LANGUAGE AND COGNITIVE PROCESSES, 2013
http://dx.doi.org/10.1080/01690965.2012.758297

Routledge
Taylor & Francis Group

Clearing the garden-path: Improving sentence processing through cognitive control training

Jared M. Novick, Erika Hussey, Susan Teubner-Rhodes, J. Isaiah Harbison, and Michael F. Bunting

Center for Advanced Study of Language, Department of Psychology, Program in Neuroscience and Cognitive Science, University of Maryland, College Park, MD, USA

How do general-purpose cognitive abilities affect language processing and comprehension? Recent research emphasises a role for cognitive control—also called executive function (EF)—when individuals must override early parsing decisions as new evidence conflicts with their developing interpretation. We tested if training on non-syntactic EF tasks improves readers' ability to recover from misanalysis during language processing. Participants completed pre/post-reading assessments containing temporarily ambiguous sentences susceptible to misinterpretation. Performance increases on a training task targeting conflict-resolution processes (*n*-back with "lures") predicted improvements in garden-path recovery. *N*-back responders—those demonstrating reliable training gains—significantly increased their comprehension accuracy across assessments. Their posttest eye-movement patterns also revealed significantly improved real-time revision following entry into disambiguating sentence regions where cognitive control is hypothesised to engage. Untrained participants and *n*-back non-responders showed no performance changes. The results provide insight into how nonlinguistic functions contribute to parsing and interpretation and suggest that certain language skills are amenable to improvement via domain-general EF training.

Keywords: Parsing; Syntactic ambiguity resolution; Language processing; Cognitive control; Executive-function training; Working memory; Reading-time measures; Conflict/interference resolution.

Correspondence should be addressed to Jared M. Novick, Center for Advanced Study of Language, University of Maryland, 7005 52nd Ave., College Park, MD 20742, USA. E-mail: jnovick1@umd.edu

The University of Maryland Center for Advanced Study of Language supported this research. The authors thank Sharona Atkins, Jeff Chrabaszcz, Carrie Clarady, Ryan Corbett, Alexei Smaliy, David Alexander and Pooja Datta for their assistance with collecting and scoring data. Additionally, we thank Kiel Christianson, Barbara Forsyth, Seth Greenberg, Jesse Snedeker, Michael Ullman and Scott Weems for valuable feedback on the design of the experiment. Thanks to John Trueswell and Gerry Altmann for key discussion and comments on earlier drafts of the manuscript. Finally, a portion of this work was presented at the 16th Annual Conference on Architectures and Mechanisms for Language Processing (York, England), the 51st Annual Meeting of the Psychonomics Society (St. Louis, Missouri) and the 24th Annual CUNY Conference on Human Sentence Processing (Stanford University, Palo Alto, California). We thank the attendees for their important suggestions.

INTRODUCTION

When reading or processing speech, individuals assign provisional analyses to words and phrases as they are perceived moment-by-moment, rather than delaying interpretation until sentences unfold entirely (Altmann & Kamide, 1999; Tanenhaus, 2007). Although efficient, processing language "on-the-fly" comes at the cost of having to deal with temporary ambiguity; initial analyses sometimes turn out wrong as newer input suggests a quite different interpretation. Consider this *New York Times* headline: "Google's computer might betters translation tool". Because "might" frequently appears as an auxiliary verb, readers may initially misunderstand that it is used here as a noun (meaning *strength*). The processing difficulty experienced following such misanalysis is known as the "garden-path effect" and requires cognitive control to countermand early processing decisions and recover alternative interpretations.

Cognitive control, also called executive function (EF), refers to the regulation of mental activity to guide and support flexible behaviour, enabling individuals to bias the selection of appropriate over inappropriate information during goal-directed tasks (Miller & Cohen, 2001). Converging data from behavioural, brain-imaging and neuropsychological studies suggest that shared mechanisms in left ventrolateral prefrontal cortex (VLPFC) support regulatory functions across a range of tasks, including working memory, attention and language processing (Thompson-Schill, Bedny, & Goldberg, 2005). In the case of interpreting sentences, one account suggests that cognitive control enables individuals to override initial mischaracterisations of the input to prevent comprehension failure (Novick, Trueswell, & Thompson-Schill, 2005). Specifically, the discovery of a misinterpretation triggers domain-general cognitive control processes to resolve incompatible representations of sentence meaning—the one originally favoured vs. the correct alternative that must be recovered.

Research examining special populations demonstrates the importance of cognitive control for syntactic ambiguity resolution. For example, young children and patients with VLPFC damage commonly exhibit an inability to revise incorrect parsing decisions during spoken comprehension tasks (e.g., Novick, Kan, Trueswell, & Thompson-Schill, 2009; Trueswell, Sekerina, Hill, & Logrip, 1999; Weighall, 2008), whereas healthy adults exert rapid control to capture the intended analysis (e.g., Novick et al., 2009; Novick, Thompson-Schill, & Trueswell, 2008; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Trueswell et al., 1999). The difficulty that young children and patients exhibit overriding initial interpretations is likely related to the putative difficulty they show implementing control across a range of non-syntactic EF measures (e.g., Davidson, Amso, Cruess Anderson, & Diamond, 2006; Hamilton & Martin, 2005; Mazuka, Jincho, & Oishi, 2009; Thompson-Schill et al., 2002). For instance, a VLPFC patient's garden-path recovery failure was recently tied to his exaggerated susceptibility to interference during the "recent probes" task, indexed by a failure to disregard highly familiar but currently irrelevant memoranda that were presented earlier (Novick et al., 2009). Similar findings have been reported in young children, perhaps related to the maturational delay of prefrontal systems (Huttenlocher & Dabholkar, 1997; Khanna & Boland, 2010; Nilsen & Graham, 2009; see also Mazuka et al., 2009; Novick et al., 2005). Neuroimaging data from healthy adults corroborate this behavioural and neuropsychological pattern, wherein ambiguity resolution and non-syntactic cognitive control processes co-localise within individuals in posterior regions of left VLPFC (January, Trueswell, & Thompson-Schill, 2009; Ye & Zhou, 2009). Altogether, this

suggests that one of the functions of cognitive control may be to override early misinterpretations during sentence processing.

A current wave of research demonstrates that EFs can be improved through extensive practice. The findings suggest that training gains extend beyond the specific tasks trained; improvements transfer (i.e., generalise) across domains to novel performance measures that also rely on EF, including intelligence, text comprehension, task-switching and interference-resolution tasks (e.g., Chein & Morrison, 2010; Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Karbach & Kray, 2009; cf. Redick et al., 2012). Yet, no studies have tested if training can effectively help individuals override well-known biases that sometimes lead to misinterpretation during language processing.

Here we investigate if enhancing regulatory functions through cognitive-control training improves garden-path recovery in healthy adults. Given the claims that EF and flexible cognition also play an important role in a range of other language-processing tasks, including lexical ambiguity resolution (e.g., Bedny, Hulbert, & Thompson-Schill, 2007; Rodd, Johnsrude, & Davis, 2010), verbal fluency (e.g., Novick et al., 2009; Robinson, Blair, & Cipolotti, 1998; Schnur et al., 2008), abstract word comprehension (Hoffman, Jefferies, & Lambon Ralph, 2010) and reference resolution (Brown-Schmidt, 2009), a positive result from the present research could open the door to exploring whether EF training may be an effective intervention tool for improving reading comprehension and verbal fluency, particularly in rare instances when multiple evidential sources do not conspire to guide or facilitate processing (i.e., when various representations compete, resulting in small but reliable consequences for production and interpretation; see Hussey & Novick, 2012). There could also be broader implications for clinical patients who suffer EF impairments that impinge on language-processing skills under conditions that place high demands on cognitive control resources (see Bedny et al., 2007; Novick et al., 2009; Robinson et al., 1998; Robinson, Shallice, & Cipolotti, 2005; Schnur et al., 2008; for a review, see Novick, Trueswell, & Thompson-Schill, 2010). Although the language deficits that such patients endure are relatively transient and arise under limited circumstances (i.e., they do not suffer from the full symptom-complex of agrammatism or other chronic aphasias), they are significant nonetheless and are characterised by a general-purpose cognitive control impairment with linguistic implications (see Novick et al., 2010). Provided the relevant findings of transfer to the language domain, EF training could become an important component of behavioural remediation for broad deficits in cognitive control that affect language use in these circumscribed situations.

Our training tasks were developed based on benchmark working memory tasks, some of which are known to elevate demands for cognitive control (e.g., by increasing the need to resolve among conflicting representations in memory). It is important to note that we administered a battery of working memory training tasks to test the possibility that performance improvements on specific ones might contribute differentially to improvements in syntactic ambiguity resolution (see Method and General Discussion). As highlighted in the work of Novick et al. (2005; see also D'Esposito & Postle, 1999; Novick et al., 2009, 2010), some working memory tasks rely on non-mnemonic capacity, which involves the EF ability to resolve conflicting (or interfering) representations—a general skill necessary for some linguistic tasks, like garden-path recovery. Here, the terms "conflict" or "interference" refer to circumstances in which an individual receives incompatible information regarding how best to characterise or respond to a stimulus when some input or new source of evidence automatically triggers an internal representation that competes with earlier

representations (see Botvinick, Braver, Barch, Carter, & Cohen, 2001). Such cases require EF—in particular *conflict resolution*—to rein-in initial mischaracterisations of the input. Not all working memory tasks necessarily share this conflict-resolution property to the same extent; thus, it is possible that only those training tasks that tax the need to resolve among competing alternatives and re-characterise information will predict performance improvements in syntactic ambiguity-resolution abilities across assessments.

We hypothesise that improved EF following training should generalise to real-time sentence processing and comprehension. Notably, transfer effects should be restricted to parsing conditions under high EF demands, namely when readers must revise an early parsing commitment after encountering new evidence that conflicts with their developing interpretation. Because EF training is hypothesised to transfer only to tasks requiring common underlying EF mechanisms, no training-related changes are expected under low EF demands, when reinterpretation is unnecessary.

Finally, we examine whether performance increases on some working memory training tasks (*n*-back, *letter–number sequencing* (LNS), *block span* and/or *running span*; see Method) contribute to changes in ambiguity-resolution performance more than others. Under a process-specific training account (see Dahlin, Neely, Larsson, Bäckman, & Nyberg, 2008), the amount of transfer to untrained tasks following intervention depends on the extent of overlap between the cognitive and neural resources shared by the training and the transfer tasks. As outlined above, garden-path recovery engages conflict-resolution processes supported by regions within left VLPFC. One of our EF training tasks, namely *n*-back with lures, has been shown previously to recruit regions within VLPFC owing to the interference generated by lure trials (see Gray, Chabris, & Braver, 2003; Jaeggi et al., 2003; Owen, McMillan, Laird, & Bullmore, 2005). Thus, a process-specific account predicts that only those EF tasks that recruit common areas in VLPFC for conflict resolution, such as the *n*-back task (see Method), should transfer to syntactic ambiguity resolution (Novick et al., 2005); EF training tasks involving exclusively other functions like maintenance and manipulation of information in verbal or spatial working memory absent conflicting representations (e.g., LNS; block span) should demonstrate little or no transfer. By including multiple training tasks that employ different components of EF to varying degrees in our training regimen, we can test whether EF training at the broadest level sufficiently improves garden-path recovery, or whether conflict-resolution training specifically is necessary to increase syntactic ambiguity-resolution abilities, thus informing a deeper understanding of the domain-general cognitive control mechanisms that contribute to sentence reinterpretation.

