Technical Details

Working Memory Training for Improving Cognition and Language
Impact and lessons learned from military and non-military populations

Alan E. Mishler, Polly L. O’Rourke, DeLaura M. Jansen, Patrick J. Cushen, Nina S. Hsu, J. Isaiah Harbison, Jared M. Novick, and Michael F. Bunting

How are two very different tasks like remembering where your car is parked and understanding a sentence related? To remember where your car is parked (assuming you park in a different spot every day), you have to suppress irrelevant memories of previous parking spots and retrieve the appropriate memory of your current spot. To understand a sentence, you have to suppress incorrect alternative interpretations of words and phrases to construct the appropriate meaning of the sentence. For example, to understand the sentence “John went to the bank, but he realized that he had forgotten his fishing pole,” you have to conclude that “he” refers to John, as opposed to other animate male entities that may be elsewhere in the text or in your environment; and you probably have to revise your initial interpretation of “bank” in order to conclude that it refers to the bank of a river rather than to a financial institution. If this sentence were in a foreign language, you would also have to suppress the words and grammar of your native language in order to access the appropriate linguistic representations.

Both these tasks are representative of countless tasks throughout the day that require you to focus your attention on the information at hand while suppressing irrelevant information that might lead you astray. As discussed in much greater detail in Chapter 1 of this report, researchers at the University of Maryland Center for Advanced Study of Language CASL have been accumulating evidence to suggest that attention, memory, and language functions share important characteristics and are enabled by common brain systems (e.g., Novick, Hussey, Teubner-Rhodes, Harbison, & Bunting, 2013; Sprenger et al., 2013; but see Chapter 1). We are using this evidence to inform ways to target multiple abilities—attention, memory, and language abilities—for improvement through a “cognitive exercise regimen.” Just as doing biceps curls in the gym might help you throw a ball farther (because ball throwing is supported by arm muscles), training certain memory functions can similarly help language performance and language learning, because they rely on common underlying support systems.

Working memory refers to the cognitive system responsible for the control, regulation, and active maintenance of information in the face of distracting information, including irrelevant stimuli from the environment and one’s own intrusive thoughts (cf. Conway, Jarrold, Kane, Miyake, & Towse, 2007). According to Cowan’s (2005) highly influential model of working memory and attention, there are three essential components to the working memory system: the focus of attention, short-term memory, and the cognitive control center.
The focus of attention is a zone of privileged and immediate access to information. Memory in the focus of attention is readily available and easily retrieved, but it is capacity-limited to a fixed number of items, or chunks (see also Haarmann & Usher, 2001). Attention is required in order to maintain information in focus.

Short-term memory (STM) is recent memory that is no longer inside the focus of attention. It includes traces of all recent sensory experiences as well as recollections retrieved from long-term memory. In principle, short-term memory is not capacity-limited; however, the contents are susceptible to interference (or confusion) and rapid decay unless attention is deployed to move contents out of STM and back into focus.

The third component of working memory, the cognitive control center (also frequently referred to as the central executive), includes those processes that focus, switch, block, and divide attention. Information can enter the working memory system from sensation and perception of the environment and also via retrieval from long-term memory, which is the set of memories for all previous experiences as well as acquired knowledge and skills. (Long-term memory is also not capacity-limited in principle, but certain memories may not be readily retrievable at any given point in time.) The cognitive control system serves to direct and shape the flow of information, helping to determine what information enters the working memory system and what gets maintained in the focus of attention and in short-term memory.¹

What processes does the cognitive control center, or central executive, comprise? Although a number of processes are widely recognized and studied, the concept of cognitive control (or executive functioning) is amorphous, and the nature and number of proposed functions varies on a study-by-study basis (Shipstead, Harrison et al., 2013; Shipstead, Trani et al., 2013). The literature also has yet to settle on a singular name for these functions, which are referred to variously as cognitive control processes, executive function processes, and executive attentional control processes, just to name a few (e.g., see the variety of uses in Miyake & Shah, 1999). Among the more influential frameworks for representing these processes is that of Miyake and colleagues (Friedman et al., 2006; Miyake et al., 2000; Miyake & Friedman, 2012), in which complex cognition is largely explained via three processes: memory updating, inhibition, and shifting. However, many other executive functions have been described prominently in the literature, and the field seems far from settling this question.

Given the diversity of proposed executive functions, we do not aim to exhaustively characterize the working memory system. Instead, we focus in this report on a single aspect of this system—namely, the ability to regulate mental activity to override biases and recharacterize input in the face of conflict. We refer to this ability in the rest of this report as cognitive control. This definition closely matches Miyake et al.’s (2000) definition of inhibition, and this capacity is recognized as essential for efficient working memory processing. We focus on cognitive control because we believe that it is also critical for a variety of tasks in the language domain, a view supported by Shipstead, Trani et al. (2013). For example, in second language learning, the ability to override initial first language lexical and syntactic interpretations is thought to require cognitive control. Cognitive control also underpins the ability to understand temporarily ambiguous sentences (O’Rourke, 2013).

Our goal, therefore, has been to determine if strengthening the cognitive control ability of the working memory system through training will improve language processing and, by extension, second language learning. In three chapters, we present and discuss CASL research showing that cognitive control is a domain-general ability that is important to many essential memory and language functions. Our research also shows that cognitive control training, administered under the proper conditions, can improve an individual’s ability to regulate attention and control their thoughts and actions across a variety of tasks, including language comprehension tasks.

In Chapter 1, Novick et al. summarize CASL research conducted at the University of Maryland in a controlled laboratory setting that shows that cognitive control training leads to improvements in untrained assessments of working memory and analyst-relevant language production and comprehension tasks. As predicted, these training effects were selective: improvements were only observed in conditions involving cognitive conflict, and people who were in a control condition and trained on low-conflict tasks did not demonstrate these improvements.

¹ A more thorough review of the working memory system is beyond the scope of this report, but the interested reader is referred to Cowan (1995, 2005), for a detailed description of his theory, or Miyake and Shah (1999), for a review of other models of working memory.
In Chapter 2, Mishler et al. report on CASL’s attempt to replicate these robust effects with U.S. government participants in their workplace using online versions of the training tasks hosted on a CASL server. Difficulties recruiting participants, high attrition rates, and low participant engagement with the training tasks render it impossible to draw meaningful conclusions about the efficacy of working memory training from this dataset. However, after gathering extensive feedback from participants who completed the study, as well as those who withdrew, CASL has undertaken a series of changes to the online tasks and the user interface which are designed to make the tasks more interesting, to increase participant motivation, and to minimize the likelihood of attrition. Once these changes are complete, CASL plans to conduct a study to validate the efficacy of the online versions of the training tasks. Chapter 2 also presents additional evidence from a large meta-analysis that improved cognitive control has a large and robust positive effect on foreign language learning and proficiency.

Finally, in Chapter 3, O’Rourke, Novick, Hsu, and Jaeggi describe current and future CASL research using both functional magnetic resonance imaging (fMRI) and encephalographic (EEG) recording to investigate the neural underpinnings of working memory training. Preliminary results support the hypothesis that overlapping brain areas are involved in a variety of both linguistic and non-linguistic cognitive control tasks, thus providing converging evidence for a domain-general cognitive control in both function and brain anatomy that has implications for the design of future training protocols.

REFERENCES


Memory and Language Improvements Following Working Memory (Cognitive Control) Training

A summary of CASL findings at the University of Maryland

Jared M. Novick, Erika Hussey, J. Isaiah Harbison, Susan Teubner-Rhodes, Alan Mishler, Nina Hsu, and Kayla Velnoskey

While observing and interacting with the environment, individuals periodically confront situations that require them to override the first or dominant cognitive reaction that comes to mind. For instance, upon recognizing a colleague in an unfamiliar setting (e.g., the grocery store), one might have to rein in a tendency to discuss work-related issues and instead engage in more context-appropriate behavior. Similarly, readers and listeners must revise an initial misinterpretation of a word or phrase that has more than one meaning, when additional input or contextual evidence disambiguates it (recall this example from the introduction: “At the bank, John got out his fishing pole and cast his line.”). On the surface, these circumstances are very different. The first entails politeness or social norms; the second involves word and sentence comprehension, where the reader must abandon the initial, more frequent interpretation of “bank” (financial institution) and recover the less common but context-appropriate meaning (river’s edge) instead. Despite obvious dissimilarities, however, both scenarios illustrate cases in which individuals must consult goal-relevant rules or new knowledge to countermand familiar or automatic responses, in the service of resolving among competing evidential sources in a highly variable environment. This regulation of thoughts and actions is known as cognitive control, a term that describes a constellation of mental functions that guide goal-directed behavior consistent with situation-specific requirements. As such, cognitive control is particularly important for performing non-routine tasks in which new information or plans of action conflict with habitual procedures. One of CASL’s core research efforts has been to determine whether focused practice engaging cognitive control mechanisms improves memory and language behavior specifically under strong demands for resolving information-conflict. In what follows, we summarize the findings and sketch some implications that are central to CASL’s mission.

BACKGROUND AND MOTIVATION

In the literature on cognitive control, conflict designates instances where an individual receives contradictory information either about how to characterize or represent some input, or how best to act on that input. Across various studies, researchers have investigated the conditions that induce conflict and the mental functions that then resolve it at both the representational and response levels. Here, we will focus almost exclusively on two types of representational conflict that can arise in memory and language. One type is
prepotent representational conflict, sketched in the above examples, in which a dominant (prepotent) way of characterizing a stimulus must be overridden. The canonical laboratory task that measures this type of conflict is the Stroop task, where subjects must name the ink color of printed color-terms. On trials that generate conflict, the word stimulus gives rise to an automatic reading response, which must be adjusted in favor of the perceptual response, thus triggering cognitive control (e.g., the word GREEN printed in blue font). Prepotent representational conflict has been differentiated from underdetermined representational conflict, which develops from competition among multiple representations of a stimulus, all of which may be uniformly acceptable candidates for selection (i.e., the stimulus itself does not create a dominant response; Botvinick et al., 2001, 2004). One well-known task that indexes this sort of conflict is a verb generation task, an important measure of language fluency, in which subjects produce a verb that is related to a given noun (Barch, Sabb, Braver, & Noll, 2000; Petersen, Fox, Posner, Mintun, & Raichle, 1988; Thompson-Schill et al., 1997). In one condition, the presented nouns are those with a clear verb associate, and thus low cognitive control demands (e.g., kite \(\rightarrow\) fly). These trial types are contrasted with a condition in which the nouns have many related verbs without an obvious or dominant one, and thus high cognitive control demands (e.g., ball \(\rightarrow\) kick, throw, toss, catch; see Figure 5). Production times are reliably longer in the high- versus low-competitor condition.

