship. In these cases, the statistical artifact

wherein correlated variables also show comparable (and related) improvement scores masks the possibility of determining whether process-specific training has occurred.

One way to disentangle these two interpretations—a statistical artifact versus a change in an underlying core cognitive process—is to carefully design a training study that minimally manipulates the process of interest, testing for *main effects* of training group and ignoring individual differences in training performance. In the next chapter (Experiment 2), I adopted this new approach to elucidate the nature of the cross-assessment effects in Experiment 1. That is, if the results of Experiment 1—training on *n*-back confers selective advantages to garden path recovery—are replicated within the context of a carefully planned design, then it may be possible to consider these new findings as a causal extension of Experiment 1.

In light of this caveat, the results discussed thus far are among the first to establish training-related transfer to the linguistic domain, and extend theoretical and empirical work highlighting the role of general-purpose cognitive functions in language processing. Certainly, the more language-specific experience readers have, the better they cope with difficult linguistic input: indeed, two studies have reported that consistent practice reading complex or ambiguous material results in (1) an improved ability to process the constructions that were routinely repeated and (2) a transfer effect, such that practice generalizes to previously unseen difficult constructions (see Long & Prat, 2007; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). I believe that the present findings complement this notion by demonstrating that domain-general cognitive abilities, even for healthy adults, *may be a causal factor* in sentence reinterpretation abilities. This conclusion is especially warranted given that the subjects had minimal exposure to the

ambiguous sentences—just 12 per assessment, embedded within several fillers—which was probably insufficient to produce a reliable practice effect. This also seems a reasonable conclusion alongside the finding that untrained controls and non-responders, who had the same amount of practice, failed to improve reliably across assessments. Moreover, given that EF plays a role in a range of other specific language processing skills including lexical ambiguity resolution, common-ground assessment, and verbal fluency, another implication of the current findings is that EF training, within a well-considered process-specific framework, could result in broader improvements beyond just garden-path recovery. Indeed, this is the goal of Experiment 2.

## Chapter 3: Experiment 2 – Process-Specific Training for Parsing and Non-Parsing Skills

Experiment 1 demonstrated that trainees' improved accuracy and faster regression-path times in disambiguating regions may reflect better controlled revision following training: upon encountering new evidence that is incompatible with developing interpretations, readers who undergo EF training (and those who respond particularly well to interference resolution training on *n*-back-with-lures) spend less time regressing to earlier material in order to recover successfully from their misanalysis before advancing. Improved sentence reinterpretation abilities may be attributed to domain-general benefits of EF training.

One remaining question, however, concerns why the *n*-back task in particular was critical in capturing training and transfer success. I reasoned that *n*-back predicted garden-path-recovery improvement because of the controlled processing needed to resolve among the conflicting representations generated by interference lures. In a standard *n*-back task, participants may rely more readily on familiarity as a cue to correctly identify which letter is a target. The presence of lure items undermines such familiarity cues, such that upon encountering a lure, conflict resolution processes are initiated to successfully override familiarity-based evidence and re-characterize the stimulus as familiar but not in the relevant *n*-back location. Prior work has highlighted such an information recharacterization function as crucial for resolving syntactic conflict. As sketched in the previous chapters, during parsing, interference resolution processes trigger when readers or listeners encounter input that is incompatible with their developing analysis (see January et al., 2009; Novick et al., 2005; 2009; 2010).



Moreover, VLPFC is routinely identified as the crucial neural underpinning of conflict/interference resolution (see Jonides & Nee, 2006), such that VLPFC resources have been implicated specifically as supporting cognitive control during sentence reinterpretation in both patient and neuroimaging studies (January et al., 2009; Novick et al., 2005; 2009; 2010; Ye & Zhou, 2009). Thus, the need for interference resolution is clearly shared across a range of tasks, including garden-path recovery and *n*-back. Related research exploring the utility of similar interference resolution tasks for training demonstrates far-transfer to other interference resolution measures that tap VLPFC-supported EF (Persson et al., 2007). The illustration of transfer to syntactic ambiguity resolution in Experiment 1 may be considered an extension of that finding.

Nevertheless, the above are merely speculations about the nature of the training benefit observed in Experiment 1. The present experiment directly tested the effects of interference lures by comparing *n*-back-with-lures training to an otherwise identical training regimen sans lure items. The goal of the present study was to identify whether *n*-back lures were necessary and sufficient to usher in improvements in cognitive control that generalize to untrained, but related, measures requiring similar information recharacterization. To test this, rather than having trainees to practice a battery of training tasks, they trained on just one task in order to isolate the EF mechanism of interest.

Directly comparing versions of *n*-back with and without lures—in the absence of other WM training tasks—allowed for the following points to be addressed: First, I tested the mechanistic locus of training and, in particular, what about the version of *n*-back used in Experiment 1—lures presence, adaptivity—allowed for far-transfer to untrained measures of sentence processing. Put differently, I asked: are *n*-back lures necessary and

sufficient features to confer transfer to sentence reinterpretation measures? By replicating the results of Experiment 1 for just the trainees assigned to practice *n*-back-with-lures (and not individuals exposed to a version absent lures), the transfer effect is likely to be driven by interference resolution abilities acquired through practicing lures items, and not some other skill honed over the course of training on another practiced process (recall, participants of Experiment 1 performed 8 different cognitive training tasks). Namely, the design of Experiment 2 minimized confounds that may have, in part, accounted for the improvements observed in Experiment 1. Indeed, comparing a group that practiced many skills to no-contact controls invites an infinite number of interpretations for any observed training-transfer result, regardless of its selectivity (for a discussion of the *Hawthorne effect*, see Shipstead et al., 2010).

A second benefit to the design of Experiment 2 involved the inclusion of active control conditions. All participants were assigned to a training condition that required them to visit the lab for 18 total sessions. Different from other training studies with active control groups, the control tasks were similar (all were versions of *n*-back), but with various elements carefully and systematically removed. Note that typical intervention designs include active control groups that perform wildly different tasks (e.g., knowledge training or trivia; e.g., Buschkuhl et al., 2008; see Brehmer et al., 2011 for a discussion) that presumably do not tap the general-purpose ability of interest. In Experiment 2, the comparison groups were different on more subtle features (presence of lures and/or performance-adaptivity; see Method), providing a unique opportunity to understand the distinct role of certain task properties that support enhanced interference-resolution skills.

Implemented in Experiment 1, one approach for identifying subjects who improved on disparate skill sets is to hierarchically cluster participants who respond to different training tasks within a battery. Comparing cross-assessment effects of responders to Training Task A (e.g., *n*-back responders) compared to responders of Training Task B (e.g., block span responders) may offer a way to test for the transferability of various EFs to untrained tasks. Subjects improving on spatial working memory (block span responders) might be expected to show different transfer effects compared to subjects improving on interference resolution (*n*-back responders). Indeed, this rationale was used when pinpointing the selectivity of the training effects in Experiment 1. Deriving “control groups” among trainees when the natural comparison condition is a no-contact control group is one approach that should be interpreted cautiously, given that every subject in the training condition was exposed to the same set of tasks. Put differently, although a participant might improve on (respond to) *n*-back training but not block span training, this subject spent a good deal of time—the same amount of time, in fact—practicing both tasks. The design of Experiment 2 bypasses such concerns by issuing a single training task to each subject, against which subjects who practiced a minimally-dissimilar training task may be compared. That is, the nature of training was non-overlapping across trained “control” groups.

By parametrically and minimally manipulating the training tasks in Experiment 2, participants tasked with one training regimen could be dubbed the interference-resolution training condition, while those assigned to another intervention absent a subtle task feature (lures) could be considered non-interference resolution controls. This design provided the chance to carefully test for the *causal* relationship of general-purpose EFs

for language processing without having to identify responders of separate tasks. As sketched above, one possible shortcoming of Experiment 1's analysis centers around the interpretation that can be gleaned from a responder/nonresponder analytic approach; when a battery of training tasks is presented, theoretically, multiple regression should define the training tasks that account for the majority of the variance of pre/post changes in untrained abilities. However, the assumption that a training task and a transfer measure share EFs (i.e., are initially related in some way) may obscure the outcome and interpretation of such regression models. If Task A is correlated with Task B, then gains on Task A will likely predict gains on Task B assuming that a shared latent variable accounts for the correlation of both tasks. That is, the correlation between gain scores, and therefore the differences between responder groups, could be true regardless of whether there were true training effects (Tidwell et al., 2013). Considering this, it may not be possible to disentangle an interpretation favoring cognitive control training as the mediating factor of *n*-back responders' improvement on reinterpretation measures from a statistical artifact influencing such a pattern.

Provided that there are clear, selective training group differences in Experiment 2 such that only subjects practicing lures demonstrate cross-assessment improvement in garden-path recovery, I would be able to conclude a *causal* relationship (rather than conflate this pattern with a statistical artifact, as was the case in Experiment 1). Another way to frame this is to ask if *n*-back lures necessary and sufficient for generalized improvement in cognitive control.

Finally, given the putative role of EF in several other measures of linguistic and non-linguistic processing, transfer might extend beyond just syntactic ambiguity

resolution. As detailed in Chapter 1, there is great overlap in the EF processes involved to carry out several language tasks successfully; these involve employing interference resolution mechanisms to recover from temporary misanalysis during sentence parsing (as demonstrated in Experiment 1) and to select the right word for production in the face of many plausible contenders (e.g., under-determined representational conflict).

Experiment 2 tested these predictions by incorporating untested far-transfer measures of verbal fluency. Moreover, a non-conflict sentence processing measure (relative clause parsing) was included to test the degree to which effortful linguistic processing overlaps with interference resolution (see next section). Additionally, non-linguistic tasks containing high-EF conditions were also completed at pre- and posttest, including Stroop (with high-EF incongruent conditions contrasted with low-EF congruent items; see Milham et al., 2001) and a recognition memory task (with a high-EF block containing interfering memoranda and a low-EF block sans conflicting features; see Oberauer, 2005). By including a host of tasks with high-EF and control (low-EF) conditions, the breadth of interference resolution training—and thus, a process-specific account—could be tested.

### **3.1 Experimental Preliminaries**

All participants performed pretest and posttest versions of tasks, which included both high- and low-EF conditions. These included: (1) A version of the Stroop task with response-eligible (representational and response conflict inherent in incongruent items) and response-ineligible (only representational conflict for incongruent items) blocks containing congruent (*green* written in green ink), incongruent (*green* written in blue ink), and neutral (*horse* written in green ink) trials; (2) A recognition memory task with

global (targets matched the identity of words from the most recent memory list) and local blocks (targets matched the identity *and* location of words from the most recent memory list) containing targets (a familiar word in the global block; a familiar word in the proper location in the local block), fillers (unfamiliar words in both blocks, regardless of position), and lures (a familiar word in the local block that appeared in a location other than where it was presented during the study phase); (3) A verb generation task with high and low competition (many or few competing alternatives for production) and high and low association (strong or weak competitors for production) nouns; (4) A garden path recovery task with ambiguous (requiring reinterpretation of an initial, default meaning) and unambiguous sentences (void of reinterpretation demands); and (5) Reading relative clauses with object- (effortful processing) and subject-extracted (easier processing) constructions. Finally, at posttest, all participants also completed an *n*-back-with-lures task as a means to verify that each training group improved where expected relative to the other groups. Two dimensions were of interest: Adaptivity (contrasting the adaptive—Lures and No-Lures—groups to the 3-Back Group in terms of performance across two *n*-levels) and Interference (contrasting the Lures group to the non-lures groups—No-Lures and 3-Back—in terms of lure accuracy).

**Stroop.** Subjects completed a Stroop task to index non-linguistic EF changes as a function of training group assignment. A classic version of Stroop requires participants to respond to the ink color of visually presented words, some of which are color words (Stroop, 1935). High-EF incongruent trials refer to cases when the ink color mismatches the semantic representation of the color word (*blue* written in green ink), thereby recruiting cognitive control resources to resolve the conflict between the competing

sources of perceptual and semantic information (see January et al., 2009; Milham et al., 2001; 2003). Low-EF congruent trials (e.g., *blue* written in blue ink) remove such conflict, and instead engender a facilitatory condition whereby either perceptual, semantic, or some combination of both representations may be used to arrive at the correct response. Neutral trials that are void of color-semantic information (e.g., a string of asterisks or the word *horse* in green ink) are often used as controls for congruent and incongruent trials.

To extract out different levels of conflict, I implemented a version of Stroop with response-eligible and response-ineligible trials. During the response eligible block, incongruent trials were restricted to only words that matched the possible response options (*blue, yellow, green*), resulting in a conflict that arose at the both representational level (prompting one to override an automatic reading bias) or at the level of the response (prompting one to override choosing the incorrect color response; see Chapter 1 for a discussion of response versus representational conflict; see also Milham et al., 2001; 2003). That is, when participants encountered a response-eligible trial, two possible types of conflict contribute to difficulty experienced when a prepotent bias to read a lexical representation must be countermanded in favor of a perceptual (color) representation. Brain-imaging findings point to separable neuroanatomical involvement for two forms of conflict: VLPFC and anterior cingulate cortex (ACC) are routinely recruited when both representational and response conflict are present (as is the case during response-eligible trials); interestingly, ACC is less active when response conflict is removed (see Chapter 1; Milham et al., 2001). Thus, a second block of Stroop trials isolated representational conflict by including color words that were not possible response options on incongruent

trials (*red, orange, brown*). Indeed, Stroop Cost on *ineligible* trials and syntactic ambiguity resolution coactivate areas of VLPFC within subjects (see January et al., 2009), providing some evidence for the role of shared interference resolution abilities for the representational component of these tasks among healthy adults (see also Badre & Wagner, 2006). Similar VLPFC resources are recruited during lures trials of the *n*-back task (Gray et al., 2003), pointing to a possible link between the task performed by individuals in the Lures Group and Stroop. Thus, response-ineligible trials were regarded as the critical items on which the Lures Group should improve.

**Recognition Memory.** Another untrained non-linguistic assessment measure completed by all participants involved recognition memory, which aimed at capturing the difference between familiarity- and interference-based recognition abilities. Similar to a classic Sternberg task (Sternberg, 1969), subjects completed two recognition blocks: global and local recognition memory. During the global recognition task, participants simply indicated whether a probe word belonged to the most recent memory set, while the local block required participants to encode a contextual feature—location—for each to-be-remembered word. This additional feature introduced the unique opportunity to incorporate high-interference probe items, which matched in identity to a study-list word, but mismatched in location, rendering the item highly familiar, but not a target. This demand directly paralleled that of the version of *n*-back with lures: To successfully handle lure trials of the local block, participants had to override their default cognitive reaction to issue a positive recognition response to all familiar (repeated) items, and instead consider a secondary task demand of item location (or *n*-back level, in the case of the *n*-back task). As a result, the Lures group—and no other intervention group—is



expected to enjoy selective cross-assessment benefits on high-conflict local recognition trials.

**Verb Generation Task.** I included a canonical verb generation task to test the effects of cognitive control training for situations of elevated competition for production of linguistic representations. Indeed, much recent work focuses on conflict-control during production, namely when the accessibility of a single to-be-generated word is compromised by the presence of other semantic associates (see Chapter 1). Competition and association strength are two factors that may give rise to the challenge encountered for some words. High-competition refers to nouns with multiple competing contenders for production (e.g., *ball – bounce, throw, kick, roll, dance*), while low-competition nouns had a few appropriate options for production and often a single dominant verb associate is produced (e.g., *job – work*). High-competition nouns were expected to elicit greater conflict, thus signaling the need for interference resolution (Thompson-Schill et al., 1997). As a result, production times to these items were expected to decrease for the Lures Group, but not the No-Lures and 3-Back Groups. High-association nouns maintain strong connections to their nearest neighbors (e.g., *bed – sleep*), and thus are regarded as items with low retrieval demands, while low-association nouns have weak connections to their nearest neighbors (e.g., *valley – hike*), resulting in high retrieval demands (see Martin & Cheng, 2006; Snyder & Munakata, 2008). Although retrieval demands introduce a level of difficulty, they do not engender conflict in the same way as high-competition situations. That is, two separable sources of accessibility difficulty may be responsible for the observed production errors and latencies. The present study aims to

demonstrate that production times during high-competition cases—and not low-association items—are improved with conflict-control training.

**Sentence Processing.** In addition to a production-based linguistic task (verb generation), two parsing measures were included to test for the role of cognitive control training for difficult sentence constructions, including syntactic ambiguity and embedded clause processing. The garden-path sentences used in Experiment 1 were provided to (1) replicate the patterns of the first experiment, and (2) to test if *n*-back lures were necessary and sufficient to warrant the improved cognitive control needed to handle syntactic ambiguity. Subjects also read control constructions thought to engage general-purpose abilities *separate from* the EF of interest (interference resolution)—subject- (SE) and object-extracted (OE) relative clauses (see Fedorenko et al., 2006). Despite a wealth of psycholinguistic research focused on parsing relative clauses (with most work examining the OE/SE asymmetry in processing difficulty), few reports document the role of cognitive control. That is, the slowed reading time normally accompanying OEs (compared to SEs) is typically explained in terms of how taxing syntactic integration is on the parser, and the degree to which the parser relies on linguistic versus non-linguistic verbal working memory capacities to overcome such demands. Consider (3) and (4):

3. The farmer who the expert questioned promoted the product at the fair.  
(Object-extracted)
4. The farmer who questioned the expert promoted the product at the fair.  
(Subject-extracted)

In both cases, the verb “promoted” must be integrated with “the farmer” into a verb phrase to arrive at the interpretation that the farmer was the one doing the promoting at

the fair. Differences in processing difficulty emerge, however, at the embedded clause (“who the expert questioned”/“who questioned the expert”), owed to the fact that in (3), the farmer is the extracted *object* of the embedded clause (*the expert questions the farmer*); whereas in (4), the farmer serves as the extracted *subject* of the embedded clause (*the farmer questioning the expert*). That is, the farmer is differentially integrated into the embedded clause, with object-extraction requiring more extensive syntactic movement than subject-extraction: “Who” references the farmer in both sentences, but the referent in the SE condition is much closer, or more local, than that of the OE condition (see Fedorenko, Woodbury, & Gibson, 2013; Just & Carpenter, 1992; Lewis & Vasishth, 2005; Warren & Gibson, 2002). Such locality explanations have been provided to account for OE processing difficulty—indexed by slowed reading times—experienced at the embedded clause.

Additionally, much work has focused on mnemonic (span-based) measures of cognitive ability to characterize the exaggerated OE reading times (Caplan & Waters, 1999; Just & Carpenter, 1992; King & Just, 1991). Such evidence tends to take two forms: Verbal working WM (as measured by a reading span task) predicts the difference in online sentence processing between OEs and SEs (see Just & Carpenter, 1992; King & Just, 1991, *inter alia*), and dual-task performance selectively influences real-time reading of complex OE sentences, but not simple SEs (see Fedorenko et al., 2006; Waters, Caplan, & Hildebrandt, 1995). Although a host of results point to a critical role of working memory capacity for OE processing, to my knowledge, no work has emphasized the potential role of the hypothesized trained mechanisms in the Lures training task detailed above. Put differently, the processing demands encountered during OE processing is

dissimilar to that required to *recharacterize* an initial misinterpretation while reading garden-path sentences. Namely, as others have shown, capacity-based verbal working memory explains some differences in processing ability; here, I argue that the skills gained during *n*-back-with-lures training is quite different from the capacities tapped during verbal working memory tasks like reading span (see discussions in Chapters 1 and 2).

### **3.2 Hypotheses**

Thus far, I have argued that a process-specific training account presumes appropriate linking hypotheses between the types of cognitive abilities required to perform certain assessment tasks in order to choose an effective training regimen. Again, transfer can only be expected if the cognitive skills (e.g., interference resolution) underlying certain outcome measures are targeted through training so as to affect shared mechanisms that facilitate performance; likewise, training tasks not involving these mechanisms are not expected to confer transfer. Thus, considering the results of Experiment 1, I hypothesized that the interference resolution abilities boosted through practice with an *n*-back-with-lures task would exclusively benefit the high-EF conditions of each untrained linguistic (high-competition conditions of verbal production; syntactic ambiguity resolution) and non-linguistic (incongruent Stroop trials; recognition in the presence of interfering memoranda) assessment task, compared to conditions where the need for EF was removed (unambiguous sentences and congruent Stroop trials, for example). Furthermore, for linguistic materials with heightened cognitive demands due to their difficult structures (relative clauses), I hypothesized that the effect of conflict-control training would be negligible. A positive transfer effect for these items would hint

at an improved general process (compared to a process-specific account of cognitive control).

To test these hypotheses, three versions of *n*-back were parametrically manipulated to target distinct cognitive processes hypothesized to subserve separate cognitive mechanisms. A letter version of *n*-back-with-lures (henceforth referred to as the Lures Group) was almost identical to the *n*-back task administered in Experiment 1, such that it was performance-adaptive (*n*-level varied with accuracy) and designed to tap interference-resolution abilities (due to the presence of lure items). A letter version of *n*-back-without-lures (No-Lures Group) was also performance-adaptive, but did not include lure items; thus, this version of the task may be expected to promote a heightened familiarity bias. A multi-stimulus version of 3-back-without-lures (3-Back Group) was minimally different from the No-Lures Group: I removed the performance-adaptive component from the 3-Back task, given that much work suggests that difficulty re-thresholding is a critical component to ensure that cognitive abilities are actually boosted (see Brehmer et al., 2011). Thus, I hypothesized that the 3-Back Group would underperform relative to the No-Lures Group where adaptivity is advantageous. Specifically, transfer benefits for the No-Lures Group—compared to the 3-Back Group—might manifest under conditions capitalizing on performance-adaptivity (akin to detecting a target in an *n*-back position greater than 3), given that participants in the 3-Back Group were only ever required to maintain three items at a time.

Participants were assigned to a condition where they practiced just one of these training tasks across 16 training sessions, totaling 8 hours of exposure. I expected to observe a process-specific dissociation across the performance-adaptive training groups,

such that interference resolution training should selectively benefit performance on untrained task conditions with elevated interference demands. Specifically, I hypothesized:

1. If Lures are necessary and sufficient to amp up cognitive control abilities that are shared across training and assessment measures, then individuals assigned to the Lures Group should improve selectively on high-EF conditions embedded within each of the pre/post assessment tasks (including faster real-time reanalysis and better accuracy to comprehension questions following syntactically ambiguous sentences, faster production time for verbs given high-competition nouns, faster response time to incongruent trials of the Stroop task, and faster recognition time on the local memory block), whereas those in the No-Lures and 3-Back Groups, who did not practice interference resolution functions, should *not* demonstrate test-retest improvements in the abovementioned high-EF conditions. The complementary low-EF conditions of all assessment tasks should result in no selective changes for individuals assigned to the Lures Group.
2. Finally, I tested the hypothesis that the Lures Group—practicing the most cognitively-demanding training condition—improved selectively on just the most complex stimuli by evaluating participants’ pre/post performance while processing relative clauses. Namely, much psycholinguistic evidence suggests that object-extracted relative clauses introduce processing difficulty relative to subject-extracted control sentences. If training on *n*-back with lures confers a general advantage for trainees to manage complex and difficult stimuli, then

the Lures Group should show benefits on object-extracted clauses; however, if practice with *n*-back lures improves interference-resolution skills *alone*, then I would expect to see no selective benefits for this group when reading difficult (object-relative) sentences.

## 3.2 Method

### 3.2.1 Subjects

Healthy native-English-speaking subjects were recruited from the University of Maryland community to participate in this experiment for pay (totaling \$200 for 10 total hours of participation across 18 lab sessions). All were randomly assigned to one of three training groups (Lures, No-Lures, or 3-Back). Thirty-five participants were excluded from analyses (5 from Lures; 12 from No-Lures; 10 from 3-Back) for either failing to complete all study phases ( $n=19$ ) or for allowing at least 2 weeks to lapse between any two consecutive sessions ( $n=16$ ).<sup>8</sup> The final participant group comprised 81 individuals (Lures:  $N=30$ , 22 women,  $M_{\text{age}} = 19.8$  years, age range = 18-27 years,  $M_{\text{education}}: 14.53$  years; No-Lures:  $N=23$ , 17 women,  $M_{\text{age}} = 20.0$  years, age range = 18-23 years,  $M_{\text{education}}: 14.09$  years; 3-Back:  $N=28$ , 19 women,  $M_{\text{age}} = 20.0$  years, age range = 18-22 years,  $M_{\text{education}}: 14.36$  years). None of the subjects had a history of neurological disorders, stroke, or learning disabilities, and no one reported taking medications to correct problems related to neuropsychological or neuropsychiatric impairment. All subjects had normal or corrected-to-normal vision and hearing.

---

<sup>8</sup> The excluded participants were no different from the include participants in terms of demographics ( $N=35$ ; 11 women;  $M_{\text{age}}= 20.26$  years; age range = 18-29;  $M_{\text{education}} = 14.57$  years).

### 3.2.2 Design

A double-blind pretest/posttest design was used; accordingly, neither subjects nor experimenters knew subjects' condition assignments. Different moderators held training and assessment sessions in separate labs, so that the experimenter who collected the assessment data was blind to the condition to which each subject had been assigned. Additionally, because subjects in the experimental and control conditions never interacted, they were in principle blind to each other's condition and unaware of the differences between them. All participants visited the training lab for 16 thirty-minute sessions in the three-to-six weeks ( $M=4.8$  weeks) intervening pretest (Assessment 1) and posttest (Assessment 2) (see Figure 6). To combat attrition and promote active participation, each subject was notified of an incentive program following the eighth training session. In this e-mail notification, participants were provided a figure depicting their training performance with high scores (average  $n$ -back scores) clearly marked with a star. Participants were told that for each high training score, their names would be put into a drawing to earn a prize worth up to \$200 (effectively doubling the pay-out of the study); thus, they were encouraged to put forth their best efforts when performing  $n$ -back each day. Only participants who completed the entire training study were eligible for this award.

During each assessment, participants completed 4 short tasks, including a reading task (including syntactic ambiguity resolution, relative clause processing, and reading illusions of ungrammaticality), a verb generation task testing under-determined representational conflict, the Stroop task, and a recognition memory task with interfering memoranda. These latter two tasks were included to test for boosts in non-linguistic



interference resolution abilities. A near-transfer measure (*n*-back-with-lures) was presented only at posttest. The purpose of the post-training administration of this version of *n*-back enabled me to verify the effects of training; that is, *n*-back with lures served as a manipulation check to ensure that only those training groups exposed to certain task characteristics improved on those practiced components. For example, just the Lures Group—not the No-Lures and 3-Back Groups which did not see lures—should improve on lure items of the posttest *n*-back task (see specific predictions below). Each assessment battery was completed within one 2-hour session, with task order Latin-squared, such that participants saw tasks in different orders at each assessment and with respect to other participants. All participants, however, began the first assessment battery by completing a demographic questionnaire and ended the second assessment with *N*-back with Lures followed by an exit survey and a debriefing statement.

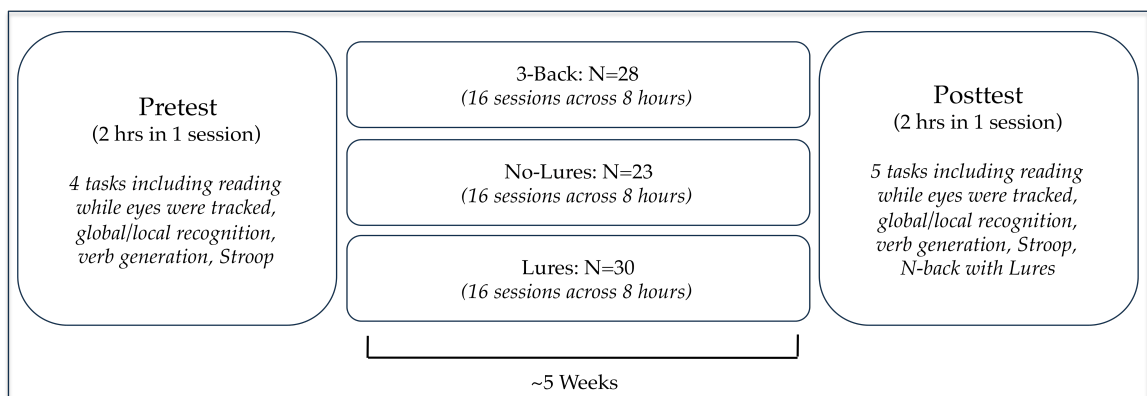


Figure 6. Longitudinal design of Experiment 2.

### 3.2.3 Training Tasks

In the interval between assessments, participants completed 8 hours of practice on a single version of one of three *n*-back tasks. The assigned version of *n*-back was administered for a total of 30 minutes during each of 16 in-lab training sessions. The

three versions of  $n$ -back (see Table 7 and details below) included: 1) Lures training (interference-resolution training task), 2) No-Lures training (adaptive training task), and 3) 3-Back training (non-adaptive control task). The Lures and No-Lures groups received adaptive training, akin to the training tasks administered in Experiment 1, while the 3-Back task was *not* adaptive (i.e.,  $n=3$  regardless of task performance). Adaptivity was parametrically manipulated to allow for the development of an active control version of  $n$ -back, given that performance-adaptive  $n$ -back tasks have been shown to confer greater transferability to untrained measures compared to non-adaptive versions (see Brehmer et al., 2011; Holmes et al., 2009; Klingberg et al., 2005). A second parametric difference that I implemented involved the presence or absence of lures, maintaining the assumption that encountering lure items would force participants to train their interference resolution skills. The Lures version of  $n$ -back was almost identical to that used in Experiment 1 (but see subtle task differences below); the No-Lures version was entirely identical to the present (updated) Lures task, but absent lure items. I operationally defined a lure item as any repeating non-target item that appeared within a set, but  $n$ -level-dependent buffer (see details below).

**Lures Training.** The Lures version of  $n$ -back was identical to that used in Experiment 1, such that letters were displayed serially and participants indicated by button press whether the current letter had appear  $n$  items previously (see the first panel of Table 7). The newer version and that used in Experiment 1 differed in the following ways: First, all sequences contained  $20+n$  items, rather than a 25 items. Thus, every sequence included  $8+n$  fillers among 6 targets and 6 lures, ensuring that all sequences had the same number of eligible target responses (i.e., 20) regardless of  $n$ -level. This change

Version	Adaptive	Lures	Stimuli	Example Sequence
3-Back	No	No	Letters, Words, Non-words, Symbols	<p><i>For 3-back task:</i></p>
No-Lures	Yes	No	Letters	<p><i>For 4-back task:</i></p>
Lures	Yes	Yes	Letters	<p><i>For 4-back task:</i></p>

Table 7. Explanations of the 3 versions of  $n$ -back used in Experiment 2.

attenuated the response bias asymmetry throughout a sequence (i.e., later sequence positions are more likely to contain targets at higher  $n$ -levels in the 25-item sequence condition). For example, in the version administered for Experiment 1, a 7-back resulted in 19 eligible response positions, with the initial 6 items always requiring a “no” response; whereas, a 3-back task had 23 possible target positions and fewer initial items requiring negative responses, 2. Although a response bias asymmetry exists as an inherent limitation of all versions of the  $n$ -back task, adjusting each sequence length to contain the same number of eligible target positions across  $n$ -levels minimizes the potential for a response-bias strategy at higher  $n$ -levels.

A second difference between the newer Lures version of  $n$ -back and that used in Experiment 1 involves the nature of the performance adaptivity. The maximum  $n$ -level was boosted from 8-back to 13-back to allow for participants to reach natural asymptotic performance; note that in Experiment 1, some participants mastered an 8-back task

relatively quickly. A second change was that the lure level was kept constant, rather than occurring as an adaptive feature. Here, all possible lure positions ( $n+1$ ,  $n+2$ ,  $n-1$ ,  $n-2$ ) were presented in every sequence regardless of  $n$ -level (see Gray et al., 2003).

A final difference between the newer version of  $n$ -back compared to that employed in Experiment 1 was the criterion that dictates changes in  $n$ -level. In lieu of using total accuracy of responses to all items in a sequence (e.g., in Experiment 1, 85% or greater accuracy on a sequence led to an increase in difficulty level, while less than 65% rendered a reduction in difficulty level), the updated Lures version considered the *number* of correct target items (see Jaeggi et al., 2011). Here, if less than three target errors occurred, difficulty was increased by one  $n$ -level, while more than five target errors led to a reduction in one  $n$ -level. This change emphasized the relevance of target detection for successful task completion.

Similar to Experiment 1, the Lures task was designed to provide participants with ample practice overriding a familiarity bias that arises when a recent item repeats, but is not in a relevant target position. Repeated exposure to such instances should boost interference resolution skills, considering neural evidence suggesting VLPFC recruitment when lures are encountered (see Method and Discussion of Experiment 1).

**No-Lures Training.** The No-Lures  $n$ -back task was identical to the Lures  $n$ -back with the exception that lures were removed from all trials, resulting in sequences of 6 targets and  $14+n$  fillers. The “lure buffer” was designed to identify highly-confusable positions (presumably, those which would contain the most effective lure items), and was operationally defined as the sequence of items within the scope of  $n$ —not inclusive of the target position—plus items appearing in the two positions prior to the  $n^{\text{th}}$  back item (see

panel 2 of Table 7). Within the No-Lures task, no item repeated within the lure buffer, ensuring that only targets constituted highly-familiar items in a sequence. Note, however, that items could repeat within a sequence that were *not* targets, but that these cases could occur in  $n+3$  or later positions (e.g., during a 5-back task, an item could repeat in the No-Lures task in the 8-back position, but never sooner unless it was a target). This task characteristic is not dissimilar from classic versions of the  $n$ -back task, in that repeating items would be purely incidental and thus do not occur at high frequencies (contrary to the Lures task above, which always included 6 lure items). Thus, targets were the only repeating recent (within  $n+3$ ) items in a sequence. This element of the design fosters a strategy favoring familiarity detection (however, see discussion in Chapter 4), such that participants could feasibly complete the task by simply responding whenever the current item exceeded some familiarity/recency threshold. Thus, I predicted that participants receiving this training task would improve in their ability to recognize recent, familiar items, while not necessarily improving interference resolution skills that are acquired during the Lures  $n$ -back task.

**3-Back Training.** Participants monitored serially-displayed sequences of 23 items (6 targets and 17 fillers), and were asked to indicate via button press whether the current item appeared 3 trials previously. Task difficulty was changed as a function of performance, thereby making the current task a non-adaptive version of the No-Lures task. In addition to the adaptivity difference between the 3-Back and No-Lures tasks, the 3-Back task included various stimulus sets in addition to letters. The purpose of this change was to minimize attrition of participants assigned to this condition. Although all  $n$ -back tasks provided feedback after each sequence (accuracy and average response

time), this may not have been sufficient to keep subjects in the 3-Back condition engaged for 16 training sessions. That is, the participants assigned to the adaptive conditions may be motivated to reach the next  $n$ -level, knowing that achieving this hinges on good performance on the current sequence. With this element removed for the 3-Back task, I included multiple stimulus sets to keep subjects engaged over the course of training.

Stimulus sets included letters, high-imageability single-syllable words, pronounceable single-syllable nonwords, and two sets of popular symbols (taken from webdings; see Appendix B for stimulus sets). Sets were cycled across sessions in the same order for all participants; for example, all subjects performed a letter 3-back at training session 1, a word 3-back at session 2, one version of a symbol 3-back at session 3, a nonword 3-back at session 4, and a second version of a symbol 3-back at session 5 before repeating the same sequence for sessions 6-10 and 11-15. All subjects finished their final (16<sup>th</sup>) session with a letter 3-back.

Provided that adaptivity introduces a task demand that benefits performance, by removing this feature of the  $n$ -back training task, I anticipate participants in this condition to show little to no transfer, even under conditions that hinge on the ability to detect targets. Thus, the 3-Back condition may be viewed as an active control to the No-Lures task.

### *3.2.4 Transfer Tasks*

**Posttest  $N$ -back-with-Lures.** To verify the differential effects of each of the three training tasks, all participants completed one block of a 3-back-with-lures and one block of a 6-back-with-lures at posttest only. Two levels of  $n$  were included to verify that the performance adaptive Lures and No-Lures Groups showed a relative advantage at  $n$ -

levels greater than 3 (here, on 6-back) compared to the 3-Back Group. Both blocks included lure items, which allowed me to verify that the Lures Group outperformed the No-Lures and 3-Back Groups on lure accuracy.