EXPERIMENTAL PRELIMINARIES

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

1. While the thief hid the jewelry that was elegant and expensive sparkled brightly. (temporarily ambiguous)
2. The jewelry that was elegant and expensive sparkled brightly while the thief hid. (unambiguous)

In (1), the ambiguity springs from the verb "hid", which can be used either reflexively (individuals can hide themselves), or transitively (individuals can hide objects). Here, the transitive interpretation is strongly supported due to the absence of a comma following "hid", which would impose the reflexive analysis (Ferreira, Christianson, & Hollingworth, 2001). The presence of a plausible object ("the jewelry") further supports the transitive interpretation (see Garnsey, Pearlmutter, Myers, & Lotocky, 1997). Hence, readers rapidly interpret the sentence to mean the thief is hiding the jewelry. This analysis, however, is ultimately unviable because "the jewelry" turns out to be the subject of a new clause ("the jewelry sparkled . . ."), not a direct object. Upon encountering late-arriving disambiguating evidence that conflicts with one's developing interpretation (". . . sparkled brightly"), readers must initiate cognitive control processes in order to re-characterise their initial representation of sentence meaning, i.e., to resolve the conflict and revise their misinterpretation (Novick et al., 2005, 2009). In (2), the reversed clause order unambiguously signals the reflexive analysis; consequently, reinterpretation is unnecessary and conflict-resolution and cognitive-control processes need not deploy. As noted above, another way to disambiguate (1) would be to simply add a comma following the verb ("While the thief hid, the jewelry . . ."). However, in order to provide sufficient room for accuracy improvement across assessments, we adopted the reversed clause order disambiguation in (2) to maximise the ambiguity effect between conditions, following prior work indicating nominally higher error rates when comparing (1) to this unambiguous construction, vs. the comma-disambiguation construction (see Exp. 3 in the study by Christianson, Hollingworth, Halliwell, & Ferreira, 2001).[1]

Participants answered questions that probed for lingering effects of misinterpretation, for example, "Did the thief hide himself?" Full reanalysis does not always occur in ambiguous cases, resulting in erroneous "no" responses (Christianson & Luke, 2011; Christianson, Williams, Zacks, & Ferreira, 2006). Importantly, all comprehension questions queried the correct (reflexive) interpretation, and *not* the initially conceived and consequently incorrect analysis (the transitive interpretation; e.g., "Did the thief hide the jewelry?"). This was designed as such to avoid vulnerability to memory effects, where error commissions (i.e., a "yes" response to "Did the thief hide the jewelry?") could be influenced by familiarity of the memory trace of the initial misinterpretation. In other words, even if a reader did correctly revise the sentence, they might respond erroneously because the question itself restated the incorrect transitive analysis, thereby reactivating the initial misinterpretation. Error commissions to such sentences could be furthermore shaped by plausible inferences or general world knowledge, since thieves are likely to hide jewelry. Instead, in order to correctly respond "yes" to our questions—for instance, verifying that the thief was hiding himself—readers actually had to override the initial, incorrect transitive interpretation

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

(that the thief was hiding the jewelry) to recover the alternative reflexive interpretation, which is a more straightforward indicator of garden-path recovery. Similarly, an incorrect "no" response to our questions signifies a lingering commitment to the early direct-object analysis and, thus, recovery failure.

Our domain-general EF training regimen may support controlled revision. Hence, we hypothesise that interpretation recovery—reflected by comprehension accuracy for ambiguous sentences—should improve following training. Such performance increases might be especially related to training tasks aimed at enhancing conflict-resolution abilities. No changes are expected in unambiguous cases where the need for cognitive control is removed.

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

Furthermore, training-related improvements in garden-path recovery processes—indexed by both online and offline measures sketched above—may depend to a greater extent on performance increases on some training tasks than others. Theoretically, the extent of improvement on a training task targeting conflict-resolution mechanisms might be especially likely to predict gains in garden-path recovery, because conflict-resolution processes are thought to help recovery of alternative parsing options when other sources of evidence have guided the parser towards an incorrect syntactic characterisation of the input (see Novick et al., 2005, 2009, 2010). In other words, enhanced conflict-resolution abilities may help readers better avoid misinterpretations by more rapidly countermanding early parsing decisions in real-time, reflected by more efficient changes in controlled revision processes (i.e., regressions) once a misanalysis has been discovered.

Finally, as highlighted by previous findings in the cognitive training literature, the ability to observe reliable effects of training across assessments hinges on whether individuals in the treatment group *actually improve* on the task(s) completed throughout the training regimen (see Chein & Morrison, 2010; see also Jaeggi, Buschkuehl, Jonides, & Shah, 2011). That is to say, there may be decisive differences within the group of trainees regarding the extent to which individuals respond positively to the regimen. Here, only those who successfully improve performance during conflict-resolution training are expected to transfer these benefits to untrained measures of garden-path recovery. An important approach to analysing this kind of study, therefore, is to differentiate training "responders" from "non-responders" (Chein & Morrison, 2010; Jaeggi et al., 2011). We do this in two ways: (1) by using training responsiveness to the various tasks as a continuous variable to test the relation between the amount of improvement during the regimen and the amount of pre-/post-improvement in ambiguity resolution (i.e., transfer)—a multiple regression analysis that is consistent with prior training work; and (2) by treating responsiveness to the various training tasks as a discrete variable, separating responders from

non-responders via well-established statistical clustering methods (see Fraley & Raftery, 2002, 2011) and comparing these groups' performance on the sentence-processing task (using accuracy and reading time data as dependent variables) to the untrained group, which had no inter-assessment training data to evaluate. We expected that only the responders would exhibit cross-Assessment garden-path recovery gains that are significantly greater than the other two groups. This latter analytic approach is novel for training studies. Generally speaking, the responders are the people of most theoretical and practical interest here.

To summarise, we hypothesise that:

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

METHOD

Subjects

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

Design

A double-blind pre/posttest design was used; accordingly, neither subjects nor experimenters knew subjects' condition assignments. Different moderators held training and Assessment sessions in separate labs so that the experimenter who collected the Assessment data was blind to the condition to which each subject had been assigned. Additionally, because subjects in the experimental and control

conditions never interacted, they were in principle blind to each other's condition and unaware of the differences between them. The experimental group visited the training lab for 20 one-hour sessions in the three-to-six weeks ($M = 4.9$ weeks) intervening pretest (Assessment 1) and posttest (Assessment 2). Importantly, training did not involve practising syntactic ambiguity resolution or reading of any kind. Thus, any demonstrated effects of transfer might reasonably be attributed to improvements in domain-general processes, rather than to extra experience practising linguistic- or syntactic-specific processes. Control participants received no contact during this interval (see Chein & Morrison, 2010; Jaeggi et al., 2008), but the interim between their assessments was also three to six weeks ($M = 5.1$ weeks), matched to the training group.

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

Training tasks

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

[2] All subjects also completed a third assessment, which occurred three months following Assessment 2 without additional training for the experimental group. Assessments 1 and 2 were of primary interest, as performance at Assessment 2 measured the immediate effects of training vs. Assessment 1 (Assessment 2 was completed approximately one week after trainees finished the regimen). Assessment 3 was included to evaluate maintenance of training effects primarily for the non-syntactic measures of cognitive function (i.e., the assessments of fluid and crystallised intelligence). We do not include Assessment 3 data, as they do not bear on our central hypotheses.

"match-it" (targeting the ability to match auditory and visual representations of a phoneme), "sound-replay" (targeting phoneme categorisation and discrimination) and "listen-and-do" (targeting the ability to follow auditory instructions).[3]

Four tasks were administered per training session for approximately 15 minutes each. Over the 20 sessions, each task was repeated 10 times, and task difficulty adapted dynamically to individual levels to keep participants continually on the threshold of their best performance. Task order was the same for all participants. We describe the four EF tasks briefly below.

N-back

Sets of 25 single letters were displayed serially and participants indicated by button press whether the current letter had appeared n items previously. For example, if given the sequence H-B-K-H in a three-back condition, the second H would be a "target"; in the sequence H-B-K-T, the T would be a "non-target" because it does not match the three-back stimulus, H. Our version of n-back was intended to train conflict/interference-resolution mechanisms by including "lure" trials—recently presented letters that occurred either immediately before $(n-1)$ or after $(n+1)$ the nth-back item (Kane, Conway, Miura, & Colflesh, 2007; see also Burgess, Gray, Conway, & Braver, 2011; Gray et al., 2003). For example, if given the sequence H-B-H-D-K in a three-back condition, the second H would be a lure (an $n-1$ lure) because it was a two-back, not a three-back, stimulus. Thus, subjects would have to respond "non-target" to this item. Because lures did not appear in the specified n-back location, participants had to override a tendency to respond based on familiarity alone and resolve the conflict between the correct representation and a familiar, but incorrect one (see General Discussion). Participants encountered three lure levels before n increased: no lures, $n+1$ lures only, and both $n+1$ and $n-1$ lures. Task difficulty increased when participants achieved at least 85% accuracy by first increasing lure level incrementally and then by increasing n. Task difficulty decreased if participants fell below 65% accuracy, again by first decreasing the lure level and then by decreasing n. Difficulty values reflected both the value of n and the lure level.

Running span

Anywhere from 12 to 20 letters were presented auditorily in a continuous stream. Each string ended unpredictably, after which participants immediately had to recall the last n items from a fleeting auditory memory store. Initially, $n=2$ and n increased after participants successfully satisfied the criteria for progression at each of three presentation rates: 1000, 750 and 500 ms. If a participant achieved a 100% accuracy on four successive trials, then presentation rate increased in increments of 250 ms to a maximum rate of 500 ms. If the presentation rate was already 500 ms, then n increased by 1, and the presentation rate slowed to 1000 ms. If mean accuracy dropped to 25% or below, then task difficulty decreased by slowing the presentation rate by 250 ms; if

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

the presentation rate was already 1000 ms, then *n* decreased by 1. Difficulty values reflected both the value of *n* and the presentation rate.

Letter–number sequencing (LNS)

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

Block span

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

Transfer task: Syntactic ambiguity resolution

We developed separate but complementary versions of the ambiguity-resolution task so that participants never saw the same materials across assessments. Twenty-four verbs that could be used both transitively or reflexively (e.g., "hid") were borrowed from Christianson et al. (2006) and were used to create 12 ambiguous and 12 unambiguous sentences per assessment (see examples 1 and 2 above). At each assessment, these 24 items were embedded within 90 filler sentences (borrowed directly from Christianson et al., 2006, personal communication), which did not contain syntactic ambiguities and sampled a variety of constructions to draw attention away from the ambiguity manipulation. This variety included transitive structures that resembled the experimental items but removed any critical temporary indeterminacy (e.g., "While the father prepared the burgers he covered them with pepper"; "The exterminator entered the school while the cockroaches scurried"). For each assessment, two lists were created: if an item in one list was ambiguous, it was unambiguous in its counterpart list. List administration was pseudorandom and counterbalanced across participants and assessments. Thus, each reflexive/transitive verb appeared only once per assessment. If a participant saw a particular verb in an ambiguous construction at Assessment 1, that verb appeared in an unambiguous construction at Assessment 2 in a different sentence. Hence, while verbs repeated, they did so only

across assessments and appeared in new contexts and ambiguous/unambiguous frames.

A comprehension question about the correct reflexive interpretation was presented following every sentence. For ambiguous items, this meant that the questions probed for lingering effects of ambiguity and thus failure to revise and arrive at the correct interpretation (Christianson et al., 2006). For instance, in order to correctly answer "Did the thief hide himself?", readers were forced to override the initially favoured transitive analysis. The same question was presented for the unambiguous versions of an item. Thus, for all ambiguous and unambiguous items, the correct response was "yes". However, correct "yes" and "no" responses were balanced across the 114 total items (ambiguous, unambiguous, filler) at a given assessment.

Apparatus

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

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

Each trial began with a fixation box in the position of where the leftmost character of the sentence would appear. Once a subject fixated this box, the sentence appeared automatically, replacing the fixation box; this procedure served as a trial-by-trial calibration check. Each sentence was presented in its entirety on a single line. Participants were instructed to read each sentence at a comfortable pace and press a button when finished to advance to the comprehension question, to which they responded "yes" or "no" via button press. Before the experiment, participants completed 10 practice trials to ensure that they understood the procedure. Total task time averaged 40 minutes (range = 25–50 minutes), including recalibration and a scheduled break.

ANALYSIS AND RESULTS

Analysis of the training data revealed that participants showed the expected improvement on the four in-house training tasks (average effect size, Cohen's $d = 1.7$). However, did training gains transfer to syntactic ambiguity resolution?

Accuracy data

Analysis

Using multiple regression, we examined the relation between individual training gains and cross-assessment improvement in syntactic ambiguity resolution with training task as a factor, to understand the nature of the *continuous* relation between training improvement and transfer on a subject-by-subject basis. Crucially, this analysis allowed insight into whether trainees' gains on certain intervention tasks significantly predicted performance gains in garden-path recovery.