The resolution of both types of conflict has been associated with a common control function in memory and language tasks alike (Kan & Thompson-Schill, 2004; Novick, Trueswell, & Thompson-Schill, 2005), and has been linked to neuroanatomical regions within the lateral prefrontal cortex (PFC) (Botvinick et al., 2001; Fletcher & Henson, 2001; Miller & Cohen, 2001; Thompson-Schill, Bedny, & Goldberg, 2005). For example, when individuals complete Stroop conflict trials and a language processing task that requires syntactic ambiguity resolution, there is co-localized brain activity within lateral prefrontal areas (January, Trueswell, & Thompson-Schill, 2009; Ye & Zhou, 2009). The linking assumption here is that syntactic ambiguity is a type of prepotent representational conflict: in the natural course of language comprehension, individuals must occasionally rely on cognitive control functions to revise misinterpretations due to initially erroneous characterizations of the input (Novick et al., 2005; see Box 1). In the memory domain, the same cortical regions are recruited in a wide variety of item-recognition tasks where people encounter recent stimuli that are not part of the current memory set, thus creating conflict between highly familiar but non-target memoranda (Jonides & Nee, 2006; Gray et al., 2003; Oberauer, 2005). Finally, patients who suffer focal damage to these neurobiological structures demonstrate specific conflict-resolution deficits within memory (Hamilton & Martin, 2005; Thompson-Schill et al., 1998), which are coupled behaviorally with language processing impairments under high cognitive control demands, including failure to generate a word from many versus few competitors and an inability to recover from misanalysis during comprehension (Novick, Kan, Trueswell, & Thompson-Schill, 2009, 2010; see also Robinson et al., 1998, 2005, 2010; Vuong & Martin, 2011). Together, these findings suggest a shared cognitive control system that supports the ability to resolve information-conflict within memory, revise default interpretations in language comprehension, and produce a single word from among multiple competing alternatives—again, a key gauge of fluency.

CASL research has examined the range of memory and language conditions that create both underdetermined and prepotent representational conflict. In particular, CASL has tested the extent to which focused practice engaging cognitive control during a short-term training paradigm might fine-tune behavior across domains, reflected in performance improvements specifically under high conflict-resolution demands. Importantly, the assessment and training tasks that have been administered in this work are well studied tasks in the memory and psycholinguistic literatures and have all been associated with conflict-control. In this way, CASL has been able to address the question of whether conflict-control mechanisms per se can be sharpened through training. Specifically, to what extent is cognitive control reflected in cross-domain transfer, such that performance increases in memory generalize to novel language and memory tasks under conditions of information-conflict? The overall goal has been to test if short-term training that is expressly designed to target people’s ability to resolve conflict improves cognitive control functioning generally, particularly on unpracticed tasks (i.e., a “transfer effect”). Significant performance improvements in transfer assessments would (a) add to mounting evidence for a general-purpose conflict-control system, (b) shed light on how cognitive control abilities outside the syntactic domain interact with and perhaps even shape language processing in both production and comprehension, (c) suggest that this domain-general system is to some extent plastic (see Hussey...
& Novick, 2012), and (d) suggest that training this domain-general system improves a number of analyst-
relevant functions that rely on it such as conflict resolution in language and memory tasks.

Crucially, language and intelligence analysts have confirmed that they are frequently faced with ambiguous
text or speech, as well as with discrepant sources of input that give rise to conflicts in memory between familiar
but irrelevant information (Jamie Hester, personal communication). Thus, they need good cognitive control to
do their jobs well. CASL’s training research therefore has been aimed at determining whether conflict resolution
abilities are malleable; if so, positive effects of this work could open the door to exploring whether job
performance might be improved in high-stakes circumstances where information-conflict must be resolved.
Namely, improved cognitive control may decrease the likelihood that ambiguous input will result in a
comprehension error, and that familiar but irrelevant information will be categorized incorrectly as relevant.

<table>
<thead>
<tr>
<th>Box 1.</th>
<th>Consider examples 1 and 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. While the thief hid the jewelry that was elegant and expensive sparkled brightly.</td>
<td></td>
</tr>
<tr>
<td>(Temporarily Ambiguous)</td>
<td></td>
</tr>
<tr>
<td>2. The jewelry that was elegant and expensive sparkled brightly while the thief hid.</td>
<td></td>
</tr>
<tr>
<td>(Unambiguous)</td>
<td></td>
</tr>
</tbody>
</table>

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. The presence of a plausible object (“the jewelry”) further supports the transitive
interpretation. 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 one’s developing interpretation (“. . . sparkled brightly”),
readers must initiate cognitive control processes to re-characterize their initial representation of
sentence meaning, i.e., to resolve the conflict and revise their misinterpretation. In (2), the reversed
clause order unambiguously signals the reflexive analysis; consequently, reinterpretation is
unnecessary and conflict-resolution and cognitive control processes need not deploy.

At the pretest and posttest assessments in the CASL studies, participants read sentences such as
these while their eye-movements were recorded, and 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. To correctly respond “yes”,
readers must override the initial, incorrect transitive analysis.

SUMMARY OF RESULTS FROM TWO CASL STUDIES

Study 1. CASL’s first study revealed tantalizing evidence for domain-general plasticity in cognitive control.
As outlined above, recent research emphasizes that one important cognitive control function in the domain of
language is to override early parsing decisions as new evidence conflicts with a developing syntactic
interpretation. CASL tested if training on a non-syntactic conflict-control task in the memory domain improved
readers’ ability to recover from misanalysis during language processing. Participants completed pre/post-reading
assessments containing temporarily ambiguous sentences susceptible to misinterpretation (Box 1), and trainees
completed a battery of working memory training tasks (for 20 hours over several weeks), only one of which
tapped conflict-control functions supported by lateral PFC (the N-back task with lures; see Figure 4).
Importantly, none of the training tasks involved reading sentence material of any kind.
**Figure 1.** Example 4-back condition in the N-back training task. During N-back, subjects viewed single letters sequentially, indicating whether items appeared n trials previously. N-level increased adaptively throughout the regimen according to individual increases in performance; also, at each N-level, when trainees reached a certain performance criterion they encountered recent items appearing in non-n positions (lures), thus honing conflict-control between highly familiar but non-target stimuli. In this 4-back example, the second R stimulus is a lure because it appeared previously in a 3-back, not a 4-back, position.

Performance increases throughout the regimen on only the N-back-with-lures training task, the sole task targeting conflict-resolution processes, predicted improvements on two assessments of interpretation recovery. N-back responders (N = 13)—those demonstrating reliable training gains (see Figure 2)—significantly increased their comprehension accuracy across assessments (Figure 3). Their posttest eye-movement patterns also revealed significantly improved real-time revision following entry into sentence regions where cognitive control is hypothesized to engage (i.e., when information-conflict is encountered; Figure 4). That is, upon discovering material that signaled an incompatibility with their developing interpretation, those who performed well on the N-back task were reliably faster to revise their misanalysis on the posttest assessment. Untrained participants (N = 22) and N-back non-responders (N = 7) showed no performance changes. These results provided critical insight into how non-syntactic functions contribute to parsing and interpretation, and suggest that certain language skills are amenable to improvement via domain-general cognitive control training. Thus, in healthy adults, conflict resolution training within memory transfers performance benefits to sentence processing under select conditions—that is, when readers must revise once they have detected a misanalysis (Novick, Hussey, Teubner-Rhodes, Harbison, & Bunting, 2013).
Figure 2. $N$-back performance curves by training session for responders and non-responders.\(^2\)

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). Following each ambiguous sentence that subjects read, they were asked a question that probed for lingering effects of misinterpretation. The large positive difference for $N$-back responders (A) reflects that this subgroup had significantly better accuracy at Assessment 2 than Assessment 1 for ambiguous items only, namely those that require cognitive control. This increase was significantly different from untrained subjects’ and non-responders’ performance changes. No changes were observed for any group on unambiguous items (B) as expected, as

these sentences did not require ambiguity resolution and thus cognitive control. Training therefore conferred selective benefits in comprehension contexts that involved conflict-control.³

Figure 4. Trainees’ and untrained controls’ regression-path times across assessments launched from each sentence region for ambiguous (top panel) and unambiguous (bottom panel) items. The regression-path measure reflects the total time it takes a reader to go past a particular region of a sentence (the regions of interest are indicated on the x-axis; reading time is plotted on the y-axis). The conflict region (e.g., “sparkled brightly”; circled in red) of ambiguous sentences was the only region where N-back responders (top panel; leftmost graph) spent reliably less time “looking back” to earlier material in the sentence at Assessment 2 before moving on. Reading time analyses were conducted on correctly answered items only; therefore, when responders are getting it right at posttest, they are doing so faster than in Assessment 1. N-back non-responders and untrained participants demonstrated no significant change across assessments for ambiguous items in any region. Crucially, for unambiguous items, N-back responders, non-responders, and untrained controls demonstrated no change in reading time from any region between assessments as expected, since these items contained no information-conflict and thus no need for cognitive control. As with the accuracy data in Figure 3, training conferred selective benefits.⁴

However, subjects in CASL’s first study completed a battery of working memory tasks during training, and trainees were compared to a no-contact control group. The findings clearly indicated that performance improvements on only the conflict-control training task (N-back with lures) significantly increased readers’ ability to revise early misinterpretations in real time, following the detection of syntactic conflict. In contrast, performance increases on the non-conflict tasks (i.e., working memory tasks in which conflict demands were not parametrically manipulated) throughout the memory-training regimen did not generalize to garden-path recovery in any way. Although highly suggestive, it remains possible that the other training tasks were necessary as part of a combined suite to confer transfer.