*Procedure.* The *n*-back-with-lures comprised the final cognitive task performed by all participants at posttest. The task began with a block of 3-back sequences, followed immediately by a block of 6-back sequences. Similar to the Lures task detailed above, letters were displayed serially and participants indicated by button press whether the current letter had appeared *n* items previously. During the 3-back block, sequences contained 6 targets, 6 lures, and 11 fillers, while the 6-back sequences included 6 targets, 6 lures, and 14 fillers. Following each sequence, participants were provided with accuracy and average response time feedback. *N*-level was not varied within a block regardless of performance, and was carefully controlled to be either 3 or 6 depending on the current block. Subjects were explicitly notified of the task change with an instruction screen when the task transitioned from the 3-back to the 6-back block.

**Stroop.** At pretest and posttest, participants completed a modified version of Stroop (see Milham et al., 2001). Participants were asked to indicate the ink color of each word presented on a computer monitor via button press. There were three possible color responses—blue, yellow, and green—across each of two blocks: a response-eligible and a response-ineligible block. During the response eligible block, incongruent trials were restricted to only words that matched the possible response options (*blue, yellow, green*), resulting in a conflict that arose at the both representational level (prompting one to override an automatic reading bias) or at the level of the response (prompting one to override choosing the incorrect color response). Response-eligible and -ineligible

incongruent conditions were compared to neutral trials of non-color words written in blue, yellow, or green ink. These neutral words were matched in length and syllable count to the eligible and ineligible color words. For example, *blue*, *yellow*, and *green* were matched with *deal*, *plenty*, and *horse*, while *red*, *orange*, and *brown* had the neutral counterparts of *tax*, *farmer*, and *stage* (items borrowed from January et al., 2009). Congruent trials always included the eligible response color words matched in the proper ink color response (*blue*, *yellow*, *green*). Neutral trials were used as controls for both cases to compute a Stroop Cost (incongruent response time minus neutral response time) and a Stroop Benefit (neutral response time minus congruent response time).

*Procedure.* At each assessment, participants performed two blocks of the modified Stroop task, one with only response-eligible trials and one containing just response-ineligible trials. Block order was randomized and counterbalanced across participants. The overall task, however, began with simple instructions, orienting the participants to the response buttons and the possible types of items to-be-presented. All subjects began by performing six baseline practice trials (non-color words), followed by six trials with color words (3 congruent and 3 incongruent). Following each practice trial, participants were provided feedback of their response accuracy. Relevant instructions and practice trials preceded each block, regardless of order. Participants were then prompted to begin the actual experiment, wherein no feedback was provided.

On each trial, a fixation point appeared in the center of the screen for 750 milliseconds, followed by a word written in blue, yellow, or green ink, to which a color response was provided. The lexical representation of this word rendered the trial congruent, incongruent, or neutral. Congruent trials were more common than incongruent



or neutral trials, occupying 50% or 72 trials of each block; incongruent and neutral trials each occurred 25% of the time (36 trials per). Although incongruent trials result in markedly slower response times compared to neutral and congruent items, prior work suggests that the proportion of incongruent trials further influences response time: A smaller proportion (akin to 25% of all trials) elicits slower times compared to cases when incongruent trials are more common (50% of all trials; see Kane & Engle, 2003). Conflict adaptation accounts—which cite that the density of high-EF trials causes reactive adjustments in cognitive control—provide theoretical support for this shift (Larson, Kaufman, & Perlstein, 2009; West, 2004).

All trials within a block were randomized, such that any trial type (e.g., congruent) could precede any other trial type (say, neutral). For each trial type, participants saw equal numbers of each stimulus word, such that the three eligible color responses were evenly distributed (e.g., one-third of all incongruent responses required the ‘yellow’ button to be pressed, with half of these cases including each stimulus word *green*, *blue*). Participants used the middle row on a number pad to make responses; each color response was mapped to a finger (key): blue-right pointer finger (numpad 4), yellow-right middle finger (numpad 5), green-right ring finger (numpad 6). As soon as a response was made, the fixation point would appear again for 750ms followed by another word. Each block took approximately 6 minutes to complete, and a short schedule break separated blocks.

**Recognition Memory Task.** At each assessment, participants were presented with a modified version of the recognition memory task implemented by Oberauer in the

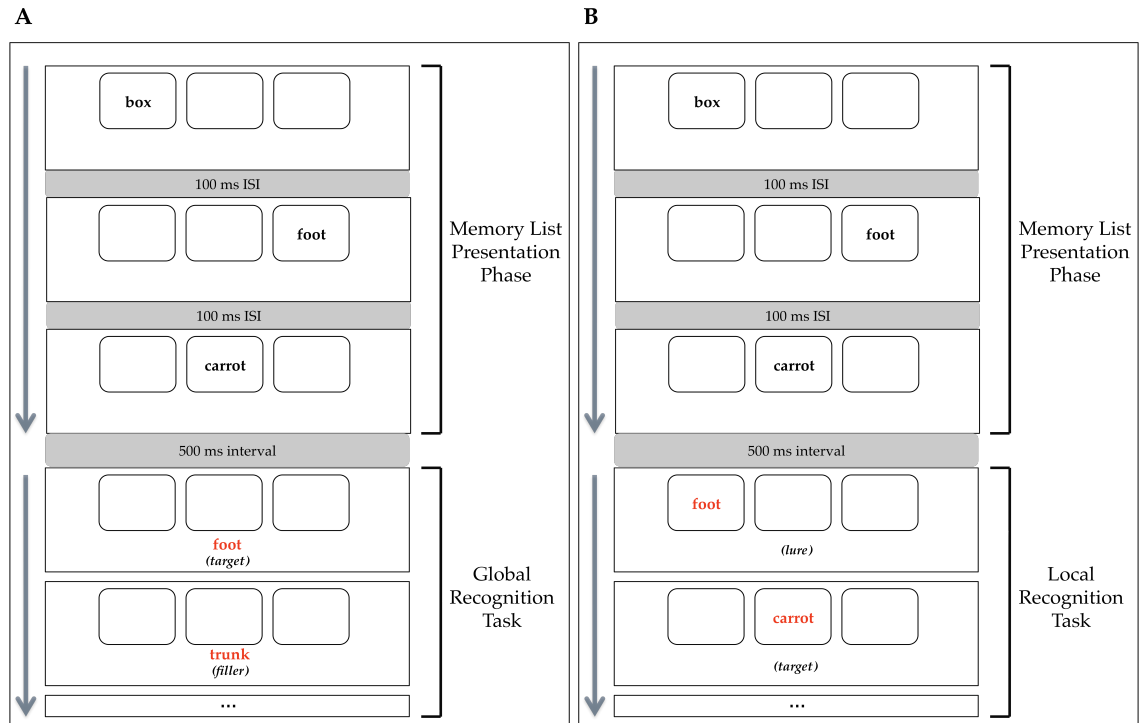


Figure 7. A sample trial of the global (A) and local (B) recognition task.

second experiment of his 2005 paper, during which they performed a block of low-EF global recognition trials followed by a block of high-EF local recognition trials.

Regardless of block, during each trial, a memory list of words appeared serially in one of several frames arranged in a row on a computer screen, followed by a recognition test phase. The words were selected to never repeat within or across assessments to prevent *recent-no* interference effects (see Chapter 1). Anywhere from two to five words were presented on a given trial, such that when two words were presented, two empty frames appeared on the screen, each of which was serially populated with a word; likewise, when three words were presented on a trial, three empty frames appeared which were later serially populated with each random word (see Figure 7). Words were randomly selected from a set of candidate stimuli (non-overlapping sets were developed for each assessment) and presented for 900 milliseconds within frame locations with a 100-

millisecond inter-stimulus interval. Word location varied from trial to trial, such that words could appear in any unused location within a trial. For example, in the case of a 3-word list, the first word might appear in the first (leftmost) frame, the second word in the third (rightmost) frame, and the final word in the second (center) frame (see also Figure 7, wherein word 1—*box*— appears in frame 1, word 2—*foot*—in frame 3, and word 3—*carrot*—in frame 2). The number of trials containing each list length was balanced evenly within each block, along with the location order in which words were presented. Differing across blocks, however, was the nature of the probes during the recognition test. Since the local block included an additional probe-type, it occupied twice as much time as the global block, given by double the number of memory sets.

*Global Recognition Block.* Subjects performed 9 practice trials before beginning the experimental component of the global block, which included a total of 80 trials during the global recognition block, 20 trials of each list length (2, 3, 4, and 5). After the presentation of the final word of a memory list, a probe items were serially displayed in red font color in a centralized location on the screen, directly below the empty row of frames. Participants were instructed to judge whether each probe was a member of the most recent memory list by pressing computer keys mapped onto yes/no responses. The number of probes was equal to the list length of each trial, such that when a memory list included 2 items, only 2 probes were presented during the recognition test. A target was marked as any word that appeared in the memory list (e.g., *carrot* in Figure 7A), and a filler/non-target item was a novel word that had not been encountered on any previous (or subsequent) trial in the task (e.g., *trunk* in Figure 7A). Target and filler probes were evenly distributed across all trials of the same list length, such that within the global

block, participants always encountered equal numbers of target and filler probes for each list length—a total of 70 targets/70 fillers within an entire global block. Participants took approximately 10 minutes to complete the global block, including one scheduled break at the halfway point of the block.

*Local Recognition Block.* A total of 160 trials were presented during the local recognition block (40 trials of each list length), not including the 9 practice trials that all subjects completed prior to beginning the block. Double the amount of trials were included to increase the number of observations for a second type of non-target probe—lures—that did not appear in the global recognition task. Similar to the global block, following the presentation of the final word of a memory list, probe items were serially displayed in red font color, to which participants made yes/no judgments about whether each probe was a member of the most recent memory list. What was different from the global block was the position in which the probes were presented. Rather than being serially presented in an unchanged centralized location beneath the row of frames, probes were presented within the frames themselves. As a result, an item was deemed a target if and only if it matched both the identity and location of the word presented in the memory list. Therefore, two forms of non-targets were possible: novel words that did not match the identity of any memory list words (fillers) and words that did match in identity, but mismatched in terms of location (lures; see *foot* of Figure 7B). The number of target and non-target probes was equal across each list length, resulting in a total of 140 targets/140 non-targets within the entire local block. Among the non-targets, half were lures and half were fillers, with each of the three probe-types (targets, lures, fillers) evenly distributed across all trials of the same list length. Similar to the global block, the number of probes

presented during the recognition phase matched that of the list length; thus, a memory list containing 4 items (and 4 frames) would have a probe appear in each of the four frames in a random order. Probe items did not repeat within or across trials, removing *recent-no* items (wherein a probe matched an item on a recent, but not the current list) and the possibility of a target also appearing as a lure in the same trial. Participants took approximately 20 minutes to complete the local block, including three breaks scheduled every 5 minutes during the block.

*Materials.* Consistent with Oberauer's word constraints, a list of 1-2 syllable nouns high in imageability, concreteness, and familiarity, and high-to-moderate written frequency (Kucera-Francis frequency=114.3) were drawn from the MRC psycholinguistic database ([http://www.psy.uwa.edu.au/mrcdatabase/uwa\\_mrc.htm](http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm); Wilson, 1988). Words were randomly divided in half to create two versions of the experiment. Subjects were assigned to one version for pretest and completed the complimentary version at posttest. Words did not repeat on trials within or between blocks or across assessments. That is, subjects never saw the same nouns twice within this task.

**Verb Generation.** At both assessments, participants were asked to generate a single verb in response to a noun cue. The nouns that were presented belonged to one of four groups, which varied along two continua of Competition and Association.

*Procedure.* Participants were instructed to think of and produce the first verb that came to mind given a noun cue. On each trial, a noun was visually presented on a computer screen for up to 3400ms (following the presentation rate used by Persson et al., 2006). If the 3400ms expired prior to a response, then the trial was tagged as a failed

retrieval. Participants were first instructed to press the spacebar when they thought of a verb associated with the current noun cue. They were, then, asked to verbalize the word that was generated into a microphone for later scoring. Prior to beginning the task, participants were provided extensive instructions, defining precisely what a verb was and providing examples of verbs that might be generated to a set of sample nouns. They were then given 4 practice trials before beginning the actual experiment.

*Materials.* A total of 100 nouns (26 HCHA, 24 HCLA, 24 LCHA, 26 LCLA; see Appendix C) were borrowed from the materials used in Snyder et al., 2010 (personal communication). Half of each noun-type was randomly assigned to a set, one of which was then randomly selected for pretest presentation; the remaining set was presented at posttest. This resulted in a total of 50 unique nouns to which verbs were generated at each assessment (in a non-repeating fashion across assessments). Of these 50 nouns, half were high-competition and half were high-association, such that there were 13 HCHA, 12 HCLA, 12 LCHA, and 13 LCLA at each assessment. Nouns were presented randomly over the course of the task, and no break was provided, as the task took no more than 10 minutes to complete.

**Sentence Processing: Garden-Path Recovery.** To replicate the findings of the first training experiment, I included the same sentence-processing task that was used to assess garden-path recovery in Experiment 1. To recap, participants read ambiguous and unambiguous sentences embedded within a series of filler items while eye movements were recorded. They then answered comprehension questions that probed for misinterpretations (*Did the thief hide himself?*). I hypothesized that recovery from misinterpretation (reflected by comprehension accuracy) should improve for individuals

receiving interference resolution training (Lures Group) and no other group.

Unambiguous cases were not expected to benefit, because the need for controlled revision is absent in these constructions. As for eye-movements, I hypothesized that real-time recovery efforts (as reflected by regression path time) should improve following training for just the Lures Group. Further, changes in reading patterns should be confined to the disambiguating sentence region where EF is hypothesized to trigger. Changes were not expected in other regions of ambiguous sentences, or anywhere in unambiguous sentences, where the need to revise (and use EF) is removed. Such selective changes in real-time revision were not expected for subjects assigned to the No-Lures or 3-Back groups.

*Materials.* Similar to Experiment 1, separate but complementary versions of the ambiguity resolution task were developed so that participants never saw the same materials across assessments (see Appendix A for exact sentences used). The same pseudorandomization and counterbalancing principles were implemented to ensure that participants only saw each critical verb once per assessment, and that the sentential context changed across assessments. At each assessment, 12 ambiguous and 12 unambiguous constructions were presented and embedded within 120 filler sentences (some borrowed directly from Christianson et al., 2006, along with other constructions that served as control sentences; see next section).

**Sentence Processing: Parsing Relative Clauses.** Separate but complementary versions of the relative clause sentences were developed so that participants never saw the same materials across assessments (see Appendix D for all experimental sentences). Forty-eight unique sentences were borrowed from Fedorenko et al. (2006) to create 12

object-extracted and 12 subject-extracted sentences per assessment (see examples 3 and 4 above). At each assessment, these 24 items were embedded within filler sentences (including the abovementioned syntactically ambiguous and unambiguous constructions among 120 other filler items). For each assessment, two lists were created: if an item in one list appeared as a subject-extraction, it was an object-extraction in its counterpart list. List administration was pseudorandom and counterbalanced across participants and assessments. Thus, a participant never saw the same verbs and actors within or across assessments. A comprehension question followed every sentence as a way to verify that participants were indeed processing the sentence for its meaning. Contrary to the questions probing for interpretations of syntactically ambiguous sentences, the OE/SE questions were not designed to test for the meaning of the critical embedded clause (for the examples above, the question, *Was the product promoted on TV?* was posed). The number of yes/no responses was evenly balanced across the OE/SE set at each assessment.

### **3.3 Analyses and Results**

#### *3.3.1 Training Task Performance*

Analysis of the training data revealed that participants showed marked improvement on their respective *n*-back training tasks (average effect size, Cohen's  $d=1.38$ ). Performance was indexed by *n*-back scores, which were calculated by multiplying the average *n*-level by average accuracy achieved by each subject at each training session. Generally, *n*-back scores provide a measure of mean difficulty level achieved—akin to what was used in Experiment 1—to track training-related performance gains. Figure 8A illustrates that average *n*-back score varies significantly across training



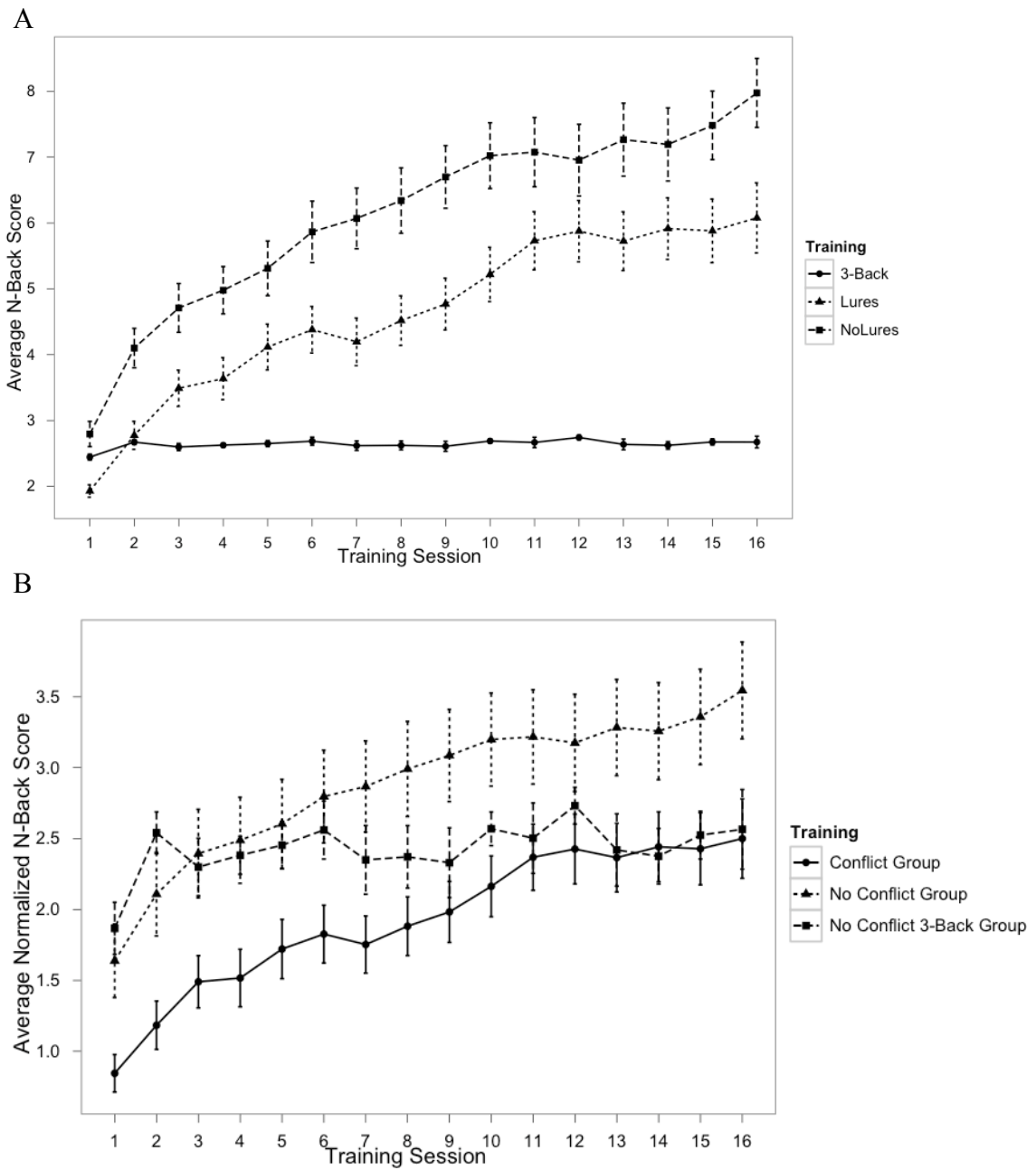


Figure 8. Training performance over the course of 16 training session for each training group indexed by (A) n-back score—average accuracy multiplied by average n-level for each subject at each session—and (B) normalized (z-scored and scaled on initial training session score) n-back score—within each training group, but collapsed across each session. Error bars =  $\pm 1$  standard error of the mean.

groups, in part due to the average *n*-level achieved on each: The No-Lures Group reaches higher average *n*-levels (7.09) than the Lures Group (5.44), while the 3-Back Group only has experience monitoring 3 items at a time, restricting the range of their *n*-back scores.

Two alternative approaches were taken to compare performance across training groups, while alleviating this confound: First, *n*-back scores at each session were normalized (z-scored) within each training group with the mean of the initial training session added to the score to scale the measure according to baseline performance—see Figure 8B—wherein I observe comparable training gains for the two adaptive groups (Lures and No-Lures; mean standardized change from first to final session is 1.6625 and 1.8011, respectively). The 3-Back Group, on the other hand, shows more variable performance across sessions, ultimately improving (mean standardized change from first to final session is 0.6767), but doing so much less systematically than the adaptive groups. This pattern may be suggestive of the importance of adaptivity to observe general gains in training performance.<sup>9</sup> However, one caveat to this analytic approach is that even normalized measures are subject to conflation. *N*-level is used to compute *n*-back score—the measure over which participants were normalized—which may have imposed a restriction on the performance range of the 3-Back Group. Compared to the variance of the Lures and No-Lures Groups, the variance of the 3-Back Group is fairly large and constant over the course of training. The adaptive groups, on the other hand, demonstrate increased variance with training session, reflecting individual differences in subjects' asymptotic levels.

---

<sup>9</sup> Note, however, that these session-by-session changes may be reflective of either improved training or more precise indices of baseline cognitive ability; see Discussion in Chapter 4

To overcome this issue, I implemented a second approach that captured training-mediated differences in task performance, deferring to a single version of the  $n$ -back task that all subjects completed at posttest. This task was used as a proxy to test for group differences in adaptivity (by comparing performance on low- and high- $n$ -levels across groups) and interference resolution ability (by comparing performance on  $n$ -back lure items across groups). Although I did not include a pretest version of the same task to measure baseline efforts, a posttest  $n$ -back task is liable to still convey selective training improvements.

### 3.3.2 Index of Training Effects: Posttest $N$ -back-with-Lures

**Analysis.** Posttest  $n$ -back data were excluded from two participants who performed below 80% accuracy on filler items averaged across both blocks (1 subject in the Lures Group; 1 in the No-Lures Group). As sketched above, the purpose of the posttest  $n$ -back task was to verify that, relative to the other conditions, each training group performed where expected on a common measure of  $n$ -back that all groups experienced.<sup>10</sup> I sought to demonstrate two important effects: First, relative to the adaptive groups (Lures and No-Lures), the non-adaptive 3-Back Group should be *less accurate* when performing  $n$ -back tasks where  $n$ -level is larger than the practiced 3-level. Since all members of the adaptive groups had ample practice at  $n$ -levels greater than 3 (most reaching levels greater than 6), the Lures and No-Lures Groups should demonstrate superior accuracy on items appearing in sequences where  $n$  exceeds 3, namely, in the 6-back block of the posttest version of  $n$ -back. Second, compared to the non-lures groups (No-Lures and 3-Back), the Lures Group should be *more accurate* when encountering

---

<sup>10</sup> Although all groups practiced some version of  $n$ -back, each was fundamentally different, making it difficult to compare training performance of these tasks across training groups

lure items regardless of  $n$ -level, given that subjects in this condition regularly practiced responding to lures during every training sequence. To test these hypotheses, I conducted analyses of variance (ANOVAs). Where appropriate, I also conducted JZS Bayes-factor tests to both verify the results of any t-tests and one-way ANOVAs reported and to provide an index of effect size.

**Test of Adaptivity.** Groups practicing adaptive training tasks (Lures and No-Lures) were combined to form an Adaptive Group against which the Non-Adaptive Group (3-Back) was compared. I observed an interaction of Adaptivity (Adaptive vs. Non-Adaptive) and  $N$ -level (3-Back vs. 6-Back) on overall accuracy ( $F(1,155)=4.13$ ,  $p=0.044$ ;  $BF=2.19$ ). The left panel of Figure 9A illustrates that on the 3-back block, Adaptive ( $M=0.935$ ,  $SD=0.059$ ) and Non-Adaptive ( $M=0.907$ ,  $SD=0.065$ ) training resulted in the same total accuracy ( $F(1,78)=3.86$ ,  $p=0.053$ ,  $BF=1.053$ ); whereas, on the 6-back block (right panel of Figure 9A), Adaptive training groups ( $M=0.874$ ,  $SD=0.081$ ) outperformed the Non-Adaptive group ( $M=0.798$ ,  $SD=0.075$ ) in terms of total accuracy ( $F(1,77)=16.686$ ,  $p<0.001$ ,  $BF>100$ ). Together, these findings verify that the non-adaptive 3-Back Group did not perform as well on more-challenging, unpracticed  $n$ -levels, suggesting that adaptive training afforded the adaptive groups with the ability to process (maintain and update) more than three units of serially-presented information. Indeed, practicing at a higher difficulty level throughout training seems to have been advantageous for performing at a greater difficulty level during a posttest assessment of  $n$ -back performance.

**Test of Interference-Resolution Ability.** To test for interference resolution advantages, the groups practicing training tasks void of lure items (No-Lures and 3-Back

Group) were combined to form a Non-lure group against which the Lures Group was compared. I observed an interaction of Lure Presence (Lure vs. Non-lure) for accuracy to  $n$ -back lure items regardless of  $n$ -level ( $F(1,78)=4.768$ ,  $p=0.032$ ,  $BF=12.334$ ). Figure 9B illustrates that Lure training ( $M=0.907$ ,  $SD=0.071$ ) resulted in better lure accuracy compared to Non-lure training ( $M=0.874$ ,  $SD=0.090$ ). This pattern verifies that subjects

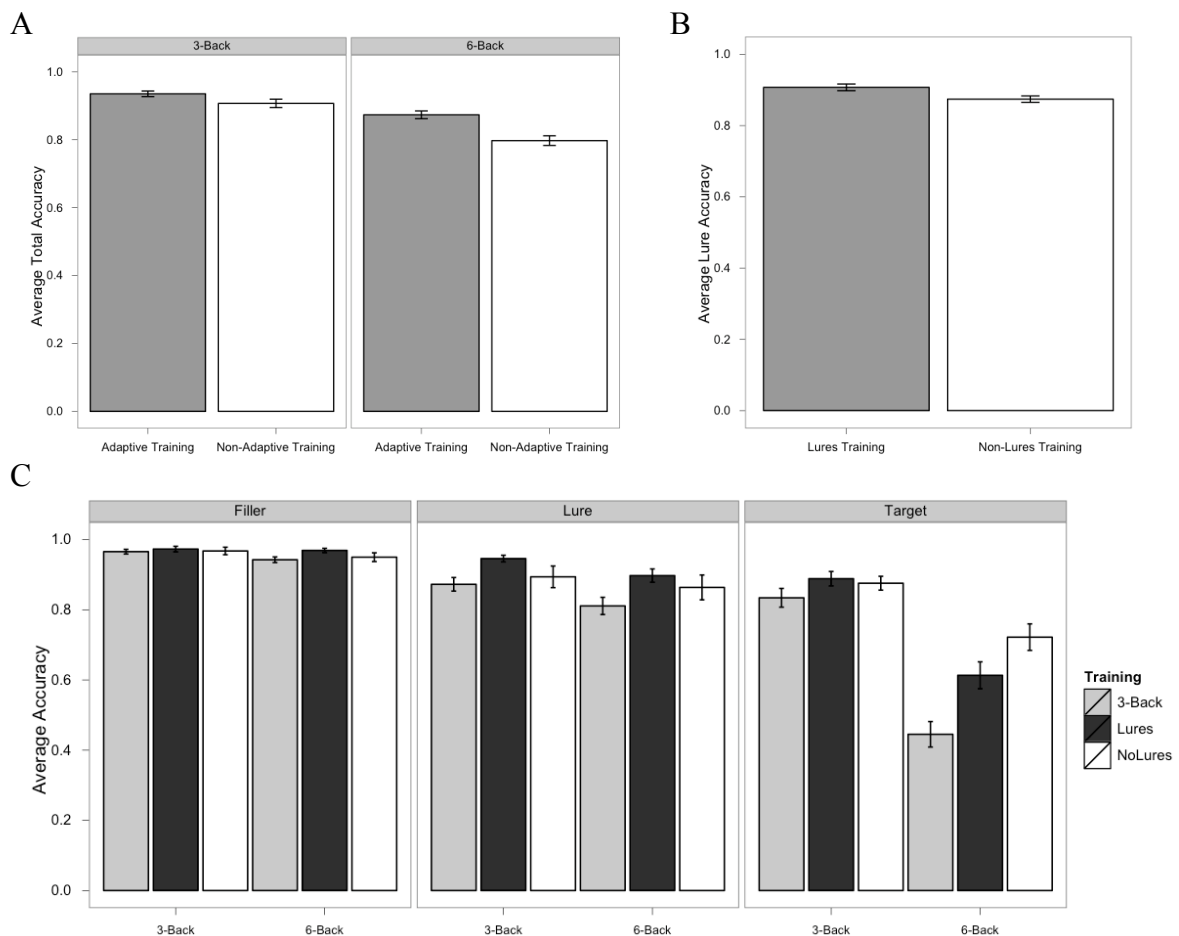


Figure 9. Posttest  $n$ -back task performance given by (A) overall accuracy on each block—3-back versus 6-back—for adaptive (Lures and No-Lures groups combined) and non-adaptive (3-Back group only) groups, indexing adaptivity benefits; (B) lure accuracy performance collapsed across block for lures (Lure group only) and non-lures (No-Lures and 3-Back groups combined) groups, indexing cognitive control benefits target accuracy on each block; (C) accuracy is decomposed by training group and block for target items (leftmost panel), non-target filler items (center panel), and non-target lure items (rightmost panel). Error bars =  $\pm 1$  standard error of the mean.

who practiced lure items demonstrated better accuracy on lures trials compared to those who do not see lures during training.

In addition to verifying the effects of adaptivity and interference resolution, I also examined the effect of Training Group (not in composite form, as above) on accuracy for each Item Type (targets, non-target fillers, non-target lures) on each Block (3-back and 6-back) in order to decompose the effects of the abovementioned analyses. Figure 9C depicts the average accuracy for each of these six conditions (3 items types by 2 *n*-levels). Main effects of Training Group were observed for the lure items in the 3-back block ( $F(2,77)= 3.709$ ,  $p=0.029$ ,  $BF=0.699$ ) and lure ( $F(2,76)= 3.09$ ,  $p=0.050$ ,  $BF=0.406$ ) and target items ( $F(2,76)= 13.117$ ,  $p<0.001$ ,  $BF= 43.53$ ) in the 6-back block (all other conditions:  $F$ 's $<2.67$ ,  $p$ 's $>0.07$ ,  $BF$ s $<0.331$ ). Ad hoc comparisons indicated that the effects among lure items (at both *n*-levels) were based on a distinction between the Lures and 3-Back Groups (Welch two-sample  $t$ 's $>2.786$ ,  $p$ 's $<0.006$ ,  $BF$ s $>4.739$ ), even though numeric differences existed such that the Lure Group outperformed the No-Lures Group. The target item effects on the 6-back block were driven by accuracy differences between all of the training groups (Welch two-sample  $t$ 's $>2.018$ ,  $p$ 's $<0.05$ ,  $BF$ s $>1.005$ ), indicating that the No-Lures Group ( $M=0.722$ ,  $SD=0.177$ ) was more accurate to respond to 6-back targets than the Lures Group ( $M=0.613$ ,  $SD=0.206$ ). Superior target accuracy for the No-Lures Group may be viewed as evidence for this group's superior target-detection skills acquired over the course of training on a version of *n*-back that favors a familiarity/recency bias. One may have expected the 3-Back Group to show a similar advantage, given that subjects in this condition could presumably also default on a familiarity bias to perform their version of the *n*-back task. That 3-Back trainees did not show a pattern

consistent with this expectation may be suggestive of their use of a strategy other than one that capitalized on a familiarity bias (see discussion at the end of this chapter).

Together, the posttest *n*-back task largely verified that practiced tasks conferred relative advantages among subjects based on their group assignments. The effect of Adaptivity was evidenced by overall accuracy of the adaptive groups being greater than that of the non-adaptive 3-Back Group at higher *n*-levels (i.e., 6-back block).

Interestingly, the No-Lures Group showed better accuracy to target items compared to the Lures Group, perhaps reflecting these trainees' boosted familiarity bias/target detection abilities. Finally, the Lures Group outperformed the Non-lures groups in terms of lure accuracy performance at both high and low *n*-levels, confirming an effect of selective interference-resolution abilities (at minimum, as they pertain to *n*-back) for this group. With each training condition demonstrating relative performance advantages where expected, the next question was one of whether these selective improvements transferred to untrained conditions of interference resolution. I explored this question by examining four untrained far-transfer tasks, two non-linguistic measures (Stroop and recognition memory) and two linguistic measures (verb generation and sentence processing).

### *3.3.3 General Analyses for Pre/Post Measures*

For all tasks presented at both pre- and post-test, I used analyses of covariance to examine the relation between training condition and cross-Assessment improvement. To tap Assessment-by-Training Group effects, I predicted posttest performance as a function of Training Group while covarying out pretest performance. Importantly, this approach emphasized posttest differences among Training Groups, directly addressing the question of whether any group outperforms other groups following intervention (at posttest), while

assuming all groups are performing equivalently at pretest. Training Group was included as a 3-level factor to examine the contribution of all groups with respect to one another. Where appropriate, ad hoc comparisons treated Training Group as a 2-level factor through which two sets of trainees were contrasted to probe minimal group differences.

For the sentence processing measures (comprehension accuracy and eye movements), I reported multilevel mixed-effects models testing for the fixed effects of Training and Assessment (see Jaeger, 2008). Mixed-effects models were used to statistically evaluate pre/post improvements among the training groups while nesting fixed factors within the random variables of Subjects and Items. Importantly, crossing both random variables was not a possible feat within the context of an ANCOVA (factors were nested within Subject only for these analyses), supporting the use of mixed-effects models for such data. I also included JZS Bayes-factor tests to verify the presented results, where appropriate; however, note that BFs are omitted where mixed-effects models are reported because such models are often unbalanced and include more than one factor of interest (see elaboration of BFs in Method section of Chapter 2).

#### 3.3.4 Stroop Task

**Analysis.** The average accuracy for subjects performing the Stroop task was 96.45% (range=91.67-99.82%); thus, no participants were excluded from analyses on the basis of poor performance. Response times were analyzed for correct trials only, and were normalized with respect to neutral (control) trials. Below I examined these normalized response times for high-interference incongruent trials in the form of a Stroop Cost by computing differences between median<sup>11</sup> response times on incongruent (*blue*

---

<sup>11</sup> Medians were used to correct for extreme response time values in lieu of Winsorizing.



written in green ink) and neutral trials (*deal* written in green ink). An index of low-interference Stroop trials (Stroop Benefit) served as a control measure for Stroop Cost. Stroop Benefit was calculated as the difference between median response times on neutral and congruent trials (*green* written in green ink). Both Stroop Cost and Stroop Benefit measures were computed for each subject separately for response-eligible and response-ineligible blocks.

Recall that response-ineligible trials isolate representational conflict, while response-eligible trials include both representational and response conflict. As a result, a cross-Assessment effect on response-eligible trials in the absence of such an effect on response-ineligible trials would suggest a change in *response-level* interference resolution, alone; whereas, a pre/post change on response-ineligible trials would be consistent with a shift in interference resolution at the *representational level*. Finally, a comparable change on both blocks irrespective of response eligibility would, too, suggest an effect at the representational level, as this component is shared across both blocks; note, however, that effects on both blocks are likely to also be driven by response-level interference in the event that the response-eligible trials elicit a quantitatively larger improvement than response-ineligible trials. Evidence for this representational/response-level distinction comes from prior work indicating that response-eligible trials engender more overall conflict evidenced by slowed response times which are accompanied by greater ACC activation (an area implicated in response-level conflict across a multitude of tasks; see Milham et al., 2001; 2003).