Following the multiple regression analysis, we report cluster analyses that identified responders and non-responders on each of the four training tasks; we then entered these discrete responsiveness variables into multilevel mixed-effects models to test for Group-by-Assessment interactions, to determine if the responders' ambiguity-resolution improvements differed reliably from both the non-responders *and* the untrained controls, who provided no training data between reading assessments. We conducted multilevel mixed-effects models using R's lmer function (lme4 library, Bates & Sarkar, 2007) due to their appropriateness for handling categorical data (see Jaeger, 2008). All accuracy data were first transformed using an empirical (e)logit function to correct potential problems related to heterogeneity of variance (see equation 5 in the study of Barr, 2008). For clarity, untransformed data are reported and illustrated in the figures. (Transformations did not result in any change in data patterns or significance values.) Such mixed-effects models were used both to statistically evaluate any test–retest improvement in the conditions of interest and to examine whether any reliable differences emerged among the groups (responders, non-responders, controls; see below) in test–retest changes. For all statistical models, subjects and items were crossed as random intercepts (Baayen, Davidson, & Bates, 2008; Jaeger, 2008; Quené & van den Bergh, 2008). In each analysis reported, we evaluated whether both random slopes and intercepts improved the fit of the models. Corrected Akaike information criteria (AIC_C; see Burnham & Anderson, 2004) were used to determine whether the best-fitting model included random slopes. In every case, only random intercepts improved model fit (see AIC_Cs in Tables A1 and A2 in Appendix A); therefore, all models that we report in the main text exclude random slope terms (but see Appendix A for models where slope terms are included for comparison).

Results

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

Relating garden-path recovery improvement to training responsiveness: Multiple regression results. Because participants in a training group typically achieve different levels of training performance (cf. Chein & Morrison, 2010; Jaeggi et al., 2011), we investigated whether training gains on the four in-house intervention tasks—computed by subtracting subjects' first-session performance from their final-session performance—were related to individual levels of garden-path recovery improvement, an approach consistent with prior training studies. Given that the regimen targeted a range of EFs (see Method), this analysis also permitted scrutiny of the specific training tasks that reliably forecasted gains in ambiguity resolution, thereby providing insight into whether practising particular EFs contributed to increased sentence-reinterpretation abilities vs. others.

Entering training task as an independent factor while controlling for training gains, we conducted a multiple regression analysis testing for the continuous relationship

between performance increases on the four intervention tasks and post-intervention improvement in accuracy to comprehension questions following syntactically ambiguous sentences (the dependent variable). We ran separate models for ambiguous and unambiguous data because we maintained the a priori hypothesis that training-mediated differences should occur only in the high-conflict, ambiguous condition, whereas no such effects were expected in the unambiguous condition. Interestingly, as hypothesised, the *n*-back task was the only training task to result in a main effect of comprehension accuracy improvement on ambiguous items ($b = -.37$, $t(71) = -2.11$, $p < .05$; all other main effects: $ps > .16$). No training tasks accounted for such an effect in unambiguous items ($ps > .38$). Crucially, the interaction of training gains-by-training task nearly reached significance for the *n*-back task only ($b = .19$, $t(71) = 1.92$, $p = .05$; the analysis of covariance interaction terms for the remaining training tasks: $ps > .11$), indicating that accuracy improvement on ambiguous sentences depends on performance increases on this task in particular. (Notably, there was sufficient variance in how responsive the trained group was for the three non-*n*-back tasks; see cluster analyses below for LNS, block span and running span. Nevertheless, gains-by-task interactions still were not observed). An interactive relationship did not emerge for unambiguous items for any training task ($ps > .38$). Taken together, these results suggest that the greater improvement achieved through consistent practice with the *n*-back task, the more improvement achieved on a far-transfer task of syntactic ambiguity resolution. As hypothesised, we believe that this selective correspondence is due to shared processing attributes (i.e., conflict resolution) across *n*-back-with-lures and garden-path recovery (see General Discussion).

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

Because the multiple regression analysis revealed important individual differences in responsiveness on the *n*-back training task, we employed hierarchical cluster analyses to statistically separate individual trainees who showed gains on this task from those who did not. This served to confirm the relation between *n*-back training gains and improvements in garden-path recovery and ultimately allowed us to compare both subgroups to the untrained subjects (thus further probing the interaction observed in the regression analysis for this task separately). Cluster analyses are a novel approach to examining individual differences in training responsiveness, as prior studies have either tested for correlations between training gains and transfer effects (Chein & Morrison, 2010), or used a median split to define training responders and non-responders (Jaeggi et al., 2011) when evaluating how training variability relates to transfer success. In previous research, responders and non-responders have usually been analysed separately in terms of transfer effects, rather than being statistically evaluated against each other. Here, we entered the cluster-defined responsiveness variable into a larger mixed-effects model to allow the model to determine whether a Group-by-Assessment interaction provided the best fit of the garden-path accuracy data, where the Group factor contained three levels: responders, non-responders and untrained controls. In other words, the mixed effects comparisons allowed us to ascertain whether the responders' accuracy reliably

improved and if this improvement was significantly different from the other two groups.

N-back cluster analysis results: Identifying responsive trainees and non-responsive trainees. We identified individuals who responded well to the *n*-back task with a model-based cluster analysis using R's mclust function (mclust library, Fraley & Raftery, 2011), which implements maximum likelihood estimation and Bayes criteria to identify the number of naturally occurring clusters of subjects given the distribution of an outcome measure (see Fraley & Raftery, 2002). This analysis was conducted for *n*-back performance using training gains as the primary index of training responsiveness, where gains were computed by subtracting each participant's initial training-session performance from his or her final training-session performance (see also Jaeggi et al., 2011). This analysis identified two clusters of subjects—13 "responders" (6 women, $M_{age} = 22.4$ years; age range = 19–39 years, $M_{education}$: 14.6 years) and 7 "non-responders" (6 women, $M_{age} = 20.5$ years, age range = 18–34 years, $M_{education}$: 14.2 years). As shown in Figure 1, this particular responder/non-responder distinction demonstrates wide variability in terms of subjects' performance curves throughout the course of *n*-back training, an illustration that was confirmed by an analysis of variance (ANOVA). One-sample ANOVAs were conducted to compare the two clusters in

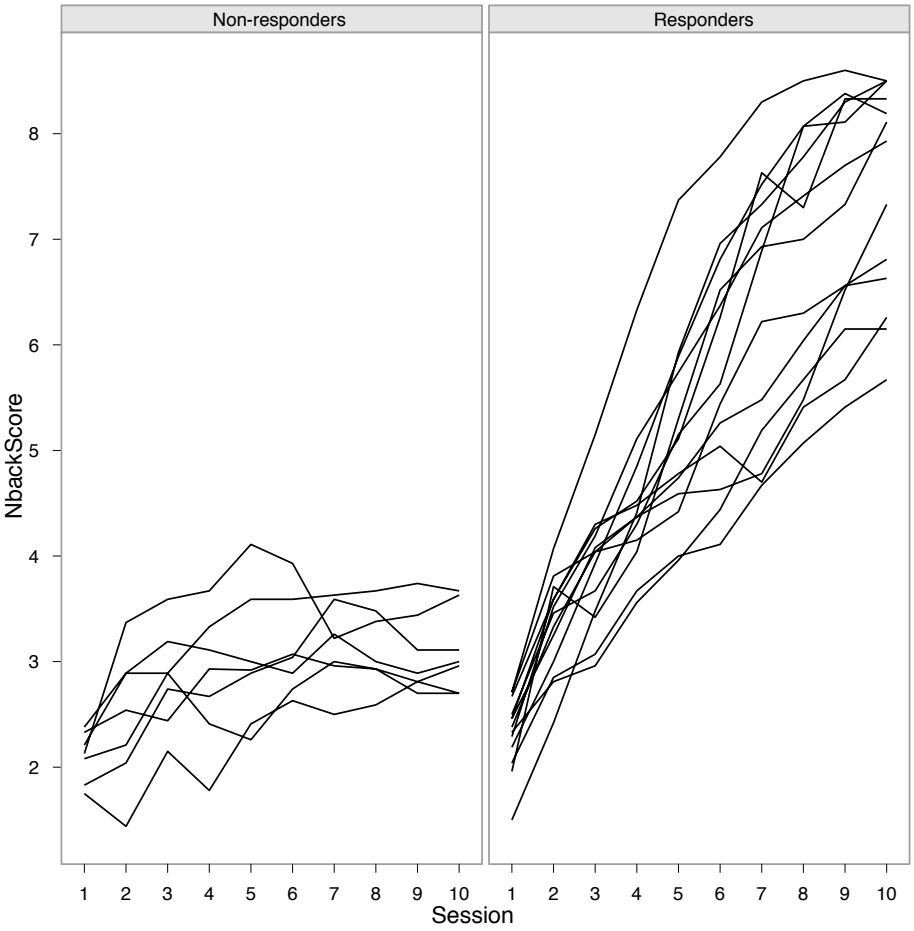

Figure 1. *N*-back performance curves by training session for responders and non-responders.

terms of improvement; in general, the clusters reliably diverged on a range of dependent measures, including (1) mean *n*-back training score over all 10 training sessions, (2) *n*-back gains from session-to-session as indexed by slope, (3) *n*-back gains from session 1 to session 10 (i.e., the measure by which the clusters were defined) and (4) final *n*-back session score. Importantly, responders did not differ from non-responders at the onset of training as reflected by session 1 *n*-back score (see upper panel of Table 1). Taken together, this suggests that the cluster analysis segregated subjects into two meaningfully, systematically and significantly different groups.

Garden-path recovery improvement in responders vs. non-responders vs. untrained controls: Mixed-effects model results. We compared test–retest performance across Assessments 1 and 2 to evaluate if the "responders" showed selective improvement in pre-/post-garden-path recovery performance that was statistically different from that of untrained subjects and "non-responsive" trainees who did not demonstrate gains on *n*-back. To do this, we fit the data for ambiguous and unambiguous materials in separate mixed-effects models with subjects and items as crossed random effects and both Assessment (1 vs. 2) and Group (*n*-back responders vs. non-responders vs. untrained controls) as fixed categorical factors, using the results of the above cluster analysis to determine the levels of each fixed Group factor. As we did for the multiple regression analysis earlier, we ran separate models for ambiguous and unambiguous data because we hypothesised a priori that training-mediated differences should occur

TABLE 1
Performance measures of responders and non-responders across the four training tasks. Groups were defined by a two-component cluster analysis (see text)

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

Notes: Block span responders and non-responders differ at training-session 1, and LNS responders and non-responders show only a marginal difference in average training performance. $^{\dagger}p < .06$, $*p < .05$, $**p < .01$, $***p < .001$.

only in the high-conflict, ambiguous condition, whereas no such effects were expected in the unambiguous condition. Again, a categorical independent variable was used in lieu of a continuous measure of training responsiveness because untrained controls had no analogous measure of inter-assessment gains (i.e., this group received no contact between assessments and therefore had no training data to contribute).

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

When analysing cross-Assessment changes in accuracy, there were main effects of both Assessment and Group and a significant Group-by-Assessment interaction for the

TABLE 2
Significant fixed effects from the best fitting mixed-effects models of comprehension accuracy data, testing for an Assessment (1 vs. 2) by Group (responders vs. non-responders vs. untrained controls) interaction separately for ambiguous and unambiguous items on each of the four training tasks

Significant model parameters	*Beta estimate*	*SE*	*Z-value*
	Task: n-back		
Ambiguous			
Intercept	0.79	0.31	2.59*
Assessment	−0.44	0.20	−2.16*
Group (responders)	1.05	0.51	2.07*
Assessment × group (responders)	−0.73	0.37	−1.98*
Unambiguous			
Intercept	2.82	0.32	8.92***
	Task: LNS		
Ambiguous			
Intercept	0.78	0.30	2.58**
Assessment	−0.42	0.20	−2.06*
Unambiguous			
Intercept	2.82	0.31	8.96***
	Task: running span		
Ambiguous			
Intercept	0.78	0.30	2.58**
Assessment	−0.42	0.20	−2.05*
Unambiguous			
Intercept	2.81	0.31	9.047***
Group (non-responders)	1.27	0.57	2.23*
	Task: block span		
Ambiguous			
Intercept	0.78	0.31	2.56*
Assessment	−0.42	0.20	−2.08*
Unambiguous			
Intercept	2.81	0.32	8.89***

Notes: When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and items were input into the models as crossed random effects. Excluding random slopes yielded better fits of every model, as indexed by lower AIC_C values for models without random slopes as compared to those with random slopes (see Table A1 in Appendix A). Thus the best-fitting models *without* slopes are reported here. *$p < .05$, **$p < .01$, ***$p < .001$.

ambiguous (*p*s < 0.05) but not the unambiguous sentences (see shaded panel of Table 2). To investigate this interaction further, we fit the data for each group separately, crossing subjects and items as random effects and including assessment (1 vs. 2) as a fixed factor. This revealed a significant main effect of Assessment for successfully trained subjects (i.e., *n*-back responders; $z = -3.68$, $p < .001$), but not for subjects in the untrained control condition ($z = -1.51$, $p > .13$) or the non-responder subgroup ($z = 0.83$, $p > .40$), such that only the intercept was reliable in the models for these two latter groups.