Study 2. In the second study, CASL was explicitly interested in the malleability of conflict-control procedures across domains and thus aimed to extend earlier findings in important ways. First, CASL included a larger set of transfer tasks, this time both memory and language, all of which have been previously established to rely on cognitive control to resolve information-conflict under some conditions but not others (see Background and Motivation, and Figure 5). These two design components—more tasks measuring the same construct and within-task manipulations of conflict-control—offer important advantages: CASL could test the extent of transfer to new tasks as well as the specificity of these effects to the conditions that share conflict-processing demands (Li et al., 2008; Shipstead, Redick, & Engle, 2010). Second, CASL employed active control groups that underwent a training regimen involving the same task (N-back), but with critical features of the task removed (e.g., the presence of lures in the High-Conflict training group but the absence of lures in the No-Conflict training group). Eighty-one native English-speaking adults were randomly assigned to one of three N-back training groups: High-Conflict, No-Conflict, or static 3-Back. During N-back, as in Study 1, subjects viewed single letters sequentially, indicating whether items appeared n trials previously (Figure 1). N-level increased adaptively in the High- and No-Conflict groups. Only High-Conflict trainees encountered recent items appearing in non-n positions (lures), honing conflict-control between highly familiar but non-target stimuli. No-Conflict and 3-Back trainees never encountered lures, and therefore did not train cognitive control. Each training group practiced their N-back training regimen for just eight total hours (30 minutes per day over 3-5 weeks), a much shorter duration than in Study 1.
Figure 5. Example transfer tasks in CASL’s second experiment. Subjects in Study 2 completed three assessments. In addition to the same ambiguity resolution task that subjects in Study 1 completed (not pictured), they also completed the verb generation task (A) and a recognition-memory task (B). In the verb generation task, a critical measure of language fluency, subjects were given a noun (e.g., scissors) and were instructed to produce an associated verb. This task involves both high- and low-conflict conditions; in some cases, a noun (e.g., ball) has many verbs associated with it and subjects must resolve the conflict among the competing options using cognitive control (see Background and Motivation in main text; the dotted arrows and greyed out responses indicate weaker associations, or greater competition). In the recognition-memory task, subjects were given a set of items in three different spatial positions, one at a time, to encode in memory (e.g., box, foot, carrot in left, right, or center positions on a horizontal axis). After a brief delay (500 milliseconds), they were given probes and asked to indicate whether they appeared in the memory set. In the low-conflict condition (left panel), any repetition of a word would be a target; subjects could therefore ignore the position of where the item occurred in the memory list and use familiarity alone to respond correctly (foot is a target; trunk is not because it did not previously appear). By contrast, in the high-conflict condition (right panel), a probe was a target only if its position during the recognition test matched the position of its original occurrence in the memory list. In this example, the probe foot in the leftmost position would be a non-target lure even though it was presented during the study phase (where it was presented in the rightmost position). Thus, subjects had to override a familiarity bias and use cognitive control to resolve conflicting information, i.e. recognize familiar but irrelevant information as non-targets. Response times and error rates rise significantly in this condition.

Employing these alternative intervention paradigms allowed a focused evaluation of whether practicing cognitive control specifically modified processing under high cognitive control demands regardless of task or domain, namely in comprehension, production, and within memory despite conflicting memoranda. A broad convergence of selective transfer effects would strongly indicate common conflict-control procedures and, importantly, plasticity of these procedures.
All trainees improved on N-back with training; however, significant group-by-assessment interactions were observed such that only High-Conflict trainees improved in high-conflict assessment conditions. At posttest (Assessment 2), those who practiced N-back with lures: (1) spent reliably less time re-reading earlier regions of ambiguous sentences after encountering conflicting evidence (‘sparkled brightly’ as in Study 1); (2) demonstrated significantly faster verb-production times for nouns with many associates; and (3) improved recognition memory when familiar stimuli interfered with target memoranda (the high-conflict condition in Figure 5). No benefits were found for low-conflict conditions, as expected, since these conditions did not require cognitive control. The benefits conferred by high-conflict training were therefore selective, resulting in performance increases only when information-conflict had to be resolved, be it in the language or memory domain. Crucially, the 3-back and No-Conflict training groups did not demonstrate any of these effects.

DISCUSSION AND IMPLICATIONS

The combined findings from Studies 1 and 2 suggest that cognitive control abilities may be plastic and that training can positively influence language production and comprehension as well as memory abilities, when individuals must resolve information-conflict. Moreover, Study 2 in particular demonstrated that the benefits can be obtained with relatively shorter training (20 hours versus 8 hours). These are critical results as language is rife with ambiguity, and analysts often encounter multiple sources of evidence that they must rapidly coordinate and interpret, sources that regularly point to conflicting memories or solutions to a task.

Crucially, the training and assessment tasks in the research described herein are ostensibly different; therefore any observation of transfer cannot be ascribed to changes in how subjects represent particular information in the long term. That is to say, the task goals and stimulus characteristics are, on the surface, very unalike—e.g., verb generation, sentence processing and comprehension, recognition memory—with the key overlapping feature being the need to resolve conflict in select conditions. As such, effects of generalized performance to certain conditions involving high cognitive control demands suggest behavior-specific tuning (reflected in an enhanced ability to deal with conflict), rather than task-specific strategies or changes in the nature of the conflict experienced in each task. For example, because subjects in the experimental group in Study 2 practiced N-back with lures during the regimen—that is, managing conflict within memory—and not verb generation, improved production times in high-conflict conditions could not be due to a change in noun representations and their competitive interaction since subjects’ experience with this task per se had not been altered. Instead, the common conflict-control procedure across tasks despite superficial differences is a critical component of the present investigations, enabling a true examination of the malleability of this procedure without task-similarity confounds.

In sum, cognitive control may be a malleable skill that individuals (including analysts) can improve with practice. The benefits generalize to a range of analyst-relevant memory and language tasks where information-conflict must be resolved. Critically, one take-home message of the CASL findings is that in order for transfer effects to be observed, the training and assessment tasks must share behavioral procedures that engage the targeted cognitive component functions (here, conflict-control) and perhaps their neurobiological substrates (all tasks studied here are known to be subserved by common regions within lateral PFC; see also current and future directions below). A second is that all subjects in the two studies committed to the training, completing the prescribed sessions dutifully and consistently in the lab. Without these important design features of the experiments and the faithful treatment of the training protocol by the subjects, it is not clear that such wide-ranging effects could be expected.

REFERENCES


Chapter 2

Working Memory Training Applied to Classroom and Workplace

Working memory training with U.S. government participants

Alan E. Mishler, DeLaura M. Jansen, Patrick J. Cushen, Polly O’Rourke, J. Isaiah Harbison, and Michael F. Bunting

This study examines the impact of cognitive control-based working memory training on measures of cognitive control capacity and linguistic ambiguity resolution. The specific purpose of the current study was to evaluate the effectiveness of CASL’s web-based working memory training program for employees in the U.S. government.

RESEARCH GOAL AND RELEVANCE

In the controlled laboratory studies described in Chapter 1 of this report, CASL researchers found evidence that working memory training can improve cognitive control and boost performance on non-linguistic tasks as well as analyst-relevant linguistic tasks. The goal of the research described in this chapter was to replicate these results in a different population and under less rigid and more real-world conditions; namely, with U.S. government foreign language professionals and language learners trying working memory training in the workplace or in a language learning classroom.

STUDY DESIGN

This study utilized a design similar to the laboratory studies described in the first section of this report. The study consisted of three phases: pre-test, training, and post-test. In the pre-test phase, participants completed a battery of five cognitive assessments designed to measure a variety of working memory abilities, including cognitive control and the ability to resolve linguistic ambiguities (See Figure 6). The pre-test was administered in person by CASL researchers. Following the pre-test, participants were randomly assigned to either the active training condition or the control condition. Participants in the active training condition completed a training regimen designed to strengthen cognitive control (as defined in the introduction to this report), while participants in the control condition completed a regimen designed to strengthen vigilance, an attention-related ability that is not related to cognitive control (see Table 1 and Table 2). Each participant was given login credentials to access a set of working memory training tasks hosted on CASL servers, with instructions to complete 8-10 hours of performance-adaptive cognitive training in 30-minute sessions spread across 4-5 weeks. After this training was completed, participants completed an in-person post-test assessment consisting of...
different versions of the tasks they saw at pre-test. The assessment and training tasks are described in the subsequent section.

Given the robust effects of training that we observed in the studies described in Chapter 1, we hypothesized that participants in the active training condition would show greater improvements in cognitive control following training than participants in the control condition. We expected this effect to emerge in all the assessment tasks, since there is an extensive literature showing that performance on all five tasks requires cognitive control. Three of the assessment tasks—Garden Path, N-back, and Stroop—contain both “conflict” trials that require cognitive control as well as “non-conflict” trials that do not engage cognitive control to the same degree. In these tasks, we predicted that the training effect would be selective; that is, we expected the effect to emerge only on conflict trials. The two remaining assessment tasks—Operation Span and Symmetry Span—do not have conflict versus non-conflict trials per se, but aggregate performance on these tasks serves as a measure of cognitive control, as explained in the descriptions of the tasks below. On these tasks, we expected participants in the active training group to show greater overall improvements in performance following training than participants in the control group.

Again, this is consistent with our previous results (cf. Novick et al., 2013; and Experiment 1 in Sprenger et al., 2013): our goal in this study was to replicate those results with U.S. government participants and under less constrained, more real-world conditions than in the laboratory.

Participants

The study was administered at five locations, including four U.S. government locations and CASL. Participants were foreign language professionals, primarily language analysts, and all were employees of the U.S. government. Approximately 20 participants were provided the opportunity to obtain workplace learning credit for their participation. The remaining participants were provided no compensation. A total of 110 participants completed the study.

Assessment Tasks

All participants in both conditions completed the same five tasks at both pre-test and post-test (see Figure 6).

Operation Span (OSpan). This task is a complex working memory capacity task that places demands alternately on the storage and manipulation of items in memory. In each trial, participants are presented with an arithmetic statement (e.g., “3 + 12 = 15”) and must make a rapid yes-no judgment as to whether that statement is correct. After responding to this statement, participants are presented with a to-be-remembered letter. After three to seven iterations of equations and letters, participants are prompted to report all of the letters that they have seen in the order in which they appeared. Both recall performance and arithmetic accuracy are recorded. Bunting
(2006) showed that individual differences in performance on this task are due to differences in cognitive control, specifically the ability to resist automatic (proactive) responses and to avoid confusing the current to-be-remembered items on a trial with memoranda from previous trials.

**Symmetry Span (SSpan).** This task is also a complex working memory capacity task. In this task, participants are presented with a visual array in response to which they must make a yes-no judgment as to whether that array displays vertical symmetry. After making symmetry judgments, participants see a memory item (a red box located somewhere on a 4x4 grid). After a variable number of symmetry judgments and memory items, participants are prompted to report the location of all memory items in the order that they were originally presented. Again, both recall performance and symmetry judgment accuracy are recorded. Kane et al. (2004) showed that this task requires cognitive control.

**Garden Path Sentences.** This is a self-paced reading task. On each trial, a sentence is presented on the screen in which every letter is replaced with a dash. Participants read the sentence by pressing the space bar to reveal the sentence one word at a time. The sentences consist of three types: (A) ambiguous items containing temporary syntactic ambiguities (e.g., “While the thief hid the jewelry sparkled brightly”); (B) content-matched unambiguous items (e.g., “The jewelry sparkled brightly while the thief hid”); and (C) syntactically unambiguous filler items (e.g., “As the researcher compiled the data he created graphs”). Each sentence is followed by a true/false comprehension probe. On each trial, the reading time for each word of the sentence is recorded, as is the accuracy of the response to the comprehension probe. Ambiguous sentences (type A) are designed such that initial parsing biases lead to an incorrect interpretation. Cognitive control is required to overcome these biases and correctly resolve the syntactic ambiguity, while unambiguous and filler items do not engage cognitive control to the same degree (cf. Novick et al., 2013).