Here, I considered response-eligible and response-ineligible trials separately while assessing the effects of Interference (Stroop Costs vs. Stroop Benefits), Training

Group (Lures vs. No-Lures vs. 3-Back), and Assessment (Pretest vs. Posttest). Before running this analysis, however, I first performed a manipulation check to ensure that (1) incongruent trials elicited longer response times compared to neutral trials, which in turn had longer response times relative to congruent trials, and (2) response-eligible trials resulted in overall longer response times than response-ineligible trials, given that two levels of conflict are present and compounded in the former case.

**Manipulation Check.** To test for the classic Stroop effect and the influence of response-based conflict, I examined correct pretest response times as a function of Trial Type (incongruent, congruent, neutral) and Eligibility (response-eligible, response-ineligible) at collapsing across Training Group. A repeated-measures analysis of variance (ANOVA) revealed an Eligibility-by-Trial Type interaction  $F(2,474)=3.778$ ,  $p=0.024$ ,  $BF=0.24$ ), supported by main effects of Trial Type ( $F(2,474)=43.161$ ,  $p<0.001$ ,  $BF>100$ ) and Eligibility ( $F(1,474)=20.752$ ,  $p<0.001$ ,  $BF>100$ ; see Figure 10).<sup>12</sup> Overall, subjects were slowest to respond to incongruent trials ( $M=700.56\text{ms}$ ,  $SD=170.25\text{ms}$ ) relative to neutral ( $M=609.23\text{ms}$ ,  $SD=101.27\text{ms}$ ) and congruent trials ( $M=577.84\text{ms}$ ,  $SD=91.43\text{ms}$ ), verifying that high-interference incongruent items are costly to process, while performance is facilitated on congruent trials when information converges on a single response (i.e., matched lexical and perceptual information). Moreover, subjects were slower in the face of multiple forms of conflict when responding to incongruent items on

---

<sup>12</sup> Participants excluded from the present analyses (see Method for more detailed exclusion criteria) were no different from the current subset of included participants in terms of baseline response times on Stroop ( $F(1,458)=0.0472$ ,  $p=0.82$ ).

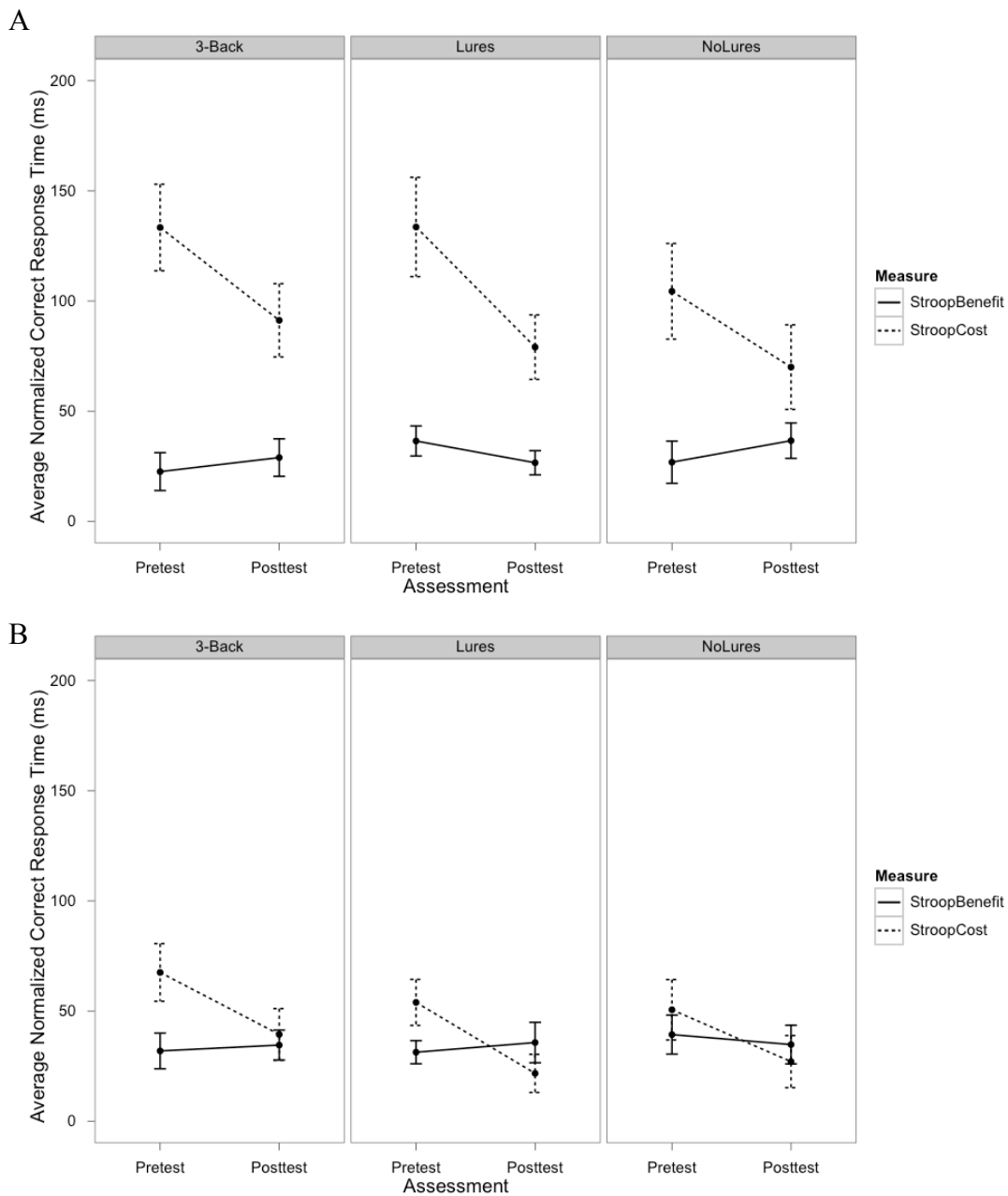


Figure 10. Response time improvements on conditions of the Stroop task. Pre/post correct response times in terms of Stroop Cost (incongruent minus neutral trials, reflecting high-interference resolution demands) and Stroop Benefits (neutral minus congruent trials, reflecting low-interference resolution demands) for each training group (3-Back in leftmost panels, Lures in center panels, and No-Lures in rightmost panels) on (A) a block where responses were eligible—reflecting response and representational conflict—and (B) a block where responses were ineligible—isolating representational conflict. Error bars =  $\pm 1$  standard error of the mean.

the response-eligible block ( $M=747.82\text{ms}$ ,  $SD=193.87\text{ms}$ ) compared to cases requiring the resolution of conflict at just the representational level ( $M=653.31\text{ms}$ ,  $SD=127.41\text{ms}$ ;  $F(1,158)=13.278$ ,  $p<0.001$ ,  $BF>100$ ). Consistent with this, I observed a main effect of Eligibility for Stroop Cost ( $F(1,158)=22.559$ ,  $p<0.001$ ,  $BF>100$ ), but not Stroop Benefit ( $F(1,158)=0.578$ ,  $p=0.44$ ,  $BF<0.001$ ). This pattern is evident by contrasting pretest performance in Figures 10a (eligible block) and 10b (ineligible block).

**Assessment-By-Training Interaction.** Given the verified effects of Eligibility and Trial Type, I next explored the effect of Training Group by conducting an analysis of covariance of posttest performance, with fixed effects of Training Group (Lures vs. No-Lures vs. 3-Back), Eligibility (Eligible vs. Ineligible), and Congruency (Stroop Cost vs. Stroop Benefit) while controlling for pretest (baseline) performance of each subject. Although I did not observe a significant interaction of all variables, I observed main effects of Assessment ( $F(1,626)=11.116$ ,  $p<0.001$ ,  $BF>100$ ), Eligibility ( $F(1,626)=26.825$ ,  $p<0.001$ ,  $BF>100$ ), and Congruency ( $F(1,626)=60.933$ ,  $p<0.001$ ,  $BF>100$ ), as well as Eligibility-by-Congruency ( $F(2,626)=39.513$ ,  $p<0.001$ ,  $BF>100$ ) and Assessment-by-Congruency ( $F(1,626)=13.665$ ,  $p<0.001$ ,  $BF>100$ ) interactions. That a main effect of Training Group is absent suggests that individual training manipulations had no selective effects on Stroop performance; however, all subjects demonstrated improved cross-Assessment performance in terms of Stroop Cost ( $F(1,144)=18.414$ ,  $p<0.001$ ,  $BF>100$ ), with selective reliable improvements on eligible trials ( $F(1,72)=12.670$ ,  $p<0.001$ ,  $BF>100$ ; ineligible trials:  $p=0.09$ ; see Table 8). Training did not yield cross-assessment changes in Stroop Benefit did not change from pre- to posttest suggesting that practice effects were isolated to only high-conflict incongruent trials

<b>Conflict Condition</b>	<b>F-Value</b>
<b>Stroop Task (Response Time)</b>	
<b>Response-Eligible Stroop Cost</b>	
Pretest Stroop Cost	F(1,72)=12.6695***
Training	F(2,72)=0.4463
Pretest Stroop Cost x Training	F(2,72)=0.7236
<b>Response-Ineligible Stroop Cost</b>	
Pretest Stroop Cost	F(1,72)=2.8066
Training	F(2,72)=0.7576
Pretest Stroop Cost x Training	F(2,72)=0.3555
<b>Recognition Task (Response Time)</b>	
<b>Global-Targets</b>	
Pretest Response Time	F(1,73)=5.1201*
Training	F(2,73)=1.2446
Pretest Response Time x Training	F(2,73)=0.0058
<b>Global-Fillers</b>	
Pretest Response Time	F(1,73)=8.8608**
Training	F(2,73)=0.3744
Pretest Response Time x Training	F(2,73)=5.4970**
<b>Local-Targets</b>	
Pretest Response Time	F(1,71)=87.5994***
Training	F(2,71)=1.2779
Pretest Response Time x Training	F(2,71)=0.3855
<b>Local-Fillers</b>	
Pretest Response Time	F(1,71)=14.4176***
Training	F(2,71)=1.4977
Pretest Response Time x Training	F(2,71)=0.2754
<b>Local-Lures</b>	
Pretest Response Time	F(1,71)=38.9777***
Training	F(2,71)=0.9357
Pretest Response Time x Training	F(2,71)=6.0179**
<b>Verb Generation Tasks (Production Time)</b>	
<b>High-Competition</b>	
Pretest Production Time	F(1,147)=204.965***
Training	F(2,147)=4.565*
Pretest Production Time x Training	F(2,147)=26.614***
<b>Low-Competition</b>	
Pretest Production Time	F(1,146)=67.7707***
Training	F(2,146)=2.8709
Pretest Production Time x Training	F(2,146)=1.9003
<b>High-Association</b>	
Pretest Production Time	F(1,146)=165.1020***
Training	F(2,146)=5.1648**
Pretest Production Time x Training	F(2,146)=2.2600
<b>Low-Association</b>	
Pretest Production Time	F(1,147)=88.4158***

Training	F(2,147)=2.4053
Pretest Production Time x Training	F(2,147)=5.5021**
<b>Offline Garden-Path Recovery (Accuracy)</b>	
<b>Ambiguous</b>	
Pretest Accuracy	F(1,69)=41.2038***
Training	F(2,69)=0.5502
Pretest Accuracy x Training	F(2,69)=1.2651
<b>Unambiguous</b>	
Pretest Accuracy	F(1,69)=2.9542
Training	F(2,69)=0.4226
Pretest Accuracy x Training	F(2,69)=0.9898

Table 8. Summary of all analyses of covariance conducted for the Assessment measures of Experiment 2.

( $F$ 's < 2.86,  $p$ 's > 0.06,  $BF$ 's < 0.27). Paired with the Eligibility effect (cross-assessment changes emerged *only* for response-eligible trials), this pattern lends support for improved response-conflict level improvements for all  $n$ -back trainees. Perhaps, this improvement is diagnostic of a critical function of practicing  $n$ -back urges subjects to continuously represent relevant stimulus chunks such that they must choose among representations at the response level throughout training (i.e., when issuing yes/no recognition judgments on  $n$ -back).

Next, I conducted a series of planned comparisons to evaluate pre/post gains for each Training Group, with the expectation that if the Lures Group inherited an interference-resolution advantage by virtue of practicing conflicting lure items, then they should demonstrate a quantitatively larger pre/post change in Stroop Cost compared to the remaining Training Groups, accompanied by no change in Stroop Benefit (indicative of a selective gain). Indeed, pairwise comparisons of Assessment for the Lures Group revealed pre/post changes in Stroop Cost on both eligible ( $t(28)=2.695$ ,  $p=0.01$ ,  $BF=2.619$ ; 55ms speed-up) and ineligible blocks ( $t(28)=2.828$ ,  $p=0.008$ ,  $BF=3.486$ ; faster by 32ms). This pattern was not observed for Stroop Benefit on either block (Eligible:

$t(28)=1.256$ ,  $p=0.21$ ,  $BF=0.226$ ; Ineligible:  $t(28)=-0.373$ ,  $p=0.71$ ,  $BF=0.112$ ). Moreover, paired comparisons testing for pre/post Stroop Cost differences among the remaining Training Groups bore no reliable change on either block (Eligible:  $t$ 's $<1.934$ ,  $p$ 's $> 0.07$ ,  $BFs<0.606$ ; Ineligible:  $t$ 's $<1.614$ ,  $p$ 's $>0.11$ ,  $BFs<0.364$ ); the No-Lures Group improved by 34ms on eligible trials and 24ms on ineligible trials, while the 3-Back Group was faster by 42ms on eligible and 28ms on ineligible trials. As would be expected given the above analysis of covariance, Stroop Benefit did not vary across assessments for the No-Lures ( $p$ 's $> 0.41$ ) or 3-Back Groups ( $p$ 's $> 0.53$ ). Despite no clear effect of Training Group for Stroop Cost in the above test for an interaction, planned comparisons yielded some support for a process-specific account such that the Lures Group showed a reliable pre/post change in Stroop Cost regardless of Response Eligibility, an effect that no other Training Group demonstrated. This pattern may be driven by Lures trainees' quantitatively faster response times from pre- to posttest compared to the training groups that did not practice resolving interference present in lure items. That this difference failed to emerge in the overall ANCOVA is likely due to the small magnitude of pre/post change (e.g., a small effect size), an issue perhaps owed to subjects' adept ability to perform Stroop prior to training (leaving little room to improve). To investigate the widespread nature of process-specific interference-resolution training (and to attempt to gather converging evidence across multiple tasks), I next examined Assessment-by-Training Group interactions in a second non-parsing measure tapping recognition memory under varying cognitive control demands.

### 3.3.5 Recognition Memory Task

**Analysis.** Data from two participants was excluded from all subsequent recognition memory task analyses on the basis of poor performance (filler accuracy did not exceed 80%; n=1, Lures Group) and missing posttest data (n=1, No-Lures Group). The recognition memory task included two blocks (global and local), each of which contained anywhere from 2-5 to be remembered list items. Recognition probes could either be targets and non-targets, with two forms of non-targets occurring in the local block (lures and fillers). Typically, recognition accuracy declines as a function of list length, and relative to global blocks, performance on blocks containing highly-confusable lure trials is worse. I conducted a manipulation check of these effects by looking at average accuracy at pretest (collapsed across all other factors) as a function of List Length (2, 3, 4, 5), Block (local, global), and Probe Type (target, filler, lure). Response times were analyzed for correct trials only, wherein I evaluated effects separately for low-interference global and high-interference local blocks by conducting analyses of covariance of posttest response time given fixed effects of Training Group (Lures, No-Lures, 3-Back) and Probe Type (Targets, Fillers, Lures), while controlling for pretest performance. A selective cross-Assessment effect for the Lures Group on the local block—but not the global block—would lend support for a process-specific training account. No change is expected among members of the No-Lures and 3-Back Group.

Finally, similar to the planned comparisons conducted for the Stroop task (above), I hypothesized that the Lures Group should show significant pre/post changes on the local block (especially so for lures trials of this block), given the heightened need for



cognitive control on these trials, but not the global block where interfering to-be-remembered items are non-existent.

**Manipulation Check.** I implemented a series of ANOVAs on pretest accuracy to verify that the recognition task was performed in conjunction with prior work. In particular, I tested the effects of List Length, Block, and Probe Type, and observed main effects for all variables: List Length ( $F(1,310)=303.49$ ,  $p<0.001$ ,  $BF>100$ ), Block ( $F(1,155)=223.27$ ,  $p<0.001$ ,  $BF>100$ ), and Probe Type ( $F(2,233)=169.59$ ,  $p<0.001$ ,  $BF>100$ ).<sup>13</sup> Specifically, accuracy declined as list length increased ( $M_s=98.40\%$ ,  $95.63\%$ ,  $92.35\%$ ,  $87.74\%$ , respectively for ascending lengths 2-5), performance was superior on the global block ( $M=98.38\%$ ) compared to the local block ( $M=89.06\%$ ), and average filler accuracy ( $M=98.60\%$ ) was greater than target accuracy ( $M=92.53\%$ ), which exceeded lure accuracy ( $M=78.30\%$ ). All subsequent analyses were conducted on correct response times. Comparable main effects of List Length ( $F(1,310)=41.657$ ,  $p<0.001$ ,  $BF>100$ ), Condition ( $F(1,155)=96.397$ ,  $p<0.001$ ,  $BF>100$ ), and Probe Type ( $F(2,233)=72.993$ ,  $p<0.001$ ,  $BF>100$ ) were observed for median reaction times: Response time increased with list length ( $M_s=685\text{ms}$ ,  $786\text{ms}$ ,  $876\text{ms}$ ,  $927\text{ms}$ , respectively for ascending list lengths 2-5); reactions were also slower in the local block ( $M=936\text{ms}$ ) relative to the global block ( $M=672\text{ms}$ ), and were slowest for lures ( $M=1174\text{ms}$ ) compared to target ( $M=823\text{ms}$ ) and filler items ( $M=743\text{ms}$ ). Together, this indicates that participants did not commit a speed-accuracy tradeoff on this task. To test for the effect

---

<sup>13</sup> Participants excluded from the present analyses (see Method for exclusion criteria) were no different from the current subset of included participants in terms of baseline response times on either block of the recognition memory task ( $F's<1.696$ ,  $p's>0.19$ ).

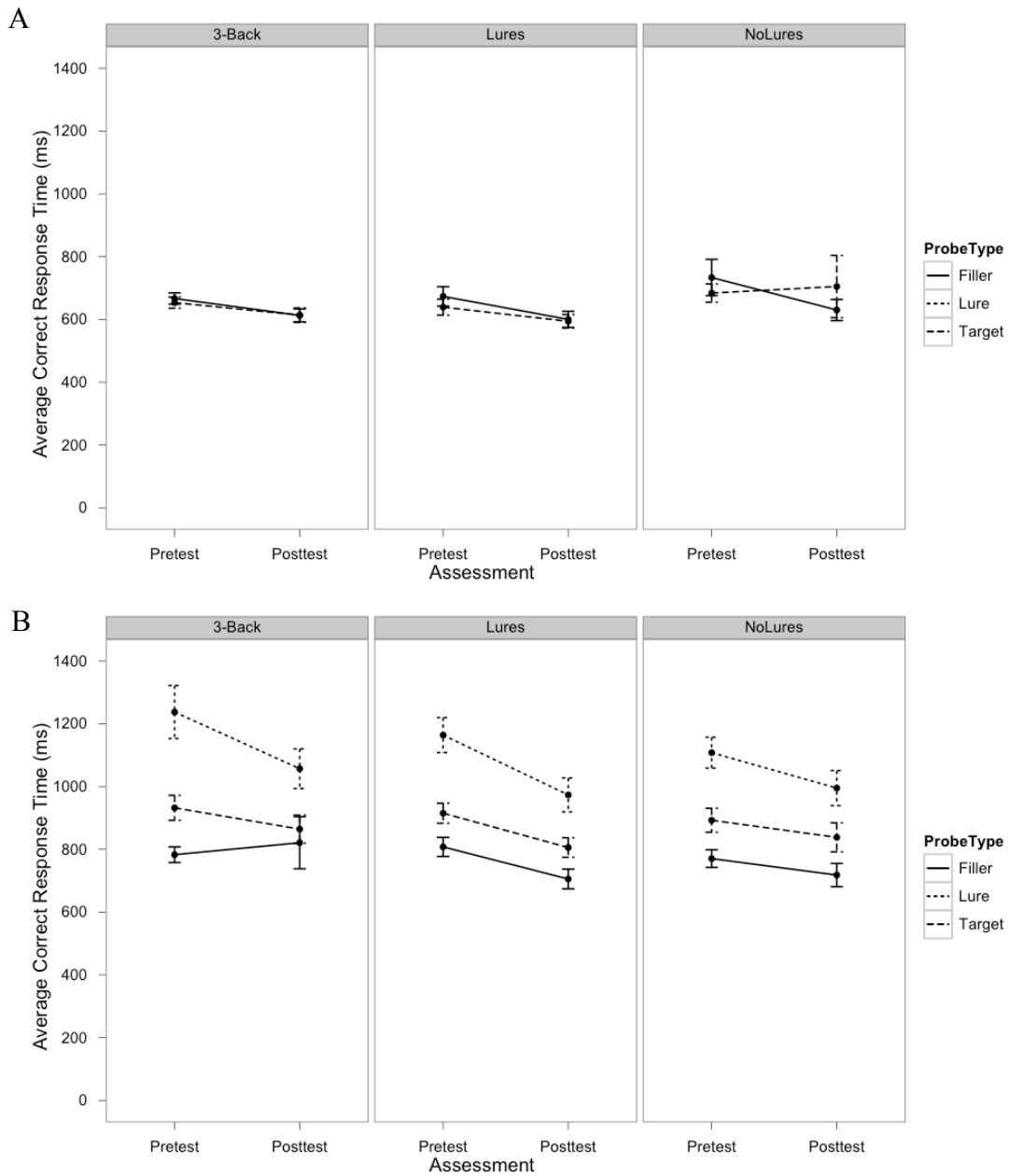


Figure 11. Response time improvements on conditions of a recognition memory task. Pre/post correct response times by item type (targets, non-target fillers, non-target lures) for each training group (3-Back in leftmost panels, Lures in center panels, and No-Lures in rightmost panels) on (A) the global recognition block—low interference-resolution demands—and (B) the local recognition block—high interference-resolution demands. Note that because the global block does not contain lure trials, only targets and fillers are plotted. Error bars =  $\pm 1$  standard error of the mean.

of Training Group on response time, I conducted separate ANCOVAs for each Condition (local, global) and each Probe Type (target, filler, lure) testing for posttest correct response time as a function of Training Group, controlling for pretest response time. This resulted in a total of five analyses for each combination of Probe Types that could appear in each Condition.<sup>14</sup>

**Assessment-by-Training Interaction on the Local Block.** I expected to observe an Assessment-by-Training Group interaction for just lure trials in the local recognition block, an effect driven by improved performance of the Lures Group. Indeed, the ANCOVA of posttest response time to Local-Lure items resulted in a reliable interaction ( $F(2,71)=6.018$ ,  $p=0.003$ ,  $BF=3.997$ ; see Table 8), accompanied by a main effect of Assessment ( $F(1,71)=38.978$ ,  $p<0.001$ ,  $BF>100$ ). No other items of the local block showed this pattern; however, all tests, regardless of item type, demonstrated robust main effects of Assessment ( $F$ 's $>14.417$ ,  $p$ 's $<0.001$ ,  $BF>100$ ; see Table 8). Upon further examination of the Local-Lure interaction, I observed numerically larger pre/post improvements for the Lures Group (191ms) relative to the No-Lures (113ms) and 3-Back Groups (180ms; see Figure 11A). This finding was bolstered by a series of planned comparisons testing for the difference between pre and posttest performance for each of the Training Groups, revealing that only the Lures Group was faster to respond to Local-Lure items at posttest relative to pretest response time ( $t(29)=6.350$ ,  $p<0.001$ ,  $BF>100$ ; No-Lures:  $t(19)=1.898$ ,  $p=0.07$ ,  $BF=0.656$ ; 3-Back:  $t(26)=1.776$ ,  $p=0.08$ ,  $BF=0.481$ ). Interestingly, corresponding pairwise comparisons for Local-Target and Local-Filler

---

<sup>14</sup> No lure probes were present on the global recognition task; thus, there were three levels of Probe Type within the Local Condition and two levels of Probe Type within the Global Condition.

items patterned similarly: The Lures Group, alone, was faster to respond to these items, as well ( $t$ 's $>4.369$ ,  $p$ 's $<0.001$ ,  $BF$ s $>128$ ; other Groups:  $p$ 's $>0.07$ ). Although the pre/post improvement for the Lures Group does not appear to be selective for just Local-Lure items (targets and fillers of the local block also improve), for this pattern to be consistent with a process-specific account of interference-resolution training, no such improvements among Lures trainees should be observed in the *global* condition, as it is void of any interference-resolution demands. To test for this, I conducted ANCOVAs of Global-Filler and Global-Target items, testing for the effect of Training Group while controlling for pretest response times.

**Assessment-by-Training Interaction on the Global Block.** I noted a significant Assessment-by-Training Group interaction for Global-Filler items, paired with a main effect of Assessment ( $F$ 's $>5,497$ ,  $p$ 's $<0.005$ ,  $BF$ s $>34.99$ ; see Table 8). Analysis of Global-Target items resulted in no reliable interaction, but also revealed a main effect of Assessment (see Table 8). Upon evaluating the interaction among Global-Filler items, I observed the greatest cross-Assessment improvement for the No-Lures Group (104ms), followed by the Lures (73ms) and 3-Back Groups (53ms; see Figure 11B), although none of these pre/post changes reached statistical reliability ( $p$ 's $>0.13$ ). Beyond this, I conducted planned comparisons to test any Training Group improved in target response time following training to find that no groups improved from pre- to posttest in terms of response time to Global-Targets ( $t$ 's $<1.77$ ;  $p$ 's $>0.08$ ,  $BF$ 's $<1.187$ ). In sum, the Lures Group was exclusive in showing significant pre/post improvement, an effect that emerged only for items in the local recognition block, namely those conditions where interference demands are elevated. No pre/post changes were observed in the global block for any

Training Group, an effect which may be driven by floor effects in this condition. Nevertheless, cases where reliable cross-Assessment benefits were observed occurred solely for high-EF conditions among members of the group that practiced conflict control over the course of training.

Additionally, correct response time effects provide one avenue for understanding the mechanisms shared between  $n$ -back and the untrained recognition memory task; accuracy profiles could complement such interpretations. Thus, I explored the possible mechanisms underlying each of these effects in Chapter 4 using signal detection analyses. To preview, the Lures Group demonstrates a large cross-Assessment shift in response criterion, such that these trainees become more conservative, committing fewer false alarms to lure items. The No-Lures and 3-Back Groups demonstrated no change in response threshold from pretest to posttest. Paired with the above response time analyses, the recognition memory task provides further support for a process-specific account of cognitive control training: Only trainees with experience dealing with interfering representations gain on measures tapping similar resources in the non-linguistic domain. Next, I explored the notion of process-specificity on two linguistic tasks—containing high- and low-EF conditions.

### *3.3.6 Verb Generation Task*

**Analysis.** All subjects' responses were coded for accuracy using the following criteria: Trials that resulted in the production of non-verb items, a significant delay in producing a verb following a button press (indicating a premature spacebar press), or the generation of an unrelated, but repeating verb (e.g., “have” for several trials) were coded as incorrect answers. These cases were used to identify subjects who failed to perform the

task as instructed. Data were excluded from six participants on the basis of poor accuracy—less than 60% correct at pretest (4 subjects in the Lures Group, 1 in the No-Lures Group, 1 in the 3-Back Group). Generation times were computed on correctly-generated verbs only. I performed a manipulation check at pretest collapsing across Training Group to verify baseline effects of Competition (many versus few competing contenders) and Association (strong versus weak nearest associates), with the expectation that High-Competition items would result in slower generation times compared to Low-Competition words because a greater number of competing verbs for a noun generates competition among all possible items when selecting just one to produce. Low-Association nouns, with weak verb associates, should be slower to produce than High-Association nouns with easy-to-retrieve (strong) verb associates; moreover, High-Competition/Low-Association (HCLA) items should result in the most exaggerated generation times given that these items compound retrieval and interference demands (see Snyder et al., 2010). Following the manipulation check, I conducted ANCOVAs to test for the effects Training Group (Lures vs. No Lures vs. 3-Back), Association (High vs. Low), and Competition (High vs. Low) on posttest generation time, controlling for pretest time. The Lures Group—and no other group—is expected to demonstrate selective pre/post improvements on High-Competition items, irrespective of Association level. The rationale for this hypothesis follows from work indicating that association level indexes retrieval demands, while competition level provides a proxy for under-determined conflict among plausible contenders (see Snyder et al., 2011, 2012). Since Lures training is geared toward resolving among interfering representations rather than retrieving items that are less active, all process-specific effects are anticipated for High-

Competition nouns. Planned comparisons were conducted to test for such an effect of Assessment on generation time following High- and Low-Competition nouns in each training group.

**Manipulation Check.** An ANOVA testing for the effects of Association and Competition on pretest generation time, irrespective of Training Group, revealed mean effects of both Association ( $F(1,314)=19.79$ ,  $p<0.001$ ,  $BF>100$ ) and Competition ( $F(1,314)=5.567$ ,  $p=0.01$ ,  $BF=2.354$ ), but no interaction of the two factors ( $p=0.83$ ).<sup>15</sup> As expected, generation times were slower for Low- ( $M=2554\text{ms}$ ,  $SD=227\text{ms}$ ) compared to High-Association items ( $M=1672\text{ms}$ ,  $SD=115\text{ms}$ ) and for High- ( $M=2241\text{ms}$ ,  $SD=172\text{ms}$ ) compared to Low-Competition items ( $M=1886\text{ms}$ ,  $SD=194\text{ms}$ ). It was also the case that HCLA items resulted in the longest generation times ( $M=2838\text{ms}$ ) compared to any other condition (next slowest  $M=2346\text{ms}$  for LCLA nouns). With the standard Competition and Association effects in place, I next examined the contribution of Training Group on cross-Assessment changes in generation time, testing for separately for effects related to Competition and Association.

**Assessment-By-Training Interaction for Competition.** I observed a reliable Training Group-by-Competition effect for posttest generation time when controlling for pretest time ( $F(2,142)=10.898$ ,  $p<0.001$ ,  $BF>100$ ), accompanied by main effects of both factors ( $F$ 's $>7.656$ ,  $p$ 's $<0.01$ ,  $BF$ s $>19.96$ ). This prompted me to examine the Training-by-Assessment interaction at each level of Competition, with the expectation that only the High-Competition items would bear a reliable result. Indeed, High-

---

<sup>15</sup> Participants excluded on the basis of drop-out or extended training time (see Method section) were no different from the current subset of included participants in terms of baseline production times on the verb generation task ( $F(1,438)=1.257$ ,  $p=0.26$ ).

( $F(2,147)=26.614$ ,  $p<0.001$ ,  $BF>100$ ) but not Low-Competition nouns ( $F(2,146)=1.900$ ,  $p=0.15$ ,  $BF=0.07$ ) demonstrated a significant Assessment-by-Training Group interaction (see Table 8). Main effects of Assessment were observed in both conditions ( $F$ 's $>67.771$ ,  $p$ 's $<0.001$ ), while a main effect of Training Group only occurred for High-Competition items ( $F(2,147)=4.565$ ,  $p=0.01$ ,  $BF=0.968$ ; see Table 8). Planned comparisons testing for cross-Assessment changes for the Lures Group revealed a selective effect for High-Competition nouns ( $t(27)=2.000$ ;  $p=0.05$ ,  $BF=0.687$ ; Low-Competition:  $t(27)=-1.07$ ;  $p=0.29$ ,  $BF=0.187$ ). No other Training Groups demonstrated this selectivity in terms of competition-sensitive pre/post improvement. Instead, the No-Lures Group did not improve reliably for either item type ( $t$ 's $<1.43$ ,  $p$ 's $>0.16$ ,  $BF$ s $<0.286$ ) and the 3-Back Group improved reliably for both High- and Low-Competition nouns ( $t$ 's $>2.51$ ,  $p$ 's $<0.02$ ,  $BF$ s $>1.796$ ). That is, the 3-Back Group contributes to the interaction term of the high-competition test, but this pattern is not selective to just these high-EF items. Indeed, as illustrated in Figure 12A, the Lures Group is quicker by 361ms to generate verbs for High-Competition nouns, and only 33ms faster for Low-Competition items; the pre/post change for the No-Lures Group was 249ms and 78ms for High- and Low-Competition items, respectively. Finally, the 3-Back Group improves the most of all of the Training Groups on both item types, demonstrating a generation time reduction of 385ms word with many competitors and 496ms for those with few contenders for production, showing a larger improvement for cases when competition is attenuated (see Figure 12A). The lack of selectivity for the 3-Back Group favors an interpretation of a generalized—not a process-specific—benefit.



**Assessment-By-Training Interaction of Association.** Similar to the Competition

effects reported above, I conducted an ANCOVA testing for a Training Group-by-Association effect for posttest generation time while controlling for pretest time.

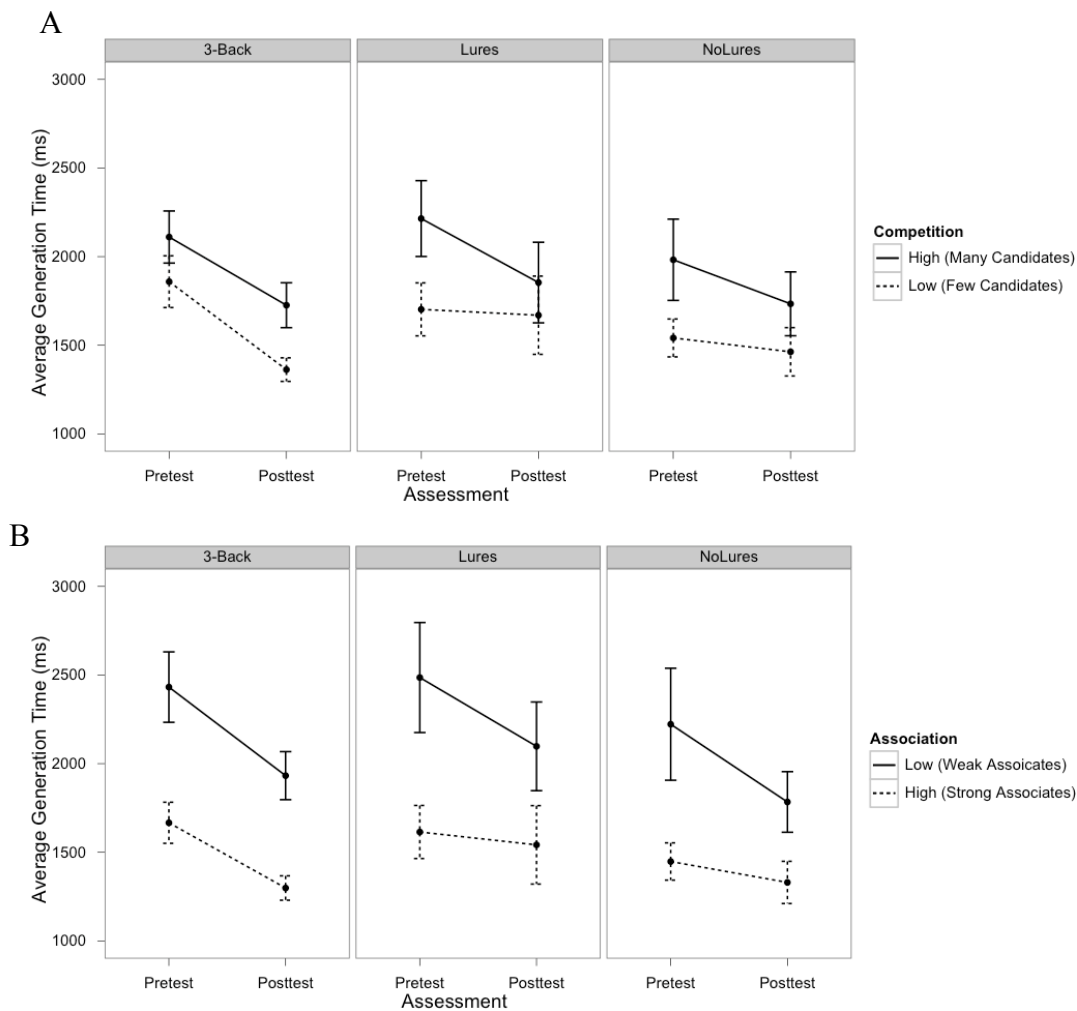


Figure 12. Pre/post verb generation latency for nouns in terms of (A) competition level (high-interference items with many production candidates vs. low-interference items with limited contenders for production) and (B) association level (low-retrieval demand items with strong near neighbors vs. high-retrieval demand items with weak near neighbors) for each training group (3-Back in leftmost panels, Lures in center panels, and No-Lures in rightmost panels). Error bars =  $\pm 1$  standard error of the mean.