Crucially, there was no such interaction for unambiguous sentences across sessions; indeed, the best-fitting model included only the intercept suggesting that the fixed factors did not account for accuracy patterns in unambiguous items (see shaded panel of Table 2). Figure 2 illustrates the magnitude of accuracy change across assessments for each group on ambiguous and unambiguous sentences; as can be seen, responders' accuracy increases most on comprehension questions following ambiguous sentences (M = 16.67%), compared to non-responders (M = -3.57%) and untrained controls (M = 7.20%), who do not differ reliably in performance to ambiguous items across assessments. As expected, none of the groups demonstrate a cross-Assessment change in accuracy for unambiguous items; however, it is important to note that there was little room for change in this condition given that accuracy performance was near ceiling at Assessment 1.

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

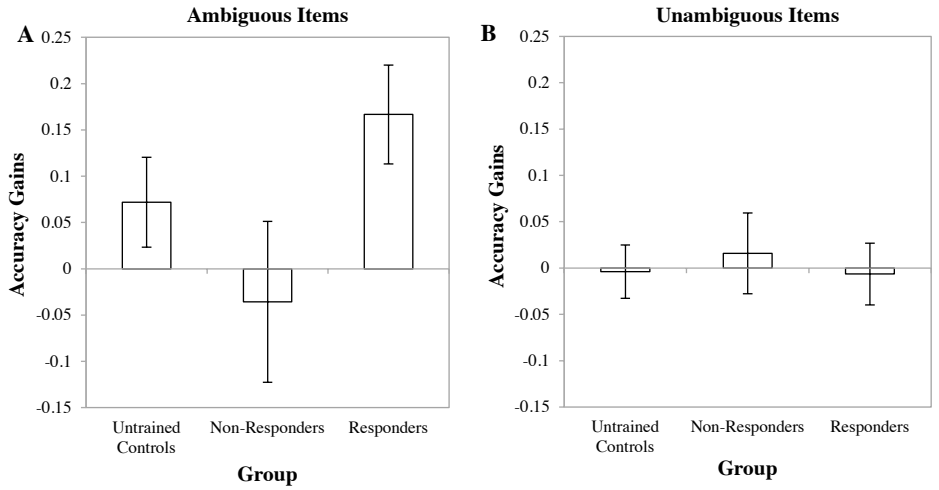

Figure 2. Change from Assessment 1 to Assessment 2 in comprehension accuracy rates split by Group (untrained controls, *n*-back non-responders and *n*-back responders). The large positive difference score for *n*-back responders (see text) reflects that this subgroup had significantly better accuracy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

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

To evaluate this possibility, we conducted additional cluster analyses to identify responders and non-responders on LNS, running span and block span. The results showed that there was in fact significant performance variability on these three training tasks, but that *n*-back responders did not necessarily also respond well to them, indicating that this subgroup was not selected for being better learners in general. Moreover, entering this responsiveness variable for the three other training tasks into mixed-effects models revealed that, as expected, increases on those tasks did not result in a Group-by-Assessment interaction regarding garden-path-recovery improvements, patterning with the non-significant contributions of each in the multiple regression model.

Responsiveness to other training tasks: Additional cluster analyses and mixed-effects models. We identified responders and non-responders to the three other training tasks (LNS, block span and running span) using the same model-based cluster analysis outlined previously for *n*-back. The goal of this analysis was twofold: (1) to determine if the group of responders identified for the *n*-back task *necessarily* comprises the same individuals who responded well to the other training tasks; and (2) to test whether responders on the other training tasks demonstrated a reliably greater improvement in sentence re-interpretation ability across assessments as compared to non-responders and untrained subjects. The second goal is particularly critical when entertaining a process-specific account of the present results, such that other tasks not designed to tap the EF of interest (conflict resolution) should confer little pre-/post-benefit to syntactic reanalysis, or resolution of incompatible interpretations.

A model-based cluster analysis yielded only a single cluster for the LNS and running span tasks (Ns $= 21$), and three separate clusters for the block span task (non-responder group A: $n = 2$; non-responder group B: $n = 8$; responder group: $n = 9$). Because the model-based cluster analyses of LNS and running span gains did not reveal distinct groups of subjects, we employed an alternative cluster-analytic method whereby the model must create a specific number of distinct clusters by maximising the distance between them, such that the most similarly performing individuals coalesce within a cluster. Since we aimed to identify two broad clusters of individuals (responders and non-responders), we used a two-component clustering approach. The bottom three panels of Table 1 show that, by and large, the responder/non-responder groups did not differ in performance scores at the first training session (except for block span), but did reliably diverge in terms of average training performance, final-session score, training gains and performance slope. Together this indicates that the

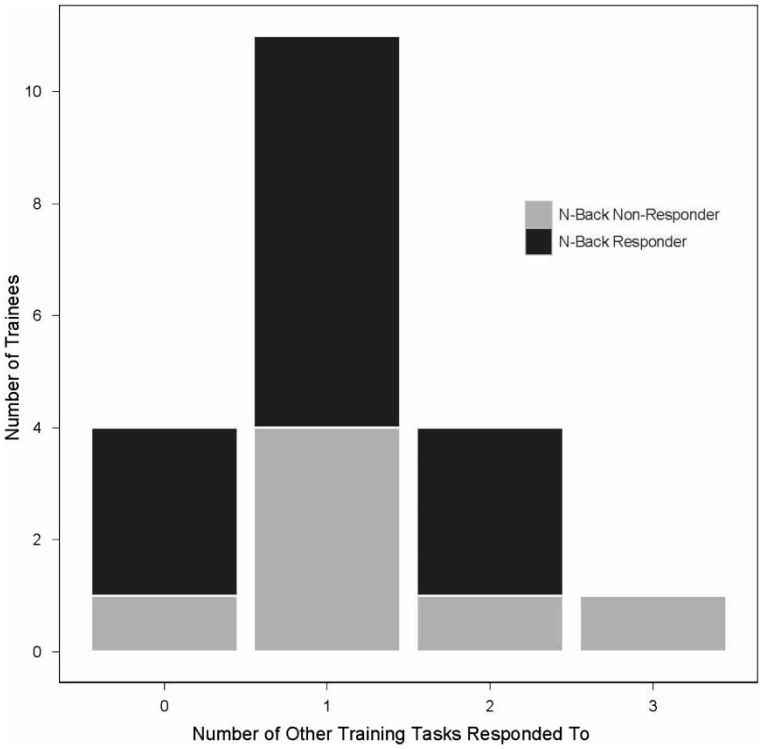

Figure 3. Responsiveness on the three other training tasks assuming *n*-back responsiveness**.** Responsiveness to each training task was defined by the output of a two-component cluster analysis (see text). Subjects varied in their ability to improve consistently across the other training tasks, irrespective of *n*-back gains, suggesting that *n*-back responders are not necessarily general-purpose learners. The *x*-axis depicts the number of other, non-*n*-back tasks that a trainee responded to based upon *n*-back responsiveness. It does not depict the *total* number of tasks to which a trainee responded. The "0" column, for example, shows that three *n*-back responders responded to 0 other tasks, and (only) 1 *n*-back non-responder did not respond to anything else.

two-component clustering approach consistently defined two groups of subjects for each task that differed significantly on various measures of training responsiveness.[4]

To address whether *n*-back responders were also classified as "responders" on other tasks in the training regimen, we tallied the number of training tasks on which each participant was considered a responder. Figure 3 illustrates trainees' propensity for general "responder" status given their *n*-back performance. First, it is important to note in the figure that, indeed, the clusters of responders and non-responders identified across the four training tasks *do not systematically overlap*, that is, subjects showing improvements on the *n*-back training task may not have performed well on the other training tasks. In fact, most *n*-back responders (54%) were considered responders on only *one* of the other training tasks (out of three), and *none* of the

[4] When applying this forced two-cluster approach to *n*-back performance data, the model identified the same individuals as responders ($n = 13$) and non-responders ($n = 7$) as previously categorised, replicating the model-based approach sketched in the main text above. This was also true for Block Span, except that the two subjects identified in the lowest-performing group were clustered here with the other non-responders into a single non-responder group, resulting in the following two clusters: responders ($n = 9$) and non-responders ($n = 10$).

n-back responders were responders on *all* training tasks (Figure 3). Moreover, three *n*-back responders (23%) were actually *non*-responders on LNS, block span and running span. This pattern suggests that the *n*-back/garden-path recovery relation does not simply index a superior ability to learn, reflected commonly across improved performance on these two tasks. The reason is that one would expect those with a better or faster capacity to learn to demonstrate this capacity across *all* tasks. Instead, as hypothesised and demonstrated below, the relationship observed between *n*-back gains and a significantly improved ability to resolve temporary syntactic ambiguity was selective, which we believe reflects a positive response to practising the EF functions common to *n*-back and garden-path recovery, rather than a sharper capacity to learn in general.

Second, we conducted multilevel mixed-effects models to test whether responders to the three other training tasks we developed (which tapped different EFs than *n*-back by design; see Method and General Discussion) showed reliably better garden-path-recovery improvement across assessments than untrained controls and non-responders to those tasks. If not, then this would confirm the selectivity of—and suggest a special status for—*n*-back training in terms of its ability to tap and improve those domain-general EFs that are shared with sentence re-interpretation. As before, we fit the data for ambiguous and unambiguous materials separately with subjects and items as crossed random effects and entered both Assessment (1 vs. 2) and Group (responders vs. non-responders vs. untrained controls) as fixed categorical variables for each of the three other training tasks, as identified by the various task-specific two-component cluster analyses.[5]

As can be seen in the three lower panels of Table 2, there was a reliable fixed effect of Assessment for ambiguous items for the three other training tasks (LNS, block span and running span), but no Group-by-Assessment interactions. Moreover, the best fitting models for unambiguous sentences included only the intercept for all tasks with the exception of running span (wherein non-responders differed from untrained controls and responders on accuracy to unambiguous sentences). Furthermore, as mentioned earlier, treating gains on these tasks as continuous variables in a multiple regression analysis—instead of forcing two categorical clusters of subjects—revealed that no task apart from *n*-back reliably predicted a relationship between garden-path gains and training gains. Taken together, the mixed-effects models and regression analysis may further reveal the importance of *n*-back training with lures, due to the conflict-resolution processes that are common to resolving temporary syntactic ambiguities.

Although we find this selectivity for *n*-back with lures, it is likely that this training task includes other working memory and EFs besides conflict resolution that are also shared with garden-path recovery, such as attention maintenance and memory updating. We merely wish to highlight that, whereas the other training tasks also involve attention maintenance and memory updating, *n*-back is the only task designed to target conflict resolution, which is why it is our task of interest. This is not intended to imply that *n*-back with lures recruits no other cognitive processes of relevance.

[5] First, we tested if the three groups (responders, non-responders, controls) for each of the remaining three training tasks differed in syntactic ambiguity-resolution performance at Assessment 1. Critically, in mixed-effect models that tested for a Group-by-Sentence-Type interaction at Assessment 1 (similar to what is reported above for the *n*-back task), only Sentence-Type improved the fit of each model ($ps < 0.001$). That the Group factor was absent from each best-fitting model ($ps > 0.26$) indicates no differences in garden-path recovery between responders, non-responders, and controls for LNS, Running Span, and Block Span prior to intervention.