**N-back.** In this task, single letters are displayed serially, with participants indicating via button press whether the current letter is the same as the item that appeared \( n \) items previously. For example, if a participant were in a 2-back condition and were presented with the stimuli \( L-F-Q-F-F \), they would make a “no” response to the first three letters, because none of them matches the letter that occurred two items previously (e.g., \( Q \) does not match \( L \)). They would make a “yes” response to the second \( F \), since \( F \) also occurred two positions prior; and they would make a “no” response to the final \( F \), since it does not match the \( Q \) that occurred two positions prior. This task includes three item types: fillers, which are letters that have not been seen before (such as the first three letters in this sequence); targets, which are items that match the \( N \)-back letter (the second \( F \)); and lures, which are recently presented letters that occur near the \( n \)-back item (the final \( F \)). Distinguishing between lure and target items requires participants to override a tendency to simply respond “yes” to familiar items and “no” to unfamiliar items. Both item types thus may induce conflict that must be resolved through cognitive control. Participants receive multiple sequences of trials in a 2-back and 4-back condition. Kane, Conway, Miura, and Colflesh (2007) showed that performance in this task can be ascribed to differences in cognitive control ability, and CASL’s cognitive training study with University of Maryland participants showed that this ability can be improved (see Chapter 1).

**Stroop.** This task is used to assess inhibition and facilitation in attention. In the task, participants are asked to indicate the color (red, green, or blue) of a word or a neutral series of “X”s presented on the screen. In those trials in which a word is present, the word itself is the name of a color. For example, the word “red” may appear colored in green. In order for the participant to correctly respond “green” to the word’s color, the participant must engage cognitive control in order to inhibit the typically automatic tendency to simply read the word, “red.” If, however, the word “red” appeared colored in red, then the participant would not need to inhibit the reading of the word in order to respond correctly. Participants typically respond more quickly and more accurately to trials of the latter type (“congruent” trials) than to trials of the former type (“incongruent” trials). Performance on both of these trial types can also be compared to “neutral” trials in which a colored series of “X”s are shown, with faster reactions to color-word matches reflecting facilitation, and slower reactions to color-word mismatches reflecting inhibition. Kane and Engle (2003) showed that cognitive control is required in order to respond correctly on incongruent trials.
Training tasks: Active training condition

Participants in the active training condition completed the same two performance-adaptive tasks during every training session (See Table 1). This regimen was designed to strengthen cognitive control.

Table 1. Training tasks for the active training condition.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Task</th>
<th>Session duration (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proactive interference resolution</td>
<td>N-Back</td>
<td>15</td>
</tr>
<tr>
<td>Spatial working memory</td>
<td>Symmetry span</td>
<td>15</td>
</tr>
</tbody>
</table>

*N-back Training*. This task followed the pattern of the assessment version of the N-back task, but the task difficulty increased adaptively: both sequence length and n-level increased as participants improved. Participants started with a sequence of 22 trials of $n = 2$ without any lures present. If a participant achieved an accuracy of 85% on that sequence, lures were added to the next sequence. When participants reached 85% accuracy on a sequence with lures, the n-level was increased and the sequence length was increased, up to a maximum length of 34 items and a maximum n-level of 21. If participant performance fell below 65% accuracy on a sequence, then difficulty was decreased, first by removing lures, and then by decreasing the n-level, down to a minimum length of 21 items and a minimum n-level of 1.

*Symmetry Span Training*. This task followed the pattern of the assessment version of the symmetry span task, but the task difficulty increased adaptively, with the sequence length (of both symmetry judgments and memory items) increasing as participants improved. Participants started with sequences of 2 symmetry judgments and memory items. After every fourth sequence, participants saw a feedback screen indicating the number of memory items that were recalled correctly and the percentage of accurate symmetry judgments. If, at this point, participants had successfully recalled all memory items from all 4 preceding sequences and had at least 80% accuracy on symmetry judgments for the preceding 4 sequences, the number of memory items per trial increased by one. There was no maximum sequence length. If the participant had successfully recalled all memory items from 3 of the preceding 4 sequences and had at least 80% accuracy on symmetry judgments for the preceding 4 sequences, difficulty did not change. If performance fell below this level, then the number of items per trial decreased by one, to a minimum sequence length of 2 symmetry judgments and memory items.

Training tasks: Control condition

Participants in the control condition completed the same two tasks during every training session (See Table 3). These performance-adaptive tasks were designed to have similar characteristics to the active training tasks, but *without* taxing cognitive control. Instead, the control tasks required participants to continuously attend to incoming stimuli to identify and respond to target items among distractor items.

Table 2. Training tasks for the control condition.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Task</th>
<th>Session duration (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vigilance (i.e, alert watchfulness for a target to appear)</td>
<td>Vigilance</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Symmetry Vigilance</td>
<td>15</td>
</tr>
</tbody>
</table>

*Vigilance Training*. The Vigilance training task was modeled after the N-back training task. In this task, participants were instructed to pay attention to a stream of letters, presented one at a time in sequences of 50 on the computer screen. Before each trial, participants were given a set of target letter item(s) that varied from 1 to 4 items. The set of target items always remained on the screen so that there was no working memory load for the task. Participants responded to each letter as it appeared on the screen, pressing the right arrow key when the current stimulus matched any item in the target list, and pressing the left arrow key when the current stimulus did not match any items on the target list. When participants reached 85% accuracy, the task was made more difficult in three ways: by increasing the number of target items to attend to, by reducing the available response time between trials, and by increasing the level of phonological similarity between the target and distractor items.
Symmetry Vigilance Training: The Symmetry Vigilance training task was modeled after the Symmetry Span training task. Before each trial, participants were given a set of target picture item(s) that varied from 1 to 4 items. The set of target items always remained on the screen so that there was no working memory load for the task. Participants viewed a 4 X 4 grid and observed as pictures appeared one at a time in various grid locations. For all trials (grouped into sequences of 50 pictures), participants responded to each picture as it appeared in a grid location on the screen, pressing the right arrow key when the current stimulus matched any item in the target list, and pressing the left arrow key when the current stimulus did not match any items in the target list. In between each vigilance judgment, participants made two symmetry judgments. When participants reached 85% accuracy, the task was made more difficult in three ways: by increasing the number of target items to attend to, by reducing the available response time between trials, and by increasing the level of visual similarity between the target and distractor items.

RESULTS

Training Compliance and Training Task Performance

Since participants completed the training on their own time, researchers monitored the CASL training website to make sure participants logged in and completed the prescribed sessions. Participants who went several days without logging into the CASL training website were contacted by email and encouraged to log in and complete the training. Of 206 participants who took the pre-test, only 110 returned for the post-test. Participants cited a number of reasons for withdrawing from the study, including lack of time to complete the training, boredom with the training tasks, and skepticism about the efficacy of training. Attrition was roughly equal across the two groups, with 58 active training participants and 52 control participants completing the study. Nevertheless, the high attrition rate raises the possibility that participants who remained in the two groups differed in terms of baseline working memory ability. To address this concern, researchers analyzed all the pre-test data to check for differences between the two groups. These results are reported in the subsequent section.

Although researchers contacted participants to remind them to complete the training, compliance to the training schedule among those who returned for the post-test was low. Of the 20 prescribed training sessions, the actual number of sessions these participants completed varied from 0 to 27, with a mean of 9.2 and a standard deviation of 7.7 (see Figure 7). Participants in the active training condition completed on average 2.8 sessions more than participants in the control condition. Anecdotal reports from participants suggest that this may be due to the fact that the active training tasks were more engaging than the control tasks. To factor out the variability in training compliance, researchers included compliance, defined as the number of training sessions completed, as a covariate in a number of the analyses that follow.
To demonstrate that training leads to improved performance on cognitive control tasks, it is vital to show that training actually took place. Despite the variability in training compliance, performance on all four tasks improved over time (see Figure 8). Paired-sample t-tests that compared participant scores on the final training session to their scores on the initial training session revealed that these improvements were significant for all four tasks, across both conditions ($p$’s < .001). (Note that scores are computed in a task-specific way and are not comparable across tasks or across the active training and control conditions. This suggests that the training in both training groups did indeed serve to strengthen participants’ task-related cognitive abilities.
Figure 8. Training curves for control and active training groups over the course of 20 sessions of working memory training. The "n-type" tasks are Vigilance (control) and N-Back (active training). The “symmetry” tasks are Symmetry Vigilance (control) and Symmetry Span Training (active training). Scores are computed task-internally and are not comparable across tasks or conditions. All four groups show significant gains in performance over the course of training.

Assessment Results

Performance on OSpan and SSpan was indexed by the proportion of items recalled from memory. Performance on the other three assessment tasks was indexed by both accuracy and reaction time to different item types. As is standard for these tasks, reaction times were calculated for correct trials only, since incorrect responses may not reflect the cognitive process of interest. (For example, incorrect trials may result from participants’ not paying attention to the task rather than from a failure to resolve a syntactic ambiguity.)

Results from OSpan and SSpan, the two complex span tasks, are plotted in Figure 10. Both groups’ performance improved from pre-test to post-test (p’s < .05); however, the active training group did not show significantly larger improvements than the control group.

Accuracy results for the other three tasks—Garden Path Sentences, N-back, and Stroop—are plotted in Figure 11, while reaction time results for N-back and Stroop are plotted in Figure 12. For graphical purposes, item types are group as follows. “Conflict” items consist of ambiguous sentences in the Garden Path task, lures in the N-back task, and incongruent trials in the Stroop task. “Base” items consist of the unambiguous sentences in the Garden Path task, targets in the N-back task, and congruent items in the Stroop task. “Filler” items consist of fillers in the Garden Path and N-back tasks and neutral items in the Stroop task. In the Garden Path and Stroop tasks, only the conflict items are hypothesized to engage cognitive control; while in the N-back task, as described earlier, both conflict and base items are hypothesized to do so. The N-back task contained both 2-back and 4-back sequences, which are analyzed separately.
In the Garden Path task, reaction times were calculated across four sentence regions (see Figure 9). The primary region of interest is Region 4 in the ambiguous items, since it is in this region that participants must revise their initial misinterpretation of the sentence. Reaction times for Garden Path sentences are plotted in Figure 13. Descriptions of the analysis procedures and statistical tests used can be found in Appendix 1.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Region 1</th>
<th>Region 2</th>
<th>Region 3</th>
<th>Region 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous</td>
<td>While the thief hid</td>
<td>the jewelry</td>
<td>that was elegant and expensive</td>
<td>sparkled brightly</td>
</tr>
<tr>
<td>Unambiguous</td>
<td>The jewelry</td>
<td>that was elegant and expensive</td>
<td>sparkled brightly</td>
<td>while the thief hid</td>
</tr>
</tbody>
</table>

Figure 9. Analysis regions in ambiguous and unambiguous items in the garden path task. Region 4 in the ambiguous items is the primarily region of interest, since it is in this region that participants must revise their initial misinterpretation of the sentence.