Although this did not reach significance ( $F(2,141)=0.964$ ,  $p=0.38$ ,  $BF=0.314$ ), I observed main effects of Training Group, Association, and Assessment ( $F$ 's  $> 11.397$ ,  $p$ 's  $< 0.001$ ,

BFs>100). I investigated the nature of these main effects by separately testing the effect of Training and Assessment on High- and Low-Association items: All groups improved on Low-Association words, responding faster to such items at posttest relative to pretest performance (Lures: 893ms, No-Lures: 388ms, 3-Back: 499ms; see Figure 12B) evidenced by a reliable main effect of Assessment among Low-Association nouns ( $F(1,147)=88.416, p<0.001, BF>100$ ). However, probing the main effects of Training and Assessment ( $F$ 's>5.16,  $p$ 's<0.007, BFs>1.688) for High-Association nouns revealed that only the 3-Back Group demonstrated a reliable pre/post change on such items—a 368ms reduction in generation time (Lures: 73ms, No-Lures: 118ms; see Figure 12B and Table 8). Paired with the non-selective benefits experienced by the 3-Back group regardless of Competition level, it is evident that 3-Back training had wide-ranging benefits on word production times. One possibility for this general improvement might involve improved general speed-of-processing, as this group's only index of training performance feedback was accuracy and response time (contrary to the adaptive groups who observed their  $n$ -levels dynamically changing with performance).

Finally, that the Lures Group exclusively improves on high-effort conditions (High-Competition and Low-Association) in the absence of comparable boosts on low-effort cases (Low-Competition and High-Association) might be representative of their newfound competence to deal with such high-complexity scenarios. This explanation seems unlikely given that other difficult items do not enjoy such benefits (see object-extracted relative clause results below). Moreover, much work (see Martin & Cheng, 2006; Snyder, Banich, & Munakata, 2011; Wagner, Paré-Blagoev, Clark, & Poldrack, 2001) emphasizes the relevance of cognitive control for effortful retrieval from semantic

memory, making it possible that the benefits experienced by the Lures Group for Low-Association items are driven by different sub-processes than those demonstrated by the No-Lures and 3-Back Groups. This seems quite feasible in light of the Lures Group's reliable pre/post changes on other untrained high-conflict measures alongside absent effects among the No-Lures and 3-Back Groups.

### *3.3.7 Lingering Garden-Path Recovery (Comprehension Accuracy)*

**Analysis.** Offline comprehension data from six subjects was not included due to poor accuracy to comprehension questions following non-critical filler items (1 subject in the Lures group; 1 in the No-Lures group; 4 in the 3-Back group). For the remaining participants, I first used a mixed-effects model to conduct a manipulation check of Ambiguity at pretest, collapsed across Training Group. I anticipated that ambiguous materials would provoke difficulty in interpretation-recovery, evidenced by worse accuracy to comprehension questions relative to unambiguous control sentences. Next, I included Assessment (pre vs. post) and Training Group (Lures vs. No-Lures vs. 3-back) as fixed factors in the mixed-effects models for each level of Ambiguity (ambiguous vs. unambiguous) with the expectation that an interaction would only emerge for ambiguous sentences, and not unambiguous items. To confirm the mixed-effects model outcomes, I then ran ANCOVAs separately for each Ambiguity level testing for the effect of Training Group on posttest accuracy, controlling for pretest accuracy. These models normalize all training groups' baseline performance and test for differences at the Group level at posttest. Finally, I conducted planned contrasts of cross-Assessment performance for the Lures Group at each level of ambiguity. Similar to the tests conducted for Stroop and the recognition memory task, these paired comparisons were used to understand the nature of

high- and low-interference task conditions for the subjects who practiced interference-resolution training. The prediction I held was consistent with Experiment 1's findings: The Lures Group should only demonstrate improved pre/post accuracy to questions following high-interference ambiguous constructions; no change was expected for unambiguous sentences, where the demand to resolve among competing interpretations is removed.

**Baseline Ambiguity Results.** To determine first whether the ambiguous materials imposed the hypothesized difficulty compared to unambiguous items, I fit mixed-effects models of the accuracy data for pretest only, crossing Subjects and Items as random effects and including Sentence-Type (Ambiguous, Unambiguous) and Training Group (Lures, No-Lures, 3-Back) as fixed factors. The best-fitting mixed-effects model included a reliable effect of Sentence-Type, revealing significantly more errors in ambiguous (31%) than unambiguous (9%) conditions ( $z=6.718, p<0.001$ ). Moreover, that Training Group was not a significant contributor to this model suggests that the Training Groups did not perform differently from each other at pretest ( $z's<1.068, p's>0.28$ ).<sup>16</sup> This pattern indicates that these sentence items elicited a lingering garden path effect, and did so comparably for participants assigned to each training group. Accordingly, I tested if Training Group predicted improvements in garden-path recovery from pre- to posttest using both mixed-effects models and ANCOVAs.

**Assessment-By-Training Interaction.** When analyzing cross-Assessment changes in accuracy, there was a main effect of Assessment for the ambiguous ( $z=-3.892,$

---

<sup>16</sup> A mixed-effects models testing for an effect of participation-exclusion did not lead to a reliable difference between participants excluded on the basis of drop-out or extended training time from those included in terms of baseline ambiguity effects ( $z's<1.176, p's>0.23$ ).

$p < 0.001$ ) but not the unambiguous sentences ( $z$ 's  $< 0.812$ ,  $p$ 's  $> 0.41$ ). Indeed, the best-fitting model included only the intercept suggesting that the fixed factors did not account for accuracy patterns in unambiguous items (see Table 9). An analysis of covariance verified the outcome of these mixed-effects models. To test for posttest accuracy difference among Training Groups while controlling for pretest performance, I ran two separate ANCOVA models for each level of Ambiguity (ambiguous and unambiguous), and replicated the above mixed-effects outcomes. I failed to demonstrate an Assessment-by-Training Group interaction for either sentence type ( $F$ 's  $< 1.27$ ,  $p$ 's  $> 0.28$ ,  $BFs < 0.001$ ), but observed a main effect of Assessment for ambiguous items ( $F(1,69) = 41.204$ ,  $p < 0.001$ ,  $BF > 100$ ); no such effect was seen for unambiguous sentences (see Table 8).

Significant Model Parameters		Beta			AIC <sub>C</sub> with / without slopes
		Estimate	SE	z-value	
<i>Ambiguous</i>	Intercept	2.0349	0.3833	5.309***	1594/1582
	Assessment	-0.9087	0.2335	-3.892***	
<i>Unambiguous</i>	Intercept	3.05321	0.3566	8.561***	989/960

Table 9. Significant fixed effects from the best fitting mixed-effects models of garden-path sentence comprehension accuracy data, testing for an Assessment (1 vs. 2) by Group (Lures vs. No-Lures vs. 3-Back) interaction separately for ambiguous and unambiguous items. When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and Items were input into the model as crossed random effects. Excluding random slopes yielded better fits of every model, as indexed by lower AIC<sub>C</sub> values for models without random slopes as compared to those with random slopes. Thus the best-fitting models *without* random slopes are reported here. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$

Planned comparisons aimed at evaluating the pre/post change for each Training Group bolstered these findings: On ambiguous items, the Lures Group improved by 10.9% ( $t(28) = 3.494$ ,  $p = 0.001$ ,  $BF = 16.09$ ), the No-Lures Group by 15.5% ( $t(21) = 2.481$ ,  $p = 0.02$ ,  $BF = 1.77$ ), and the 3-Back Group by 12.6% ( $t(23) = 3.009$ ,  $p = 0.006$ ,  $BF = 5.117$ );

while on unambiguous items, no group demonstrated significant changes ( $t$ 's $<1.351$ ,  $p$ 's $>0.19$ ,  $BFs<0.254$ ). Figure 13 illustrates this pattern, such that all Training Groups demonstrated robust cross-Assessment improvements in comprehension accuracy to questions following ambiguous sentences, while no effect appeared for unambiguous sentences across sessions. That is, every training group demonstrated a selective effect, favoring boosts on high-conflict ambiguous items, but not low-conflict sentences. Such a pattern supports the notion that all training groups' performance boosts were due simply to practice effects. This possibility is unlikely, however, alongside the pre/post accuracy effects for the untrained controls of Experiment 1 (Chapter 2). Recall that these participants showed no significant improvement upon encountering ambiguous items a second time at posttest (7.2% change, compared to a 16.7% change among  $n$ -back responders of that study). Together, this suggests that the  $n$ -back task, *in general*, may be regarded as necessary and sufficient to confer benefits in offline comprehension to questions probing for misinterpretation of temporarily ambiguous sentences. Indeed, a component inherent to all versions of  $n$ -back involves using memory to update representations as new information is presented (see Chatham et al., 2011). This ability may be critical for reinstating the sentential information encoded on the previous screen while attempting to answer comprehension questions, a demand which may usurp interference-resolution abilities, aiding reinterpretation processes. Perhaps a more appropriate index of reanalysis that taps interference resolution involves eye movements that occur when conflicting information is initially encountered in real-time. That is, similar to Experiment 1, I assessed pre/post changes in regression path time, or the time

associated with fixations launched to earlier regions of the sentence when new input is encountered before revisiting that information.

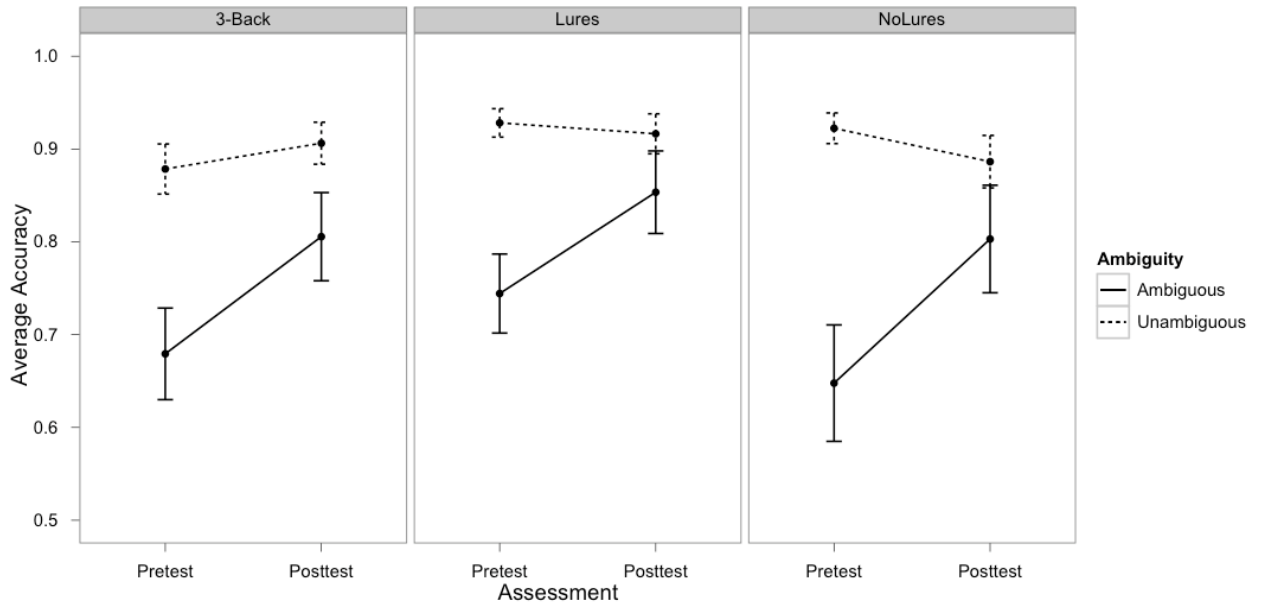


Figure 13. Accuracy to comprehension questions probing for lingering effects of misinterpretation of garden-path sentences. Pre/post accuracy by sentence type (high-interference ambiguous and low-interference unambiguous items) for each training group (3-Back in leftmost panels, Lures in center panels, and No-Lures in rightmost panels). Error bars =  $\pm 1$  standard error of the mean.

### 3.3.8 Real-time Reanalysis of Garden-Path Sentences

**Analysis.** Eye-movement data were excluded from eight participants who could not be calibrated at either pre- or post-test (2 subjects in the Lures group; 4 in the No-Lures group; 2 in the 3-Back group). Following from Experiment 1 (Novick et al., 2013), sentences were divided into four regions of interest (see Table 4), with the region of interest being sentence-final, where conflicting evidence first arrives in ambiguous sentences (Region 4; “sparkled brightly”). In unambiguous sentences, Region 4 was examined as a comparison, to match the region of analysis to the position of the critical

region in ambiguous sentences. Due to the reversed clause order, this region necessarily contains different semantic content; despite this, I compared both regions, given that regression path time is liable to capture some component of wrap-up effects. By comparing the sentence-final region in both sentences, I neutralized the potential contribution of wrap-up. Namely, the semantically comparable region of unambiguous sentences (Region 2) is relieved of wrap-up effects, given its position (see Table 4); thus, if such effects do appear, this should occur regardless of the construction (i.e., ambiguous and unambiguous equally), rather than posing the risk of being an additional potential confound when comparing across sentence types.

I conducted analyses on correct trials only as a means of measuring eye-movement patterns during successful garden-path recovery, that is, when one would expect readers to make leftward saccades in search of information to help them revise and ultimately arrive at the correct interpretation. I conducted multilevel mixed-effects models for ambiguous and unambiguous materials separately with Subjects and Items as crossed random effects. Initially, I conducted these models for just Pretest data to test for Training Group effects, expecting to find no baseline differences among groups. I then fit models to include both Assessment and Group as potential fixed factors. MCMC simulations were used to assess the effects of each fixed parameter within the mixed-effects models.

To test for the individual contribution of interference resolution for real-time recovery efforts, I designated contrasts within the Group level to capture on this difference; that is, the intercept of these models was the Lures Group to ensure that I could test for effects between this critical group and every other group. By this contrast,



any Training-by-Assessment interaction would indicate that the Lures Group has a significantly different cross-Assessment change compared to the contrasted group. That is, a reliable interaction term for the No-Lures (or 3-Back) Group would indicate a significant difference between Lures and No-Lures (or 3-Back) training in regression path time, while a non-significant interaction term would suggest no training-mediated difference between groups. Thus, by a process-specific account, I expected to find two reliable interaction terms for ambiguous sentences and no interactions for unambiguous items. Finally, I verified the multi-level mixed-effects models with ANCOVA models testing for posttest regression path time differences among Training Groups, covarying out pretest regression-path time.

**Baseline Regression-Path Time Results.** I fit a mixed-effects model that crossed the random effects of Subjects and Items for each level of Ambiguity to test for the fixed effect of Training Group for regression-path time in the sentence-final (critical) region. I observed no baseline differences as a function of Group for either level of ambiguity ( $t$ 's < 1.124,  $p$ 's > 0.26).<sup>17</sup> Furthermore, to replicate the patterns in Experiment 1, I also tested for the fixed effect of Region for ambiguous items, expecting to find a main effect indexed by longer regression path time for “sparkled brightly.” Indeed, I observed a main effect of Region ( $F(3,64)=419.123$ ,  $p<0.001$ ,  $BFs>100$ ), such that the final region of ambiguous sentences elicited longer regression path times at pretest than any other region (Region 1  $M=713$ ms, Region 2  $M=488$ ms, Region 3  $M=1121$ ms, Region 4  $M=2686$ ms). This was also the case for unambiguous sentences ( $F(3,64)=330.78$ ,  $p<0.001$ ,  $BFs>100$ ;

---

<sup>17</sup> Mixed-effects models testing for an effect of participation-exclusion did not lead to a reliable difference between participants excluded on the basis of drop-out or extended training time from those included in terms of pretest regression path time in all regions of ambiguous ( $z$ 's < 1.289,  $p$ 's > 0.19) and unambiguous sentences ( $z$ 's < 0.831,  $p$ 's > 0.40).

Region 1 M=266ms, Region 2 M=962ms, Region 3 M=721ms, Region 4 M=1886ms), indicating two important possibilities: First, every region of unambiguous items resulted in faster regression-path time relative to ambiguous items, suggesting that all regions of

Significant Model Parameters		Beta Estimate	SE	t-value	AIC <sub>C</sub> with / without slopes
<i>Ambiguous</i>	Intercept	2181.51	221.35	9.855***	14000/13993
	Assessment	871.16	173.61	5.018***	
	Assessment x Training (No-Lures)	-526.40	282.14	-1.866 <sup>†</sup>	
<i>Unambiguous</i>	Intercept	1664.90	143.15	11.630***	11127/11100
	Assessment x Training (3-Back)	395.35	210.67	1.877 <sup>†</sup>	

Table 10. Significant fixed effects from the best fitting mixed-effects models of regression-path time following entry into the final region only for garden-path materials testing for an Assessment (1 vs. 2) by Group (Lures vs. No-Lures vs. 3-Back) interaction separately for Ambiguous and Unambiguous items. The intercept for both models is the Lures Group at Posttest. Markov Chain Monte-Carlo (MCMC) simulations were conducted to test for the significance of each fixed effect, through which I generated 10,000 samples from the posterior distribution. When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and Items were input into the model as crossed random effects. Excluding random slopes yielded better fits of every model, as indexed by lower AIC<sub>C</sub> values for models without random slopes as compared to those with random slopes. Thus, the best-fitting models *without* slopes are reported here. <sup>†</sup>p<.06, \*p<.05, \*\*p<.01, \*\*\*p<.001

ambiguous items have room to improve. Thus, if only the final region of these items resulted in cross-Assessment improvement (replicating Experiment 1’s patterns), then this can be considered a selective region effect. Second, regression-path times of the final region of both constructions was exaggerated relative to all other regions within each sentence, suggesting that a selective benefit for just ambiguous sentences would rule out any explanations of regression to the mean. That is, the region in which there is the most

room to improve should be the region demonstrating greatest improvement if regression to the mean is indeed the driving factor of any cross-Assessment effects.

**Assessment-By-Training Interaction of Regression-Path Time.** The top panel of Table 10 shows the results of MCMC simulations for all mixed-effects models testing for fixed effects of Assessment and Training Group that fit the total regression-path data from the sentence-final region (Region 4). The model for ambiguous items yielded a marginal Assessment-by-Training Group interaction for the No-Lures Group ( $t=-1.866$ ,  $p=0.06$ ), but not 3-Back Group ( $t=-0.699$ ,  $p=0.48$ ), indicating that the cross-Assessment change between the Lures and No-Lures Groups was trending toward being different, but that the Lures and 3-Back Groups were not different in terms of pre/post change. No other region demonstrated this pattern ( $t$ 's $<1.131$ ,  $p$ 's $>0.25$ ). Unpacking this effect with paired comparisons of pre- and posttest performance for each group, I observed reliable improvements for the Lures ( $t(25)=-3.1241$ ,  $p=0.004$ ,  $BF=6.649$ ) and 3-Back ( $t(18)=-3.107$ ,  $p=0.006$ ,  $BF=5.934$ ) Groups, but not the No-Lures Group ( $t(14)=-1.5584$ ,  $p=0.14$ ,  $BF=0.448$ ). Figure 14A illustrates this effect, such that the cross-Assessment change for the Lures Group in the center panel (a 785ms improvement) trumps the No-Lures Group (274ms), with the 3-Back Group demonstrating a smaller quantitative improvement (680ms) that is not different from that observed by the Lures Group. Given that all training groups improved on comprehension accuracy to questions following ambiguous sentences, the lack of an effect for the No-Lures Group in terms of regression path time when disambiguating information is first encountered may, instead, be the aberrant case; that is, the No-Lures trainees' failure to demonstrate faster real-time reanalysis following training may be the byproduct of a Type II error. One source of evidence suggesting

otherwise is the No-Lures Group's lack of cross-assessment improvement on any other untrained tasks (Stroop, recognition memory, and verb generation) aside from an offline assessment of misinterpretation.

Surprisingly, the model of unambiguous items also revealed a marginal Assessment-by-Training Group interaction for the 3-Back Group ( $t=1.877$ ,  $p=0.06$ ).<sup>18</sup> That this pattern was not observed when comparing the Lures and No-Lures Groups suggests that the marginal effect for ambiguous items may be selectively due to the minimal intervention difference separating these two Groups, namely, practice with interference lures. I examined this marginal interaction by comparing pre- and posttest performance for each training group. This revealed a significant cross-Assessment effect for 3-Back trainees ( $t(18)=-4.0028$ ,  $p<0.001$ ,  $BF=36.077$ ), but no other group ( $t's>-0.8327$ ,  $p's>0.41$ ,  $BFs<0.166$ ). Figure 14B depicts this effect among unambiguous sentences, where the 3-Back Group in the leftmost panel sped up by 841ms from pre to posttest, while the Lures (190ms) and No-Lures Groups (177ms) showed no analogous change.

Alongside the cross-Assessment comparisons for ambiguous items (as well as the verb generation effects in the previous section), these patterns provide support for a *general* improvement profile for 3-Back trainees: They spend significantly less time returning to earlier regions following entry to the final sentence region *regardless of the presence of conflict* at posttest relative to pretest. Moreover, the Lures Group possessed an improvement profile consistent with a process-specific account; namely, Lures trainees were faster to revisit earlier regions only after encountering conflicting

---

<sup>18</sup> No interactions were observed for any other unambiguous sentence regions ( $t's<0.982$ ,  $p's>0.32$ ).

information in ambiguous sentences; unambiguous sentences were unaffected (see Figure 14). The No-Lures Group enjoyed no benefits on either item type, as expected.

A series of ANCOVA models were used to verify these effects, wherein an analysis of regression path time at the sentence-final—“sparkled brightly”—region of ambiguous items yielded a reliable Assessment-by-Training Group interaction. No other region of ambiguous sentences showed this pattern ( $F's < 1.8232$ ,  $p's > 0.17$ ,  $BF = 0.082$ ). Put differently, the pre/post correlations for each training group were significantly different. The Lures Group demonstrated a slope equal to 0.259, while the No-Lures and 3-Back Groups showed higher test-retest reliability with respective slopes of 0.958 and 0.418. I compared slopes with t-tests to learn that the Lures Group's slope coefficient was significantly different from the No-Lures Group's slope ( $t(37) = -2.414$ ,  $p = 0.01$ ,  $BF = 2.235$ ), but not 3-Back Group's slope ( $t(41) = -1.005$ ,  $p = 0.16$ ,  $BF = 0.271$ ). Interestingly, the comparison of the No-Lures and 3-Back Group's slopes was also reliably different ( $t(30) = -1.839$ ,  $p = 0.03$ ,  $BF = 0.832$ ). This suggests that the Lures Group demonstrated the smallest test-retest reliability (though, no different from the 3-Back Group), a pattern perhaps consistent with Lures trainees' greatest cross-Assessment change in regression path time in the disambiguating region of ambiguous sentences (785ms speed-up). That the pre/post slope of the 3-Back Group is no different from the Lures Group follows from their comparable, yet quantitatively smaller regression-path gains of 680ms. Finally, the 3-Back Group's deviation from the No-Lures Group is consistent with the No-Lures Group's lack of a pre/post change, on average, compared to both the remaining two training groups.

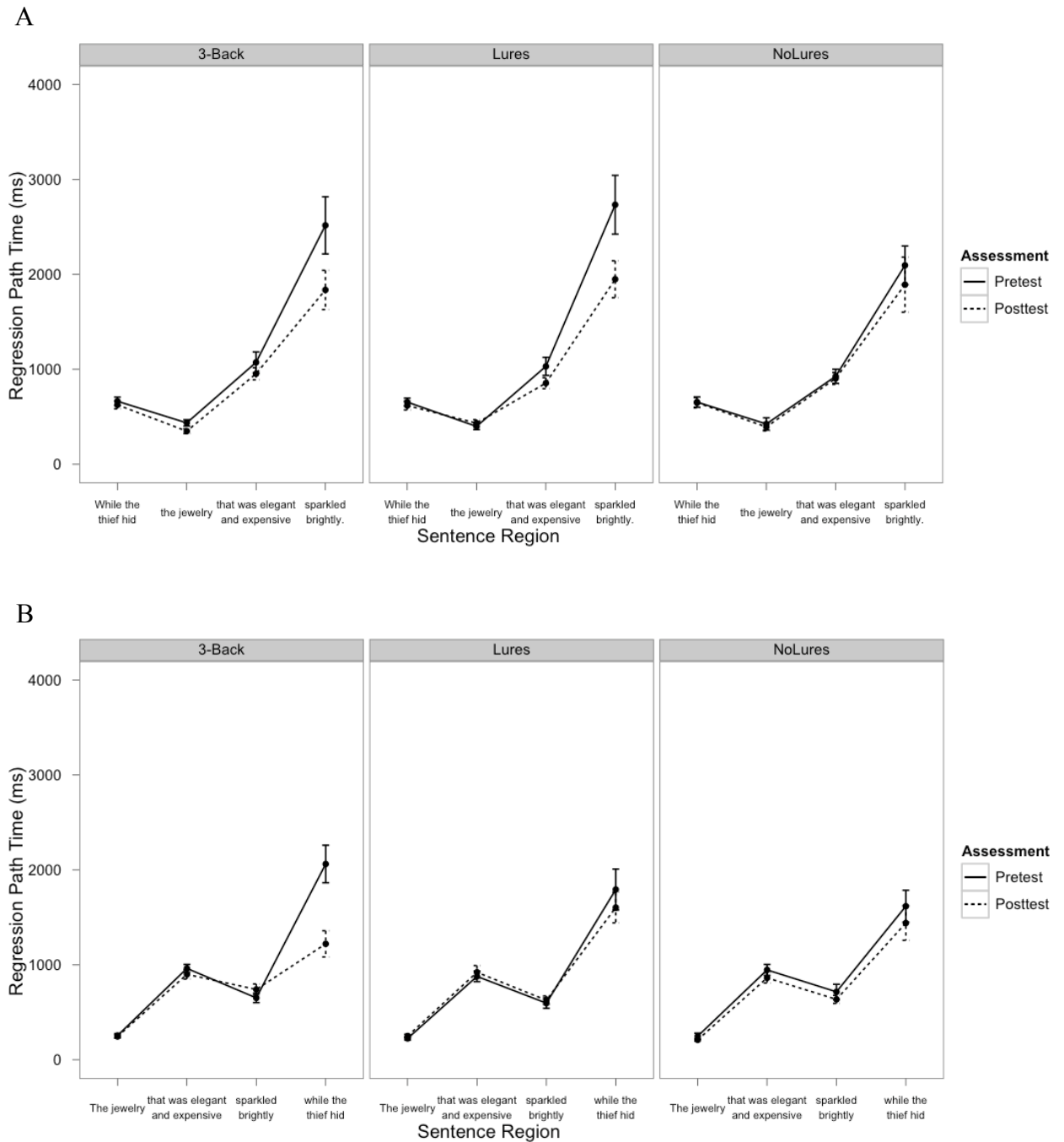


Figure 14. Regression-path time—the total time associated with fixations from the point that a reader first enters a region from the left before exiting it to the right—for garden-path sentences. Regression path time by sentence region (note: “sparkled brightly” of ambiguous sentences is the primary region where conflict arises) at each assessment for each training group (3-Back in leftmost panels, Lures in center panels, and No-Lures in rightmost panels) for (A) ambiguous and (B) unambiguous sentences.

To further probe this effect, I conducted ANCOVAs emphasizing minimal Training Group contrasts; that is, I tested for the Assessment-by-Training Group interaction with consideration for *pairs* of groups (e.g., Lures vs. No-Lures, Lures vs. 3-Back, and No-Lures vs. 3-Back). By this method, a reliable interaction emerged when I examined the effect of Group (Lures vs. No-Lures) for Region 4 of ambiguous posttest regression path time controlling for pretest reading times ( $F(1,37)=5.7892$ ,  $p=0.02$ ,  $BF=1.885$ ). A similar model honing in on a comparison of the No-Lures and 3-Back Groups yielded a marginal interaction term ( $F(1,33)=3.7743$ ,  $p=0.05$ ,  $BF=0.788$ ), while the analysis assessing Group with respect to a Lures Group/3-Back Group contrast did not reach significance ( $F(1,41)=0.9370$ ,  $p=0.33$ ,  $BF=0.31$ ). This pattern is consistent with the outcome of the mixed-effects models reported above: the Lures Group improves to a greater degree on ambiguous sentences following training relative to the No-Lures Group. Importantly, the comparable analysis for Region 4 of unambiguous sentences resulted in no such effects ( $F(2,56)=0.0592$ ,  $p=0.94$ ,  $BF=0.018$ ); the same null effect held for all other regions ( $F$ 's < 2.4056,  $p$ 's > 0.09,  $BF$ s < 0.167). This pattern is not entirely consistent with the mixed-effects models of regression-path time for the sentence-final region of unambiguous items, which demonstrated a marginal Assessment-by-Training Group interaction for the 3-Back Group ( $F(2,54)=3.2672$ ,  $p=0.04$ ,  $BF$ s = 0.385).

One possible reason for this divergence may involve the assumptions of ANCOVA analyses versus mixed-effects models. The hierarchical models treated Assessment as a repeated measure, an approach that allows cross-Assessment change of a variable to be assessed, while the ANCOVAs capitalize on test-retest reliability to glean something about the Training Group effects *at posttest* given normalized pretest

performance. Moreover, the mixed-effects models nested the random factors of Subjects and Items, making it possible that an effect at the Item level for even a subset of subjects yielded a marginal interaction.

Even if a general cross-Assessment boost is assumed for the 3-Back trainees, such a pattern is not inconsistent with an interference-resolution process-specificity account; the Lures Group pattern (selective improvement on just high-conflict ambiguous items) aligns with such an account. Put differently, the group that was prompted to practice interference-resolution abilities throughout training enjoyed selective improvements when confronted with conflict in real-time sentence processing. To further unpack this effect, I conducted analyses on residual second pass times with the assumption that re-reading should more directly index the revision component of regression path time. As a result, if the regression-path time improvement of the Lures Group—and not the 3-Back Group—is one driven by real-time *revision*, then the Lures Group should be the only group to demonstrate improvement in second-pass time in earlier sentence regions following training. If a process other than revision is the locus of the 3-Back Group's regression-path time improvement (e.g., wrap-up), then no pre/post change in second-pass time is expected in earlier regions of ambiguous sentences for these trainees. Although total second-pass (re-reading) time may be one index of real-time revision properties, a more optimal eye movement measure would conditionalize second-pass time on first entering the disambiguating region. Indeed, re-reading during the initial regions prior to entering the region of conflict is included in a total second-pass measure. In light of this, alongside the regression path time patterns wherein no other regions (aside from region 4) demonstrate cross-assessment changes, it is unlikely that much of



the total second-pass time is due to rereading of earlier regions before encountering the sentence-final critical region. Thus, the next section focuses on total second-pass time as a means to isolate the revision component of regression-path time to test the hypothesis that the Lures Group improves in terms of reanalysis, while the 3-Back Group's pre/post gains are due to some other reading time component.

**Assessment-by-Training Interaction of Second Pass Time.** Raw re-reading times were corrected for variation in string length for the four specified sentence regions of ambiguous and unambiguous sentences. For each subject at each assessment, I computed a regression equation of re-reading time as a function of region length (in number of characters) for all ambiguous, unambiguous, and a subset of filler items; the residual reading times used for the present analysis encompassed the difference between raw and predicted reading times (see Trueswell, Tanenhaus, & Garnsey, 1994). Table 11 presents the results of MCMC simulations for all mixed-effects models testing for fixed effects of Assessment and Training Group that fit residual reading times for all sentence regions. The intercept reflects performance of the No-Lures Group at posttest; I chose to contrast the other groups to the No-Lures Group to test whether the Lures and 3-Back Groups demonstrated similar improvements to these trainees who did not change in their regression-path time. Again, if Lures trainees' regression-path time gains reflects improved real-time revision and 3-Back trainees' patterns reflect a process separate from revision, then the mixed-effects models should only yield a reliable Assessment-by-Training Group interaction for the Lures Group (not the 3-Back Group). Indeed, in earlier regions of ambiguous sentences, I observed a pattern consistent with this hypothesis: The Lures Group was different from the No-Lures Group in Region 1 ( $t=-2.193$ ,  $p=0.02$ ) and

Region 3 ( $t=-2.033$ ,  $p=0.04$ ), evidenced by significant Assessment-by-Training Group interaction terms. The 3-Back Group, on the other hand, did not show this pattern ( $p's>0.53$ ). No assessment-by-group differences emerged in Regions 2 or 4 of ambiguous sentences for any group contrasts ( $p's>0.14$ ).

Pairwise comparisons were used to unpack the reliable interactions by testing for pre/post changes in each region for each training group. Following training, the Lures Group demonstrated significantly less second pass time in Regions 1 ( $t(25)=-2.94$ ,  $p=0.006$ ,  $BF=6.195$ ) and 3 ( $t(24)=-2.43$ ,  $p=0.02$ ,  $BF=2.755$ ); the remaining training groups did not show these effects ( $p's>0.16$ ). Moreover, despite no Assessment-by-Training Group interaction, the Lures Group also showed faster re-reading times in Region 2 ( $t(24)=-2.85$ ,  $p=0.008$ ,  $BF=5.537$ ), while the remaining groups did not ( $p's>0.07$ ). The sentence-final region was void of any such effects ( $p's>0.13$  for all Groups). Figure 15A illustrates these findings, such that the pre/post change for the Lures Group in the center panel is exaggerated compared to that of the No-Lures and 3-Back Groups. Namely, trainees practicing resolving among competing representations demonstrate 215.60ms speed-up on average in Region 1, 114.66ms in Region 2, and 160.28ms in Region 3.

Critically, no Assessment-by-Training Group interactions emerged on unambiguous items. The best-fitting mixed-effects models probing for this effect for second pass time in Regions 1 and 2 included no significant terms, and the models for Regions 3 and 4 included only the intercept (see Table 11). This null effect is evident in Figure 15B, wherein no groups show pre/post changes in any regions of unambiguous items. Indeed, these selective patterns were not observed across a host of commonly-

Significant Model Parameters		Beta Estimate	SE	t-value	AIC <sub>C</sub> with / without slopes
Region 1					
<i>Ambiguous</i>	Intercept	170.60	61.95	2.754**	13034/13022
	Assessment x Group (Lures)	-180.00	82.08	-2.193*	
<i>Unambiguous</i>	n.s.	-	-	-	5844/5832
Region 2					
<i>Ambiguous</i>	Intercept	149.68	32.99	4.537***	11473/11449
<i>Unambiguous</i>	n.s.	-	-	-	8966/8937
Region 3					
	Assessment x Group (Lures)	-176.20	86.25	-2.033*	12443/12416
<i>Unambiguous</i>	Intercept	-83.55	36.65	-2.28*	8179/8156
Region 4					
<i>Ambiguous</i>	n.s.	-	-	-	7821/7801
<i>Unambiguous</i>	Intercept	-80.50	36.63	-2.198**	6662/6633

Table 11. Significant fixed effects from the best fitting mixed-effects models of residual re-reading time for garden-path materials testing for an Assessment (1 vs. 2) by Group (Lures vs. No-Lures vs. 3-Back) interaction separately for Ambiguous and Unambiguous items. Markov Chain Monte-Carlo (MCMC) simulations were conducted to test for the significance of each fixed effect, through which I generated 10,000 samples from the posterior distribution. When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and Items were input into the model as crossed random effects. Excluding random slopes yielded better fits of every model, as indexed by lower AIC<sub>C</sub> values for models without random slopes as compared to those with random slopes. Thus, the best-fitting models *without* slopes are reported here. \*p<.05, \*\*p<.01, \*\*\*p<.001

reported eye movement measures, each of which may index processing other than reanalysis. I conducted mixed-effects models to test for the fixed effects of Assessment and Training Group to find no reliable Assessment-by-Training Group interactions for (1) Residual first-pass time for ambiguous ( $t$ 's<1.956,  $p$ 's>0.05) or unambiguous items ( $t$ 's<1.147;  $p$ 's>0.25) for any regions; (2) Residual total time for ambiguous ( $t$ 's<1.548,  $p$ 's>0.12) and unambiguous items ( $t$ 's<1.487,  $p$ 's>0.12) for any regions; or, (3)

Probability of regressing out of any region of ambiguous ( $t$ 's $<0.952$ ;  $p$ 's $>0.34$ ) or unambiguous sentences ( $t$ 's $<1.820$ ,  $p$ 's $>0.06$ ).