Moreover, despite the lack of transfer from the three other training tasks, we cannot exclude the possibility that they did not contribute anything to the observed transfer effects. We return to these important issues in the General Discussion.

Eye movement data

Analysis

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

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

Because regression-path time is susceptible to exaggeration from eye movements to the left side of the screen—for instance, in preparation for a subsequent comprehension question—we truncated analysis at the final word of the sentence for any trial during which participants launched a leftward eye movement from that point, but did not then launch any rightward eye movements to continue re-reading the sentence. As our disambiguating region was always sentence-final (see Table 3, which defines the sentence regions), this truncation method was implemented to exclude extraneous regressions that were not associated with returning to later regions to continue processing the sentence.

TABLE 3
Sentences were divided into four regions for fine-grain analysis. Note that for ambiguous items, Region 4 is the critical disambiguating region

Sentence type	*Region 1*	*Region 2*	*Region 3*	*Region 4*
Ambiguous	While the thief hid	the jewelry	that was elegant and expensive	sparkled brightly.
Unambiguous	The jewelry	that was elegant and expensive	sparkled brightly	while the thief hid.

Given the evidence thus far that the *n*-back training task was the only one to yield performance differences that resulted in the relevant Group-by-Assessment interaction for the accuracy data, we mirrored the mixed-effects analysis for the eye-movement data to determine if regression-path durations patterned similarly. In other words, we report an analysis that compares reading behaviour from the performance subgroups on the *n*-back task, responders and non-responders, to the untrained controls. As with the accuracy data, we expected that responders' reading latencies (following entry into a conflict region of ambiguous items) would shorten reliably, whereas the other groups' reading behaviour would remain unchanged (the critical Group-by-Assessment interaction). Finally, responders to the three non-*n*-back tasks should demonstrate no significant change relative to non-responders and controls, concomitant with the accuracy findings.

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

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

Results

Relating real-time processing changes to training responsiveness: Mixed-effects models. Table 4 shows the results of MCMC simulations for all mixed-effects models that fit the total regression-path data from Region 4, which is the disambiguating region in ambiguous sentences. In unambiguous sentences, Region 4 was examined as a comparison, to match the region of analysis to the position of the critical region in ambiguous sentences, which necessarily contains different semantic content.

For ambiguous items, the model that included *n*-back responsiveness and Assessment as fixed effects (shaded panel of Table 4) revealed that an Assessment-by-Group interaction emerged ($t = 2.55$, $p < .05$), such that only *n*-back responders spent reliably less time passing this region at Assessment 2 vs. Assessment 1 (a difference of 640 ms), as compared to *n*-back non-responders (a 41-ms difference) and untrained subjects (a 57-ms difference) (see also Figure 4).

Additionally, in the comparable model of unambiguous items—i.e., for the final region (Region 4) of an unambiguous construction—no reliable fixed effects or interaction terms emerged ($p > 0.37$ for the Group-by-assessment interaction). More-

TABLE 4
Significant fixed effects from the best fitting mixed-effects models of regression-path time following entry into the final region of each sentence (which is the disambiguating region for ambiguous items, e.g., ''sparkled brightly''). The model tests for an Assessment (1 vs. 2) by Group (responders vs. non-responders vs. untrained controls) interaction separately for ambiguous and unambiguous items on each of the four training tasks

Significant model parameters	*Beta estimate*	*SE*	*t-value*
	Task: n-back		
Ambiguous			
Intercept	1675.24	219.12	7.645***
Assessment × group (responders)	616.70	241.44	2.554*
Unambiguous			
Intercept	809.102	55.397	14.606***
	Task: LNS		
Ambiguous			
Intercept	1656.7	230.7	7.182***
Unambiguous			
Intercept	804.890	58.918	13.661***
	Task: running span		
Ambiguous			
Intercept	1664.2	223.6	7.442***
Unambiguous			
Intercept	802.136	55.406	14.478***
	Task: block span		
Ambiguous			
Intercept	1650.91	218.38	7.56***
Unambiguous			
Intercept	809.461	55.231	14.656***

Notes: Markov Chain Monte-Carlo (MCMC) simulations were conducted to test for the significance of each fixed effect, through which we generated 10,000 samples from the posterior distribution. When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and items were input into the model as crossed random effects. Excluding random slopes yielded better fits of every model, as indexed by lower AIC_C values for models without random slopes as compared to those with random slopes (see Table A2 in Appendix A). Thus, the best-fitting models *without* slopes are reported here. *$p < .05$, **$p < .01$, ***$p < .001$.

over, only the intercept was included for the other models that examined regression-path data launched from *all* other regions of both ambiguous *and* unambiguous sentences, including Region 3 of unambiguous sentences ($ps > .17$), which contained the same semantic content as the critical region of ambiguous sentences (e.g., ''sparkled brightly''). (We did not create a table for the results of the mixed-effects models for the data in all other regions, but see Figure 4). Together this pattern indicates that the Group-by-Assessment interaction was selective for regression-path times stemming from the disambiguating region of ambiguous items, the region where conflict-resolution and controlled revision processes are hypothesised to engage. Importantly, at Assessment 1, mixed-effects models with Group as a fixed factor contained only the intercept as a reliable term for Region 4 of both ambiguous ($t = 7.08$, $p < .001$) and unambiguous sentences ($t = 13.97$, $p < .001$). That Group as a fixed effect did not significantly improve model fit suggests equivalent regression-path times across groups prior to training in this critical region (as well as in all other regions of both sentence types: $ps > .39$).

To further investigate the Group-by-Assessment interaction for ambiguous items, we fit the data for each group individually with subjects and items as crossed random

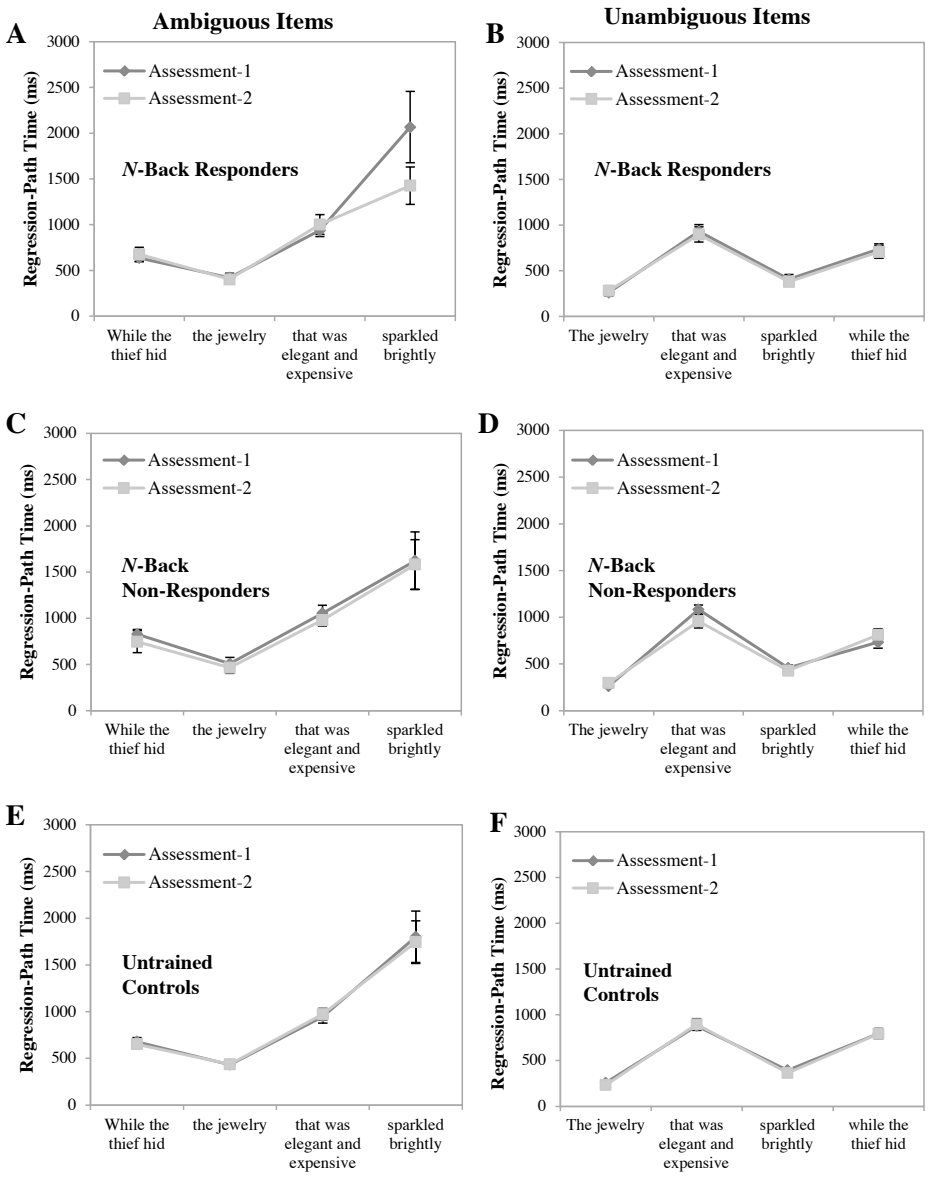

Figure 4. Trainees' and untrained controls' regression-path times across assessments launched from each sentence region for ambiguous and unambiguous items**.** To remove the potential impact of outliers, we eliminated from analysis any per-region regression-path times that fell 2.5 standard deviations above or below a subject's mean across all conditions. (This Winsorization procedure affected less than 2.4% of the overall data across all regions and sentence types.) Only correct trials were analysed (see text). The disambiguating region (e.g., "sparkled brightly") was the only region in temporarily ambiguous sentences from which *n*-back responders spent reliably less time regressing to previous material at Assessment 2 (A). *N*-back non-responders (C) and untrained participants (E) demonstrated no significant change across assessments for ambiguous items in any region. Assessment 2 regression-path durations for responders are nominally shorter than the other groups following entry into the "sparkled brightly" region (154 ms shorter than non-responders and 317 ms shorter than untrained controls). Crucially, for unambiguous items, *n*-back responders (B), non-responders (D) and untrained controls (F) demonstrated no change in regression-path time from any region between assessments. Error bars reflect ±SEM.

effects and Assessment (1 vs. 2) as a fixed factor. This revealed a main effect of Assessment for responsive *n*-back trainees ($t = 4.27$; $p < .001$), due to a drop in regression-path time associated with reading behaviour post-entry into the disambiguating region (and only the disambiguating region) across assessments; in contrast, only the intercept was reliable in models testing for the effect of Assessment in the disambiguating region among the untrained controls ($t = 1.69$; $p > .09$) and trainees who did not demonstrate improvement on the *n*-back task ($t = 0.31$; $p > .75$), suggesting that neither group showed reduced times at Assessment 2.

Responsiveness to other training tasks: Additional mixed-effects models. As can be seen in Table 4 (see also Table A2 in Appendix A), only the intercept was included in the comparable models evaluating performance of responders, non-responders and untrained controls on LNS, running span and block span for both ambiguous and unambiguous data. Thus, the Group-by-Assessment interaction failed to emerge for responders on the other training tasks, again indicating selective improvement for *n*-back responders and corroborating the interaction pattern we observed for the accuracy data (see three lower panels of Table 4). Said another way, the responders to each of the three other training tasks did not demonstrate any reading-time changes associated with the disambiguating region (or any other region) across assessments that were significantly different from the non-responders and untrained control subjects.

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

INTERIM DISCUSSION OF EYE-MOVEMENT DATA

Because regression-path analyses were conducted on correct trials only, the pattern of results suggests that when *n*-back responders arrive at the correct interpretation, they are, as a group, doing so in less time than they did before training, decreasing the duration associated with regressing out of the region of conflict and eventually reading past it (but, see subject-by-subject analyses in Appendix B, which test for relations between accuracy and regression-path durations on an individual level). *N*-back responders' improved accuracy and shorter regression-path times after encountering disambiguating evidence may reflect better controlled revision following training: when confronted with new evidence that conflicts with developing interpretations, readers who undergo EF training (specifically, those who respond well to training on the *n*-back-with-lures task) spend less time regressing to earlier material in order to recover successfully from their misanalysis, effectively gathering information more quickly to arrive at the correct sentence meaning. Although it would be difficult, in our opinion, to attribute *n*-back responders' cross-Assessment changes launched from the disambiguating region to a better capacity to learn in general—because the changes occur following entry into the sentence region where information re-characterisation is precisely hypothesised to deploy—this learning interpretation is further discounted by the specificity of the interaction to *n*-back responders alone.