As predicted, accuracy improved significantly for all three tasks from pre-test to post-test, and reaction times generally decreased, indicating greater ease of cognitive processing. In the 4-back sequences in the N-back task, the active training group showed significantly larger gains in accuracy on target items than did the control group. In the Stroop task, the active training group showed marginally higher accuracy at post-test on incongruent and neutral items than did the control group ($p's = .06$), though this appeared to be driven by a significant decrement in performance in the control group from pre-test to post-test ($p < .05$). In the Garden Path task, significant slow-downs in reading times were found in Regions 1 and 2 in the ambiguous items and Region 1 in the unambiguous items for both groups. Significant speedups in reading times were found in Region 4 in the unambiguous items for both groups and Region 2 in the unambiguous items for the control group only ($p's < .05$). No effects were found in Region 4 of the ambiguous items, the primary region of interest, and no accuracy differences were observed between the two conditions.
Figure 10. Performance on the OSpan and SSpan assessments for the active training and control groups. Error bars are standard errors. As predicted, both groups improved from pre-test to post-test; however, there were no differences in improvements between the two groups.
Figure 11. Mean accuracy for each of three item types in the Garden Path, Stroop, and N-back assessments. Error bars are standard errors. As predicted, accuracy was generally lower in trial types that engage cognitive control (conflict items in Stroop and Garden Path and conflict and base items in N-back) than in non-conflict trials. Performance generally improved from pre-test to post-test across the control and active training groups. In the 4-back sequences in the N-Back task, the active training group showed significantly larger gains in accuracy on target items than the control group.
Figure 12. Mean reaction time in milliseconds for each of three trial types in the Stroop and N-back assessments, with results for 2-back and 4-back sequences plotted separately. Error bars are standard errors. As predicted, reaction times decreased across all three tasks; however, the active training group did not show significantly larger decreases in conflict items than the control group. Differences in baseline reaction times in the N-back task between the two groups may be a result of different patterns of attrition across the two groups.
DISCUSSION

We predicted that the active training group would show selective improvements from pre-test to post-test on components of the assessment tasks that required increased cognitive control. Indeed, in the 4-back sequences on the N-back assessment, the active training group showed significantly larger gains in accuracy on targets from pre-test to post-test than the control group. As predicted, this effect did not emerge on filler trials, which do not require cognitive control. Although the effect did not emerge on lure trials, this is not necessarily surprising. Ongoing CASL research suggests that performance on the N-back task is mediated by a number of factors, including the threshold of familiarity that participants set for making a “yes” or “no” response (that is, how strong the feeling of “familiarity” must be in order for an item to be considered a potential target). Lowering the familiarity threshold would cause participants to make more “yes” responses, which would increase their accuracy on targets but decrease their accuracy on lures. Strengthening cognitive control, however, should increase accuracy on both item types. It is possible that active training participants on average lowered their familiarity threshold from pre-test to post-test, but that through strengthened cognitive control they raised their accuracy on lures back to a level indistinguishable from their pre-test performance. This would account for this pattern of results. CASL is currently conducting research to identify and distinguish different components that contribute to N-back performance in order to refine these analyses.

On the other assessment tasks, we did not find equivalent effects to those found in our previous studies (such as in the Garden Path task: Novick et al., 2013). This is not surprising, however, given the low compliance
to the training regimen. The large body of evidence for the efficacy of working memory training is balanced by substantial evidence that this training is only effective given proper dosing (cf. Jaeggi et al., 2008; Owen et al., 2010). Participants in this study completed on average less than half the prescribed training, and there were differences in compliance between the control and active training groups that make it difficult to locate the source of differences in task performance.

The high rate of attrition also makes it difficult to interpret these results with confidence. Random assignment to experimental conditions—here, the control and active training groups—is designed to satisfy the assumption that the two groups are statistically equivalent prior to the experimental treatment, such that any differences that emerge between the two groups are due to the treatment. Attrition renders that assumption suspect. It is possible, for example, that participants who dropped out from the control group did so because they found the tasks too easy, while participants who dropped out from the active training group did so because they found the tasks too hard. In that case, the cognitive attributes of the remaining groups may differ independently of the training, in which case it becomes impossible to disentangle those differences from the effects of the training.

The poor participation highlights a number of important differences between this study and our previous, laboratory-based studies. In the current study, we were able to arrange for some participants to receive classroom credit for their participation, but the majority did not receive any kind of compensation. Participants in our previous studies, by contrast, were primarily undergraduate students who were paid for their time.

Many participants in the current study reported that they had difficulty finding time to complete the tasks. These participants generally had full-time jobs and a variety of obligations. Many had families to take care of; some were in the midst of preparations for overseas deployment. Many participants expressed that they were unconvinced of the value of working memory training. Additionally, some participants were concerned that their performance on the assessment and training tasks might be taken as indicative of their ability to perform their jobs. Even though the results were confidential, fear of the researchers’ perceptions or their own self-perceptions should they perform poorly or fail to produce training effects may have caused many of these participants to drop out.

These differences in incentive levels and life circumstances are apparently vital. In Study 2 reported in Chapter 1, for example, fewer than 10% of participants dropped out or failed to complete the prescribed training regimen, as compared to the 9 out of 206 participants in the current study who remained in the study and completed the prescribed 20 training sessions. Clearly, the prospect of deriving some kind of cognitive benefit from the training was insufficient incentive for the participants in this population.

Our previous studies also took place in a laboratory setting in which participants were monitored as they completed the prescribed training tasks. The current study, by contrast, relied on participants’ completing the training on their own time. Although CASL researchers monitored participant progress online, there was no way to ensure that users’ environments were free of distractions and that users were fully engaged with the tasks. Ideally, participants would be sufficiently motivated to remain focused on the tasks outside laboratory conditions, but the poor compliance and high attrition raise the concern that participants may have completed the tasks without fully engaging—and training—the relevant cognitive faculties.

These issues highlight the challenges involved in transitioning working memory training from a controlled, laboratory setting to an operational setting such as a classroom or workplace. For working memory training to be effective, participants must be engaged with and committed to the training, and they must complete a sufficient amount of training to realize the benefits. Since it is impossible for researchers or managers to fully control users’ interactions with working memory training outside the laboratory, the motivation to remain on task must come from participants.

As part of a plan to improve participant motivation, CASL collected extensive feedback on the training both from participants who completed the study and, where possible, from participants who dropped out. The primary complaint about the training was that the tasks were boring and repetitive. In response, CASL consulted with Human-Computer Interaction (HCI) specialists and designed a multi-phase plan to make the tasks more game-like and engaging. Steps in this plan include (1) allowing participants to track their high scores across sessions; (2) providing positive feedback whenever participants get a trial correct, and showing a continuous indicator of points earned for correct trials; (3) creating a leaderboard with high scores from all participants, so that individuals can set score goals and compete with their friends; and (4) providing “badges” or other rewards...
when participants pass certain score thresholds. Although some attrition is inevitable in any study, we hope that these improvements, which will be completed during fiscal year 2014, will reduce attrition in similar future studies to a degree sufficient to enable us to replicate our earlier findings. CASL will validate the improved tasks in fiscal years 2014-2015.

We emphasize that the evidence both from CASL’s research and from the field at large still strongly supports the efficacy of working memory training and its value for foreign language learning and processing. Indeed, a recently published meta-analysis of 79 studies of working memory and foreign language outcomes, the first of its kind, found a strong correlation between working memory capacity and both foreign language learning and processing (Linck et al., 2013). Working memory training, therefore, maintains its promise as a low-cost way to enhance foreign language professionals’ learning and performance. This study highlights the importance of proper dosing and participant engagement for effective training, and it raises a number of challenges that must be addressed in order for working memory training to be implemented outside a laboratory setting.

REFERENCES

Current and Future Directions

Polly O’Rourke, Jared M. Novick, Nina S. Hsu, and Suzanne M. Jaeggi

Thus far, the research program described herein has assumed that language and memory tasks both involve cognitive control. When parsing sentences, cognitive control enables the revision of initial misinterpretations (Novick et al., 2005); when remembering items, cognitive control facilitates target selection despite familiar-but-irrelevant memoranda (Jonides & Nee, 2006). Indeed, this has been the theoretical guiding principle that informed the design of the work summarized above.

Is there a common neurobiological basis for syntactic and non-syntactic conflict-control?

Current research

Here, we ask: Do these conflict-control procedures share neurobiological underpinnings? CASL’s current work is designed to achieve the following objectives: (1) replicate and extend previous results demonstrating that cognitive control is important for a range of memory and language tasks, including tasks used in previous CASL training research; (2) investigate whether the implementation of cognitive control recruits a common neurobiological system regardless of whether the conflict arises in memory (e.g., recognizing a familiar but irrelevant object as a non-target) or language (e.g., ambiguous input); and (3) validate prior CASL work by establishing that the memory-training and language-assessment tasks used in the training studies share a neuroanatomical basis that can be linked to the need for cognitive control, corroborating the possibility that training-transfer effects result from a functional-structural association between the tasks.

Indeed, neuropsychological and neuroimaging evidence suggests the use of common neural resources across a range of cognitive control tasks: patients with cognitive control deficits within memory fail to recover from parsing misanalysis (Novick et al., 2009), and healthy subjects completing Stroop-conflict trials and a syntactic ambiguity task show co-localized activity within lateral prefrontal cortex (the left inferior frontal gyrus, or LIFG; January et al., 2009; see also Ye & Zhou, 2009). Yet, two unresolved issues remain. First, what counts as “co-localized” activity across distinct tasks? Group analyses afford the statistical power to observe a predicted effect, but single-subject analyses provide increased anatomical sensitivity to better identify functional specificity. Moreover, it remains untested whether multivariate neuroimaging analyses could complement traditional univariate findings within a study, providing convergent evidence through complementary approaches.