Taken together, the second-pass time results (crucially, in the absence of other eye movement results) suggest that the Lures Group improved on an index of the revision component of regression-path time, while the 3-Back Group—who also enjoyed regression-path time gains on ambiguous items—did not. Moreover, recall that the 3-Back Group improved in their regression-path time on unambiguous control sentences. These trainees showed no second-pass gains (on ambiguous and unambiguous sentences), suggesting that their boosts were, in part, due some other mechanism. Regardless of this changed process, what is important to note is that the 3-Back Group's regression-path effect does not appear to be one driven by improvement on a mechanism akin to interference-resolution (assuming that regression-path time aptly captures this ability); the patterns exhibited by the Lures Group, in contrast, does fall in line with such a process-specific account of cognitive control.

The regression-path and second-pass time patterns have important implications for the interpretations of the verb generation task results. Specifically, in terms of word production on the generation task and regression-path time, the Lures Group improved selectively on just high-conflict cases, the No-Lures Group failed to improve on any condition, and the 3-Back Group's benefits were blind to conflict demands. Second-pass time served as a proxy to draw out the revision component of regression-path time, and proved to verify the presence of a process-specific effect of cognitive control training for real-time revision of garden-path sentences. An appropriate measure sensitive to competition-based interference for the verb generation task could serve a similar role to

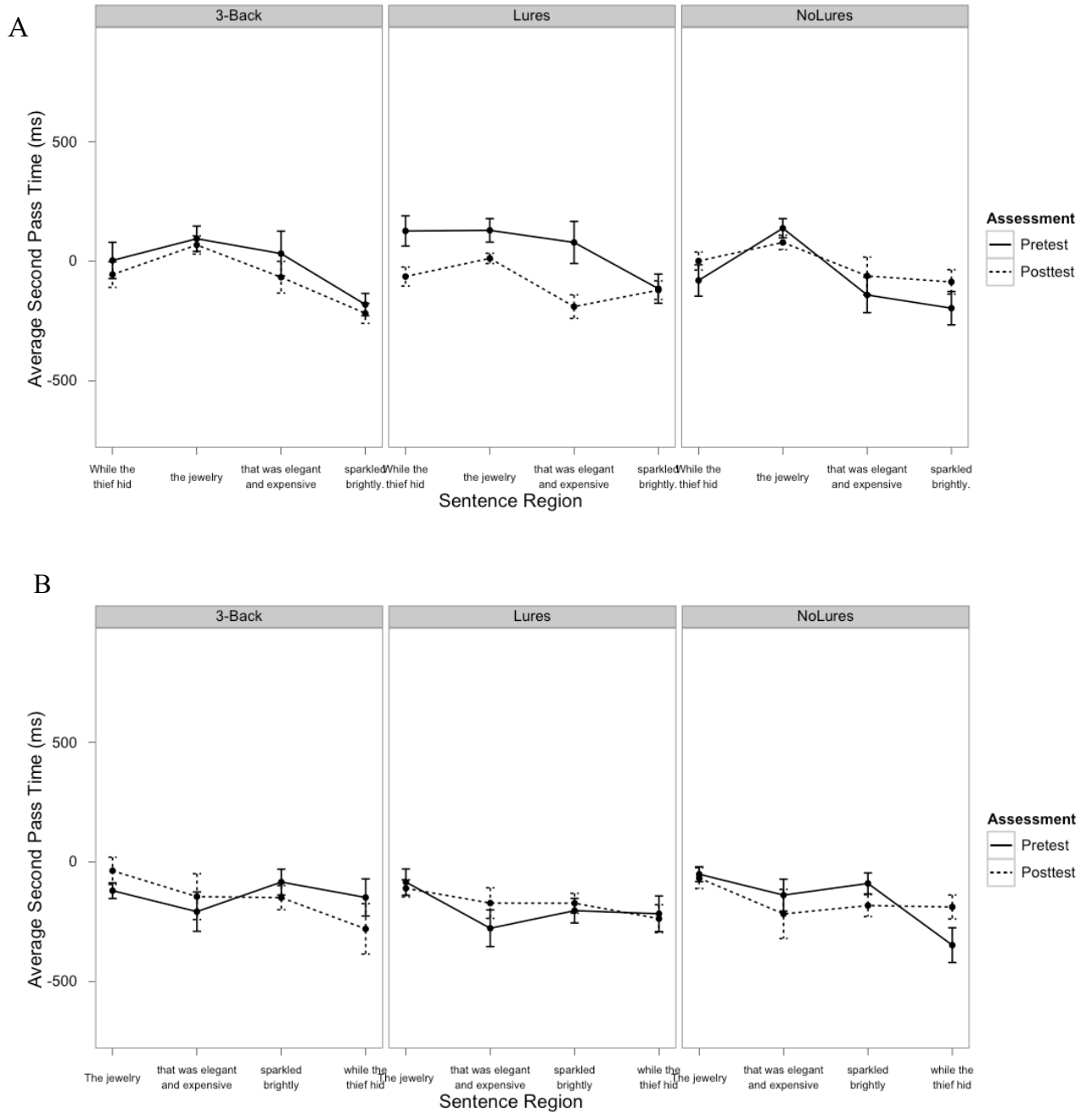


Figure 15. Residual second-pass time—the total re-reading time in a region corrected for string length of that region—for garden-path sentences. Second-pass time by sentence region at each assessment for each training group (3-Back in leftmost panels, Lures in center panels, and No-Lures in rightmost panels) for (A) ambiguous and (B) unambiguous sentences.

elucidate the mechanistic loci of the Lures and 3-Back Groups' word production improvements. In the meantime, it seems likely that a common underlying process is responsible for this repeated pattern across both tasks.

### 3.3.9 *Real-time Reanalysis of Relative-Clause Sentences*

In light of these patterns, one important observation—perhaps even a caveat—is that the most cognitive-demanding training task (Lures training) confers benefits to the most taxing parsing scenarios, i.e., when evidence conflicting with one's developing interpretation is first encountered. Other researchers have acknowledged the importance of general-purpose cognitive abilities for cases of heightened parsing difficulty, perhaps regardless of the presence of conflict (Fedorenko, Nieto-Castañón, & Kanwisher, 2012; Hagoort, Baggio, & Willems, 2009; but see Novick et al., 2009; Thothathiri, Kim, Trueswell, & Thompson-Schill, 2012). To test this hypothesis, I included syntactically-complex object-extracted relative clauses (compared to subject-extracted relatives), as processing difficulty with these constructions is well-documented (Caplan, Alpert, & Waters, 1998; Fedorenko, Gibson, & Rohde, 2006; Rogalsky, Matchin, & Hickock, 2008; *inter alia*). Despite such difficulty, most prior work emphasizes the importance of verbal working memory, broadly construed, as critical for processing object relatives. Thus, the role of interference resolution for relative clause processing is largely unknown. It is probable (as I assume here) that the nature of the conflicting representations arising during garden path recovery and relative clause parsing are quite unlike, rendering different degrees of cognitive control demand in each.

That is, in the case of garden-path recovery, the reader must first commit to an incorrect interpretation to be 'garden-pathed' (the state under which cognitive control is

triggered to allow an initial default interpretation to be overridden in favor of a new, correct one). Although some degree of conflict appears in relative-clause parsing, such that two eligible subject nouns are present (i.e., *farmer* and *expert* in Example [3]), readers do not encounter a comparable situation forcing them to rein-in and recharacterize a default cognitive reaction. Although object-relative clauses may introduce a form of syntactic ambiguity (subject-relatives are more common, and often expected when an embedded clause is encountered), the degree to which this conflict arises is not as exaggerated as that of garden-path sentences. Instead, I argue here that readers must make a real-time interpretive decision upon encountering the embedded clause (*who the expert questioned*) when parsing relativized constructions, a demand quite different from *revising* and *recharacterizing* an initial commitment. I argue here that a *revision* element involves interference-resolution skills more than such decision components. As a result, object relatives—being low-probability constructions—introduce parsing difficulty likely distinct from that requiring the level of cognitive control that is needed to countermand an initial processing commitment to recover from a garden-path scenario.

If practicing *n*-back lures confers boosted interference-resolution abilities rather than producing trainees skilled at dealing with difficult (low-probability) items, then the Lures Group should *not* demonstrate improvements under low-cognitive-control demands. Object relatives provide a case where complexity is amped up in the absence (or minimization) of conflict; thus, if Lures trainees are improving in their ability to manage high-effort items, then they should show faster second-pass time in the embedded clause of object relatives—where initial processing demands appear. No

effects should appear in the same region of subject relatives. No selective posttest benefits for the Lures Group would suggest a process-specific account: Only high-conflict scenarios (garden-path recovery) benefit from interference-resolution training, irrespective of a heightened need for general cognitive demands in the face of complexity and effortful processing.

**Analysis.** Following Fedorenko et al. (2006), I partitioned object- and subject-extracted relative clauses into four regions (see Table 12), with the region of interest housing the embedded clause (Region 2, where processing difficulty arises; *who the expert questioned/who questioned the expert*). To examine cross-Assessment changes in eye-movement patterns, I fit mixed-effects models of residual second-pass time—the measure used above to index real-time reanalysis in syntactically ambiguous sentences—with the fixed effects Assessment and Training Group nested in random effects of Subjects crossed with Items. Models were fit separately for subject- and object-relatives. I designated the intercept of these models to be the Lures Group to assess the degree of change given the presence of lure items during training (see rationale in the regression-path time discussion in the previous section). Any reliable Assessment-by-Training Group interaction would indicate that Lures trainees' pre/post change was different from that of another group. Two patterns would be necessary to infer that the Lures Group is the only group to demonstrate improvement in terms of re-reading time in the embedded clause region, a finding that would debunk a process-specific account of the garden-path results: First, the Lures Group's improvement would have to be different from the No-Lures and 3-Back Groups in terms of object-extracted relative clause performance, evidenced by two Assessment-by-Training Group interactions. Second, the Lures Group



would have to be identical to the No-Lures and 3-Back Groups on subject-extracted items. This latter pattern would have to follow from the notion that the Lures Group did not improve on these items, but did improve on the object relatives, where difficulty is enhanced. Finally, I conducted analyses on *correct trials only* as a means of measuring eye-movement patterns during cases when sentences were read and understood (this resulted in a loss of 13.2% of the total data at pretest and 15.6% at posttest).

<b>Sentence Type</b>	<b>Region 1</b>	<b>Region 2</b>	<b>Region 3</b>	<b>Region 4</b>
<i>Object-extracted relative clause</i>	The farmer	who the expert questioned	promoted the product	at the fair.
<i>Subject-extracted relative clause</i>	The farmer	who questioned the expert	promoted the product	at the fair.

Table 12. Relative clauses were divided into regions for fine-grain analysis based on regions specified in earlier work (see Fedorenko et al., 2006). Note that for object relatives, region 2 (“who the expert questioned”) is the critical region wherein the most processing difficulty is encountered.

**Assessment-by-Training Interactions of Second-pass Time.** Table 13 presents the results of MCMC simulations for mixed-effects models of subject- and object-extracted relative clauses testing for fixed effects of Assessment and Training Group that fit residual reading times for the critical sentence region (Region 2). Regardless of the Sentence-Type, I observed no reliable Assessment-by-Training Group interactions or main effects of Assessment of Group; the only reliable contributor in both cases was the intercept ( $t$ 's > 2.878;  $p$ 's > 0.004). No other region of these sentences demonstrated effects ( $p$ 's > 0.18 for subject relatives and  $p$ 's > 0.34 for object relatives). Put differently, the Lures Group was no different from the No-Lures and 3-Back Groups in terms of re-reading times in the critical embedded clause region regardless of sentence type. Further,

by examining other eye movements, which may index processes other than conflict-control, I observed no reliable Assessment-by-Training Group interactions: Mixed-effects models of regression-path time did not reveal such effects for object- ( $t$ 's<1.150:  $p$ 's>0.25) or subject-extracted sentences ( $t$ 's<1.906:  $p$ 's>0.05) in any regions, nor did residual total time ( $t$ 's<1.592:  $p$ 's>0.11), residual first-pass time ( $t$ 's<1.470:  $p$ 's>0.14), and probability of regressing out of any region ( $t$ 's<1.436:  $p$ 's>0.15).

Planned comparisons testing for pre/post changes in Region 2 of object-relative sentences for each Training Group bolstered this finding: Following training, the Lures Group demonstrated no change in second-pass time ( $t(25)=0.4384$ ,  $p=0.66$ ,  $BF=0.461$ ); the remaining training groups also did not show cross-Assessment changes ( $p$ 's>0.08). Figure 16A shows this pattern, such that the pre/post change for all trainees, regardless of group assignment, show no reliable difference. However, it is interesting to highlight the

Significant Model Parameters		Beta Estimate	SE	t-value	AIC <sub>C</sub> with / without slopes
Total Time (Region 2)					
<i>Object-Extracted</i>	Intercept	276.37	76.99	3.589***	12632/12618
<i>Subject-Extracted</i>	Intercept	143.27	49.78	2.878**	12928/12900

Table 13. Significant fixed effects from the best fitting mixed-effects models of total residual reading time in the critical (second) region of relative clauses, testing for an Assessment (1 vs. 2) by Group (Lures vs. No-Lures vs. 3-Back) interaction separately for Object- and Subject-extracted items. Each panel depicts the effects in the models for each sentence region. Markov Chain Monte-Carlo (MCMC) simulations were conducted to test for the significance of each fixed effect, through which I generated 10,000 samples from the posterior distribution. When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and Items were input into the model as crossed random effects. Excluding random slopes yielded better fits of every model, as indexed by lower AIC<sub>C</sub> values for models without random slopes as compared to those with random slopes. Thus, the best-fitting models *without* slopes are reported here. \* $p<.05$ , \*\* $p<.01$ , \*\*\* $p<.001$

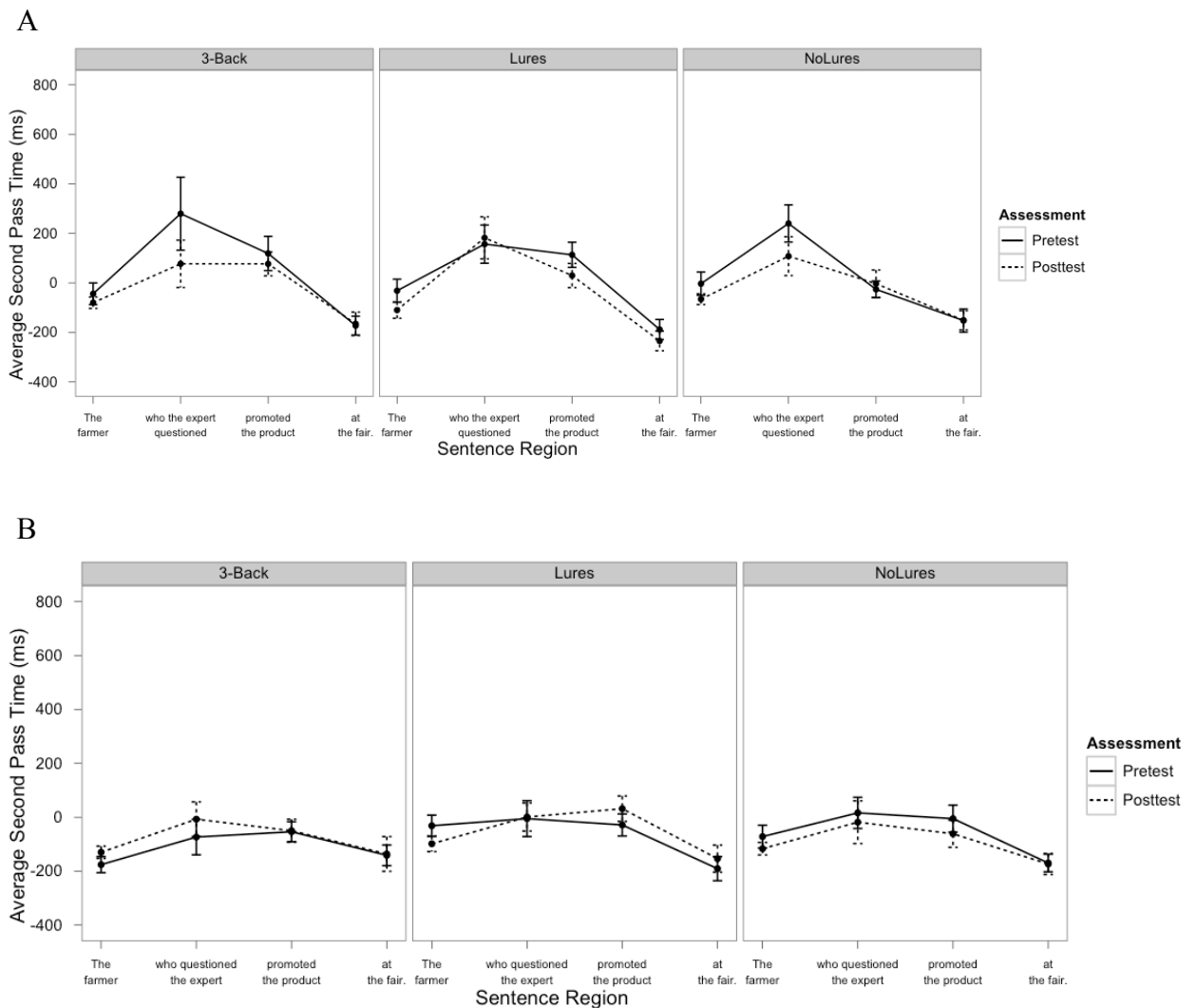


Figure 16. Residual second pass time—the total re-reading time in a region corrected for string length of that region—for relative clause sentences by sentence region (note: “who the expert questioned” of object relative sentences reflects the region where processing difficulty is often encountered) at each assessment for each training group (3-Back in leftmost panels, Lures in center panels, and No-Lures in rightmost panels) for (A) object-extracted and (B) subject-extracted relative clause sentences.

developing pre/post change for the No-Lures (171ms improvement) and 3-Back Groups (303ms speed-up) in the critical region. That these differences do not reach significance, however, may be indicative of a large degree of variance among trainees in each group. Thus, this emerging effect is something worth probing with additional research.

Regardless of how the No-Lures and 3-Back Groups may ultimately pattern, the critical component of the present findings is one that focuses on the Lures Group. Trainees who honed their interference-resolution skills demonstrated selective gains on a revision measure while processing ambiguous—but not unambiguous—sentences; this was met with no benefits on a measure of parsing difficulty (object-relative constructions). Together, these patterns support a process-specific account. Specifically, interference-resolution training did not confer a general benefit on all conditions with exaggerated cognitive demands, only cases when elevated conflict is present.

### **3.4 Discussion of Experiment 2**

#### *3.4.1 Summary*

Participants who routinely encountered lures during an intervention task demonstrated improvements on a host of untrained measures (as revealed through planned pairwise comparisons, some of which were bolstered by Assessment-by-Training Group interactions; see Table 14 for a summary of all transfer findings). Important for a process-specific account of cognitive control training, Lures participants' performance gains were exclusive to cases containing conflicting representations; by and large, task conditions void of conflicting representations resulted in little to no improvement for Lures trainees. This pattern was present across a range of ostensibly different cognitive control tasks within parsing and non-parsing domains:

1. On a canonical Stroop task, Lures trainees demonstrated attenuated Stroop Costs (a measure of incongruent trial performance), but not on Stroop Benefits (a measure extracting out congruent trial performance). This suggests that following repeated practice overriding biases (through encounters with *n*-back

lure items), Stroop conditions requiring similar recharacterization processes—when a default lexical representation must be ignored in favor of task-relevant perceptual information—are boosted; cases where conflicting representations are no longer present, removing the need to engage information recharacterization, are not changed with training.

2. Practice with  $n$ -back lures during training elicited faster response times to high-conflict local recognition trials of a recognition memory task, but did not facilitate performance on any trials of a global recognition task where conflict demands were mitigated. That is, all probe types of the high-conflict local block that contained interfering memoranda—and no probe types of the low-conflict global block, absent conflicting information—revealed faster recognition response times following Lures training.
3. Conditions of a verb generation task promoting under-determined representational conflict (High-Competition nouns) had faster generation time as a function of Lures training, while nouns with fewer verb associates (Low-Competition) did not speed up following training. Further, the Lures Group improved under conditions with elevated retrieval demands (Low-Association nouns), indicating that accessing a weak nearest neighbor may require similar interference-resolution abilities to those gained with exposure to  $n$ -back lures.
4. Lures trainee's enjoyed selective benefits in online and offline indices of syntactic ambiguity resolution (garden-path recovery), such that no change was seen on control (unambiguous) conditions, where the need to rework an incorrect, default interpretation is diminished.

This improvement profile supports a process-specific account: Individuals practicing interference-resolution improve only on untrained measures relying, in part, on similar skills. However, it also invites an interpretation favoring effortful cognition as the common thread across the improved task conditions. Evidence disentangling conflict-control demands and cognitive difficulty exists in the lack of a positive-transfer effect for Lures trainees' processing of notoriously-complex object-extracted relative clauses; the Lures Group does not improve in the face of exaggerated syntactic complexity, perhaps because the need for cognitive control is obviated for object-relative constructions (relative to simple subject-relative sentences). Thus, that Lures trainees do not improve in all cases of heightened cognitive effort provides considerable support for a process-specific account. This group only improves when presented with conflicting representations that must be resolved/recharacterized, as is the case of Stroop incongruent items, trials of a local recognition memory task (due to the presence of interfering memoranda), high-competition noun items, and syntactically ambiguous sentence structures.

Alongside these selective training-transfer patterns, the Lures Group was pitted against two minimally contrastive (active control) training groups—the No-Lures and 3-Back Groups—to test whether lure items and adaptivity were necessary and sufficient to render generalized improvement to unpracticed tasks requiring cognitive control. These non-interference-resolution groups failed to show comparable selective patterns on any of the untrained measures, lending further support for a process-specific account of cognitive control training. Instead, the No-Lures and 3-Back Groups showed one of two general performance profiles for untrained transfer measures: Improvement on *neither*

<b>Conflict Condition</b>	<b>Training Groups with Significant Pre/Post Change</b>
<b>Stroop Task (Response Time)</b>	
High-Conflict (Stroop Cost) – Response & Representational Conflict	Lures Group
High-Conflict (Stroop Cost) – Representation Conflict Only	Lures Group
Low-Conflict (Stroop Benefit)	No Groups
<b>Recognition Task (Response Time)</b>	
High-Conflict (Local Block)	Lures Group
Low-Conflict (Global Block)	No Groups
<b>Verb Generation Tasks (Generation Time)</b>	
High-Conflict (High-Competition)	Lures & 3-Back Groups
Low-Conflict (Low-Competition)	3-Back Group
Low-Conflict/High-Difficulty (Low-Association)	All Groups
Low-Conflict/Low-Difficulty (High-Association)	3-Back Group
<b>Offline Garden-Path Recovery (Accuracy)</b>	
High-Conflict (Ambiguous)	All Groups
Low-Conflict (Unambiguous)	No Groups
<b>Real-time Garden-Path Recovery (Regression-Path Time)</b>	
High-Conflict (Ambiguous)	Lures & 3-Back Groups
Low-Conflict (Unambiguous)	3-Back Group
<b>Real-time Garden-Path Recovery (Second-Pass Time)</b>	
High-Conflict (Ambiguous)	Lures Group
Low-Conflict (Unambiguous)	No Groups
<b>Real-time Relative Clause Parsing (Second-Pass Time)</b>	
Low-Conflict/High-Difficulty (Object-Extracted)	No Groups
Low-Conflict/Low-Difficulty (Subject-Extracted)	No Groups

Table 14. A summary of the training groups demonstrating reliable cross-assessment effects given by planned comparisons for each condition (high/low conflict/difficulty) of the seven measures of transfer discussed in Experiment 2. These measures included response times on Stroop and recognition memory, production times for the verb generation task, comprehension accuracy following garden-path sentences, and eye

high- nor low-conflict conditions within a task, or improvement on *both* high- *and* low-conflict conditions within a task. Only the Lures Group consistently improved in accordance with a process-specific theory, demonstrating selective pre/post benefits on movement measures while readers parsed garden-path and relative-clause constructions. *just* high-conflict conditions across a range of tasks. Despite this, it is important to note

that this conclusion was supported by an Assessment-by-Training Group interaction for three measures—response times to high-conflict trials of a recognition memory task, production times for high-competition nouns in a verb generation task, and real-time reanalysis of temporarily ambiguous sentences specifically when information inconsistent with a developing interpretation arises. Such an interaction was absent when considering response times of incongruent trials of a Stroop task; however, planned comparisons indicated that following training, the Lures Group was faster for these high-conflict trials (given by Stroop Cost), showing no change on low-conflict congruent trials (indexed by Stroop Benefit).

To bolster the claim that the training groups actually improved on their respective practiced elements, a posttest version of the  $n$ -back task verified the relative advantages of adaptivity and exposure to lure items: First, the Lures Group responded more accurately to lure items compared to subjects in the groups that did not see lures throughout training (*Test of Cognitive Control*). Second, the groups that received performance-adaptive training were more accurate to all items on a 6-back task compared to the non-adaptive 3-Back Group, indicating that practicing a task with higher  $n$ -levels confers a benefit for maintaining more items (*Test of Adaptivity*). This was even the case for a group that practiced a different task over the course of training—the No-Lures Group—such that these subjects never practiced an  $n$ -back task with lures, but still managed to outperform the 3-Back Group, perhaps because of their experience managing more than three items during training.

Related to non-selective improvement profiles, across several measures, the 3-Back Group showed equally-good improvement on high- and low-conflict cases, pointing



to a general boosted ability (perhaps akin to speed of processing). Interestingly, these cases were observed only for linguistic measures (verb generation and sentence processing): 3-Back trainees were faster to produce words on *all* conditions of the verb generation task, irrespective of conflict/retrieval (competition/association) demands. Similarly, this group was faster to regress to earlier regions of ambiguous *and* unambiguous sentences following training, an effect likely owed to an improvement on an ability *other than reanalysis*. This is supported by this group's lack of a cross-assessment boost on second-pass time in early sentence regions, a measure I argued here to directly index the revision component of regression-path time.

The No-Lures Group, on the other hand, was consistent in demonstrating little to no cross-assessment variation on any conditions of the current transfer measures, despite their clear increased performance over the course of training (see Figure 8). Considering the selective improvement pattern of the Lures Group and the generalized boosts (with respect to language measures) of the 3-Back Group, it is possible that the No-Lures Group saw negligible training-transfer effects as a function of developing a non-transferable task-specific strategy that failed to transfer to new measures (see Discussion below). Performance-adaptivity may have played a key role in this, considering that participants in the 3-Back Group, who demonstrate transfer effects, could have created similar strategies given their equivalent removal of any cognitive control demands.

In addition to performance-adaptivity, a second feature that differed between the No-Lures and 3-Back Groups involved stimulus type: The No-Lures Group practiced a letter *n*-back at every training session, while stimulus type (letters, words, nonwords, symbols) cycled across training sessions for the 3-Back Group. Multiple stimuli were

included to keep participants in the 3-Back Group engaged over the course of training in an attempt to combat attrition (see additional rationale in Method section). Although all stimulus sets included simple, pronounceable items, changing the nature of the information being encoded from session to session may have resulted in unanticipated effects for transfer. That is, this second difference between the two non-lures groups might be confounded with adaptivity when explaining any cross-assessment differences that arise between these two groups. Even in light of this, the potential challenge accompanying various stimuli might be evenly offset by the reduction in difficulty that comes with non-adaptive versions of *n*-back.

To reiterate, the Lures Group's selective improvement on high-conflict assessment task conditions hints at a common underlying mechanism (interference resolution) supporting conflict-control in certain conditions of unpracticed tasks. Although the 3-Back Group also demonstrated better performance on some cases of high-conflict (high-competition nouns in the verb generation task and regression-path time following entry into the sentence region of garden-path sentences containing information conflicting with a default interpretation), these were accompanied by equivalent improvements in control items lacking interference-resolution demands (low-competition nouns and regression path time for the sentence-final region of unambiguous items). The No-Lures Group, on the other hand, showed no significant pre/post change on any assessment measures, suggesting that an adaptive *n*-back-without-lures may remove certain processing demands that otherwise constitute critical shared properties with the chosen assessment measures.

### 3.4.2 *N-Back Strategies*

One other plausible explanation for this collection of results involves the nature of the strategies implemented by each group over the course of training; version-specific tactics may have been developed, with some *n*-back tasks eliciting a higher probability of strategy-switching than others. The importance of strategies is not a new consideration, as several training researchers have considered training efficacy as a byproduct of adopted strategies (Chase & Ericsson, 1982; see Morrison & Chein, 2011 for a discussion), while others have developed training regimens geared toward enhancing task-specific routines (Carretti, Borella, & DeBeni, 2007; McNamara & Scott, 2001). If the presence of lures and/or adaptivity biases trainees to consider a different subset of strategies, this could account for the general variation in pre/post performance across training groups.

Adaptivity, for example, might invite subjects to consider new tactics more readily than cases where *n*-level remains constant. Indeed, the participants in the present study were queried about their strategies during an exit-survey at posttest; subjects in the Lures (60%) and No-Lures (57%) Groups were most likely to report a change in strategy over the course of training compared to subjects in the 3-Back Group (37%). Preferred strategies are liable to change when demands shift over the course of training (a scenario more likely when new *n*-levels are encountered). Lower *n*-levels of an *n*-back task without a high-density of lures items might prompt a strategy consistent with maintaining the most recently-presented items; that is, processing might be more *active* at *n*-levels within the bounds of a subject's WM capacity (e.g., less than 4; see Cowan, 2001). At higher *n*-levels, however, a strategy consistent with *passive* processing may be more advantageous. This method could hinge on using a familiarity bias to detect repeating

items (targets) as those that exceed a recency threshold (see, for example, Kane et al., 2007; Oberauer, 2005; Oberauer & Kliegl, 2004). This is considered a passive strategy because participants may use a process akin to feeling-of-knowing to tag targets instead of actively maintaining all new incoming items. At higher  $n$ -levels, maintaining just-seen items in a buffer or short-term memory store might be pointlessly taxing, especially if an equally-profitable, less demanding passive strategy is feasible. Inspection of training performance for the No-Lures Group indicates that the average  $n$ -level at the second training session for this group was 4.09 (see Figure 8A), indicating that a passive strategy may have become an option for most participants of this group quite early in training. Moreover, recall that participants in the present study were highly-motivated to achieve higher  $n$ -levels, as their likelihood of earning a prize was contingent upon new high-scores (see Method), making it even more likely that subjects may have considered a multitude of strategies before settling on the least effortful option.

Interestingly, some evidence suggesting that the No-Lures Group may have adopted a target-detection strategy emerged in the posttest  $n$ -back task: No-Lures trainees showed better accuracy to  $n$ -back targets compared to other training groups (on the 6-back block). One outstanding concern for this account of target-detection involves the No-Lures Group's *lack* of selective performance benefits for targets on the recognition memory task. This suggests that target-detection strategies may be task-specific (i.e., only targets within the context of  $n$ -back are better detected). If this is the case, then one potential explanation for the No-Lures Group's *lack* of pre/post change on most assessment measures might be that rather than improving general-purpose abilities, these trainees learned an  $n$ -back-specific strategy.

In contrast, versions of  $n$ -back that restrict  $n$ -level (e.g., 3-Back task) might encourage *active* strategies throughout training. It is not unreasonable that such active strategies (capitalizing on maintenance) would confer the type of generalized benefits observed for the 3-Back trainees across two transfer measures. That is, one potential side effect of implementing a strategy that does not waver over the course of training may involve well-learned abilities associated with that strategy, assuming it is not task-specific. Thus, in the case of the 3-Back Group, maintenance abilities may have been incidentally practiced throughout training. Put differently, by switching between strategies, less time is devoted to any single one; perhaps developing a well-established memory strategy after performing the same task for 8 hours conferred a benefit to the 3-Back Group that may have eluded the minimally-different No-Lures Group.

Finally,  $n$ -back interference lures may prompt the use of active strategies regardless of  $n$ -level, in that adopting a passive familiarity bias would result in increased false alarms to lure items, preventing subjects from achieving gains in performance (here,  $n$ -back score). This paired with the heightened processing demand inherent to the perpetual need to resolve among competing representations likely dissuades subjects from adopting such passive strategies. Indeed, honed interference-resolution skills over the course of training are consistent with the selective benefits demonstrated by the Lures Group across several different untrained measures of cognitive control. Further, the Lures Group's null effect for complex parsing cases discounts a general strategy favoring dealing with effortful information, in general. Although  $n$ -level changes may elicit the consideration of new strategies,  $n$ -back-with-lures strategies may continue to be active to prevent subjects from issuing false alarm errors to highly-familiar, but irrelevant stimuli.

In sum, taking into account the plausible strategies that subjects might be tempted to implement offers one possible explanation for the profiles of pre/post transfer effects of each group. By implementing a passive (less cognitively demanding) strategy at higher  $n$ -levels, the individuals practicing an adaptive  $n$ -back-without-lures may have consequently relinquished the benefits of training. Performing a 3-Back task, on the other hand, might have biased active maintenance of the most recently-presented items, seeing that remembering 3 items appears to be a manageable feat for most subjects (see training curves in Figure 8A, wherein adaptive trainees achieve  $n$ -levels greater than 3 fairly early in training—by the third session). The combination of lure items and adaptivity may have generated the cognitive demands necessary to improve cognitive control abilities, even by a strategy-based account.

## Chapter 4: General Discussion

### 4.1 Tying Together Experiments 1 and 2

I presented training data from two initial experiments demonstrating reliable transfer from a general-purpose cognitive control task (*n*-back-with-lures) to syntactic ambiguity resolution in healthy adults, where individuals who have undergone extensive cognitive control training fare significantly better at revising early misinterpretations than their untrained counterparts. Alongside the results of an additional linguistic task (verb generation) and non-parsing tasks in Experiment 2, these patterns seem consistent with the idea that the ability to recover from misinterpretation can be enhanced by training domain-general cognitive control skills that are common to some tasks of language processing and some tasks of memory. Further, the findings indicate that within the right framework, and having appropriate linking hypotheses, cognitive training may be a viable way to improve language use under certain conditions. Despite this, there are some important distinctions between the two experiments that warrant additional discussion. Before contrasting the two experiments, I will first discuss and synthesize the measures common to both (comprehension accuracy and real-time recovery efforts). I will also highlight a series of signal detection measures that provide evidence in favor of a distinct process-based profile among the critical trainees in each experiment (*n*-back responders of Experiment 1 and Lures trainees' of Experiment 2).

#### *4.1.1. Comprehension Accuracy*

In Experiment 1, only *n*-back training responders improved in accuracy to questions probing for the lingering effects of garden-path recovery. Separate subsets of trainees corresponding to responders of other practiced tasks did not demonstrate

comparable improvement. Likewise, untrained controls and *n*-back nonresponders showed no significant pre/post gains on this measure. Contrasts to performance on low-conflict unambiguous items revealed no change following successful *n*-back training. This necessitated the question of whether these findings were due to improved cognitive control mechanisms inherited through practice on *n*-back lure items.