This again suggests that selective improvement in reading behaviour following entry into the disambiguating region is likely attributable to a positive response to consistent training on the *n*-back task, which shares EF processes with syntactic ambiguity resolution (see General Discussion).

Finally, given that the disambiguating region is always the final region of our ambiguous experimental sentences, another potential explanation for the present regression-path-time findings is that training attenuates "wrap-up" effects. These are typically marked by longer sentence-final or clause-final reading times and have been attributed to clausal integration rather than reanalysis (Just & Carpenter, 1980; Rayner, Kambe, & Duffy, 2000; but see Warren, White, & Reichle, 2009 who argue that construction difficulty does not drive such effects). We acknowledge that the eye-movement patterns associated with the final region might be affected by wrap-up as well as reanalysis. However, even if the eye-movement patterns are affected by sentence-final wrap-up, our critical results are specifically related to ambiguity, as the Group-by-Assessment interaction was not found for the sentence-final region of the unambiguous items (see text above as well as Table 4 and Figure 4). In other words, if *n*-back training affected readers' ability to initiate sentence-final wrap-up, as opposed to their ability to deal with temporary ambiguity, then presumably such effects would have been found in unambiguous items as well.

GENERAL DISCUSSION

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

Overall, we provide further supporting evidence that syntactic ambiguity resolution depends on domain-general cognitive control mechanisms, even in non-clinical populations. Our findings are a particularly strong addition to this line of research because most of the previous evidence has merely correlated neural activation patterns in response to cognitive control and garden-path tasks, or depended on pre-existing neuropathology to demonstrate a common deficit in syntactic and non-syntactic cognitive control abilities. Here, we directly manipulated EF through training to test its effect on syntactic ambiguity-resolution processes in neurologically intact adults. For those subjects in whom we successfully increased conflict resolution and cognitive control, we observed improvements in online and offline measures of garden-path recovery. For the first time, we show that garden-path recovery abilities are malleable in healthy young adults such that they can be improved by training the underlying EFs critical for revising misinterpretations, not just via practice with specific instances of ambiguous sentences. Notably, as hypothesised, performance increases on working memory tasks without the critical conflict-resolution feature are apparently insufficient to produce gains in syntactic ambiguity resolution (but may be necessary; see limitations and caveats below). The increases in syntactic ambiguity-resolution performance are closely related to *n*-back gains, importantly, and not due to preexisting differences in either garden-path recovery ability (all three groups performed equivalently at Assessment 1) or *n*-back ability (responders and non-responders performed equivalently at the first session of *n*-back training). Finally, our results speak to the role of regressive eye movements in garden-path recovery,

corroborating the notion that the regression-path measure in eye tracking may reflect the initiation of revision processes during real-time processing of syntactic ambiguity. Unique in this contribution is the result that readers may become less dependent on re-reading upon encountering conflicting evidence as their cognitive control becomes more efficient.

In particular, we demonstrated that subjects in the training group who achieved the greatest gains on the *n*-back training task, but not the other training tasks, subsequently showed better success at Assessment 2 (vs. Assessment 1) in recovering the correct alternative interpretation of temporarily ambiguous sentences susceptible to misanalysis. Furthermore, compared to the untrained group, and those in the training group who did not demonstrate consistent gains on *n*-back, the cross-Assessment performance change was reliably larger for the most successful trainees (responders), suggesting significantly increased accuracy to ambiguous items following the training regimen.

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

We have noted the selective nature of our training results several times, specifically that transfer benefits are observed (1) only for ambiguous sentences when examining comprehension accuracy and (2) only where revision processes are expected to be triggered when looking at regression-path time. Of course, with error rates of approximately 10% at Assessment 1 for unambiguous items, one could argue quite reasonably that such high performance was near ceiling (i.e., 90% accuracy), and therefore no change would be anticipated—or perhaps even possible—at Assessment 2. Moreover, although changes in eye-movement patterns following training occurred selectively after entering the disambiguating region of ambiguous sentences, it could be argued that training mediated the efficiency of handling processing difficulty, which would be greater in ambiguous than unambiguous sentences, rather than reinterpretation abilities per se. We must therefore be cautious in concluding that selective improvement in conflict resolution was the reason that our training regimen benefitted comprehension and real-time processing of ambiguous but not unambiguous sentences.

However, there are several reasons leading us to believe that the selectivity of our results is not due to ceiling effects. First, closer inspection of the error rates for unambiguous items at Assessment 2 shows some degree of individual variability. Average error proportions ranged from .059 to .11 across groups, with non-responders having the lowest error proportions after training, suggesting that individuals vary in their performance for these items at least nominally and may have some room to improve. Although one could potentially make the "ceiling" argument against our regression-path time results, it is nevertheless somewhat difficult to evaluate what ceiling performance might be. Certainly readers are bound to a lower reading-time

limit by how fast their eyes can move. But, judging by the sometimes-long regression-path times in Figure 4 (in both ambiguous *and* unambiguous sentences, depending on the region), it is uncertain that readers could not have read anything in less time at Assessment 2. Indeed, the figure reveals that aside from "sparkled brightly" (the disambiguating region in ambiguous sentences), reading times were quite similar between ambiguous and unambiguous sentences, when considering regions with the same content (as opposed to regions with the same position in the sentence). Notably, average regression-path times in the adjectival clause (e.g., "that was elegant and expensive") were as long as one second in both sentence types, which seems to allow ample room for improvement, especially given that regression-path time is computed only when subjects regress out of the region, which need not occur at all. We believe, consequently, that it is rather informative that there were no test–retest differences in any region of unambiguous items regarding regression-path times for any group. It is equally informative that there were no test–retest differences in any region *except* following entry into the disambiguating region for *n*-back responders. Together, we believe these data patterns suggest an element of sensitivity and selectivity, such that performance improves only in regions associated with syntactic conflict.

The potential importance of a process-specific training approach

Why was the *n*-back task particularly critical in capturing training and transfer success across both dependent measures of interest, namely (1) accuracy to questions probing for persistent effects of misanalysis and thus a failure to revise; and (2) regression-path times launched selectively from the disambiguating region? One explanation is that *n*-back gains, and only *n*-back gains, were related to garden-path-recovery improvement because of the controlled processing needed to resolve among the conflicting representations generated by interference lures. In a standard *n*-back task, participants can rely on familiarity to correctly identify which letter is a target. However, the introduction of lures after participants reached a certain performance criterion forced trainees to rein in such a familiarity bias; when encountering a lure, they instead had to initiate conflict/interference-resolution processes to successfully override familiarity-based evidence and re-characterise the stimulus as familiar but not in the relevant *n*-back location. Prior work has highlighted such an information re-characterisation function as crucial for resolving syntactic conflict as well; during parsing, domain-general conflict-resolution processes engage when individuals encounter input (e.g., "sparkled brightly") that is incompatible with their developing analysis (see January et al., 2009; Novick et al., 2005, 2009, 2010). When a reader comes across such conflicting evidence and discovers the misinterpretation, he or she must "slam on the brakes" and deploy conflict-resolution processes that allow for a re-characterisation of the current representation of sentence meaning, and for finding the correct, intended alternative.

Moreover, the *n*-back task has been shown to recruit posterior regions of the left inferior frontal gyrus (LIFG; BA 44/45) within VLPFC during high-conflict (lure) trials (Gray et al., 2003); this patch of cortex is routinely identified as the crucial neural underpinning of conflict/interference resolution in working memory (see Jonides & Nee, 2006) and has been implicated in cognitive control during sentence reinterpretation in both patient and neuroimaging studies (January et al., 2009; Novick et al., 2005, 2009, 2010; Ye & Zhou, 2009). In fact, the lure version of our *n*-back task is quite reminiscent of the working memory assessment—the "recent probes" item recognition task—that was used to diagnose a conflict-resolution

impairment in the patient with VLPFC (indeed, LIFG) damage described in the Introduction (see also Hamilton & Martin, 2005). This patient's deficit extended to a failure to override syntactic misanalysis and recovery from misinterpretation (Novick et al., 2009). In the recent-probes task, subjects responded to a probe (e.g., *D*) regarding whether it appeared in an immediately prior memory set (e.g., *s f d m*) (see also Jonides, Smith, Marshuetz, Koeppe, & Reuter-Lorenz, 1998; Monsell, 1978; Thompson-Schill et al., 2002). Although subjects could frequently use stimulus familiarity to judge correctly—yes or no—whether the probe had appeared or not, a small subclass of "no" trials introduced conflict and therefore susceptibility to error if one relied on a familiarity bias alone. On such "conflict" trials, the probe (e.g., *H*) did not appear in the directly preceding memory set (e.g., *k p w n*), so the correct response was "no", but it had been seen one trial earlier (e.g., *h l w p*). As such, these so-called "recent-no" trials, akin to our lures in the current study, exploited lingering familiarity of the probe owing to its recent presentation; subjects therefore had to override a dominant familiarity bias because it might yield an incorrect "yes" response, and instead re-characterise the probe representation as "familiar-but-irrelevant". The patient's unusually high error rate under such conditions, compared to "non-recent-no" trials (where there was no interference from the preceding memory set), identified a selective conflict-resolution impairment, which affected his parsing abilities under similarly circumscribed conditions. In particular, he demonstrated a failure to revise (or re-characterise) early parsing misanalyses and recover an alternative interpretation of sentence meaning when there was conflict between two incompatible syntactic representations (Novick et al., 2009; for convergent neuroimaging data see January et al., 2009; Ye & Zhou, 2009; for a review see Novick et al., 2010).

Thus, the linking assumption is that the need for conflict resolution seems to be shared across a range of tasks, including garden-path recovery and working memory tasks that manipulate such demands, for instance the "recent-no" trials of the item-recognition task and the *n*-back task with lures. Consequently, in our study, subjects who showed consistent performance increases on the *n*-back task—reflecting an enhanced ability to resolve among conflicting representations—demonstrated concomitant increases in garden-path recovery performance, likely because of the common conflict-resolution process that was targeted through training. We reiterate that the other training tasks (except possibly running span, see below) were designed explicitly not to tap this function by excluding any manipulation of demands for information re-characterisation. Finally, prior research exploring the utility of similar conflict-resolution training tasks demonstrates far-transfer to other language measures that tap cognitive control resources supported by the posterior LIFG (e.g., in a transient manipulation of cognitive "fatigue"; see Persson, Welsh, Jonides, & Reuter-Lorenz, 2007). The illustration of transfer to syntactic ambiguity resolution in the current study may be considered an extension of that finding. Overall, our results are consistent with earlier studies that tie specific language-processing abilities to domain-general cognitive control skills that putatively recruit regions within posterior LIFG (though see caveats below for further discussion).

Additionally, our account is consistent with other psycholinguistic models that associate language-processing difficulty with the need to adjust the activation of multiple representations (or interpretations) when various sources of evidence temporarily conflict. Certainly, several explanations of syntactic complexity effects hinge on ambiguity-resolution functions that could be couched in cognitive control terms (Gennari & MacDonald, 2008; Van Dyke & Lewis, 2003). In one model (Van Dyke & Lewis, 2003), for instance, an individual's failure to recover from an initial

misinterpretation is linked to retrieval interference in memory—indeed, the field wherein Jonides and colleagues (Jonides et al., 1998; Jonides & Nee, 2006) first discovered the role of posterior LIFG in cognitive control, by asking how frontal lobe activation changes when interference/conflict enters the working memory system. Similar parsing accounts relate memory interference (or "cue overload") to difficulty in language comprehension, particularly when retrieval cues are unable to differentiate among competitors (see Van Dyke & McElree, 2006). As we have suggested, our *n*-back training task involves recognition cues that cannot easily distinguish between competitors (i.e., lures/targets), and thus give rise to interference effects. An improved ability to deal with such competition might account for responders' attenuated processing difficulty, because of a transferred ability to handle the activation of multiple conflicting representations of the linguistic input (though see limitations and caveats below for further discussion).