CASL has been using functional magnetic resonance imaging (fMRI) to investigate the commonality of conflict-control mechanisms across ostensibly different memory and language tasks (such as those administered in the Chapter 1 studies). We hypothesized that cognitive control tasks commonly recruit regions within
posterior LIFG, regardless of differences in task goals and stimulus characteristics. Twenty participants completed four tasks while undergoing fMRI:

1. In the Stroop task, participants indicated the ink color of color-terms while ignoring word meaning; we included conflict, non-conflict, and neutral trials. Word meaning matched the font color in non-conflict trials (“red” written in red ink); word meaning and font color mismatched on conflict trials (“red” written in blue ink); and word meaning was unrelated to color on neutral trials (“horse” written in green ink). This is a classic and well-established conflict-control task, and is often used as an LIFG “localizer.”

2. In the recent-probes task, participants indicated whether a letter-probe (e.g., F) was among a memory set (e.g., k, b, e, p). Occasionally, the item was recent and familiar (e.g., on the prior trial, c, d, f, q), but not among the current set, thus creating conflict relative to non-recent trials also eliciting a “no” response.

3. In the 3-back memory task (which importantly is a static version of the N-back training task used in prior CASL training research), subjects viewed single words sequentially, indicating whether the current one matched the word three trials ago. We included recently presented but incorrect lures in nearby positions (i.e., 2- and 4-back), creating conflict between familiar yet irrelevant memoranda.

4. Finally, participants read garden-path sentences (e.g., “While the thief hid the jewelry that was elegant sparkled brightly”) to assess syntactic conflict-control versus a comma-disambiguation condition (“While the thief hid,...”). Critically, these are the same language-assessment materials used in the CASL training studies to evaluate transfer effects of cognitive control training.

For each task, we compared the conflict condition to its baseline. Group whole-brain analyses revealed distinct but overlapping brain activity across all tasks within LIFG. Then, in each subject’s brain, we identified active voxels (the 3-dimensional equivalents of a pixel in a volume image) in the Stroop task (i.e., the canonical conflict-control task) that were also anatomically located within posterior LIFG. Within these individually identified regions of interest, the other three tasks showed reliably or marginally reliable conflict effects (3-back: \( t = 2.57, p = 0.02 \); Recent-probes: \( t = 1.99, p = 0.06 \); Parsing: \( t = 3.15, p = 0.006 \)). We did not find co-localization in other regions. (The primary visual cortex, a comparison region, was engaged during each of the [visual] tasks but not during conflict resolution.) Together, these results suggest that common LIFG regions may selectively mediate cognitive control in syntactic and non-syntactic tasks. Moreover, to extend this univariate co-localization approach, CASL is beginning to employ multi-voxel pattern analysis to examine whether conflict detection elicits a common brain state reflected by cross-task pattern similarity that a machine classifier can reliably identify (going beyond information obtained with univariate analyses; see below). Preliminary data suggest above-chance classification in LIFG within-task (43% classification accuracy for Stroop conditions; chance = 33%; 53% classification accuracy for 3-back conditions; chance = 33%). We are currently testing whether a classifier trained on brain patterns associated with conflict in one task (e.g., Stroop) can accurately classify conflict in the three other tasks (e.g., recent-probes, 3-back memory, syntactic ambiguity), despite superficial differences. This result would suggest common mind and brain states associated with conflict, consistent with a domain-general theory of cognitive control that can inform neurobiological theories of language processing as well as what training and assessment tasks to pair in new regimens (i.e., those with greatest overlap and pattern similarity), in hopes of maximizing success.

**Future directions**

Proposed work plans to expand upon prior CASL research by continuing to “map” the common neural basis of language processing and cognitive control abilities in novel (and applicable) ways. The overall goal is to identify key neural markers that may inform and optimize the development of future training paradigms, perhaps (and ideally) even tailored to the individual based on one’s brain-behavior profile.

Specifically, CASL aims to use cutting-edge neuroimaging methods to elucidate the neural mechanisms underlying cognitive and mind states. As just mentioned, CASL is investigating whether neural similarity of task-specific mind states derived from a machine-learning, pattern-analysis approach can predict mind states across ostensibly different language, memory, and cognitive control tasks. With this information in hand, we describe two objectives that will test (1) how this measure of neural pattern analysis can map onto dependent behavioral measures and (2) how analysis of both brain activity and brain connections may serve as neural
markers of these cognitive abilities. To address these questions, CASL will employ a set of neuroscience methods including analysis of multivariate neural pattern similarity and functional connectivity measures of fMRI data. Broadly speaking, the goal of this work is to marshal these techniques to clarify the brain-behavior relationship that may yield training-transfer effects, thus serving as a neural marker of potential transfer success.

Multivariate pattern analysis (MVPA) of fMRI data is a novel neuroimaging technique that provides greater sensitivity over traditional univariate analyses. In MVPA, informative patterns of activity are identified in a few selected ways. These patterns are subsequently fed into a multivariate pattern classification algorithm. Finally, the trained classifier is given unseen patterns (importantly, from a statistically independent dataset) and predicts the task conditions associated with these patterns. Some refer to this as “mind reading,” in that the machine-classifier can correctly recognize when a subject is engaged in a particular task condition from brain activity patterns alone. Importantly, the high spatial frequency information detected by MVPA is conducive to performing within-subject analyses. An MVPA approach, then, affords greater sensitivity toward predicting unseen neural patterns (i.e., specific patterns that have not been “shown” to the classifier), and thus has the potential to cater cognitive training regimens to subjects on the basis of individual differences. CASL is currently piloting the utility of an MVPA approach in detecting brain-pattern similarity across an array of cognitive control, memory, and language tasks previously used in CASL training research, to identify the common neurobiological underpinnings that support language and memory performance. This work is summarized in the section above.

One objective of this new work, concretely, is to investigate how neural pattern analysis approaches can map onto dependent behavioral measures of language and memory. Previous CASL research demonstrates that cognitive training improves the ability to resolve temporary ambiguities during language processing. For this objective, CASL will examine the extent to which neural markers (as measured by MVPA described earlier) accompany—and may predict—behavioral effects. We expect that this approach may yield more sensitive and individualized information about the brain-behavior relationship across, language, memory, and cognitive control performance, beyond a typical univariate approach.

A second objective is to test how regional network connections and neural patterns serve as neural markers of potential transfer success. The cognitive abilities underlying successful processing of language and memory input may be reflected both in regional activity differences, but also in functional-anatomical connections across a network of brain regions (Sundermann & Pfleiderer, 2012). That is, cognitive control and language abilities may be supported, at least in part, by a combination of activity and global brain network properties. Using a functional-connectivity approach, CASL aims to investigate the role that functional-connectivity may play as an underlying neural system that supports these abilities.

Understanding the neurobiological correlates of working memory and language processing allows us to root our approach to training in large literatures that have examined the functionality of different regions and networks in the brain. This will help us refine our training regimens as we identify more precisely the mechanisms that we are targeting. In the long term, we hope that this approach will lead to training regimens perhaps tailored to individual cognitive and learning profiles.

**How do working memory mechanisms contribute to language processing?**

**Current research**

The goal of this research program was to determine how different working memory mechanisms contribute to sentence processing, both in terms of comprehension accuracy and neural response as indexed by event-related potentials (ERP). In two experiments, we examined how the two measures of sentence processing interacted with performance on a variety of working memory assessments. This synopsis will focus on the N-back task. Although the N-back task is well established in working memory and cognitive training research, it has not been examined in terms of its relationship to sentence processing. The results of these studies support the strong connection between linguistic and non-linguistic conflict resolution described by Novick in Chapters 1 and 3, and validate the use of N-back as a working memory training task.

ERP, a non-invasive method of recording brain activity, has been used to examine the neural basis of language processing extensively during the last several decades. Using electrodes placed on the scalp, the voltage changes associated with neural activity in different populations of neurons in the cerebral cortex can be
recorded. This recording of neural activity is the electroencephalogram (EEG). In ERP research, the EEG is time locked to specific experimental events (e.g., the presentation of a stimulus). When the ERPs are categorized by experimental condition and averaged across trials, patterns in brain activity emerge showing that certain negative or positive deflections (i.e., potentials) are associated with specific cognitive events. In ERP research, the EEG is sampled at 250 Hz, meaning that there are 250 data points per second. This fine temporal resolution allows the examination of processes that occur very quickly, like language processing.

While there are many ERPs related to different aspects of language processing (Kutas, van Petten & Kluender, 2006), the event-related potential of interest in this research is the P600. This component is a positive shift that emerges 500 to 800 ms after stimulus onset over posterior sites and is considered to be an index of difficulty in syntactic integration, reflecting a late controlled process in contrast to earlier, more automatic mechanisms (Friederici, 1995, 2002; Hahne & Friederici, 1999). It is elicited by garden-path sentences at the point of the resolution of the temporary syntactic ambiguity compared to unambiguous controls (Kaan & Swab, 2003; Osterhout, Bersick, & McLaughlin, 1997; Osterhout, Holcomb, & Swinney, 1994; van Berkum, Koorneef, Otten, & Nieuwland, 2007), and by sentences containing object relative clauses compared to simple sentences (Gouvea et al., 2010; Kaan, Harris, Gibson, & Holcomb, 2000; Münte et al., 1997; Phillips, Kazanina, & Abada, 2005; Ueno & Garnsey, 2007).

In the current experiments, participants were presented with sentences containing temporary syntactic ambiguities (garden-path sentences) as well as sentences containing increased syntactic complexity (object relatives). See Table 3 for examples. In the garden-path sentences, one might initially believe that the patient met the doctor and the nurse, but upon seeing the verb “showed,” it becomes obvious that this interpretation is incorrect, and must be reanalyzed. The object relatives are unambiguous but represent increased syntactic complexity. In both experiments, participants read sentences and answered comprehension questions while the EEG was recorded, after which they completed a battery of working memory assessments, including an N-back task with lures (see Chapter 1 for task description). In Experiment 1, the comprehension questions did not directly target correct resolution of the temporary syntactic ambiguity (i.e., successful conflict resolution). In Experiment 2, comprehension questions targeting that key structure were included, as were an increased number of filler/distractor sentences to reduce the predictability of the two target sentence types. The ERPs were time-locked to the underlined verb.

Table 3. Examples of Sentence Types.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garden-Path</td>
<td>The patient met the doctor and the nurse with the white dress showed the chart during the meeting.</td>
</tr>
<tr>
<td>Object Relative</td>
<td>The patient met the doctor to whom the nurse with the white dress showed the chart during the meeting.</td>
</tr>
<tr>
<td>Control</td>
<td>The patient met the doctor while the nurse with the white dress showed the chart during the meeting.</td>
</tr>
</tbody>
</table>

In both experiments, garden-path and object relative sentence types elicited P600 effects. Comprehension accuracy for control sentences was higher than for garden-paths in both experiments. The effect only reached significance for object relatives in Experiment 2.