Consistent with the general result of the first experiment, all trainees in Experiment 2—regardless of the version of *n*-back that was practiced—improved their offline garden-path recovery, indexed by better accuracy following just high-EF ambiguous sentences; similar to Experiment 1, comprehension of unambiguous items was no different from pretest performance for all trainees. This suggests that *n*-back, *in general*, may be sufficient to improve offline reinterpretation, perhaps due to the memory component inherent to *n*-back. For instance, when subjects answer comprehension questions, they must reflect on the meaning of the sentence that appeared on a previous screen. *N*-back training may have enhanced the ability to reinstate this information to answer comprehension questions. Critically, a practice effect may be ruled out when these patterns are considered alongside the untrained no-contact control group of Experiment 1; recall that this control group showed no reliable cross-assessment boost. Despite discounting a practice effect by drawing on the patterns across two experiments, it is certainly possible that motivation (or a Hawthorne effect) is responsible for the observed pre/post changes in comprehension accuracy, as one important function distinguishing all training groups from both experiments from the no-contact control group is active participation in the lab during the weeks separating pre- and posttest assessments.



Even though all participants exposed to some form of conflict-control training improved selectively on comprehension questions following ambiguous sentences, one possible inconsistency between the two experiments involves the presence of an effect for *n*-back training conditions absent lures (e.g., No-Lures and 3-Back Groups), but no effect for responders of the training tasks presented in Experiment 1 other than *n*-back; presumably, some of the non-conflict training tasks of the first experiment tap similar memory demands (e.g., LNS and running span) as those trained during *n*-back without lures. Specifically, the in-house training tasks of the first experiment require subjects to practice memory maintenance skills, which may be expected to also confer benefits in reinstating memory representations. That is, if the updating component of *n*-back was critical for improved comprehension accuracy for all conditions in Experiment 2, then the memory demands of tasks like LNS and running span might also be expected to render transfer. Such an effect would hold if the responders of these tasks had demonstrated greater pre/post gains in comprehension accuracy to questions following ambiguous sentence compared to their non-responder counterparts; this, however, is not the case (see Table 2). One possible reason accounting for this may be rooted in the nature of the responder/non-responder clusters of these other in-house tasks: Two distinct clusters were identified for *just* the *n*-back task, indicating that perhaps the clustering for the remaining in-house training tasks was unnatural, thus failing to capture a true bifurcation of training abilities on these measures (see footnote 5).

Additionally, the potential discrepancy in underlying abilities resulting in improvements in comprehension accuracy might be driven by other factors that differentiated Experiments 1 and 2, including the amount of time spent practicing

individual tasks and the demands associated with performing a battery of tasks versus a single measure over the course of training. Regardless, in light of Experiment 2's findings, it seems to be the case that the *n*-back task is sufficient for the goal of improved offline comprehension accuracy.

#### *4.1.2. Real-time Recovery Efforts*

In both experiments, participants' eye movements were used to index online revision of syntactically ambiguous sentences, with hypotheses about the reading time patterns following entry to the disambiguating (high-EF) region of ambiguous items. I expected individuals with enhanced cognitive control abilities to spend less time recovering from an incorrect default interpretation upon encountering late-arriving information suggesting a new meaning. Indeed, in Experiment 1, only *n*-back training responders were faster to regress to and read earlier sentence material following training; untrained controls, *n*-back non-responders, and responders of all other in-house training tasks did not show any pre/post change in regression-path time from the sentence-final region. Additionally, no change was observed for any other sentence region or for any regions of low-EF unambiguous sentences for any of the compared groups. In Experiment 2, this pattern replicated among Lures Group trainees across two reading time measures (regression-path time and second-pass time): Only trainees practicing a version of *n*-back that bolstered conflict-resolution abilities demonstrated selective improvement on high-EF ambiguous items, with no change on unambiguous sentences. Unexpectedly, the 3-Back trainees also improved in terms of regression-path time in the disambiguating region of ambiguous sentences, an effect accompanied by a comparable cross-assessment improvement in the sentence-final region of unambiguous items.

Selective second-pass time boosts for the Lures Group helped to pinpoint that revision is likely the locus of the Lures Group's effects. Indeed, the sub-measure of regression-path time tapping revision was the only component to improve as a function of successful  $n$ -back training in Experiment 1 (see Table 6).

Considering both measures of the syntactic ambiguity resolution task, the results of Experiment 2 build off of Experiment 1, promoting the notion that different general-purpose processes govern offline and online sentence reinterpretation. A component general to several versions  $n$ -back resulted in cross-assessment gains in offline measures (accuracy), while a component specific to interference-resolution garnered changes in real-time revision efforts (regression-path and second-pass time). That is, lure items may be necessary to alter real-time revision properties, while  $n$ -back is sufficient when readers must remember a recovered interpretation (see Patson, Darowski, Moon, & Ferreira, 2009 for effects in a paraphrasing design, when this demand is presumably exaggerated). Indeed, at least one account (Christianson et al., 2006) has tied the ability to overcome a lingering garden-path effect to verbal working memory (not interference-resolution, per se), by demonstrating that subjects with better performance on a reading span task were more accurate to answer questions like "*Did the thief hide himself?*" relative to those with poorer reading span scores.

Since others have suggested that this memory/ambiguity resolution connection is driven by a skill to inhibit an initial interpretation (see Hasher & Zacks, 1988; Novick et al., 2005), I tested this within the context of cases where interference-resolution was boosted ( $n$ -back with lures) compared to instances when this ability was unaffected ( $n$ -back versions without lures; untrained controls and non-responders to  $n$ -back). Practice

with interference lures throughout training did not play a unique role for comprehension accuracy improvement. This is liable to be the case because *n*-back (like running span and other working memory tasks used in prior work; see Christianson et al., 2006) may force subjects to actively manipulate information in a variety of ways. Indeed, others before me have suggested many properties of the *n*-back task (sans lure items) that could result in changes in other EFs, including updating (Chatham et al., 2011; Dahlin et al. 2008), maintenance and monitoring (Jaeggi et al., 2008), and familiarity recognition (Kane et al., 2007; Oberauer, 2005). Thus, any one of these factors might point to the locus of general *n*-back training when identifying its role for offline comprehension accuracy in the face of ambiguity. Future work should aim to minimally compare extensive *n*-back practice to training tasks tapping other cognitive abilities. For example, by implementing a streamlined training design like that of Experiment 2, wherein subjects are assigned to practice just a single training task, I may be able to elucidate the unique contributions of *n*-back compared to well-characterized tasks like reading span or operation span for untrained measures like lingering garden-path recovery.

#### *4.1.3 Contrasting Experiments 1 and 2*

Although the present experiments converge on a pattern indicating the validity of a process-specific account of cognitive control for certain conditions of language processing (namely, when users must resolve among conflicting representations). This evidence, however, emerged from two wildly different training designs. Here, I discuss the factors distinguishing Experiments 1 and 2, each of which I argue are critical to consider when synthesizing the findings of both studies.

First, the subjects were exposed to a different number of training tasks in each study: Trainees in Experiment 1 practiced a battery of tasks, which introduced a design element that confounds the role of improved cognitive control with gains on other measures. Experiment 2, on the other hand, was designed to identify the unique causal role of *n*-back with lures for the pre/post gains observed in Experiment 1. Even in light of this, a perfect replication of Experiment 1 still leaves open the question of what the remaining training tasks were contributing to the observed transfer effects.

Related to the nature of training involves the measurements of trainability used to verify the efficacy of practiced interventions. By performing a battery of training tasks in Experiment 1, indices of trainability were derived by identifying responders and non-responders on each in-house measure. That is, subjects who respond to a task might index a subset of trainees who can be denoted as participants with improved abilities on the core process hypothesized to underlie that task. Although this method is useful for identifying clusters of subjects expected to improve on some assessment tasks and not others, in many cases splitting trainees based on group performance introduces interpretation issues: First, restricting clustering approaches to just a subset of individuals may not generalize to parameters of the population; thus, trainees who perform well may not index true responders in the population, but rather responders with respect to a small sample. Additionally, as sketched in the discussion of Experiment 1, correlated tasks (like *n*-back and garden path recovery) will necessarily give rise to correlated gains (e.g., *n*-back training gains and pre/post garden-path gains). Experiment 2 avoided these potential issues by introducing three-minimally different tasks, the trainability quotients of which were measured through a posttest version of *n*-back with

lures that all participants performed. By contrasting each training group's performance on a single common task, I was able to evaluate the relative efficacy of training on any given version of  $n$ -back. Specifically, in Experiment 2, I verified group differences in terms of 1) lure accuracy to examine relative differences cognitive control ability and 3) total accuracy on items at each  $n$ -level (3-back and 6-back) to assess the contribution of adaptive training.

Another critical feature distinguishing between Experiments 1 and 2 involves the control conditions against which the trainees of interest were compared. Being that Experiment 1 trainees' performance was contrasted to a no-contact control group, I extracted responders of tasks to test a process-specific account of cognitive control for garden-path recovery. The rationale for this arose from the presence of a single training task hypothesized to tap conflict-resolution ( $n$ -back with lures). By identifying subjects who showed training improvements on  $n$ -back, I was able to create "active control groups." Each engineered group served to provide an important contrast: cognitive control non-responders helped me to elucidate the role of improvements  $n$ -back for garden-path recovery gains, and responders on a host of other within-subjects non-conflict measures (LNS, Running Span, and Block Span) allowed for an adequate test of process-specificity (i.e., that  $n$ -back was the only training task to confer advantages to untrained measures relying, in part, on cognitive control). Performance of these subsets was compared to  $n$ -back responders and untrained controls to make inferences about the specific role of cognitive control for garden-path recovery. The Lures Group of Experiment 2 was compared to two between-subjects active control groups—No-Lures and 3-Back Groups. The variation in control group types offered separate avenues for

testing the distinct role of interference resolution for language conditions containing competing representations.

To recap, several factors influence—and in some cases restrict—how one may assess the distinct role of interference-resolution for novel measures relying on this common skill, and perhaps even extend this finding to include assumptions about the *causal* nature of cognitive control for certain task conditions. As discussed earlier, the analyses and design of Experiment 1 may leave open the possibility of either correlated or causal role of interference resolution for garden-path recovery. The findings of Experiment 2 lend support for a causal relationship of cognitive control—as it is trained via *n*-back-with-lures—for a host of linguistic (verb generation, garden-path recovery) and non-linguistic (Stroop, recognition memory) measures. When considered alongside a carefully-designed study like Experiment 2, the results of Experiment 1 become increasingly clear: There appears to be a significant role of cognitive control for garden-path recovery processes.

#### *4.1.4 Necessary and Sufficient Features for Cognitive Control Training*

The present set of findings hint at two possibilities regarding the components parametrically introduced across the various versions of the present *n*-back training tasks: First, Experiment 2's results are consistent with *n*-back *lure* items being necessary (and perhaps sufficient)<sup>19</sup> for transfer to novel measures sharing cognitive control demands.

The findings from Experiment 1 were replicated among just the Lures Group of

---

<sup>19</sup> However, note that a *true* test of the combined importance of adaptivity and lure items was not included in Experiment 2. For this to be accomplished, a non-adaptive training group with lures would need to be included. Thus, although lures might be necessary for cognitive control training, they may not be sufficient, as the present findings might hinge on the combinatorial presence of lure items and adaptivity.

Experiment 2, given that only Lures trainees demonstrated selective improvement on high-EF ambiguous conditions alongside no performance change for low-conflict control sentences. Second, adaptivity may be necessary, but is certainly not sufficient to equip trainees with honed interference-resolution skills. Because lures seem to be a necessary component for online revision (at least with respect to a pattern consistent with a process-specific account), adaptivity alone should not confer relative cognitive control advantages on untrained measures. Adaptivity, instead, may be a necessary element, seeing that the Lures Group practiced an adaptive version; note that counter-evidence for its sufficiency stems from the (performance-adaptive) No-Lures Group's lack of transfer-effects across all measures.

Moreover, the 3-Back Group's general cross-assessment gains—on high- and low-conflict items—on the verb generation task and some measures of garden-path recovery suggests that improvement is possible even in the absence of adaptivity. It is possible that some members of the 3-Back Group incidentally trained interference-resolution as a result of superior baseline cognitive abilities. For example, participants with better maintenance abilities at the onset of training may be biased to remember items occurring well before the relevant item third back in a sequence. In such cases, the incidental presence of lures items (e.g., in positions 6-Back or higher) might render unintended training on interference-resolution abilities. However, with these instances uncontrolled (and coincidental), it is unlikely that any given 3-Back trainee would demonstrate improved conflict-control abilities on par with those subjects practicing a carefully-controlled version of *n*-back containing a high-density of lure items. The categorical root of the improved underlying abilities can be examined with process-



sensitive analyses (e.g., signal detection; see next section). Indeed, by identifying processes distinct to each subset of trainees (*n*-back trainees of Experiment 1; Lures Group of Experiment 2), I could conceivably identify whether the same source gives rise to the 3-Back and Lures Groups' improved ambiguity resolution.

## **4.2 Understanding Processes with Signal Detection Models**

Signal detection analyses offer one possible way to disentangle the underlying processes affected as a function of different training task demands. Specifically, by considering target and non-target accuracy within a task, signal detection allows for the assessment of unobservable processes related to the distributions of relevant and irrelevant items by providing proxies of target/non-target discriminability and response criterion. To probe for the locus of the training group effects of Experiment 2, I conducted signal detection analyses on performance of the posttest *n*-back and recognition memory tasks. Signal detection analyses are best suited for recognition-based measures, thus no other assessment task was examined.

### *4.2.1 Analysis and Results of Posttest N-Back Task*

I expected to replicate the effects of Adaptivity and Lures Presence among trainees performing the posttest *n*-back task among signal detection measures. If distinct processes index adaptivity versus lures performance, this approach could help me verify independent mechanistic sources of the group effects reported in chapter 3.

A signal detection analysis was conducted to examine whether Training Group assignment yielded variation in trainees' internal response thresholds ( $\beta$ ) or discriminability ( $d'$ ) of target and non-target items. Measures were obtained over all

trials, such that fillers and lures were included in one category of non-target items.<sup>20</sup> Hit and false alarm rates were computed for each subject on each block, which were then used to calculate  $d'$  and beta (assuming unequal variance; Wickens, 2002; Mickes, Wixted, & Wais, 2007). I then evaluated the effects of Training Group (Lures, No-Lures, 3-Back) and Block (3-Back, 6-Back) for each signal-detection measure separately. An effect of Training Group on response criterion, discriminability, or both measures could provide evidence for separable processes gained as a function of practice with certain training tasks.

A repeated-measures ANOVA testing for the fixed effects of Group and Block on discriminability ( $d'$ ) yielded a significant interaction ( $F(2,153)=5.213$ ,  $p=0.006$ ,  $BF=17.11$ ), as well as main effects of Block ( $F(1,153)=82.893$ ,  $p<0.001$ ,  $BF>100$ ) and Training Group ( $F(2,153)=14.386$ ,  $p<0.001$ ,  $BF>100$ ). Depicted in Figure 17A, during the 3-Back block, I observed greater sensitivity for the Lures Group ( $d'=2.89$ ) relative to the non-adaptive 3-Back Group ( $d'=2.58$ ; Welch two-sample  $t=2.54$ ,  $p=0.01$ ,  $BF=2.73$ ), while the No-Lures Group ( $d'=2.74$ ) did not differ from either group ( $t's<1.18$ ,  $p's>0.22$ ,  $BFs<0.30$ ). Performance on the 6-Back block revealed differences in sensitivity between the Lures ( $d'=2.07$ ) and No-Lures Groups ( $d'=2.28$ ) compared to the non-adaptive 3-Back Group ( $d'=1.35$ ;  $t's>4.49$ ,  $p's<0.001$ ,  $BFs>100$ ; see Figure 17A). Critically, at this higher  $n$ -level, the two adaptive groups' discriminability was not significantly different from one another (Welch two-sample  $t=1.04$ ,  $p=0.30$ ,  $BF=0.26$ ), lending some support for

---

<sup>20</sup> I replicated the present patterns when the non-target distribution included only lure items. When non-targets included only fillers, the findings were not replicated, suggesting that by collapsing non-target item types, lure accuracy drives much of the effects.

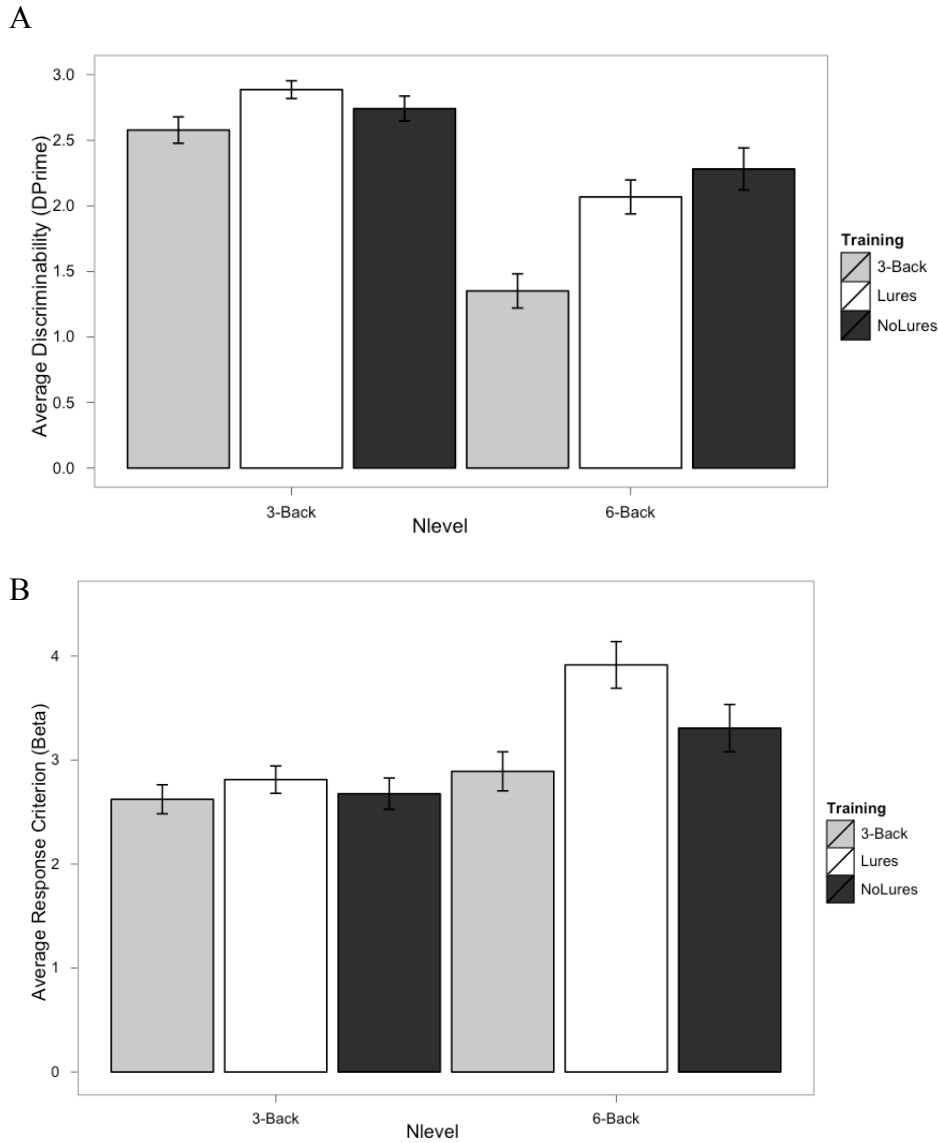


Figure 17. Signal detection measures on each block—3-back and 6-back—of the post-test  $n$ -back task for each training group indexed by (A) discriminability ( $d'$ )— and (B) response criterion ( $\beta$ ).

discriminability being the process underlying the performance-adaptivity benefits observed among the Lures and No-Lures Groups relative to the 3-Back Group. Further, this suggests that practicing higher  $n$ -levels confers a change in the distributional space over which targets and non-targets exist, such that their means are less similar, rendering them more distinguishable following adaptive training. This was especially the case for

the Lures trainees, who showed reliably better discriminability compared to the 3-Back Group regardless of  $n$ -level; the No-Lures trainees only demonstrated adaptivity-advantage on the more demanding 6-Back block.

Focusing on response criterion, a comparable analysis of Training Group and Block revealed an interaction ( $F(2,153)=2.90$ ,  $p=0.05$ ,  $BF=0.08$ ), bolstered by main effects of Block ( $F(1,153)=21.06$ ,  $p<0.001$ ,  $BF>100$ ) and Training Group ( $F(2,153)=6.21$ ,  $p=0.002$ ,  $BF=4.52$ ). Figure 17B illustrates that even though no Training Group effects were observed on the 3-Back block ( $F(2,77)=0.51$ ,  $p=0.59$ ,  $BF=0.02$ ), an effect was observed on the 6-Back block ( $F(2,76)=6.27$ ,  $p=0.002$ ,  $BF=4.77$ ), such that the Lures Group ( $\beta=3.92$ ) had a higher response criterion compared to the No-Lures ( $\beta=3.31$ ; Welch two-sample  $t=1.98$ ,  $p=0.05$ ,  $BF=0.91$ ) and 3-Back Groups ( $\beta=2.89$ ; Welch two-sample  $t=3.49$ ,  $p<0.001$ ,  $BF=27.09$ ), whose performance did not reliably differ (Welch two-sample  $t=1.41$ ,  $p=0.16$ ,  $BF=0.40$ ). A high (more conservative) response criterion among Lures trainees is representative of this group's superior ability to correctly respond to *just* target items. In other words, by being exposed to highly-confusable lure items, the Lures trainees are less likely to false alarm as a function of issuing more "non-target" judgments following training.

Furthermore, alongside the discriminability findings above, the Lures trainees constitute the only group to show both a higher response criterion *and* greater sensitivity relative to the remaining groups. No other training group demonstrated this combination of changed processes, thus providing process-based evidence favoring a unique shift in mechanisms for the Lures Group. This is especially crucial when considering the foundation of the Lures and 3-Back Groups' comparable pre/post improvements on some

assessment measures (e.g., regression-path time of ambiguous sentences and production times for high-competition nouns). The current signal detection measures indicate that different underlying processes gave rise to the training benefits enjoyed by each group, an effect that may suggest separate causes for these similar cross-assessment gains.

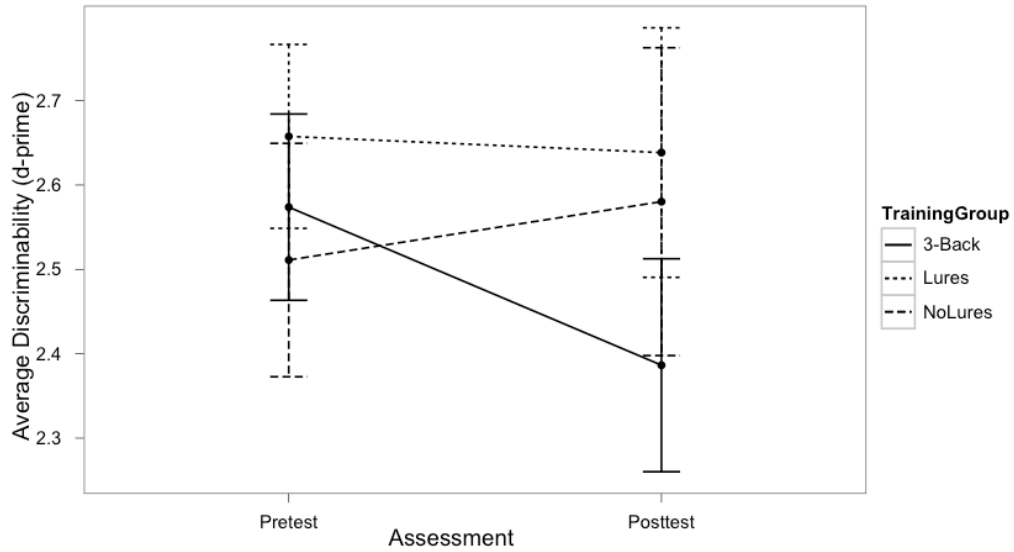
#### *4.2.2 Analysis and Results of Recognition Memory Task*

To evaluate whether these signal-detection findings were *n*-back-specific, I conducted a comparable analysis for target and non-target accuracy on both blocks of the untrained recognition memory task. That is,  $d'$  and beta measures were computed for each subject on just the high-conflict Local condition; the reason for not conducting analyses on the Global condition followed from accuracy ceiling effects on target and non-target items, which resulted in exaggerated and un-interpretable results. A similar result—Lures trainees demonstrating a separate profile from other Groups—would provide additional evidence favoring a separate mechanistic locus of improvement for the Lures Group relative to the remaining training groups.

I evaluated the effects of Training Group (Lures, No-Lures, 3-Back) by conducting an ANCOVA fitting posttest performance of each signal detection measure, while controlling for pretest performance. Considering discriminability first, I found no Assessment-by-Training Group interaction nor main effect of Training Group ( $F$ 's < 1.56,  $p$ 's > 0.21,  $BF$ 's < 0.05), but I did find a main effect of Assessment ( $F(1,71)=75.38$ ,  $p < 0.001$ ,  $BF > 100$ ), indicating large test-retest reliability regardless of training group. This was accompanied by no reliable cross-Assessment change for any training group given by pairwise comparisons ( $t$ 's < 1.78,  $p$ 's > 0.08,  $BF$ 's < 0.67; see Figure 18A). Despite

no noticeable gains in the distributional space of targets and non-targets (the means of these distributions remained constant over time), an ANCOVA testing for Training

A



B

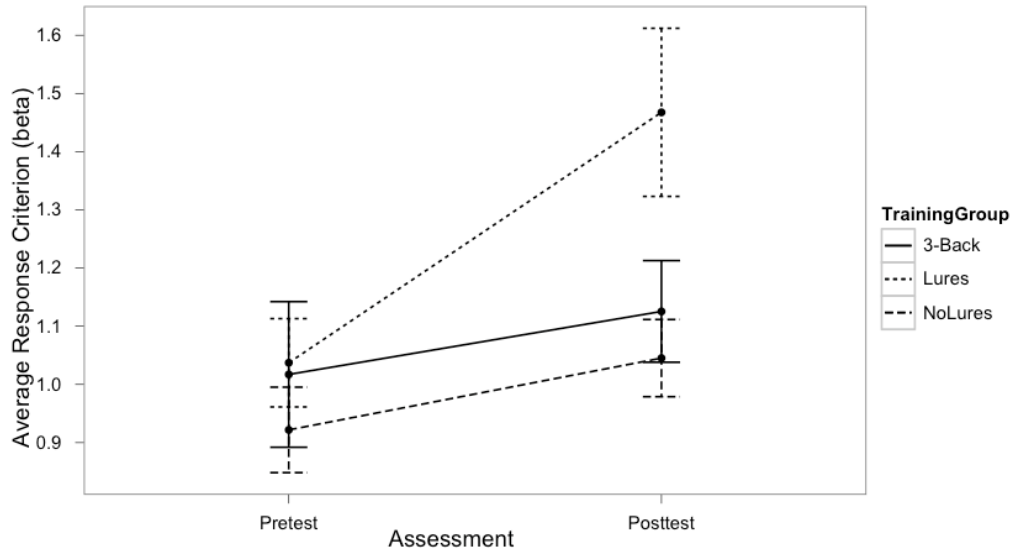


Figure 18. Signal detection measures at pre/post assessments on the recognition memory for each training group indexed by (A) discriminability ( $d'$ ) and (B) response criterion ( $\beta$ ). Note that because accuracy on the global block was at ceiling for all trial types, signal detection measures were largely uninformative; thus, only measures derived from the local block are represented.

Group effects of posttest response criterion (beta), controlling for pretest response threshold revealed main effects of Training Group ( $F(2,71)=4.11$ ,  $p=0.02$ ,  $BF=0.68$ ) and Assessment ( $F(1,71)=4.76$ ,  $p=0.03$ ,  $BF=0.92$ ), but no significant Assessment-by-Training Group Interaction ( $F(2,71)=0.94$ ,  $p=0.39$ ,  $BF=0.03$ ). Figure 18B illustrates that this effect was bolstered by a reliable cross-Assessment shift in response threshold for the Lures Group (Pretest beta=1.04; Posttest beta=1.46;  $t(29)=2.65$ ,  $p=0.01$ ,  $BF=2.37$ ). Neither the No-Lures (Pretest beta=0.92; Posttest beta=1.04) nor the 3-Back Group (Pretest beta=1.02; Posttest beta=1.12) demonstrated any such pre/post change ( $t$ 's < 1.30,  $p$ 's > 0.20,  $BF$ s < 0.24). Namely, this pattern is consistent with Lures trainees' heightened response thresholds during posttest  $n$ -back task: Only the group practicing lure items demonstrates a shift in response criterion, becoming more conservative across two tasks as a function of learning to *not* detect high-EF lure items as target. The signal detection patterns bolster a process-specific account of the Lures Group's propensity to improve on only high-conflict conditions of untrained measures. This selective profile seems to be driven by a change in how Lures trainees—and no other subjects— detect and deal with high-conflict items across multiple tasks. Namely, by encountering highly-confusable items over the course of training, individuals decreased their false alarm rates, which translated into more conservative response thresholds. Thus, response criterion may be treated as a measure of cognitive control.

#### 4.2.3 *Considering the Contribution of Separate Processes for Transfer*

One way to test for differences in response criterion and familiarity bias would be to design a follow-up study contrasting training on the Lures version of  $n$ -back with an identical task *entirely absent* recurring non-target items. Eliminating such repetitions

should, in theory, induce a strong familiarity bias in the absence of a response criterion shift, given that *all* repeating items within a sequence would be relevant targets to which a subject should respond. Likewise, removing recurrences within an entire 30-minute training session might easily induce a recency bias (perhaps a separable phenomenon from a familiarity bias; see Xiang & Brown, 1998).

If the goal is to separate familiarity and recency effects, implementing a task with novel stimuli void of pre-existing probabilistic properties might further foster familiarity biases germane to the task at hand. Most versions of *n*-back contain stimuli with which subjects have ample experience (e.g., letters). The base frequency with which these items are encountered may, for example, interact with the task-specific recency information. Such a prediction is consistent with phenomena in the recognition memory literature, indicating that high-frequency words are recognized more accurately than their less common counterparts (Scarborough, Cortese, & Scarborough, 1977), specifically with low-frequency items being falsely recognized more often than high-frequency words (Gregg, 1976). As a result, one might expect more false alarms to less common (low-probability) items compared to more frequent items. Such prior probabilities may be prone to influence performance during a letter *n*-back task, such that some stimuli will necessarily be more common (*s, t*) than others (*v, x*). Future work might vary the base frequency of to-be-remembered items in *n*-back to investigate the separable roles of familiarity versus recency effects.

Notably, in Experiment 2, some instances arose wherein prior probabilities were essentially controlled for all items. Recall that to minimize attrition in the non-adaptive 3-Back Group, participants trained on a modified multi-stimulus version of *n*-back, such



that they cycled through various stimulus sets—letters, words, non-words, and symbols—at each session. Although most sets contained highly-familiar items (e.g., the letter *b*, the word *clock*, a picture of a flower, etc; see Appendix B for a full list of all stimuli), the pronounceable non-word stimulus sets may have created unique scenarios, prompting trainees to enter the task with limited prior frequency information about the to-be-updated stimuli (e.g., *milp*, *ving*). Note that participants only encountered the non-word 3-back task three times over the course of training, which generated little data to adequately assess a familiarity/recency difference. Nevertheless, implementing a training task with novel stimuli might offer one avenue to disentangle these effects.

Finally, as implied above, the absence of a non-adaptive *n*-back with lures group in the present design leaves open the question of the importance of adaptivity for cognitive control training. Although the Lures Group demonstrates selective effects in line with a process-specific account, there is still an open question of whether the occurrence of lure items *alone* drives the observed improvements. It is entirely possible that an additive effect of lures in the presence of adaptivity is responsible for the reported transfer effects.

#### *4.2.4 A Summary of Possible Trained Mechanisms*

To summarize, the training results presented here provide evidence favoring an improvement in a core cognitive ability, such that trainees practicing a conflict-resolution training task demonstrate selective cross-assessment gains on high-conflict conditions across a range of untrained tasks. Training groups practicing tasks not thought to tap conflict-control failed to show selective improvements on high-conflict conditions; instead, these trainees show either no pre/post change or a general cross-assessment gain

for high- and low-conflict task conditions alike (in contrast to predictions supported by a process-specific account). Importantly, these gains do not appear to be driven by a superior ability to deal with difficult task conditions, evidenced by the Lures trainees' lack of a cross-assessment change for complex object-relative clause sentences. In light of this interpretation, there still exists a question of the role of other mechanisms, including changes in decision threshold and strategy choice, which may account for the selective pre/post gains observed for Lures trainees. For instance, the exaggerated response criterion shifts demonstrated by the Lures Group across two measures (*n*-back and recognition memory) lend some support for the foundation of training as one rooted in a change in decision thresholds. By becoming more conservative, participants may change the way in which conflict is *monitored*. This explanation is consistent with improved proactive cognitive control (associated with an early correction mechanism), as opposed to reactive cognitive control, a skill—much like the one studied here—that resolves conflict *after* it has been encountered (see Braver, Gray, & Burgess, 2007).

As I have hinted, response criterion shifts among Lures trainees may provide some support for improvements of a general-purpose skill beyond conflict-resolution. How this is manifested in parsing measures is unclear, yet one possibility for future study involves probing the degree to which participants commit to an interpretation following training. If practicing *n*-back lures changes proactive control mechanisms, this might result in diminished garden-path effects because the initial interpretation is not nearly as strong. Future work should focus on identifying whether training on *n*-back-with-lures contributes to changes in conflict adaptation, conflict-resolution, temporary recalibration

of linguistic biases (short-term expectation adaptation), or some combination of these factors.

As sketched in the discussion of Chapter 3, a final possibility that may account for the present selective training effects involves the nature and diversity of strategies used. Given that the Lures Group's pre/post benefits are present on very different assessment tasks, it is unlikely that subjects are learning a task-specific ( $n$ -back specific) strategy over the course of the training period. It is, however, possible that participants assigned to the Lures Group attempt more diverse (or more active) strategies compared to their No-Lures counterparts.<sup>21</sup> Disentangling a strategy-based account from a process-specific explanation may not be possible given the present designs, but may be easily achieved by developing training studies that prompt participants to implement controlled strategies throughout training. Even if certain training tasks bias subjects to implement different strategies, it is important to consider the executive functions that these strategies might tap; namely, do certain strategies target core cognitive functions?

#### *4.2.5 Training as a Method to Improve or Capture Baseline Abilities?*

Several training studies suggest that adaptive intervention tasks confer pre/post improvements beyond what non-adaptive versions offer (Brehmer et al., 2011; Holmes et al., 2009; Klingberg et al., 2005). This explanation follows from adaptive training consistently challenging participants by keeping them on the precipice of their best performance, a effortful state that forces trainees to improve (see Lövdén, Bäckman, Lindenberger, Schaefer, & Schmiedek, 2010). Of course, an alternative explanation for

---

<sup>21</sup> Self-reports of strategy use indicated that on average, fewer trainees in the 3-Back Group reported strategy changes (37%) compared to those in the Lures (60%) and No-Lures Groups (57%).

the adaptivity advantage involves a measurement advantage of such designs: Making a task adaptive provides a more expansive range under which participants *can* perform. For example, if a subject has quite good cognitive ability, then it is reasonable to expect him/her to achieve asymptotic performance on a task at a much higher difficulty (*n*-) level than a subject with poorer cognitive ability. To derive a fine measure of baseline cognitive ability, it may be necessary to provide multiple testing sessions, an element of all training studies. In the case of the present study, participants may be unable to reach asymptotic performance levels in just one 30-minute *n*-back session. This is certainly plausible given design and cognitive restrictions that subjects face: First, the number of sequences that subjects can perform in a given session is limited to what can be accomplished in a 30-minute period. Second, cognitive fatigue (see Persson et al., 2007) experienced over the course of a training session should deplete resources, resulting in degraded performance that fails to capture a subject's actual maximum abilities. Allowing subjects to participate in several sessions provides them the opportunity to demonstrate their true levels of cognitive ability. That is, from a measurement perspective, performance scores should become more stable and consistent over time (see Shipstead et al., 2012).

Contrary, the assumption guiding the present dissertation—namely, training improves cognitive abilities, rather than simply measuring baseline skills—contends that practicing a task confers benefits on underlying processes (and/or strategies that promote such processes) that result in performance gains reflecting a fundamental change in baseline cognitive abilities inherited through extensive practice. Both accounts (training improves vs. captures abilities) often separately explain the same pattern of results,

posing a challenge to interpretations of training study findings. Namely, if participants are providing a proxy for baseline levels of cognitive performance, then the causal relationship between training/transfer tasks ought to be interpreted as a relationship between cognitive measures. However, if participants' training performance gains reflect an *improvement from baseline abilities* over the course of training, then one might argue that boosting performance on an ability confers selective benefits on related, but untrained tasks.