One reason that improved performance on the other (non-*n*-back) training tasks, by contrast, failed to predict improvements in garden-path recovery may be because they did not, in their design, expressly involve conflict/interference resolution. The EFs tapped in these tasks included the manipulation and storage of visual–spatial information (block span) and the reorganisation of alphanumeric stimuli in working memory (LNS), which may be theoretically harder to connect to syntactic processing and recovery from misinterpretation specifically. In addition, demands for information re-characterisation were intentionally not parametrically or dynamically manipulated in these tasks as they were in *n*-back. Of course, verbal working memory is apt to play a role in garden-path recovery during spoken language comprehension, and even in reading studies that use a moving window paradigm where readers cannot review the input once it has past (see, for instance, Fedorenko, Gibson, & Rohde, 2006). Future work should test the relative impact of working memory training (including auditory working memory training)—with and without conflict-resolution aspects—on garden-path recovery using alternative experimental paradigms, in both the reading and spoken domains. Next, although the LNS task involved reordering verbal information in working memory, which is likely involved in garden-path recovery to some extent, this training task required the repeated application of a specified rule in a predictable manner (always sorting numbers in ascending order and letters alphabetically), which may have been too superficial to involve any deeper re-characterisation of representations that is necessary for revising misinterpretations.

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

One result, however, that may be surprising is that responders on the running span training task did not demonstrate reliable garden-path recovery improvements, as this task *can*—depending on design—involve updating and incidental proactive

interference from earlier items, processes that rely on regions within left VLPFC (cf. Postle, 2003; Postle et al., 2001). In running span (Pollack, Johnson, & Knaff, 1959), a sequence of an unpredictable number of items (e.g., letters) is presented, after which the last *n* items must be suddenly recalled. Hockey (1973) showed that presentation rate can dramatically alter the nature of the task (see also Bunting et al., 2006). When item-presentation rate is fast (e.g., three items/s), the task is conducive to a lower-effort strategy in which items are passively held until the list ends (i.e., when retrieval from a capacity-limited attentional store can occur). With a slower presentation rate (e.g., one item/s), participants can—when they are explicitly instructed—adopt a higher-effort strategy in which working memory is continually updated through rehearsal (Hockey, 1973). Thus, during retrieval in the slow-rate procedure, individuals must resolve interference from earlier stimuli that are concurrently being maintained. Questionable, however, is whether anyone would spontaneously adopt a higher-effort strategy when presentation rates change within an experiment, as in ours: based on Hockey (1973) and Bunting et al. (2006), active updating becomes increasingly difficult and almost impossible at fast presentation rates, such as our 500-ms condition. Although active updating is possible in the 1000-ms condition, it is unlikely that participants would switch strategies in alternation (see Bunting et al., 2006). The faster rate used in our task may have therefore eliminated the interference-resolution aspect of the task, explaining why running span did not predict garden-path recovery improvements. Moreover, there may be an important distinction here between the *attention control* processes needed for running span—in which listeners must rapidly *collect* information from a fleeting sensory memory store—and the *cognitive control* processes needed for *n*-back with lures—in which subjects must *re-characterise* information given conflicting internal representations.

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

Limitations, caveats and future directions

One limitation of the current study is the absence of an "active" control group that also comes into the lab and practices tasks without the crucial EF elements (namely, conflict resolution). Including such a group in future work would address the issue of whether demand characteristics or motivational factors alone are driving our effects. For instance, it would be informative to test whether those who practise *n*-back without the interference-lure component do not show improvements in garden-path recovery, thus providing greater confidence in the claim that this particular EF is at the heart of the observed increases in trainees' reinterpretation abilities. Such a contrast—*n*-back with lures vs. *n*-back without lures—would help isolate the locus of the current training results. Despite this drawback, we note the specificity of our transfer findings to (1) key individuals and (2) key features of the reading task that

require EF, specifically conflict resolution, which we believe together exclude purely placebo or motivational explanations. First, transfer was observed specifically from *n*-back responders as opposed to non-responders, untrained controls and other types of responders (i.e., LNS, block span, and running span responders). Second, *n*-back responders' overall accuracy to comprehension questions improved for ambiguous but not unambiguous items, suggesting that the offline effects were restricted to cases when readers had to revise interpretations (i.e., resolve among conflicting representations; see Novick et al., 2005). The same pattern held for regression-path times associated with regressions from key conflict regions of ambiguous sentences. Given that the control group and *n*-back non-responders (as well as the responders to the three other training tasks) did not show these effects, we are confident that the observed findings cannot be attributed solely to practice. Indeed, we observed important individual differences in training achievement, which corresponded to successful transfer to syntactic ambiguity resolution: only the *n*-back responders demonstrated significantly greater improvements in both online and offline measures than the untrained subjects, despite non-responders having the same amount of training as those who responded well.

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

As noted earlier, the convergent findings that improvements were limited to *n*-back responders essentially rules out the possibility that trainees' enhanced ambiguity resolution was the result of a generally better capacity to learn, which could have, in theory, brought about the common improvements found for *n*-back and garden-path recovery. However, additional analyses confirmed that *n*-back responders did not necessarily respond equally well to the other training tasks (clusters of responders and

non-responders did not overlap across training tasks). Thus, these individuals were not selected on the basis of a generally greater ability to learn or motivation to perform well. In addition, the groups of responders identified for LNS, block span and running span did not demonstrate improved interpretation-recovery abilities at Assessment 2 in accuracy or regression-path times compared to those tasks' non-responders and untrained subjects. Thus, we believe that the most parsimonious interpretation of these data is that *n*-back training improved conflict-resolution functions, which resulted in improved sentence re-interpretation at Assessment 2. That responders to the three other tasks, moreover, did not outperform non-responders and untrained controls argues strongly against the possibility that mere practice effects are at the heart of our findings.

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

Having said this, we reiterate a similar point made earlier: we firmly believe that traditional "span" functions such as storage, processing and maintenance factor into language interpretation irrespective of ambiguity or conflict, for instance in spoken comprehension tasks or in moving-window reading paradigms where the demands for mnemonic properties of working memory are high (see, e.g., Fedorenko et al., 2006). Thus, cognitive control—a non-mnemonic aspect of some working memory tasks—is just one explanation for what is shared across *n*-back with lures and syntactic ambiguity resolution; this does not necessarily preclude the likelihood that other working memory processes involved in *n*-back (e.g., updating) are also affecting performance. This perspective is correspondingly apt if the other non-conflict training tasks are contributing *something* to the current results, which again, we cannot rule out entirely. Indeed, we do not claim that cognitive control is the only contributor to the observed findings, but rather that conflict-resolution processes are a necessary aspect of the current training–transfer relation, given the evidence that such functions are critical to syntactic ambiguity resolution (Novick et al., 2009). Again, we selected *n*-back with lures as our task of interest because, relative to the three other training

tasks, it was the only one designed to target conflict-resolution processes. This is not meant to imply that *n*-back targets *only* conflict resolution.

In addition, we have relied on parsimony to interpret our data in terms of conflict resolution, consistent with neuroscience studies showing an important role for posterior areas within the LIFG in both syntactic and non-syntactic cognitive control (e.g., January et al., 2009; Jonides & Nee, 2006; Novick et al., 2005; Ye & Zhou, 2009). However, an important caveat is that these same conflict-responsive VLPFC regions can be involved in a number of other cognitively demanding tasks (see, e.g., Duncan, 2010). In other words, patches of cortex within VLPFC that are recruited for conflict resolution are unlikely to be involved *uniquely* in conflict-resolution aspects of cognitive control. Therefore, other cognitive functions, as discussed above, cannot be excluded as contributing to training gains on the basis of neuroimaging data alone. In sum, although a preponderance of behavioural and neuroimaging data points to a role for cognitive control in garden-path recovery, this does not mean necessarily that there exists a neural dissociation between conflict resolution and other EFs. It does suggest, however, the importance of using multiple methods to offer converging data for a particular hypothesis. Our data certainly fit with converging neurocognitive evidence from both clinical and healthy populations, providing another important approach that yields findings compatible with the cognitive control account. Nevertheless, the caveats outlined here suggest that, until both behavioural and neural dissociations are demonstrated, one can conclude only that the training–transfer results are consistent with the notion that cognitive control is the mediating ability across *n*-back-with-lures and syntactic ambiguity resolution. Other working memory and EFs certainly remain as possible contributors. To what extent they are contributing is an open empirical question and must be addressed in follow-up research using designs as those sketched above.

Before closing, we wish to reiterate that, in concert with prior research groups, we observed that individual differences in training improvement may be key to shaping successful transfer; only those who improved significantly on the *n*-back task exhibited performance increases on the various measures of syntactic ambiguity resolution, as compared to the untrained control group and *n*-back non-responders. Although training can clearly be valuable, individual differences in success might be considered a restraining ingredient. Follow-up work should study what training protocols yield the best transfer, what cognitive and motivational factors determine who is most likely to benefit (and why), and how best to maximise effectiveness in training to ensure value across a range of different groups (see Jaeggi et al., 2011 for further discussion on these issues).

Closing remarks

In sum, these results are among the first to establish training-related transfer to the linguistic domain, and extend theoretical and empirical work highlighting the role of general-purpose cognitive functions in language processing. Certainly, the more language-specific experience readers have, the better they cope with difficult linguistic input; indeed, two studies have reported that consistent practice reading complex or ambiguous material results in (i) an improved ability to process the constructions that were routinely repeated and (ii) a transfer effect, such that practice generalises to previously unseen difficult constructions (see Long & Prat, 2008; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). We believe that our findings complement this notion by demonstrating that domain-general cognitive abilities, even for healthy

adults, may be a causal factor in sentence reinterpretation abilities. This conclusion is especially warranted given that our subjects had minimal exposure to the ambiguous sentences—just 12 per assessment, embedded within several fillers—which was probably insufficient to produce a reliable practice effect. This also seems a reasonable conclusion alongside our finding that untrained controls and non-responders, who had the same amount of practice, failed to improve reliably across assessments. Moreover, given that EF plays a role in a range of other specific language-processing skills including lexical ambiguity resolution, common-ground assessment and verbal fluency, another implication of the current findings is that EF training, within a well-considered process-specific framework, could result in broader improvements beyond just garden-path recovery. Pending further research, such training might be used in behavioural remediation programmes that aim to improve language skills in situations when competitive interactions are high. Finally, EF training, especially with a conflict-resolution focus, might yield similar benefits in clinical populations—particularly left VLPFC patients—whose language production and comprehension fail under selective conditions due to poor cognitive control.

Manuscript received 31 January 2012
Revised manuscript received 8 November 2012
First published online 16 January 2013

REFERENCES

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D. J. (2008). Analyzing "visual world" eye-tracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457–474.

Bates, D. M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes. R package version 0.9975-12. Retrieved from http://cran.r-project.org/.

Bedny, M., Hulbert, J. C., & Thompson-Schill, S. L. (2007). Understanding words in context: The role of Broca's area in word comprehension. *Brain Research Special Issue Mysteries of Meaning*, *1146*, 101–114.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652.

Brain Fitness Program (Version 2.1) [Computer software]. San Francisco, CA: Posit Science.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*(5), 893–900.

Bunting, M. F., Cowan, N., & Saults, J. S. (2006). How does running memory span work? *The Quarterly Journal of Experimental Psychology*, *59*, 1691–1700.

Burgess, G. C., Gray, J. R., Conway, A. R., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General*, *140*, 674–92.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304.

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review*, *17*, 193–199.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic-roles assigned along the garden path linger. *Cognitive Psychology*, *42*, 368–407.

Christianson, K., & Luke, S. G. (2011). Context strengthens initial misinterpretations of text. *Scientific Studies of Reading*, *15*(2), 136–166.

Christianson, K., Williams, C. C., Zacks, R. T., & Ferreira, F. (2006). Misinterpretations of garden-path sentences by older and younger adults. *Discourse Processes*, *42*, 205–238.

D'Esposito, M., & Postle, B. R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia*, *37*, 1303–1315.

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science*, *320*, 1510–1512.

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

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behavior. *Trends in Cognitive Sciences*, *14*(4), 172–179.

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*, *54*, 541–553.

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, *30*, 3–20.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611–631.