With respect to the relationship between sentence comprehension effects and N-back, in Experiment 1, there was a negative relationship between N-back lure accuracy and the effects associated with increased syntactic complexity, both in the comprehension accuracy and P600 data. Individuals who had high N-back lure accuracy showed smaller P600s and smaller reductions in comprehension accuracy for sentences containing object relatives. No relationship between N-back and garden-path effects was observed. In Experiment 2, in which the target structures were less predictable and comprehension questions targeting garden-path resolution were included, there was a negative correlation between performance on N-back lures and garden-path effects in
accuracy such that individuals who did well on lures showed smaller reductions in accuracy for garden-path sentences compared to controls. For object relatives, there was a positive correlation with lure performance and overall accuracy as well as a negative correlation between lure accuracy and the P600 effect. As in Experiment 1, individuals who did well on N-back lures showed smaller P600s for object relatives.

These findings support the domain general account for conflict resolution in that, when garden-path resolution is required to correctly answer comprehension questions, the ability to successfully resolve the conflict (between the initial incorrect interpretation and the correct one) correlates with performance on N-back lures. Furthermore, this key finding of Experiment 2 validates the use of the N-back as a training task for improving the ability to interpret ambiguous language input. The findings with respect to unambiguous syntactic complexity (i.e., the object relatives) suggest that N-back training could lead to increased efficiency in purely syntactic processing (in absence of any need for reanalysis) as individuals who did well on N-back lures showed reduced P600s and smaller decrements in accuracy for the syntactically complex sentences.

**Future directions**

CASL research has validated the use of the N-back as a training task for improving the resolution of temporary ambiguity in language processing through both the training study described by Novick in Chapter 1 and the neurocognitive research described in the current chapter. The next step will be to examine the impact of N-back training on adult language learners in a controlled study done with adult learners of Spanish from the University of Maryland. This planned study will investigate how improving conflict resolution ability via N-back training will impact second language proficiency by determining how training-related improvements influence various aspects of second language proficiency and processing. The results will provide a nuanced picture of how working memory training improves second language learning outcomes.

**REFERENCES**


Chapter 2 Assessment Results

Analysis procedures and statistical tests

Alan E. Mishler

Three sets of analyses were conducted for each of the five assessment tasks. First, a pre-test analysis was conducted on the pre-test data to confirm that the tasks were functioning as intended and to check for baseline differences in task performance between the two groups. Since assignment to the two groups was random, no baseline differences were predicted, but the possibility was of concern due to the high attrition rate. In order to take into account variable compliance to the training regime, a post-test analysis was conducted on the post-test data, in which compliance, defined as the number of training sessions completed, was included as a covariate. Finally, a pre-post analysis was conducted on the combined pre- and post-test data to check for the predicted differences in improvements between the control and active training groups.

Operation Span

Performance on the Operation Span (OSpan) task was measured as the number of items that a participant remembered correctly in the correct order. Participants were excluded from these analyses if their data was missing at either pre-test or post-test or if their performance on the processing (arithmetic) component of the task fell below 85% at either pre-test or post-test. A total of 92 participants (41 control, 51 active training) were included in the analyses.

Pre-Test Results

An ANOVA was conducted to examine differences in OSpan scores at pre-test as a function of condition. As predicted, this analysis revealed no significant differences in OSpan scores between participants in the active training condition versus the control condition. These results suggest that participants in both the active training and control conditions demonstrate similar OSpan performance prior to training.

Post-Test Results

An ANCOVA was conducted to examine differences in OSpan scores at post-test as a function of condition, with training compliance included as a covariate. No significant effects were found.

Pre-Post Results

A 2 (condition) by 2 (time) repeated-measures ANOVA was conducted to examine differences in OSpan scores, with condition as a between-subjects factor and time as a within-subjects factor. This analysis identified a significant effect of time, $F(1, 180) = 4.93, p = .03$, such that post-test OSpan scores were significantly higher than pre-test scores. No other effects were significant, suggesting that the active training and control groups did not differ as a result of training.
Symmetry Span

Performance on the Symmetry Span (SSpan) task was measured as the number of items that a participant remembered correctly in the correct order. Participants were excluded from these analyses if their data was missing at either pre-test or post-test or if their performance on the processing (symmetry judgment) component of the task fell below 85% at either pre-test or post-test. A total of 102 participants (49 control, 53 active training) were included in the analyses.

Pre-Test Results

An ANOVA was conducted to examine differences in SSpan scores at pre-test as a function of condition. As predicted, this analysis revealed no significant differences in SSpan scores between participants in the active training condition versus the control condition. These results suggest that participants in both the active training and control conditions demonstrate similar SSpan performance prior to training.

Post-Test Results

An ANCOVA was conducted to examine differences in SSpan scores at post-test as a function of condition, with training compliance included as a covariate. No significant effects were found.

Pre-Post Results

A 2 (condition) by 2 (time) repeated-measures ANOVA was conducted to examine differences in OSpan scores, with condition as a between-subjects factor and time as a within-subjects factor. This analysis identified a significant effect of time, \( F(1, 200) = 14.67, p = .0002 \), such that post-test SSpan scores were significantly higher than pre-test scores. Although participants in the active training condition showed slightly larger gains in SSpan scores from pre-test to post-test, this effect was not significant.

Garden Path Sentences

Performance on the Garden Path task was measured in terms of reading times (RTs) by sentence region (in milliseconds) as well as response accuracy to comprehension probes. Performance on ambiguous items was compared to performance on unambiguous items. From pre- to post-test, the active training was group predicted to show, relative to the control group, (1) a selective improvement in accuracy scores on ambiguous items, (2) a selective reduction in reading times in the disambiguating region of ambiguous items, or (3) both. All ANOVAs and ANCOVAs conducted on the Garden Path data were mixed effects analyses with repeated measures within subjects.

Because filler items were designed to be easy, participants were excluded from analysis if their accuracy on filler items was below 80% at either pre-test or post-test, since this was taken as an indication that they were not engaged with the task or that they had mixed up the keys used to indicate “true” or “false” to the comprehension probes. A total of 91 participants (38 control, 53 active training) were included in the analyses.

Accuracy Results

Accuracy scores were elogit-transformed for each participant at each combination of the factors time (pre/post) and item type (ambiguous/unambiguous/filler). All analyses were conducted on the elogit-transformed scores. For ease of interpretation, means and error bars in the figures included in the main text are presented in terms of raw accuracy.

Pre-Test Results

A 2 (condition) by 3 (item type) repeated-measures ANOVA was conducted on the pre-test results in order to determine whether there were baseline differences in performance between the two experimental groups, and to check for the predicted differences in performance on ambiguous items as compared to unambiguous items. This analysis revealed no significant main effect of condition and no significant interaction effects, suggesting that, as expected, the groups did not differ in baseline performance. A significant effect was found for item type, \( F(2, 178) = 263, p < .001 \). Post-hoc pairwise comparisons revealed significantly higher accuracy for fillers than
Post-Test Results

In order to investigate the effect of compliance on training performance, a 2 (condition) by 3 (item type) repeated-measures ANCOVA with compliance included as a covariate was conducted on the post-test results. There was a significant effect for compliance, \( F(1, 87) = 47.04, p = .005 \); a significant effect for item type, \( F(2, 174) = 82.89, p < .001 \); and a significant interaction between item type and compliance, \( F(2,176) = 5.37, p = .005 \). Post-hoc pairwise comparisons revealed significantly higher accuracy for fillers and unambiguous items than for ambiguous items (\( p's < .001 \)), but in additional analyses compliance did not correlate with gains in accuracy from pre-test to post-test for any item type. These results suggest that there were no significant differences at post-test between the control and active training groups.

Pre-Post Results

In order to investigate the effects of training on Garden Path accuracy, a 2 (condition) by 2 (time) by 3 (item type) repeated-measures ANOVA was conducted to examine changes in Garden Path accuracy scores from pre- to post-test. This analysis identified a significant effect of item type, \( F(2, 178) = 90.02, p < .001 \); a significant effect of time, \( F(1, 89) = 13.88, p < .001 \); and a significant interaction effect of item type by time, \( F(2, 178) = 15.94, p < .001 \). Post-hoc ANOVAs conducted independently on each item type confirmed that accuracy for ambiguous items increased significantly from pre-test to post-test, \( F(1, 89) = 35.75, p < .001 \). The absence of any significant effects involving condition suggests that the active training and control groups did not differ as a result of training.

Reading Times Analysis

Reading times were first Winsorized at the word level for each level of the time factor (pre/post): reading times that fell more than 3 standard deviations outside the mean for words within the relevant sample (pre- or post-test data) were set to the 3 standard deviation threshold. This procedure affected 1.47% of the data. Next, reading times were residualized: for each participant at each level of the time factor, reading times were regressed over word length to produce residual values, i.e., deviations from expected reading time as a function of word length. Finally, ambiguous and unambiguous items were divided into regions of interest based on syntactic structure, and residualized reading times for words within each region were summed to produce by-region reading times. Analyses were conducted only on ambiguous and unambiguous items and only on trials with correct comprehension probe responses, since incorrect responses may reflect a failure to attend to the syntactic structure of the sentence and therefore an absence of cognitive conflict. Filler items were excluded from analyses due to their variable syntactic structure, which made it impossible to divide them into consistent regions. Separate analyses were carried out on each of four regions within the ambiguous and unambiguous items.

The primary region of interest is Region 4 in the ambiguous items, since this is the disambiguating region. Participants in the active training group were predicted to show, relative to the control group, selective decreases in reading time at this region at post-test.

Since the length and content of the regions differs between ambiguous and unambiguous items, analyses for the two item types were conducted independently.

Pre-Test Results

Two-way repeated measures ANOVAs were conducted on each region and item type from the pre-test results in order to determine whether there were baseline differences in performance between the two experimental groups. A significant effect of condition was found for Region 2 of the unambiguous items, \( F(1, 89) = 6.14, p = .015 \), driven by slower reading times in the control group (\( M = 21.56, SD = 50.56 \)) than in the active training group. There were no other significant effects. Crucially, the control and active training groups did not differ in reading times in the critical Region 4 of the ambiguous items; nor did they differ in whole sentence reading time on either item type, suggesting that, as predicted, the two groups were generally equivalent at baseline.
Post-Test Results

Two-way repeated-measures ANCOVAs with compliance included as a covariate were conducted on the post-test results. In the ambiguous items, a significant effect of condition was found at Region 3, driven by slower reading times in the control group, $F(1, 87) = 4.49, p = .037$. No other significant effects were found, suggesting that the two groups generally did not differ at post-test.