One piece of evidence lending support for successful training (rather than a relationship) is present in the posttest *n*-back task results. Namely, training groups show better performance under conditions that were practiced. If training performance instead indexed some measure of baseline cognitive ability, I would have expected to see no systematic differences in untrained assessment tasks across groups. That is, if practice did not confer any kind of benefit to participants, then they should have performed all levels of the assessment tasks equally well relative to the other training groups. This suggests that some forms of training led to selectively enhanced abilities under conditions congruent with trained skills (i.e., only high-conflict cases benefit from interference-resolution training).

### **4.3 Future Directions**

#### *4.3.1 Considerations for Future Process-Specific Training Studies*

As I have highlighted, a process-specific account maintains that transfer to untrained measures should only be expected only if the executive functions (here, interference resolution) underlying certain tasks are targeted through training so as to affect shared processes that facilitate performance on particular tasks (i.e., WM training

tasks not involving conflict-resolution are not expected to confer transfer). To disentangle mechanistic contributions of the present findings, future work might continue to identify the functional-anatomical overlaps across different memory and language tasks. Research examining healthy adults and patients with neurological disorders demonstrates that cognitive control hinges on the involvement of a widespread network that comprises both cortical (e.g., PFC, cingulate, and parietal) and subcortical (e.g., striatal) regions, clearly not just on prefrontal cortex alone (Burgess et al., 2011; Cools, Sheridan, Jacobs, D'Esposito, 2007; Corbetta & Shulman, 2002; *inter alia*). This pattern is bolstered by training studies documenting the underlying neural signatures accompanying post-intervention differences, including increased activation of frontoparietal regions (Olesen, Westerberg, & Klingberg, 2004); greater structural integrity evaluated by increased fiber tracts (white matter) connecting areas adjacent to intraparietal sulcus (Takeuchi et al., 2010); and an increase in the density of cortical dopamine receptors, perhaps linked to changes in striatal structures (McNab et al., 2009). Although behavioral and neuroimaging findings suggest domain-general processes in PFC that underlie cognitive-control functions across various conditions (Thompson-Schill et al., 2005), an intricate balance exists between PFC and subcortical regions that adjusts performance over different EFs (Cools et al., 2007). Such a cortical/subcortical tradeoff should be considered when investigating the relationships between training and language-transfer tasks.

It is important to note that although I chose to focus on interference-resolution functions here, this does not preclude the involvement of other EFs in the selected training and assessment tasks (Miyake et al., 2000). To this end, the lack of mutual

exclusivity of certain general cognitive processes should be considered when interpreting transfer effects within a process-specific framework, as multiple EFs might be confounded within a single training task; thus, changes in several EFs may be responsible for resultant improvements in outcome measures, a positive outcome if the goal is to show widespread transfer (e.g., Morrison & Chein, 2011; Shipstead et al., 2012). Likewise, the magnitude of the hypothesized transfer effects is sensitive to the level of cognitive control required for each task. The amount of transfer should hinge on the degree to which underlying EFs are shared between training and assessment tasks, and this mechanistic overlap is probably influenced by both the relative involvement of a single trained EF and the extent to which other EFs are recruited in the training and outcome tasks. For example, re-characterization of representations on the high-conflict lure trials of the *n*-back task likely requires other EFs beyond just conflict resolution (e.g., monitoring, updating).

Similarly, syntactic ambiguity resolution and verbal fluency (as studied here) will, of course, rely on updating processes in addition to conflict resolution. Methodologically confounding EFs is an issue that plagues training studies, rendering it difficult to extricate distinct mechanisms entirely; however, by having linking hypotheses, EF overlap across tasks maximizes chances of successful transfer. Careful design of training regimens, including tasks performed by comparison groups—for instance by maintaining minimal task differences between training and active-control tasks (see Experiment 2)—can help elucidate the contribution of distinct EFs. Indeed, the present results indicate that Lures training confers selective benefits to high-EF conditions of sentence processing and verb production tasks.

Correspondingly, the transfer conditions under which selective improvement is observed within an assessment task may mark those relying most on the trained EF. To maximize transfer, it is important to pinpoint the measures in the assessments that capture cognitive processes of interest. For instance, in the present training experiments, I argued that the strongest indices of re-interpretation ability and real-time reanalysis respectively were accuracy to comprehension questions gauging lingering effects of misinterpretation and regression-path reading time in disambiguating sentence regions and second-pass time in earlier sentences. The ability to make specific predictions for when and where transfer is selectively expected, as well as the conditions under which it is not, will ultimately lend important insight to the EFs affected during successful intervention when transfer effects are observed in studies carried out under proper linking assumptions and within a theoretically-guided process-specific account.

Also worth mentioning is the contribution of several—perhaps even overlapping—domain-general resources that may be recruited during language tasks not discussed here (e.g., mnemonic aspects of WM, maintenance, updating, task-switching, etc). This should be carefully considered upon designing outcome language assessments that will be the target of transfer benefits. In fact, I strongly believe that verbal WM ‘span’ processes, which involve maintenance, processing, and temporary storage components, must play a role in spoken language comprehension tasks in which the listener cannot review the input (as she can in normal reading) once it is spoken, without using mnemonic rehearsal strategies. This is likely true regardless of the presence of ambiguity or conflicting representations, and, indeed, verbal WM by itself has been shown to play a role in reading studies using a moving-window paradigm that does not



permit rereading (Fedorenko et al., 2006). Thus, in future work it will be important to design training protocols using tasks that maximize a theoretical match between the cognitive (and neural) processes involved in assessment and training measures, including WM tasks that do not necessarily involve the conflict-resolution aspect of cognitive control, when appropriate.

Cognitive training may also provide a novel approach to understanding whether EFs are critical for a multitude of language uses. The degree to which training improvement predicts changes in language processing can reveal the EFs involved in each condition; if no transfer is observed in selective cases, one might conclude that the trained EFs do not significantly contribute to the processing of the particular language condition (as was the case for object-extracted relative clauses in Experiment 2). This type of approach provides a powerful tool for choosing among several explanations for the same data set, where the best account of the data can be gleaned from the results of a well-designed training study that poses process-specific linking hypotheses. For example, some argue that the difficulty experienced while comprehending the meaning of abstract (compared to concrete) words hinges almost entirely on domain-general processes (Hoffman, Jeffries, & Lambon Ralph, 2010), while other accounts posit little to no contribution from EFs (Barsalou & Wiemer-Hastings, 2005; Rodríguez-Ferreiro, Gennari, Davies, & Cuetos, 2009). The opportunity exists, then, to investigate whether successful EF training permits better abstract-meaning selection.

#### *4.3.2 Applications for Other Populations*

The current findings suggest that training might be used in behavioral remediation programs that aim to improve language skills in situations when competitive interactions

are high. EF training, especially with an interference resolution focus, might yield broader improvements for patient populations—particularly left VLPFC patients—whose language production and comprehension fail under selective conditions due to poor cognitive control. Other groups demonstrating reduced conflict resolution skills that impact on language-processing abilities might benefit from training, including young children (Nilsen & Graham, 2009; Khanna & Boland, 2010; Mazuka et al., 2005; Novick et al., 2005), second-language learners (Abutalebi, 2008; Poarch & van Hell, 2012), the elderly (see Christianson et al., 2006; Hasher, Zacks, & May, 1999; Just & Carpenter, 1992), individuals with attention-deficit/hyperactivity disorder (ADHD; Blaskey, 2004; Engelhardt, Nigg, Carr, & , Ferreira, 2008; Nigg, 2006), and children with specific-language impairments (SLI; Bishop & Norbury, 2006; Tropper, 2009). Further, healthy adults may also benefit from training regimens to combat routine cases of cognitive depletion that occur as a function of stress/anxiety (Bishop, 2007; Eysenck, Derakshan, Santos, & Calvo, 2007; Pessoa, 2010), performance pressure (Beilock & Carr, 2005), and cognitive fatigue (Persson et al., 2007; Van der Linden et al., 2003). Given the high variability of individual differences in cognitive control, it is possible that targeted training regimens may benefit individuals with certain cognitive profiles more than others (Braver, Cole, & Yarkoni, 2011; Jaeggi et al., 2011).

As is highlighted in Chapter 1, much research examining the role of cognitive control for language processes focuses on young children who fail to override their initial cognitive reactions across a range of syntactic and non-syntactic EF measures (e.g., Davidson et al., 2006; Mazuka et al., 2009). Training on developmentally-appropriate versions of *n*-back (see Jaeggi et al., 2011) that include high-densities of lures items

might offset the routine errors committed by children on high-conflict conditions of linguistic and non-linguistic tasks.

In addition to remediating groups with diminished resources, this work could be particularly applicable to aptly characterize the deficits of left prefrontal patients to determine (a) if their interference-resolution performance changes on linguistic and non-linguistic tasks post-training, and (b) what new compensatory processes or brain systems they engage to support any observed performance increases (evaluated through pretest/posttest neuroimaging). Generally, this research program could suggest new inferences about the plasticity of the mind and brain, with respect to language processing especially, and the causal effects of language and cognition interactions.

#### *4.3.3 Applications for Bilingualism*

Finally, it is important to consider a growing body of research demonstrating that balanced bilinguals enjoy certain cognitive advantages relative to their monolingual peers, as this work has important implications for language education and intervention. On tasks requiring cognitive control, some findings suggest that bilinguals outperform monolinguals selectively on trials inducing conflict across a range of tasks such as the Simon task (Bialystok, Craik, Klein, & Viswanathan, 2004). Other data patterns reveal a broader effect, namely that bilinguals are better at conflict *monitoring*: they perform faster on both conflict and non-conflict trials under high, but not low, conflict-monitoring conditions, in which subjects cannot predict when a conflict-related item type (an incongruent flanker trial) might occur because their appearance is equally probable relative to non-conflict trials (Costa, Hernández, Costa-Faidella, & Sebastián-Gallés, 2009). Regardless of the specifics, it has become increasingly clear that rich linguistic

experience (akin to the rich cognitive experience achieved through training) benefits conflict-resolution and cognitive-control performance widely, perhaps due to bilinguals' consistent switching across the two language systems they know and/or their frequent suppression of one lexicon/grammar over another, thus placing a "premium" on EFs associated with updating, conflict resolution, and set-shifting (Costa et al., 2009; Martin-Rhee & Bialystok, 2008). In other words, lifelong bilingualism may be a naturalistic form of cognitive-control training. Indeed, future work should attempt to disentangle the various processing demands that are associated with being a bilingual speaker (e.g., frequent code switches) that might yield the putative cognitive-control advantage they show; such an understanding might help extract the various EFs, in addition to conflict resolution, that are at the heart of bilinguals' benefit. It will also be beneficial to know how bilinguals' cognitive-control advantage concerning conflict resolution or conflict monitoring influences this group's linguistic abilities on the conflict-related language tasks reviewed in this paper. For instance, does bilinguals' cognitive-control advantage result in a better ability to recover the correct interpretation of garden-path sentences, following a misanalysis? The answer to these questions could suggest important inferences one could draw about the prospective impact that process-specific conflict-resolution training might have on this group.

Related, much discussion challenges this general-purpose view as the underlying cause of the bilingual advantage by presenting an explanation centered on the differing language experience of monolinguals and bilinguals. Namely, rather than possessing superior cognitive control, bilinguals' superior task performance under certain task conditions may be due instead to variants in long-term representations, which in turn

changes the amount of conflict that bilinguals must monitor. That is, by having weaker linguistic representations, bilinguals may outperform their monolingual counterparts by not committing to an interpretation or having a default cognitive reaction that is quite as strong as monolinguals. Such a difference in the long-term representational structure would prevent the need to deploy cognitive control to override an initial commitment (akin to what is experienced when an unambiguous sentence is encountered). Training interventions may offer a unique opportunity to disentangle these two explanations, in that short-term cognitive control training should not change long-term representations (and the nature of conflict between these representations) but would target the underlying conflict-control EF.

Recent findings suggest that bilingualism confers protective benefits against cognitive decline: bilingual patients diagnosed with Alzheimer's disease (AD), who are matched on a range of factors (e.g., degree of cognitive impairment, symptomatic expression, demographic variables) to monolinguals with the same diagnosis, have significantly *more* brain atrophy in areas commonly examined to differentiate AD patients from healthy adults (Schweizer, Ware, Fischer, Craik, & Bialystok, 2011). The implication is that bilinguals may have greater "cognitive reserve" than would be predicted given the amount of neuropathology they exhibit; that is, the cognitive symptoms associated with AD may be delayed in this population because of their premorbid advantage. What about bilingual children and VLPFC patients? Are they "inoculated" from the cognitive control deficits they are otherwise known for (in monolinguals) in terms of their nonlinguistic and language processing abilities under high-conflict demands? If so, what behavioral mechanisms and neural systems do they

recruit to compensate?

Furthermore, will cognitive-control training over the long term yield similar protective benefits in monolinguals? Will their performance begin to approach that of (untrained) bilinguals? Will EF training confer comparable protection against normal age-related cognitive decline (Richmond, Morrison, Chein, & Olson, 2011), regardless of AD? These are open empirical questions and might be the focus of future longitudinal research. Also: To what extent does proficiency level matter in adults who have learned a second language, regarding the cognitive-control benefits they reap and the implications for intervention? Balanced bilinguals, as sketched above, enjoy certain advantages; presumably highly proficient (but unbalanced) bilinguals and those with lower proficiency levels will pattern somewhere in between the balanced group and the monolinguals regarding cognitive-control performance, depending on the relative processing demands associated with their proficiency levels. Where they pattern can provide useful insight into the design of future training studies to bring these groups' performance ranges closer to approximate the balanced population. How much room is there for balanced bilinguals to gain from EF training? If a highly proficient group shows a similar cognitive-control advantage to that of bilinguals, then it may suggest the prospect of similar benefits (in terms of effect sizes) gained from training. Conversely, if a low-proficiency group that rarely switches between linguistic systems does not demonstrate a cognitive-control advantage compared to monolinguals, this would suggest opportunity for EF training to bestow benefits. If neither high- or low-proficiency groups demonstrates a cognitive-control advantage, then perhaps learning a second language in adulthood does not enhance EF abilities similar to how early acquisition of two linguistic

systems does. EF training could therefore be beneficial to unbalanced groups across a range of proficiency levels. Ultimately, future work in this area will clarify our understanding of the interplay between bilingualism, cognitive control, and the effects of training on language and other tasks that share cognitive processes.

#### *4.3.4 Caveats*

Although cognitive training offers fruitful avenues for the remediation of deficits and the theoretical elucidation of cognitive abilities for language, there are, however, important caveats to consider. Despite several instances of successful generalization to unpracticed tasks, some reports describe research efforts failing to observe transfer (see Chooi & Thompson, 2012; Melby-Lervag & Hulme, 2013; Owen et al., 2010; Redick et al., 2012). One explanation for the absence of transfer findings may be that in at least one study, EF training was implemented casually, rather than consistently enough to actually tax trainees' EF abilities throughout the regimen (Owen et al., 2010). In this report, not all individuals in the training group received the same exposure to training, a 'dosage-dependent' factor known to confer varying levels of transfer (Jaeggi et al., 2008). Another reason for failure to show transfer may involve the use of performance-*non*-adaptive training tasks (regimens that maintain a constant level of difficulty, rather than keeping participants on the threshold of their best performance), despite evidence favoring such designs to facilitate transfer effects (Brehmer et al., 2011; Klingberg et al., 2005; Lövdén et al., 2010; but see Chapter 3 here, where non-adaptive 3-Back trainees improve on several untrained assessments). Clearly more research is needed to determine what characterizes an appropriate training regimen, as well as how dependent transfer effects are on the amount of training an individual receives (Jaeggi et al., 2008; 2011).

Finally, studies failing to show transfer might lack appropriate linking hypotheses between the types of EF required to perform certain tasks; these must be understood in order to design effective training regimens, which will ultimately inform how future intervention studies are implemented. Indeed, several cases of far-transfer have recently been called into question given failures to replicate published effects (see Cook, 2013 and Shipstead et al., 2012 for reviews). Many of these reports involve attempts at improving general cognitive abilities such as general fluid intelligence and reasoning, without considering identifiable shared underlying processes (akin to the process-specific approach adopted here). By pinpointing common mechanisms across conditions of ostensibly different tasks, it may be more likely to observe successful training-transfer effects, a requirement germane to a process-specific account. Traditional training approaches geared toward *general* task improvements do not emphasize such condition-base differences; namely, higher scores on tasks like Raven's have no carefully controlled cases that require separable EFs (a total score is computed to capture general fluid intelligence). Put differently, tasks with conditions distinguishable on the need for a certain EF (i.e., high- and low-conflict cases), allow for precise hypotheses of where certain types of training—conflict-control, for instance—ought to have an effect. Similarly, assessing transfer by computing a generalized measure on an assessment task (e.g., collapsing across all high- and low-conflict task conditions) may not be expected to confer any pre/post change, in part because control conditions (low-conflict) might mask effects in critical conditions (high-conflict). A process-specific training approach presumes that conditions tapping the EF of interest should improve following practice with another task engaging the shared ability. Indeed, by clearly identifying these cases



within and across tasks, training-mediated change may be more easily interpretable; that these conditions are often not specified for many assessment tasks could be a reason for the failure to consistently demonstrate far-transfer.

Furthermore, there appear to be important individual differences in training success (Chein & Morrison, 2010; Jaeggi et al., 2011), such that only certain individuals achieve performance increases on the training tasks over time, and thus demonstrate transfer to unpracticed measures shown through improved performance at retest. Related, one point of important discussion concerns whether conducting training responder analyses renders results that otherwise may not emerge (see Discussion in Chapter 2). Specifically, if training and transfer measures are correlated (and tap a latent variable), then change on one (e.g., training gains) will lead to a comparable shift on the other (e.g., pre/post gains; see Tidwell et al., 2013). The statistical artifact that exists as a function of identifying responders may undermine interpretations centered on training-mediated effects. Moreover, even if there are clear benefiteres of training, it is unclear if responders and non-responders can be categorized simply by baseline EF abilities, and these differences are unlikely due to motivational factors alone (Jaeggi et al., 2011; Novick et al., 2013). So, future research should address who is most likely to benefit from training, how to identify properly these individuals (if at all), and how training protocols should be modified or tailored to maximize transfer across a range of groups and populations (see Shipstead et al., 2012).

#### **4.4 Closing Remarks**

Executive function training holds promise to result in gains in cognition and language use in both production and comprehension domains, easing processing

difficulty when dominant biases must be reined-in. Such interventions could potentially mitigate problems in language use under generally high conflict/interference demands, not just in special populations (e.g., nonfluent aphasics with interference-resolution deficits), but also in healthy individuals, including developing children, who experience occasional difficulty in reading, listening, or speaking due to heightened demands for cognitive control (in some cases perhaps due to resource depletion).

As sketched in Chapter 1, language is rife with instances of conflicting representations, such that one might expect cognitive control training to extend to other measures of syntactic and lexical processing where the language user must resolve among multiple representations or overcome a default meaning, interpretation, or reaction. Similarly, training regimens that target maintenance-based (WM capacity) abilities might transfer to other cases in language processing known to hinge on verbal working memory ability (e.g., object-extracted relative clauses). Dissociating the cognitive mechanisms affecting language processing will inform the development of effective, well-specified interventions that may have differential use across populations. To close, I have emphasized study of the mental mechanisms underlying language and cognitive processes, as well as how language performance can be modulated through specialized interventions on the basis of understanding these domain-general cognitive mechanisms. Although much follow-up work needs to occur to further flesh out these relationships, the present dissertation serves as an initial attempt to investigate the causal role of cognitive control for language skills within the context of two training paradigms.

## Appendix A

List of ambiguous sentences, unambiguous sentences, and comprehension questions. All stimuli were based on or borrowed from Christianson et al., 2001; 2006.

Ambiguous	Unambiguous	Question
While Jim bathed the child that was blond and pudgy giggled with delight.	The child that was blond and pudgy giggled with delight while Jim bathed.	Did Jim bathe himself?
As the chimps groomed the baboons that were large and hairy sat in the grass.	The baboons that were large and hairy sat in the grass as the chimps groomed.	Did the chimps groom themselves?
While Frank dried off the car that was red and shiny sat in the driveway.	The car that was red and shiny sat in the driveway while Frank dried off.	Did Frank dry off himself?
As Betty woke up the neighbor that was old and cranky coughed loudly.	The neighbor that was old and cranky coughed loudly as Betty woke up.	Did Betty wake herself up?
While the thief hid the jewelry that was elegant and expensive sparkled brightly.	The jewelry that was elegant and expensive sparkled brightly while the thief hid.	Did the thief hide himself?
As Anna dressed the baby that was small and cute spit up on the bed.	The baby that was small and cute spit up on the bed as Anna dressed.	Did Anna dress herself?
While the boy washed the dog that was white and furry barked loudly.	The dog that was white and furry barked loudly while the boy washed.	Did the boy wash himself?
As the jockey settled down the horse that was sleek and brown stood in the stall.	The horse that was sleek and brown stood in the stall as the jockey settled down.	Did the jockey settle himself down?
While the mother undressed the baby that was bald and helpless cried softly.	The baby that was bald and helpless cried softly while the mother undressed.	Did the mother undress herself?
As the nurse shaved the patient that was tired and weak watched TV.	The patient that was tired and weak watched TV as the nurse shaved.	Did the nurse shave herself?
While the girl scratched the cat that was gray and white stared at the dog.	The cat that was gray and white stared at the dog while the girl scratched.	Did the girl scratch herself?
As the mother calmed down the children that were tired and irritable sat on the bed.	The children that were tired and irritable sat on the bed as the mother calmed down.	Did the mother calm herself down?
While Robert changed the paint that was vibrant and colorful spilled on the floor.	The paint that was vibrant and colorful spilled on the floor while Robert changed.	Did Robert change himself?

Ambiguous	Unambiguous	Question
As the secretary transferred the files that were important and messy collected on her desk.	The files that were important and messy collected on her desk as the secretary transferred.	Did the secretary transfer?
While Kristin put make-up on the model that was tall and thin put on her outfit.	The model that was tall and thin put on her outfit while Kristin put make-up on.	Did Kristin put make-up on herself?
As Chris worked out the issue that was confusing and unclear continued to worsen.	The issue that was confusing and unclear continued to worsen as Chris worked out.	Did Chris work out?
While Dave lied down the tiles that were detailed and pricey were cleaned by the maid.	The tiles that were detailed and pricey were cleaned by the maid while Dave lied down.	Did Dave lie down?
As the woman soaked the shirt that was clean and folded sat on the dresser.	The shirt that was clean and folded sat on the dresser as the woman soaked.	Did the woman soak herself?
While the woman disrobed the mannequin that was frail and shapely stood in the store.	The mannequin that was frail and shapely stood in the store while the woman disrobed.	Did the woman disrobe herself?
As the student prepared the salad that was healthy and fresh remained in the refrigerator.	The salad that was healthy and fresh remained in the refrigerator as the student prepared.	Did the student prepare herself?
While the squirrels relocated the acorns that were brown and ripe fell from the trees.	The acorns that were brown and ripe fell from the trees while the squirrels relocated.	Did the squirrels relocate themselves?
As the gardener showered the flowers that were yellow and blue were gathered by a child.	The flowers that were yellow and blue were gathered by a child as the gardener showered.	Did the gardener take a shower?
While the nanny stripped the girl that was tearful and fussy threw a tantrum.	The girl that was tearful and fussy threw a tantrum while the nanny stripped.	Did the nanny strip herself?
As the model covered up the portrait that was colorful and exact fell from the easel.	The portrait that was colorful and exact fell from the easel as the model covered up.	Did the model cover herself up?
As the servant bathed the king that was arrogant and pompous ate chocolate.	The king that was arrogant and pompous ate dark chocolate as the servant bathed.	Did the servant bathe himself?
While the jockey groomed the horse that was wild and testy paced in the stall.	The horse that was wild and testy paced in the stall while the jockey groomed.	Did the jockey groom himself?
As the trainer dried off the dog that was playful and friendly fetched the stick.	The dog that was playful and friendly fetched the stick as the trainer dried off.	Did the trainer dry himself off?
While the baby-sitter woke up the infant that was tiny and fragile cried in his crib.	The infant that was tiny and fragile cried in his crib while the baby-sitter woke up.	Did the baby-sitter wake up?

Ambiguous	Unambiguous	Question
As the mother hid the cookies that were warm and gooey baked in the oven.	The cookies that were warm and gooey baked in the oven as the mother hid.	Did the mother hide herself?
While the tailor dressed the figurine that was tall and shapely fell over.	The figurine that was tall and shapely fell over while the tailor dressed.	Did the tailor dress herself?
As the adolescent washed the dishes that were orange and greasy sat in the sink.	The dishes that were orange and greasy sat in the sink as the adolescent washed.	Did the adolescent wash himself?
While the farmer settled down the pig that was pink and squealing escaped from its pen.	The pig that was pink and squealing escaped from its pen while the farmer settled down.	Did the farmer settle down?
As Molly undressed the teddy bear that was plush and huggable lost a button.	The teddy bear that was plush and huggable lost a button as the girl undressed.	Did Molly undress herself?
While the barber shaved the customer that was hurried and impatient left the shop.	The customer that was hurried and impatient left the shop while the barber shaved.	Did the barber shave himself?
As the hero scratched the villain that was sneaky and traitorous kidnapped the blonde bombshell.	The villain that was sneaky and traitorous kidnapped the blonde bombshell as the hero scratched.	Did the hero scratch himself?
While the secretary calmed down the client that was ruined and desperate staked out the building.	The client that was ruined and desperate staked out the building while the secretary calmed down.	Did the secretary calm down?
As the maid changed the filter that was old and dirty collected dust.	The filter that was old and dirty collected dust as the maid changed.	Did the maid change herself?
While the operator transferred the caller that was eager and animated accidentally hung up.	The caller that was eager and animated accidentally hung up while the operator transferred.	Did the operator transfer?
As the artist put make-up on the actress that was famous and beautiful walked onto the set.	The actress that was famous and beautiful walked onto the set as the artist put make-up on.	Did the artist put make-up on herself?
While the manager worked out the contract that was beneficial and generous was signed.	The contract that was beneficial and generous was signed while the manager worked out.	Did the manager work out?
As the librarian lied down the book that was intellectual and depressing stayed on the shelf.	The book that was intellectual and depressing stayed on the shelf as the librarian lied down.	Did the librarian lie down?
While the bride soaked the groom that was handsome and smiling put away his tuxedo.	The groom that was handsome and smiling put away his tuxedo while the bride soaked.	Did the bride soak herself?

Ambiguous	Unambiguous	Question
As the prince disrobed the courtesan that was graceful and voluptuous poured the wine.	The courtesan that was graceful and voluptuous poured the wine as the prince disrobed.	Did the prince disrobe himself?
While the butcher prepared the meat that was tender and succulent went through the grinder.	The meat that was tender and succulent went through the grinder while the butcher prepared.	Did the butcher prepare himself?
As the CEO relocated the store that was small and unsuccessful held a sale.	The store that was small and unsuccessful held a sale as the CEO relocated.	Did the CEO relocate himself?
While the veterinarian showered the cat that was sickly and thin mewed in its cage.	The cat that was sickly and thin mewed in its cage while the veterinarian showered.	Did the veterinarian take a shower?
As the dancer stripped the curtains that were faded and musty blocked the light.	The curtains that were faded and musty blocked the light as the dancer stripped.	Did the dancer strip herself?
While the sculptor covered up the statue that was chiseled and perfect stood erect.	The statue that was chiseled and perfect stood erect while the sculptor covered up.	Did the sculptor cover himself up?
While Sally bathed the calf that was stubborn and restless head-butted the barn wall.	The calf that was stubborn and restless head-butted the barn wall while Sally bathed.	Did Sally bathe herself?
As the aid groomed the patient that was confused and forgetful wandered the hospital.	The patient that was confused and forgetful wandered the hospital as the aid groomed.	Did the aid groom herself?
While the soldier dried off the gun that was antique and rusty fired clouds of smoke.	The gun that was antique and rusty fired clouds of smoke while the soldier dried off.	Did the soldier dry himself off?
As Neville woke up the lady that was elegant and wealthy snorted in disapproval.	The lady that was elegant and wealthy snorted in disapproval as Neville woke up.	Did Neville wake up?
While the guest hid the present that was extravagant and glittering sat on the table.	The present that was extravagant and glittering sat on the table while the guest hid.	Did the guest hide herself?
As the butler dressed the salad that was fresh and green was prepared for dinner.	The salad that was fresh and green was prepared for dinner as the butler dressed.	Did the butler dress himself?
While the handyman washed the car that was red and flashy drove down the road.	The car that was red and flashy drove down the road while the handyman washed.	Did the handyman wash himself?
As the nun settled down the convent that was pious and devoted welcomed the Pope.	The convent that was pious and devoted welcomed the Pope as the nun settled down.	Did the nun settle herself down?

Ambiguous	Unambiguous	Question
While Theresa undressed the girl who was wet and embarrassed stood in the corner.	The girl who was wet and embarrassed stood in the corner while Theresa undressed.	Did Theresa undress herself?
As Marcus shaved the sheep that was fluffy and woolly bleated in its pen.	The sheep that was fluffy and woolly bleated in its pen as Marcus shaved.	Did Marcus shave himself?
While the outfielder scratched the tree that was supple and young waved in the breeze.	The tree that was supple and young waved in the breeze while the outfielder scratched.	Did the outfielder scratch himself?
As the team calmed down the coach that was stern and formidable bought a victory lunch.	The coach that was stern and formidable bought a victory lunch as the team calmed down.	Did the team calm themselves down?
As the sergeant changed the drill that was repetitive and exhausting stopped hurting the recruits.	The drill that was repetitive and exhausting stopped hurting the recruits as the sergeant changed.	Did the sergeant change himself?
While the executive transferred the assistant that was efficient and knowledgeable packed boxes.	The assistant that was efficient and knowledgeable packed boxes while the executive transferred.	Did the executive transfer?
As newscaster put make-up on the co-anchor that was chatty and likable read the prompt.	The co-anchor that was chatty and likable read the prompt as the newscaster put make-up on.	Did the newscaster put make-up on herself?
While Cora worked out the knot that was aching and sore throbbed in her thigh.	The knot that was aching and sore throbbed in her thigh while Cora worked out.	Did Cora work out?
As the agent lay down the badge that was battered and clunky broke in half.	The badge that was battered and clunky broke in half as the agent lay down.	Did the agent lie down?
While the chef soaked the greens that were nutritious and plentiful grew under heat lamps.	The greens that were nutritious and plentiful grew under heat lamps while the chef soaked.	Did the chef soak himself?
As the coach disrobed the champion that was muscular and agitated punched his opponent.	The champion that was muscular and agitated punched his opponent as the coach disrobed.	Did the coach disrobe himself?
While the lawyer prepared the witness that was nervous and fidgety started to hyperventilate.	The witness that was nervous and fidgety started to hyperventilate while the lawyer prepared.	Did the lawyer prepare himself?
As the miser relocated the treasure that was valuable and lucrative gained interest.	The treasure that was valuable and lucrative gained interest as the miser relocated.	Did the miser relocate himself?

Ambiguous	Unambiguous	Question
While the pool boy showered the deck that was grimy and dark creaked eerily.	The deck that was grimy and dark creaked eerily while the pool boy showered.	Did the pool boy shower himself?
As Angela stripped the paper that was pink and flowery began to peel.	The paper that was pink and flowery began to peel as Angela stripped.	Did Angela strip herself?
While the thief covered up the goods that were conspicuous and grand drew attention from pedestrians.	The goods that were conspicuous and grand drew attention from pedestrians while the thief covered up.	Did the thief cover himself up?



## Appendix B

Stimulus sets used for the 3-Back training task of Experiment 2.

Version	Stimuli
Letters	b, c, d, f, h, j, k, l, m, p, q, r, s, t, v, x
Words	bench, clock, dress, flag, grass, hill, jet, key, light, nurse, pipe, rope, sky, truck, van, wheel
Nonwords	blick, chut, desh, flonk, glit, hend, jurch, kasp, lerth, milp, noss, parf, rax, slirt, trean, ving
Symbol Set 1	! & @ = ◆ ♥ ♣ ☼ ✂ 🕶 🕯 🕸 🦒 🦉 🦊
Symbol Set 2	? \$ % ✓ ▲ ♠ ★ ☾ 🖐 📄 🔔 📖 🦋 🐰 🦢 🐎

## Appendix C

Noun cues provided to participants during the verb generation task used in Experiment 2. High-competition refers to nouns with multiple verb associates (e.g., *ball – bounce, throw, kick*), whereas low-competition refers to nouns with few verb associates (e.g., *scissors – cut*). High-association refers to nouns with strong verb associates (e.g., *bed – sleep*), while low-association refers to nouns with weak verb associates (e.g., *valley – hike*).

Association	Competition	Noun Cue	Association	Competition	Noun Cue
High	High	BEACH	High	Low	BINDER
High	High	BELT	High	Low	PHOTO
High	High	BOAT	High	Low	BASEMENT
High	High	BOX	High	Low	CABINET
High	High	CAT	High	Low	CORD
High	High	CLOCK	High	Low	COUNTER
High	High	FINGER	High	Low	DECADE
High	High	HAMMER	High	Low	ELEPHANT
High	High	HEAD	High	Low	GALAXY
High	High	MOUTH	High	Low	HALO
High	High	OVEN	High	Low	HEDGE
High	High	SUN	High	Low	MOLE
High	High	WOOD	High	Low	NEBULA
High	High	LOCK	High	Low	WART
High	High	BASKET	Low	High	CUP
High	High	PHONE	Low	High	FOOT
High	High	WHEEL	Low	High	KNIFE
High	High	ICE	Low	High	OAR
High	High	ROPE	Low	High	PLANE
High	High	LETTER	Low	High	POOL
High	High	PURSE	Low	High	RAZOR
High	High	ATTIC	Low	High	SINK
High	High	SCHOOL	Low	High	SOAP
High	High	SEED	Low	High	STOVE
High	High	BLANKET	Low	High	TOWEL
High	High	MONEY	Low	High	WING
High	Low	SPONGE	Low	High	DOLLAR
High	Low	GIRAFFE	Low	High	GLASS
High	Low	SAND	Low	High	MOVIE
High	Low	CRANE	Low	High	YARN
High	Low	BOARD	Low	High	BIRD
High	Low	JACKAL	Low	High	BED
High	Low	GRATE	Low	High	CRAYON
High	Low	FOLDER	Low	High	SHOE
High	Low	DRESSER	Low	High	SCALE
High	Low	OSTRICH	Low	High	NOSE

Association	Competition	Noun Cue	Association	Competition	Noun Cue
Low	High	GRAVE	Low	Low	TAPE
Low	High	LAWN	Low	Low	ANGEL
Low	Low	ANCHOR	Low	Low	LILY
Low	Low	BLOCK	Low	Low	RAY
Low	Low	COMPUTER	Low	Low	MARKER
Low	Low	FLAG	Low	Low	DISH
Low	Low	GLOVE	Low	Low	MARIGOLD
Low	Low	JAM	Low	Low	RING
Low	Low	KELP	Low	Low	TREK
Low	Low	LIZARD	Low	Low	CONE
Low	Low	MANHOLE	Low	Low	TISSUE
Low	Low	MUG	Low	Low	RABBIT
Low	Low	PIT	Low	Low	PALACE
Low	Low	SPRINKLER	Low	Low	HOSE

## Appendix D

List of object-extracted (OE) sentences, subject-extracted (SE) sentences, and comprehension questions. All stimuli were borrowed from Fedorenko et al., 2006.