Pre-Post Results

In order to examine the effects of training on Garden Path reading time, 2 (condition) by 2 (time) repeated-measures ANOVAs were conducted. For the ambiguous items, significant effects of time were found in Region 1, $F(1, 89) = 8.72, p = .004$; and Region 2, $F(1, 89) = 4.84, p = .03$. For the unambiguous items, a significant effect of time was found at Region 1, $F(1, 89) = 6.71, p = .011$; and Region 4, $F(1, 89) = 4.157, p = .044$; and a significant interaction between time and condition was found at Region 2, $F(1, 89) = 6.98, p < .01$, driven by a speedup in the control group from pre-test to post-test. The absence of any effects involving condition in Region 4 suggests that the active training and control groups did not differ in their ambiguity resolution abilities as a result of training.

N-back

Performance on the N-back task was measured in terms of accuracy and reaction times (RTs) for accurate responses. Participants were excluded if their data was missing at pre-test or post-test. All 110 participants (52 control, 58 active training) were included in the accuracy analyses. An additional eight active training participants were excluded from the reaction time analyses because they had zero accuracy on at least one item type at either pre-test or post-test. All participants in the N-back task completed both 2-back and 4-back sequences, referred to here as \{n2\} and \{n4\}. Data from the two n-levels were analyzed separately.

Accuracy Results

Accuracy scores were elogit-transformed for each participant at each combination of the factors time (pre/post) and item type (target/lure/filler). All analyses were conducted on the elogit-transformed scores. For ease of interpretation, means and error bars in the figures in the main text are presented in terms of raw accuracy.

Pre-Test Accuracy Results

A 2 (condition) by 3 (item type) repeated-measures ANOVA was conducted on each of the n-levels. Significant effects of item type were found in both the \{n2\} task, $F(2, 216) = 41.40, p < .0001$; and the \{n4\} task, $F(2, 216) = 66.09, p < .0001$. Although the active training group had lower numerical accuracies at pre-test, no effects of condition were found, suggesting that the control group and the active training group were not significantly different in terms of baseline N-back accuracy.

Post-Test Accuracy Results

A 2 (condition) by 3 (item type) repeated-measures ANCOVA with compliance as a covariate was conducted for each n-level to examine differences in N-Back accuracy. A significant effect of item type was found for \{n2\}, $F(2, 104) = 29.42, p < .001$; and for \{n4\}, $F(2, 212) = 42.13, p < .001$. In addition, a significant interaction between item type and compliance was found for \{n2\}, $F(2, 212) = 4.04, p = .018$. However, follow-up analyses failed to show a correlation between compliance and accuracy gains for any item type in \{n2\}. The absence of any effects involving condition suggests that the control and active training groups did not reliably differ in post-test N-back accuracy.

Pre-Post Accuracy Results

A 2 (condition) by 3 (item type) by 2 (time) repeated-measures ANOVA was conducted for each n-level to examine differences in N-back accuracy. In the \{n2\} task, a significant effect was found for item type, $F(2, 216) = 58.78, p < .0001$. In the \{n4\} task, significant effects were found for TrialType, $F(2, 216) = 90.30, p < .001$; and time, $F(1, 324) = 7.2, p = .008$; and a significant interaction was found for condition and time, $F(1, 324) = 5.74, p = .017$, driven by larger gains in accuracy in the active training group than in the control group. Follow-
up pairwise comparisons for each trial type revealed that the active training group showed significant improvements from pre-test to post-test on targets. No improvements on other item types were significant in either group, and there were no other significant effects.

**Reaction Time Results**

Reaction time (RT) analysis was conducted on correct trials only. Participants’ reaction times were trimmed separately for pre-test and post-test by removing any RTs over 3 standard deviations away from a participant’s mean RT for each testing session. This procedure affected 1.6% of the data. Participant data was then summarized in terms of mean RTs for each trial type at each testing session.

**Pre-Test Reaction Time Results**

A 2 (condition) by 3 (item type) repeated-measures ANOVA was conducted for each n-level to examine baseline differences in RTs. As predicted, these analyses revealed a significant effect of item type for both \( n_2 \), \( F(2, 200) = 48.09, p < .001; \) and \( n_4 \), \( F(2, 200) = 13.65, p < .001, \) driven by slower reaction times in lures as compared to targets or fillers (\( p \)'s < .01). A significant effect of condition was also found for both the \( n_2 \) task, \( F(1, 100) = 7.66, p = .007; \) and the \( n_4 \) task, \( F(1, 100) = 3.22, p = .076. \) Visual inspection of the means makes it clear that this effect is driven by slower reaction times for the active training group as compared to the control group at both \( n \)-levels. The baseline differences between the two conditions are likely due to different patterns of attrition between the two groups such that participants with relatively faster reaction times remained in the control group and participants with relatively slower times remained in the active training group.

**Post-Test Reaction Time Results**

A 2 (condition) by 3 (item type) by 2 (time) repeated-measures ANCOVA with compliance as a covariate was conducted for each \( n \)-level to examine differences in \( N \)-back RT. In the \( n_2 \) task, significant main effects were found for condition, \( F(1, 98) = 9.33, p = .003; \) and item type, \( F(2, 196) = 22.69, p < .001; \) and significant interaction effects were found for item type and compliance, \( F(2, 196) = 3.18, p = .04; \) and condition, item type, and compliance, \( F(2, 196) = 3.34, p = .04. \) In the \( n_4 \) task, significant main effects were found for condition, \( F(1, 98) = 7.75, p = .006; \) and TrialType, \( F(2, 196) = 7.62, p < .001; \) and significant interactions were found for condition and TrialType, \( F(2, 196) = 3.30, p = .04; \) and condition, TrialType, and compliance, \( F(2, 196) = 5.95, p = .003. \) These results suggest that compliance was an important mediator of change in performance in this task, but because the active training and control groups differ at pre-test, they do not necessarily reflect differences resulting from training.

**Pre-Post Reaction Time Results**

A 2 (condition) by 3 (item type) by 2 (time) ANOVA was conducted for each \( n \)-level to examine changes in RT. In both tasks, significant effects were found for item type (\( n_2 \): \( F(2, 200) = 48.09, p < .001; \) \( n_4 \): \( F(2, 200) = 13.65, p < .001), time (\( n_2 \): \( F(1, 300) = 46.34, p < .001; \) \( n_4 \): \( F(1, 300) = 24.54, p < .001), and condition (\( n_2 \): \( F(1, 100) = 11.63, p = .001; \) \( n_4 \): \( F(1, 100) = 7.44, p = .008). These effects are driven by slower reaction times in lures compared to targets and filler, slower reaction times in the active training group than in the control group, and an overall decrease in reaction times from pre-test to post-test.

**Stroop**

Performance on the Stroop task was measured in terms of accuracy and reaction times (RTs) for accurate responses. Participants were excluded if their data was missing at pre-test or post-test. A total of 110 participants (52 control, 58 active training) were included in the accuracy analyses. An additional 27 control participants and 26 active training participants were excluded from the reaction time analyses because they had less than 80% accuracy at either pre-test or post-test.

**Accuracy Results**

Accuracy scores were elogit-transformed for each participant at each combination of the factors time (pre/post) and item type (target/lure/filler). All analyses were conducted on the elogit-transformed scores. For ease of interpretation, means and error bars are presented in terms of raw accuracy.
Pre-Test Accuracy Results
A 2 (condition) by 3 (item type) repeated-measures ANOVA was conducted to examine pre-test performance. A significant effect was found for item type, $F(2, 216) = 48.93, p < .001$, reflecting, as predicted, lower accuracies on incongruent trials than on congruent trials ($p$’s < .001). No effect of condition was found, suggesting that the control group and the active training group were not significantly different in terms of baseline Stroop accuracy.

Post-Test Accuracy Results
A 2 (condition) by 3 (item type) repeated-measures ANCOVA with compliance as a covariate was conducted to examine differences in Stroop accuracy. Significant effects were found for item type, $F(2, 212) = 50.69, p < .001$; and condition, $F(1, 106) = 4.44, p = .04$. Post-hoc pairwise comparisons revealed significantly lower accuracy for incongruent trials as compared to the other two item types ($p$’s < .05), and marginally higher accuracy on incongruent and neutral trials for the active training group as compared to the control group ($p$’s = .06).

Pre-Post Accuracy Results
A 2 (condition) by 3 (item type) by 2 (time) repeated-measures ANOVA was conducted to examine Stroop accuracies as a function of time and training condition. Significant effects were found for item type, $F(2, 216) = 90.1, p < .001$; and the interaction between time and condition, $F(1, 324) = 15.9, p < .001$. Post-hoc pairwise comparisons revealed significant decrements in performance from pre-test to post-test for the control group on both neutral and congruent items ($p$’s < .05), and a marginal decrement in performance for the control group for incongruent items ($p = .0528$). The active training group did not show significant changes from pre-test to post-test on any item type.

Reaction Time Results
Reaction times were first Winsorized for each participant at each time (pre/post): RTs that fell more than 3 standard deviations outside the mean for words within the relevant sample were set to the 3 standard deviation threshold. This procedure affected .13% of the data.

Pre-Test Results
A 2 (condition) by 3 (item type) repeated-measures ANOVA was conducted to examine pre-test differences in RT. A significant effect was found for item type, $F(2, 110) = 254, p < .001$. No effect was found for condition, suggesting that the control and active training groups did not differ at baseline. Post-hoc pairwise comparisons revealed significantly slower RTs for neutral items as compared to congruent items, and significantly slower RTs for Incongruent items as compared to the other two item types. This is a replication of the expected Stroop effect.

Post-Test Results
A 2 (condition) by 3 (item type) by 2 (time) with compliance as a covariate was conducted for each to examine differences in Stroop RT. A significant effect was found for item type, $F(2, 106) = 167, p < .001$. Post-hoc pairwise comparisons confirmed the same pattern of RTs as at pre-test ($p$’s < .001). No other effects were significant, suggesting that the control and active training groups did not differ significantly at post-test.

Pre-Post Results
A 2 (condition) by 3 (item type) by 2 (time) ANOVA was conducted to examine changes in RT. Significant effects were found for time, $F(1, 165) = 31.42, p < .001$; item type, $F(2, 110) = 276, p < .001$; the interaction between time and condition, $F(1, 165) = 20.95, p < .001$; and the interaction between time and item type, $F(2, 165) = 4.37, p = .01$. Post-hoc pairwise comparisons revealed significant decreases in reaction time from pre-test to post-test in all item types in the control group ($p$’s < .05). The active training group showed no significant changes for any item type.