Object-Extracted (OE)	Subject-Extracted (SE)	Comprehension Question
The analyst who the governor queried proposed some changes to the plan.	The analyst who queried the governor proposed some changes to the plan.	Was the analyst questioned?
The celebrity who the athlete admired won the award at the ceremony.	The celebrity who admired the athlete won the award at the ceremony.	Did the celebrity suffer a defeat?
The clerk who the director disliked typed the letter to the administration.	The clerk who disliked the director typed the letter to the administration.	Would the administration be receiving a letter?
The client who the retailer contacted offered a deal of the century.	The client who contacted the retailer offered a deal of the century.	Did the client retract a deal?
The contestant who the host offended ruined the show for the audience.	The contestant who offended the host ruined the show for the audience.	Was the show a total success?
The detective who the spy recognized crossed the street at the light.	The detective who recognized the spy crossed the street at the light.	Did the detective cross at the sign?
The diplomat who the congressman insulted ended the negotiations on the spot.	The diplomat who insulted the congressman ended the negotiations on the spot.	Did the negotiations go well?
The employee who the executive praised finished the project right on time.	The employee who praised the executive finished the project right on time.	Did the employee finish on time?
The farmer who the expert questioned promoted the product at the fair.	The farmer who questioned the expert promoted the product at the fair.	Would the product be promoted on TV?
The guitarist who the band recommended recorded the song for the album.	The guitarist who recommended the band recorded the song for the album.	Did the guitarist recommend the band?
The journalist who the editor complimented revised the article for the newspaper.	The journalist who complimented the editor revised the article for the newspaper.	Did the journalist work for a newspaper?
The legislator who the senator visited falsified the documents for the trip.	The legislator who visited the senator falsified the documents for the trip.	Did the legislator use fake documents?

Object-Extracted (OE)	Subject-Extracted (SE)	Comprehension Question
The librarian who the teacher angered misplaced the book from the depository.	The librarian who angered the teacher misplaced the book from the depository.	Did the librarian keep track of the book?
The mathematician who the physicist addressed offered the proof at the conference.	The mathematician who addressed the physicist offered the proof at the conference.	Were the scientists working in someone's office?
The medic who the doctor assisted borrowed the instrument for the surgery.	The medic who assisted the doctor borrowed the instrument for the surgery.	Were the medical professionals preparing for an operation?
The officer who the murderer described told a lie about the past.	The officer who described the murderer told a lie about the past.	Did the officer say something that wasn't true?
The official who the manager harassed questioned the policy of lowering wages.	The official who harassed the manager questioned the policy of lowering wages.	Was the policy being challenged?
The pharmacist who the assistant helped placed the order for the drug.	The pharmacist who helped the assistant placed the order for the drug.	Did the pharmacist order a coupon book?
The priest who the nun thanked founded the shelter near the church.	The priest who thanked the nun founded the shelter near the church.	Were the priest and nun involved in charity?
The reporter who the cameraman followed damaged the equipment during the trip.	The reporter who followed the cameraman damaged the equipment during the trip.	Did the reporter break equipment?
The salesman who the cashier resented mislabeled the products in the brochure.	The salesman who resented the cashier mislabeled the products in the brochure.	Was there an error in the brochure?
The soldier who the enemy shot received a medal for the battle.	The soldier who shot the enemy received a medal for the battle.	Did the soldier receive an honor?
The waiter who the cook invited tasted the sauce for the meat.	The waiter who invited the cook tasted the sauce for the meat.	Did someone sample the meat?
The waitress who the bartender hugged dropped the tray on the floor.	The waitress who hugged the bartender dropped the tray on the floor.	Did the waitress drop the tray on the table?
The accountant who the statistician advised calculated the costs of the project.	The accountant who advised the statistician calculated the costs of the project.	Were the professionals working together?

Object-Extracted (OE)	Subject-Extracted (SE)	Comprehension Question
The actor who the starlet respected forgot the lines during the scene.	The actor who respected the starlet forgot the lines during the scene.	Were they getting ready for a music concert?
The babysitter who the parents liked planned a trip to Puerto Rico.	The babysitter who liked the parents planned a trip to Puerto Rico.	Was someone getting ready for a vacation?
The banker who the chairman informed invested a million in a start-up.	The banker who informed the chairman invested a million in a start-up.	Did someone refuse to invest money?
The bully who the nerd challenged began the quarrel with an insult.	The bully who challenged the nerd began the quarrel with an insult.	Did the bully start with a complement?
The burglar who the policeman wounded reloaded the revolver in a hurry.	The burglar who wounded the policeman reloaded the revolver in a hurry.	Was someone injured?
The carpenter who the electrician punched quit the job a week later.	The carpenter who punched the electrician quit the job a week later.	Did the electrician and carpenter get along?
The co-worker who the professional intimidated delayed his response to the question.	The co-worker who intimidated the professional delayed his response to the question.	Did the co-worker answer right away?
The committee who the applicant met explained the reasoning for their decision.	The committee who met the applicant explained the reasoning for their decision.	Did the committee neglect to give a reason?
The critic who the writer acknowledged discussed the strengths of the piece.	The critic who acknowledged the writer discussed the strengths of the piece.	Did the critic discuss the weakness?
The expert who the source revered wrote a commentary on natural selection.	The expert who revered the source wrote a commentary on natural selection.	Did the expert write a poem?
The hairdresser who the beautician hired transformed the salon for the better.	The hairdresser who hired the beautician transformed the salon for the better.	Did the hairdresser improve the salon?
The hero who the villain destroyed spent daylight hours in his lair.	The hero who destroyed the villain spent daylight hours in his lair.	Did the hero spend the day in the city?

Object-Extracted (OE)	Subject-Extracted (SE)	Comprehension Question
The host who the visitor engaged described the route to the attractions.	The host who engaged the visitor described the route to the attractions.	Did the host describe how to get to the attractions?
The investigator who the cop overheard closed the case without an arrest.	The investigator who overheard the cop closed the case without an arrest.	Did someone eventually get convicted?
The lecturer who the dean provoked left the university in the summer.	The lecturer who provoked the dean left the university in the summer.	Did the lecturer leave during the summer?
The mobster who the dealer attacked organized some crimes in New York.	The mobster who attacked the dealer organized some crimes in New York.	Was the dealer attacked?
The model who the artist approached signed the contract for a year.	The model who approached the artist signed the contract for a year.	Was a multi-year deal involved?
The physician who the cardiologist consulted checked the files in his office.	The physician who consulted the cardiologist checked the files in his office.	Were the files checked in the office?
The plumber who the janitor frustrated lost the key on the street.	The plumber who frustrated the janitor lost the key on the street.	Did the plumber lose the key?
The scientist who the technician aided assumed the blame for the error.	The scientist who aided the technician assumed the blame for the error.	Did the scientist blame himself for the error?
The student who the professor trusted answered the question about the experiment.	The student who trusted the professor answered the question about the experiment.	Was there a discussion about research?
The trumpeter who the drummer loved formed the band two years ago.	The trumpeter who loved the drummer formed the band two years ago.	Did the trumpeter start the band?
The violinist who the cellist flattered played a piece from the symphony.	The violinist who flattered the cellist played a piece from the symphony.	Was the violinist flattered?

## References

- Abutalebi, J. (2008). Neural aspects of second language representation and language control, *Acta Psychologica*, 128(3), 466-478.
- Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247-264.
- Baayen, R.H. (2010). languageR: Data sets and functions with *Analyzing linguistic data: A practical introduction to statistics*. R package version 1.2.
- Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
- Badre, D., & Wagner, A.D. (2007). Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* 45, 2883–2901.
- Balota, D.A., & Faust, M.E. (2001). Attention in dementia of the Alzheimer's type. In F. Boller & S.F. Cappa (Eds.), *Handbook of neuropsychology*, 2<sup>nd</sup> edition, pp. 51–80. New York: Elsevier Science.
- Barr, D. J. (2008). Analyzing “visual world” eye-tracking data using multilevel logistic regression. *Journal of Memory and Language*, 59, 457–474.
- Barsalou L.W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher and R. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought*. (pp. 129-163). New York: Cambridge University Press.
- Bates, D.M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes. R package version 0.9975-12.



- Bedny, M., Hulbert, J.C., & Thompson-Schill, S.L. (2007). Understanding words in context: The role of Broca's area in word comprehension. *Brain Research*, 1146, 101–114.
- Beilock, S.L., & Carr, T.H. (2005). When High-Powered People Fail: Working Memory and “Choking Under Pressure” in Math. *Psychological Science*, 16(2), 101–105.
- Belke, E., Meyer, A., & Damian, M.F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *The Quarterly Journal of Experimental Psychology*, 58(4), 667–92.
- Bialystok, E., Craik, F.I., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging*, 19, 290-303.
- Bilenko, N.Y., Grindrod, C.M., Myers, E.B., & Blumstein, S.E. (2009). Neural correlates of semantic competition during processing of ambiguous words. *Journal of Cognitive Neuroscience*, 21(5), 960–975.
- Bishop, S.J. (2007). Neurocognitive mechanisms of anxiety: An integrative account. *Trends in Cognitive Science*, 11(7), 307-316.
- Bishop, D.V., & Norbury, C.F. (2005). Executive functions in children with communication impairments, in relation to autistic symptomatology: I. Generativity, *Autism*. 9: 7-27.
- Blaskey, L.G. (2004). *Inhibitory language deficits in attention-deficit/ hyperactivity disorder and reading disorder: A candidate shared deficit*. Unpublished doctoral dissertation. Michigan State University.
- 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), 624–652.
- Brain Fitness Program (Version 2.1) [Computer software]. San Francisco, CA: Posit Science.
- Braver, T., Cole, M., Yarkoni, T. (2010). Vive les differences! Individual variation in neural mechanisms of executive control. *Current Opinion in Neurobiology*, 20, 242-250.
- Braver, T.S., Gray, J.R., & Burgess, G C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In A.R.A. Conway, C. Jarrold, M.J. Kane, A. Miyake, J.N. Towse (Eds.), *Variation in Working Memory*, Oxford University Press, pp. 76-106.
- Brehmer, Y., Rieckmann, A., Bellander, M., Westerberg, H., Fischer, H., & Bäckman, L. (2011). Neural correlates of training-related working-memory gains in old age. *Neuroimage*, 58(4), 1110-20.
- Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, 16(5), 893-900.
- Bunting, M.F., Cowan, N., & Saults, J.S. (2006). How does running memory span work? *Quarterly Journal of Experimental Psychology*, 59, 1691-1700.
- Burgess, G.C., Gray, J.R., Conway, A.R., & Braver, T.S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General*, 140, 674-92.
- Burnham, K.P., & Anderson, D.R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2), 261-304.

- Buschkuehl, M., Jaeggi, S.M., Hutchison, S., Perrig-Chiello, P., Däpp, C., Mueller, M., Breil, F., Hoppeler, H., & Perrig, W.J. (2008). Impact of working memory training on memory performance in old-old adults. *Psychology and Aging, 23*, 743–753.
- Caplan, D. Alpert, N., & Waters, G. (1998). Effects of syntactic structure and propositional number on patterns of regional cerebral blood flow. *Journal of Cognitive Neuroscience, 10*(4), 541-552.
- Carretti, B., Borella, E., & De Beni, R. (2007). Does strategic memory training improve the working memory performance of younger and older adults? *Experimental Psychology, 54*, 311–320.
- Chase, W.G., & Ericsson, K.A. (1982). Skill and working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 1–58). New York, NY: Academic Press.
- Chatham, C.H., Herd, S.A., Brant, A.M., Hazy, T.E., Miyake, A., O'Reilly, R., & Friedman, N.P. (2011). From an executive network to executive control: a computational model of the n-back task. *Journal of Cognitive Neuroscience, 23*(11), 3598-3619.
- Chein, J., & Morrison, A. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review, 17*, 193-199.
- Chooi, W-T., & Thompson, L.A., (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence, 40*(6), 531-542.

- Christianson, K., & Luke, S.G. (2011). Garden path vs. local coherence: Online processing and offline comprehension. Poster presented at the 17th Annual Conference on Architectures and Mechanisms for Language Processing. Paris, France.
- Christianson, K., Hollingworth, A., Halliwell, J., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42, 368-407.
- Christianson, K., Williams, C., Zacks, R., & Ferreira, F. (2006). Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Processes*, 42(2), 205–238.
- Cook, G. (2013, April 5). Brain games are bogus. *The New Yorker*. Retrieved from <http://www.newyorker.com/online/blogs/elements/2013/04/brain-games-are-bogus.html>
- Cools, R., Sheridan, M., Jacobs, E.J., & D'Esposito, M.D. (2007). Impulsive personality predicts dopamine-dependent changes in fronto-striatal activity during component processes of working memory. *Journal of Neuroscience*, 27, 5506–5514.
- Copland, D.A., Sefe, G., Ashley, J., Hudson, C., & Chenery, H.J. (2009). Impaired semantic inhibition during lexical ambiguity repetition in Parkinson's disease. *Cortex*, 45(8), 943–949.
- Corbetta, M., & Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201-215.
- Costa, A. Hernández, M., Costa-Faidella, J. & Sebastián-Gallés, N. (2009). On the bilingual advantage in conflict processing: Now you see it, now you don't. *Cognition*, 113(2), 135-149.

- Cowan, N. (2001) The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87–18.
- D’Esposito, M., & Postle, B. R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia*, 37, 1303-1315.
- Dahlin, E., Neely, A.S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science*, 320, 1510–1512.
- Davidson, M.C., Amso, D., Cruess Anderson, L., & 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, 2037–2078.
- Dimitrov, D.M., & Rumrill, P.D. (2003). Pretest-posttest designs and measurement of change. *Work*, 20, 159-165.
- Duncan, J. (2010) The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behavior. *Trends in Cognitive Sciences*, 14(4), 172-179.
- Engelhardt, P., Nigg, J.T., Carr, L.A. & Ferreira, F. (2008). Cognitive inhibition and working memory in attention-deficit/hyperactivity disorder. *Journal of Abnormal Psychology*, 117. 591-605.
- Eysenck, M.W., Derakshan, N., Santos, R., & Calvo, M.G., (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, 7(2), 336-353.
- Fedorenko, E., Gibson, E. & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific resources. *Journal of Memory and Language*, 54(4), 541-53.

- Fedorenko, E., Nieto-Castañon, A. & Kanwisher, N. (2012). Syntactic processing in the human brain: What we know, what we don't know, and a suggestion for how to proceed. *Brain and Language*, 120, 187-207.
- Fedorenko, E., Woodbury, R. & Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in long-distance structural dependencies in language. *Cognitive Science*, 1-17. doi: 10.1111/cogs.12021
- Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, 30, 3-20.
- Feuerstein, R. (1980). Cognitive modifiability in adolescence: Cognitive structure and the effects of intervention. *Journal of Special Education*, 15(2), 269–287.
- Fraley, C., & Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611-631.
- Fraley, C., & Raftery, A.E. (2011). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. R package version 3.4.10.
- Frazier, L. & Fodor, J. D. (1978). The sausage machine: a new two-stage parsing model. *Cognition*, 2 (4), 291–325.
- Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye- movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, 178-210.
- Friedman, N.P., & Miyake, A. (2004). The relations among inhibition and interference cognitive functions: A latent variable analysis. *Journal of Experimental Psychology: General*, 133, 101-135.

- Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory & Language*, 37, 58-93.
- Gennari, S. P. & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, 58, 161-187.
- Gray, J.R., Braver, T.S., & Raichle, M.E. (2002). Integration of emotion and cognition in the lateral prefrontal cortex. *Proceedings of the National Academy of Sciences*, 99, 4115–4120.
- Gray, J.R., Chabris, C.F., & Braver, T.S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, 6, 316-322.
- Gregg, V. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition*. New York: Wiley.
- Grindrod, C.M., & Baum, S.R. (2003). Sensitivity to local sentence context information in lexical ambiguity resolution: Evidence from left- and right-hemisphere-damaged individuals. *Brain and Language*, 85, 503–523.
- Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In *The new cognitive neurosciences* (Fourth Edition). Cambridge, MA: MIT Press.
- Hamilton, A.C., & Martin, R.C. (2005). Dissociations among tasks involving inhibition: A single-case study. *Cognitive, Affective, & Behavioral Neuroscience*, 5, 1–13.
- Hasher, L., & Zacks, R.T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193–225). San Diego, CA: Academic.
- Hasher, L., Zacks, R.T., & May, C P. (1999). Inhibitory control, circadian arousal, and

- age. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 653–675). Cambridge, MA: MIT Press.
- Haut, M. W., Kuwabara, H., Leach, S., & Arias, R. G. (2000). Neural activation during performance of number-letter sequencing. *Applied Neuropsychology*, 7(4), 237-242.
- Hays, R.D., & Hadorn, D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research*, 1, 73–5.
- Hockey, R. (1973). Rate of presentation in running memory and direct manipulation of input-processing strategies. *Quarterly Journal of Experimental Psychology*, 25, 104-111.
- Hodgson, C., Schwartz, M.F., Brecher, A., & Rossi, N. (2003). Effects of relatedness, repetition and rate: Further investigations of context-sensitive naming. *Brain and Language*, 87(1), 31–32.
- Hoffman, P., Jefferies, E., & Lambon Ralph, M.A. (2010). Ventrolateral prefrontal cortex plays an executive regulation role in comprehension of abstract words: Convergent neuropsychological and repetitive TMS evidence. *Journal of Neuroscience*, 30(46), 15450-15456.
- Holmes, J. Gathercole, S., & Dunning, D. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental Science*, 12(4), F9–F15.
- Hussey, E.K., & Novick, J.M. (2012). The benefits of executive control training and the implications for language processing. *Frontiers in Cognition*, 3(158). doi:



10.3389/fpsyg.2012.00158

- Huttenlocher, P.R., & Dabholkar, A.S. 1997. Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology*, 387, 167–178.
- Jaeger, F.T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59, 434–446.
- Jaeggi, S., Buschkuhl, M., Jonides, J., & Perrig, W. (2008). Improving fluid intelligence with training on working memory. *Proceedings from the National Academy of Sciences*, 105(19), 6829-6833.
- Jaeggi, S., Buschkuhl, M., Jonides, J., & Shah, P. (2011). Short and long term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, 108(25), 10081–10086.
- Jaeggi, S.M., Seewer, R., Nirkko, A.C., Eckstein, D., Schroth, G., Groner, R., & Gutbrod, K. (2003). Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependent processing in single and dual tasks: A functional magnetic resonance imaging study. *NeuroImage*, 19(2), 210-225.
- Jaeggi, S.M., Studer-Luethi, B., Buschkuhl, M., Su, Y., Jonides, J., & Perrig, W.J. (2010). The relationship between n-back performance and matrix reasoning — implications for training and transfer. *Intelligence*, 38, 625–635.
- January, D., Trueswell, J., & Thompson-Schill, S. (2009). Co-localization of Stroop and Syntactic Ambiguity Resolution in Broca's Area: Implications for the Neural Basis of Sentence Processing. *Journal of Cognitive Neuroscience*, 21, 2434–2444.
- Jonides, J. (2004). How does practice makes perfect? *Nature Neuroscience*, 7(1), 10–11.

- Jonides, J., & Nee, D.E. (2006). Brain mechanisms of proactive interference in working memory. *Neuroscience*, 139, 181-193.
- Jonides, J. Smith, E.E., Marshuetz, C., Koeppel, R.A. & Reuter-Lorenz, P.A. (1998). Inhibition in verbal working memory revealed by brain activation. *Proceedings of the National Academy of Science*, 95, 8410-8413.
- Just M.A., & Carpenter P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329–354.
- Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, 99, 122–149.
- Kan, I.P., & Thompson-Schill, S.L. (2004). Effect of name agreement on prefrontal activity during overt and covert picture naming. *Cognitive, Affective, & Behavioral Neuroscience*, 4, 43–57.
- Kane, M., Conway, A., Miura, T., & Colflesh, G. (2007). Working memory, attention control, and *n*-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3), 615-622.
- Kane, M.J., & Engle, R.W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, 132(1), 47-70.
- Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science*, 12(6), 978–990.

- Khanna, M.M. & Boland, J.E. (2010). Children's use of language context in lexical ambiguity resolution. *Quarterly Journal of Experimental Psychology*, 63(1), 160-193.
- King, J., & Just, M.A. (1991). Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language*, 30, 580-602.
- Klingberg, T., Fernell, E., Olesen, P.J., Johnson, M., Gustafsson, P., Dahlström, K., Gillberg, C., Forssberg, H., & Westerberg, H. (2005). Computerized training of working memory in children with ADHD—A randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44, 177-186.
- Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Increased brain activity in frontal and parietal cortex underlies the development of visuospatial working memory capacity during childhood. *Journal of Cognitive Neuroscience*, 14(1), 1-10.
- Kloo, D., & Perner, J. (2003). Training transfer between card sorting and false belief understanding: Helping children apply conflicting descriptions. *Child Development*, 74(6), 1823-1839.
- Larson, M.J., Kaufman, D.A., & Perlstein, W.M. (2009). Conflict adaptation and cognitive control adjustments following traumatic brain injury. *Journal of the International Neuropsychological Society*, 15, 927-937.
- Lewis, R. L. & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29, 375-419.
- Li, S., Schmiedek, F., Huxhold, O., Röcke, C., Smith, J., & Lindenberger, U. (2008). Working memory plasticity in old age: Practice gain, transfer, and maintenance.

*Psychology and Aging*, 23(4), 731–742.

- Linn, P.L., & Slinde, J.A. (1977). Determination of the significance of change between pre- and post testing periods. *Reviews of Educational Research*, 47, 121–50.
- Long, D. L. & Prat, C. S. (2008). Individual differences in syntactic ambiguity resolution: Readers vary in their use of plausibility information. *Memory & Cognition*, 36, 375-391.
- Lövden, M., Bäckman, L., Lindenberger, U., Schaefer, S., & Schmiedek, F. (2010). A theoretical framework for the study of adult cognitive plasticity. *Psychological Bulletin*, 136, 659–676.
- MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101, 676–703.
- Martin, R.C., & Cheng, Y. (2006). Selection demands versus association strength in the verb generation task. *Psychonomic Bulletin & Review*, 13, 396–401.
- Martin, A.M., Quinn, K.M., & Park, J.H. (2011). MCMCpack: Markov Chain Monte Carlo (MCMC) package. R package version 1-1.4.
- Martin-Rhee, M.M., & Bialystok, E. (2008). The development of two types of inhibitory control in monolingual and bilingual children. *Bilingualism: Language and Cognition*, 11, 81-93.
- Masson, M.E.J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, 43, 679-690.
- Mazuka, R., Jincho, N., & Oishi, H. (2009). Development of executive control and language processing. *Language and Linguistics Compass*, 3, 59-89.
- McNab, F., Varrone, A., Farde, L., Jucaite, A., Bystritsky, P. Forssberg, H., & Klingberg,

- T. (2009). Changes in cortical dopamine D1 receptor binding associated with cognitive training. *Science*, 323, 800-803.
- McNamara, D.S., & Scott, J.L. (2001). Working memory capacity and strategy use. *Memory & Cognition*, 29, 10-17
- Melby-Lervag, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, 49(2), 270-291.
- Mickes, L., Wixted, J.T., & Wais, P. (2007). A direct test of the unequal variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, 14, 858-865.
- Milham, M.P., Banich, M.T., & Barad, V. (2003). Competition for priority in processing increases prefrontal cortex's involvement in top-down control: an event-related fMRI study of the stroop task. *Cognitive Brain Research*, 17, 212–222
- Milham, M.P., Banich, M.T., Webb, A., Barad, V., Cohen, N.J., Wszalek, T., & Kramer, A.F. (2001). The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. *Cognitive Brain Research*, 12, 467–473.
- Miller, E.K., & Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167–202.
- Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., & Wager, T.D. (2000). The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: A latent variable analysis, *Cognitive Psychology*, 41, 49–100.
- Monsell, S. (1978). Recency, immediate recognition and reaction time. *Cognitive*

- Psychology*, 10, 465–501.
- Morey, R.D., & Rouder, J.N. (2010). BayesFactorPCL: Computation of Bayes factors for simple psychological designs. R package version 0.8.
- Morrison, A.B., & Chein, J.M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review*, 18, 46-60.
- Nelson, J.K., Reuter-Lorenz, P.A., Sylvester, C.Y., Jonides, J., & Smith, A.D. (2003). Dissociable neural mechanisms underlying response-based and familiarity-based conflict in working memory. *Proceedings of the National Academy of Science*, 100(19), 11171–11175.
- Nigg, J.T. (2006). *What causes ADHD? Understanding what goes wrong and why*. New York: Guilford Press.
- Nilsen, E., & Graham, S. (2009). The relations between children’s communicative perspective-taking and executive functioning. *Cognitive Psychology*, 58, 220-249.
- Norman W. & Shallice, T. (1986). Attention to action. In: Davidson RJ, Schwartz GE, Shapiro D, editors. *Consciousness and self regulation: Advances in research and theory*, vol. 4. New York: Plenum, p. 1–18.
- Novick, J.M., Hussey, E.K., Teubner-Rhodes, S.E., Harbison, J.I., & Bunting, M.R. (2013). Clearing the garden-path: Improving sentence processing through cognitive control training. *Language and Cognitive Processes*. doi: 10.1080/01690965.2012.758297
- Novick, J.M., Kan, I.P., Trueswell, J.C., & Thompson-Schill, S.L. (2009). A case for conflict across multiple domains: Memory and language impairments follow

- damage to ventrolateral prefrontal cortex. *Cognitive Neuropsychology*, 26(6), 527–567.
- Novick, J.M., Thompson-Schill, S.L., & Trueswell, J.C. (2008). Putting lexical constraints in context into the visual world paradigm. *Cognition*, 107, 850-903.
- Novick, J.M., Trueswell, J.C., & Thompson-Schill, S.L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, 5(3), 263-281.
- Novick, J.M., Trueswell, J.C., & Thompson-Schill, S.L. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass*, 4(10), 906-924.
- Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, 134(3), 368-387.
- Oberauer, K., & Kliegl, R. (2004). Simultaneous cognitive operations in working memory after dual-task practice. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4), 689-707.
- Olesen, P., Westerberg, H., & Klingberg, T. (2004). Increased prefrontal and parietal brain activity after training of working memory. *Nature Neuroscience*, 7(1), 75–79.
- Owen, A.M., McMillan, K.M., Laird, A.R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, 25(1), 46-59.
- Owen, A.M., Hampshire, A., Grahn, J.A., Stenton, R., Dajani, S., Burns, A.S., Howard,

- R.J., & Ballard, C.G. (2010). Putting brain training to the test. *Nature*, 465, 775–778.
- Patson, N., Darowski, E., Moon, N., & Ferreira, F. (2009). Lingerings misinterpretations in garden-path sentences: Evidence from a paraphrasing task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1), 280-285.
- Persson, J., Welsh, K., Jonides, J., & Reuter-Lorenz, P. (2007). Cognitive fatigue of executive processes: Interaction between interference-resolution tasks. *Neuropsychologia*, 45, 1571-1579.
- Pessoa, L. (2010). How do emotion and motivation direct executive control? *Trends in Cognitive Science*, 13(4), 160-166.
- Poarch, G.J., & van Hell, J.G. (2012). Executive functions and inhibitory control in multilingual children: Evidence from second-language learners, bilinguals, and trilinguals. *Journal of Experimental Child Psychology*, 113(4), 535-551.
- Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology*, 57, 137–146.
- Postle, B. R. (2003). Context in verbal short-term memory. *Memory & Cognition*, 31, 1198–1207.
- Postle, B.R., Berger, J.S., Goldstein, J.H., Curtis, C.E., & D’Esposito, M. (2001). Behavioral and neurophysiological correlates of episodic coding, proactive interference, and list length effects in a running span verbal working memory task. *Cognitive, Affective, and Behavioral Neuroscience*, 1, 10-21.



- Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59, 413-425.
- Rayner K., Kambe G., & Duffy S.A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology*, 53(4), 1061–1080.
- Redick, T.S., Shipstead, Z., Harrison, T.L., Hicks, K.L., Fried, D.E., Hambrick, D.Z., Kane, M.J., & Engle, R.W. (2012). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*. Advance online publication. doi: 10.1037/a0029082
- Richmond, L., Morrison, A., Chein, J., & Olson, I. (2011). Working memory training and transfer in older adults. *Psychology and Aging*, 26(4), 813-822.
- Robinson, G., Blair, J., & Cipolotti, L. (1998). Dynamic aphasia: An inability to select between competing verbal responses? *Brain*, 121, 77–89.
- Robinson, G., Shallice, T., & Cipolotti, L. (2005). A failure of high level verbal response selection in progressive dynamic aphasia. *Cognitive Neuropsychology*, 22(6), 661–94.
- Rodríguez-Ferreiro, J., Gennari, S.P., Davies, R., & Cuetos, F. (2011). Neural correlates of abstract verb processing. *Journal of Cognition Neuroscience*, 23(1), 106-118.
- Rogalsky, C., Matchin, W., & Hickock G. (2008). Broca's area, sentence comprehension, and working memory: an fMRI study. *Frontiers in Human Neuroscience*, 2: 14. doi: 10.3389/neuro.09.014.2008

- Rouder, J.N., Morey, R., Speckman, P.L., & Province, J.M. (in press). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*.
- Rouder, J.N., Speckman, P., Sun, D., Morey, R., & Iverson, G.J. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225-237.
- Sayala, S., Sala, J.B., & Courtney, S.M. (2006). Increased neural efficiency with repeated performance of a working memory task is information-type dependent. *Cerebral Cortex*, 16, 609–617.
- Schweizer, T.A., Ware, J., Fischer, C.E., Craik, F.I., & Bialystok, E. (2011). Bilingualism as a contributor to cognitive reserve: Evidence from brain atrophy in Alzheimer's disease. *Cortex*, 12(3), 8-15.
- Scarborough, D.L., Cortese, C., & Scarborough, H.S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, 3(1), 1-17.
- Schnur, T.T., Schwartz, M.F., Kimberg, D.Y., Hirshorn, E., Coslett, H.B., & Thompson-Schill, S.L. (2009). Localizing interference during naming: Convergent neuroimaging and neuropsychological evidence for the function of Broca's area. *Proceedings of the National Academy of Sciences*, 106(1). 322–7.
- Shipstead, Z., Redick, T.S., & Engle, R.W. (2010). Does working memory training generalize? *Psychologica Belgica*, 50, 245–276.
- Shipstead, Z., Redick, T.S., & Engle, R.W. (in press). Is working memory training effective? *Psychological Bulletin*.
- Smith, E.E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes.

*Science*, 283, 1657-1661.

Snyder, H.R., Banich, M.T., & Munakata, T. (2011). Choosing our words: Retrieval and selection processes recruit shared neural substrates in left ventrolateral prefrontal cortex. *Journal of Cognitive Neuroscience*, 23(11), 3470-3482.

Snyder, H.R., Hutchison, N., Nyhus, E., Curran, T., Banich, M.T., O'Reilly, R.C., & Munakata, Y. (2011). Neural inhibition enables selection during language processing. *Proceedings of the National Academy of Sciences*, 107(38), 16483–16488.

Snyder H.R., & Munakata Y. (2008). So many options, so little time: The roles of association and competition in underdetermined responding. *Psychonomic Bulletin and Review*, 15, 1083–1088.

Spivey, M.J., Tanenhaus, M.K., Eberhard, K.M., & Sedivy, J.C. (2002) Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45, 447-481

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In: G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 327–342). Oxford, UK: Oxford University Press.

Streiner, D.L., & Norman, G.R. (2006). Measuring change. *How measurement scales: A practical guide to their development and use. Fourth Edition.* (pp. 194–212). Oxford: Oxford University Press.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, 30, 276-315.

Sternberg, R.J., Ketron, J.L., & Powell, J.S. (1982). Componential approaches to the

- training of intelligent performance. In D. K. Detterman & R. J. Sternberg (Eds.), *How and how much can intelligence be increased?* (pp. 155-172). Norwood, NJ: Ablex.
- Stewart, A. J., Pickering, M. J., & Sturt, P. (2004). Using eye movements during reading as an implicit measure of the acceptability of brand extensions. *Applied Cognitive Psychology*, 18, 697-709.
- Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 28, 643-662.
- Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition*, 105:477–488.
- Sturt, P., Scheepers, C., & Pickering, M. J. (2002). Ambiguity resolution after initial misanalysis: The role of recency. *Journal of Memory and Language*, 46, 371-390.
- Takeuchi, H., Sekiguchi, A., Taki, Y., Yokoyama, S., Yomogida, Y., Komuro, N., Yamanouchi, T., Suzuki, S., & Kawashima, R. (2010). Training of working memory impacts structural connectivity. *The Journal of Neuroscience*, 30(9), 3297–3303.
- Tanenhaus, M.K. (2007). Eye movements and spoken language processing. In R.P.G van Gompel, M.H. Fischer, W.S., Murray, & R.L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 309–326). Oxford: Elsevier.
- Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

- Thompson-Schill, S.L., Bedny, M., & Goldberg, R.F. (2005). The frontal lobes and the regulation of mental activity. *Current Opinion in Neurobiology*, 15, 219–224.
- Thompson-Schill, S.L., Jonides, J., Marshuetz, C., Smith, E.E., D'Esposito, M., Kan, I.P., Knight, R.T., & Swick, D. (2002). Effects of frontal lobe damage on interference effects in working memory. *Journal of Cognitive, Affective & Behavioral Neuroscience*, 2, 109-120.
- Thompson-Schill, S.L. Ramscar, M., & Chrysikou, E.G. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science*, 18(3), 259–263.
- Thompson-Schill, S.L., Swick, D., Farah, M.J., D'Esposito, M., Kan, I.P., & Knight, R.T. (1998). Verb generation in patients with focal frontal regions: A neuropsychological test of neuroimaging findings. *Proceedings of the National Academy of Science*, 95, 15855-15860.
- Thothathiri, M., Kim, A., Trueswell, J.C., & Thompson-Schill, S.L. (2012). Parametric effects of syntactic-semantic conflict in Broca's area during sentence processing. *Brain and Language*, 120(3), 259-264.
- Tidwell, J., Chrabaszcz, J., Thomas, R., Mendoza, J., & Dougherty, M. (2013). Theoretical blinders and the illusion of significance. *Manuscript in Preparation*.
- Tropper, B. (2009). *An electrophysiological and behavioral examination of cognitive control in children with specific language impairment*. Unpublished doctoral dissertation. City University of New York.
- 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, 89–

134.

- Trueswell, J. C. & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives in sentence processing* (pp. 155–179). Hillsdale, NJ: Lawrence Erlbaum Assoc.
- Trueswell, J.C., Tanenhaus, M.K., & Garnsey, S.M. (1994). Semantic influences on parsing: use of thematic role information in syntactic disambiguation. *Journal of Memory and Language*, 33, 285–318.
- Van der Linden, D., Frese, M., & Meijman, T.F. (2003). Mental fatigue and the control of cognitive processes: Effects on preservation and planning. *Acta Psychologica*, 113, 45–65.
- Van Dyke, J.A. & McElree, B. (2006). Retrieval interference in sentence processing. *Journal of Memory and Language*, 55(2), 157-166.
- Vuong, L.C., & Martin, R.C. (2011). LIFG-based attentional control and the resolution of lexical ambiguities in sentence context. *Brain & Language*, 116, 22–32.
- Wagner, A.D., Paré-Blagoev, E.J., Clark, J., & Poldrack, R.A. (2001). Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. *Neuron*, 31, 329–338.
- Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, 85, 79–112.
- Warren, T., White, S.J., & Reichle, E.D., (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z Reader. *Cognition*, 11(1), 132-137.

- Waters, G., Caplan, D., & Hildebrandt, N. (1987). Working memory and written sentence comprehension. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading*. Hillsdale, NJ: Erlbaum.
- Weighall, A.R. (2008). The kindergarten path effect revisited: Children's use of context in processing structural ambiguities. *Journal of Experimental Child Psychology*, 99, 75–95.
- Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, 58, 250–271.
- West, R. (2004). The effects of aging on controlled attention and conflict processing in the Stroop task. *Journal of Cognitive Neuroscience*, 16(1), 103-113.
- Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Ostensson, M.L., Bartfai, A., & Klingberg, T. (2007). Computerized working memory training after stroke - a pilot study. *Brain Injury*, 21(1), 21–29.
- Wickens, T.D. (2001). *Elementary signal detection theory*. New York: Oxford University Press.
- Xiang, J.Z., & Brown, M.W. (1998). Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology*, 37(4), 657-676.
- Ye, Z., & Zhou, X. (2009). Conflict control during sentence comprehension: fMRI evidence. *Neuroimage*, 48, 280–290.