Fraley, C., & Raftery, A. E. (2011). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. R package version 3.4.10. Retrieved from http://www.stat.washington.edu/research/reports/2012/tr504.pdf.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *2*(4), 291–325.

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

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*, 58–93.

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, *58*, 161–187.

Gray, J. F., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, *6*, 316–322.

Hamilton, A. C., & Martin, R. C. (2005). Dissociations among tasks involving inhibition: A single case study. *Cognitive, Affective, Behavioral Neuroscience*, *5*, 1–13.

Haut, M. W., Kuwabara, H., Leach, S., & Arias, R. G. (2000). Neural activation during performance of number-letter sequencing. *Applied Neuropsychology*, *7*(4), 237–242.

Hockey, R. (1973). Rate of presentation in running memory and direct manipulation of input-processing strategies. *Quarterly Journal of Experimental Psychology*, *25*, 104–111.

Hoffman, P., Jefferies, E., & Lambon Ralph, M. A. (2010). Ventrolateral prefrontal cortex plays an executive regulation role in comprehension of abstract words: Convergent neuropsychological and repetitive TMS evidence. *Journal of Neuroscience*, *30*(46), 15450–15456.

Hussey, E. K., & Novick, J. M. (2012). The benefits of executive control training and the implications for language processing. *Frontiers in Cognition*, *3*(158), 1–14. doi:10.3389/fpsyg.2012.00158

Huttenlocher, P. R., & Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *The Journal of Comparative Neurology*, *387*, 167–178.

Insight (Version 1.1) [Computer software]. San Francisco, CA: Posit Science.

Jaeger, F. T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short and long term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, *108*(25), 10081–10086.

Jaeggi, S. M., Seewer, R., Nirkko, A. C., Eckstein, D., Schroth, G., Groner, R., & Gutbrod, K. (2003). Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependent processing in single and dual tasks: A functional magnetic resonance imaging study. *NeuroImage*, *19*(2), 210–225.

January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience*, *21*, 2434–2444.

Jonides, J., & Nee, D. E. (2006). Brain mechanisms of proactive interference in working memory. *Neuroscience*, *139*, 181–193.

Jonides, J., Smith, E. E., Marshuetz, C., Koeppe, R. A., & Reuter-Lorenz, P. A. (1998). Inhibition in verbal working memory revealed by brain activation. *Proceedings of the National Academy of Sciences*, *95*, 8410–8413.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Kane, M., Conway, A., Miura, T., & Colflesh, G. (2007). Working memory, attention control, and *n*-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615–622.

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science*, *12*(6), 978–990.

Khanna, M. M., & Boland, J. E. (2010). Children's use of language context in lexical ambiguity resolution. *The Quarterly Journal of Experimental Psychology*, *63*(1), 160–193.

Long, D. L., & Prat, C. S. (2008). Individual differences in syntactic ambiguity resolution: Readers vary in their use of plausibility information. *Memory & Cognition*, *36*, 375–391.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

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

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology*, *10*, 465–501.

Nilsen, E., & Graham, S. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cognitive Psychology*, *58*, 220–249.

Novick, J. M., Kan, I. P., Trueswell, J. C., & Thompson-Schill, S. L. (2009). A case for conflict across multiple domains: Memory and language impairments follow damage to ventrolateral prefrontal cortex. *Cognitive Neuropsychology*, *26*(6), 527–567.

Novick, J. M., Kim, A., & Trueswell, J. C. (2003). Studying the grammatical aspects of word recognition: Lexical priming, parsing, and syntactic ambiguity resolution. *Journal of Psycholinguistic Research*, *32*, 57–75.

Novick, J. M., Thompson-Schill, S. L., & Trueswell, J. C. (2008). Putting lexical constraints in context into the visual world paradigm. *Cognition*, *107*, 850–903.

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(3), 263–281.

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass*, *4*(10), 906–924.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59.

Persson, J., Welsh, K. M., Jonides, J., & Reuter-Lorenz, P. A. (2007). Cognitive fatigue of executive processes: Interaction between interference resolution tasks. *Neuropsychologia*, *45*, 1571–1579.

Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology*, *57*, 137–146.

Postle, B. R. (2003). Context in verbal short-term memory. *Memory & Cognition*, *31*, 1198–1207.

Postle, B. R., Berger, J. S., Goldstein, J. H., Curtis, C. E., & D'Esposito, M. (2001). Behavioral and neurophysiological correlates of episodic coding, proactive interference, and list length effects in a running span verbal working memory task. *Cognitive, Affective, & Behavioral Neuroscience*, *1*, 10–21.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425.

Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology*, *53*(4), 1061–1080.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., ... Engle, R. W. (2012). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology General*, doi: 10.1037/a0029082

Robinson, G., Blair, J., & Cipolotti, L. (1998). Dynamic aphasia: An inability to select between competing verbal responses? *Brain*, *121*, 77–89.

Robinson, G., Shallice, T., & Cipolotti, L. (2005). A failure of high level verbal response selection in progressive dynamic aphasia. *Cognitive Neuropsychology*, *22*(6), 661–694.

Rodd, J. M., Johnsrude, I. S., & Davis, M. H. (2010). The role of domain-general frontal systems in language comprehension: Evidence from dual-task interference and semantic ambiguity. *Brain and Language*, *115*(3), 182–188.

Schnur, T. T., Schwartz, M. F., Kimberg, D. Y., Hirshorn, E., Coslett, H. B., & Thompson- Schill, S. L. (2008). Localizing interference during naming: Convergent neuroimaging and neuropsychological evidence for the function of Broca's area. *Pr Proceedings of the National Academy of Sciences*, *106*(1), 322–7.

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

Stewart, A. J., Pickering, M. J., & Sturt, P. (2004). Using eye movements during reading as an implicit measure of the acceptability of brand extensions. *Applied Cognitive Psychology*, *18*, 697–709.

Sturt, P. (2007). Semantic re-interpretation and garden-path recovery. *Cognition*, *105*(2), 477–88.

Sturt, P., Scheepers, C., & Pickering, M. J. (2002). Ambiguity resolution after initial misanalysis: The role of recency. *Journal of Memory and Language*, *46*, 371–390.

Tanenhaus, M. K. (2007). Eye movements and spoken language processing. In R. P. G van Gompel, M. H. Fischer, W. S. Murray & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 309–326). Oxford: Elsevier.

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

Thompson-Schill, S. L., Bedny, M., & Goldberg, R. F. (2005). The frontal lobes and the regulation of mental activity. *Current Opinion in Neurobiology*, *15*, 219–224.

Thompson-Schill, S. L., Jonides, J., Marshuetz, C., Smith, E. E., D'Esposito, M., Kan, I. P., ... & Swick, D. (2002). Effects of frontal lobe damage on interference effects in working memory. *Cognitive, Affective, & Behavioral Neuroscience*, *2*, 109–120.

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

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier & K. Rayner (Eds.), *Perspectives in sentence processing* (pp. 155–179). Hillsdale, NJ: Lawrence Erlbaum Associates.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A retrieval interference theory of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence processing. *Journal of Memory and Language*, *55*(2), 157–166.

Warren, T., White, S. J., & Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z Reader. *Cognition*, *11*(1), 132–137.

Weighall, A. (2008). On still being led down the kindergarten path: Children's processing of structural ambiguities. *Journal of Experimental Child Psychology*, *99*, 75–95.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*, 250–271.

Ye, Z., & Zhou, X. (2009). Conflict control during sentence comprehension: fMRI evidence. *NeuroImage*, *48*, 280–290.

APPENDIX A

TABLE A1
Significant fixed effects from the best-fitting mixed-effects model with random slope terms of accuracy testing for an Assessment (1 vs. 2) by group (responders vs. non-responders vs. untrained controls) interaction separately for ambiguous and unambiguous items on each of the four training tasks. AIC_C denotes the goodness of fit of the model, and the AIC_C of the model without slopes (in the "AIC_C–w/o" column; detailed in Table 2) is shown for comparison purposes

Significant model parameters	*Beta*	*SE*	*Z-value*	AIC_C	AIC_C–*w/o*
		Task: n-back			
Ambiguous					
Intercept	0.89	0.37	2.37*	1170.38	1148.78
Assessment	−0.56	0.27	−2.08*		
Unambiguous					
Intercept	2.78	0.301	9.23***	609.19	589.42
Group (non-responders)	5.35	1.83	2.93**		
		Task: LNS			
Ambiguous					
Intercept	0.92	0.40	2.32*	1204.08	1188.58
Assessment	−0.58	0.29	−2.04*		
Unambiguous					
Intercept	2.73	0.29	9.51***	632.46	604.75
Group (non-responders)	1.24	0.54	2.29*		
		Task: running span			
Ambiguous					
Intercept	0.88	0.38	2.32*	1207.18	1187.48
Assessment	−0.54	0.27	−2.02*		
Unambiguous					
Intercept	2.78	0.29	9.46***	627.46	600.10
Group (non-responders)	2.12	0.74	2.85**		
Assessment × group (non-responders)	−2.14	0.86	−2.50*		
		Task: block span			
Ambiguous					
Intercept	0.94	0.42	2.22*	1153.69	1142.09
Unambiguous					
Intercept	2.68	0.28	9.56***	607.45	581.52
Group (responders)	1.56	0.66	2.36*		
Assessment × group (responders)	−1.63	0.82	−1.99*		

Note. When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and items were input into the model as crossed random effects. Excluding random slopes yielded better fits of every model (as indexed by lower AIC_C values for models without random slopes as compared to those with random slopes). Thus, the best-fitting models *without* slopes are considered in our in-text interpretations of these patterns. *$p < .05$, **$p < .01$, ***$p < .001$

TABLE A2

Significant fixed effects from the best-fitting mixed-effects model with random slope terms of regression-path time following entry into the final region of each sentence. The models test for an Assessment (1 vs. 2) by group (responders vs. non-responders vs. untrained controls) interaction separately for ambiguous and unambiguous items for responders/non-responders on each of the four training tasks. AIC_C denotes the goodness of fit of the model, and the AIC_C of the model without slopes (in the "AIC_C–w/o" column; detailed in Table 4) is shown for comparison purposes

Significant model parameters	*Beta*	*SE*	*z-value*	AIC_C	AIC_C–*w/o*
		Task: n-back			
Ambiguous					
Intercept	1647.23	226.2	0.0001	7398	7376.3
Assessment × group (responders)	614.03	253.1	0.006		
Unambiguous					
Intercept	807.96	54.69	0.0001	12832.1	12812.1
		Task: LNS			
Ambiguous					
Intercept	1645.68	231.7	0.0001	4480.43	4451.93
Unambiguous					
Intercept	805.28	58.24	0.0001	7932.67	7906.97
		Task: running span			
Ambiguous					
Intercept	1617.9	236	0.0001	5328.88	5303.16
Unambiguous					
Intercept	801.35	54.15	0.0001	9417.14	9386.74
		Task: block span			
Ambiguous					
Intercept	1646.02	221.1	0.0001	5570.06	5544.56
Unambiguous					
Intercept	809.16	55.87	0.0001	9884.73	9861.13

Notes: When main effects or interactions do not appear in the table, these terms did not reliably improve the fit of the model. Subjects and items were input into the model as crossed random effects. Excluding random slopes yielded better fits of every model (as indexed by lower AIC_C values for models without random slopes as compared to those with random slopes). Thus, the best-fitting models *without* slopes are considered in our in-text interpretations of these patterns. $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$.

APPENDIX B

How are n*-back responders' accuracy and regression-path times related within and across assessments?* In the main text, we demonstrated that *n*-back responders as a group were more accurate after training, reflecting improved garden-path recovery, and that they could also handle conflicting evidence/disambiguating material more efficiently in real-time, as indexed by significantly reduced regression-path durations on correct items. But are accuracy and reading time related on a subject-by-subject basis under ambiguous conditions? Is it the case, on an individual level, that an *n*-back responder increases accuracy *and* decreases regression-path time as a consequence of training? One possibility, of course, is that an individual can transfer the effects of training in different ways, namely by either becoming more accurate, taking less time to read past regions of conflict, or both. Moreover, *how* an individual realises the benefits of *n*-back training on the sentence processing task may depend in part on how much room he or she has to improve in accuracy, on the basis of their starting point in Assessment 1.

First, we tested if, at Assessment 1, higher accuracy to ambiguous items was associated with longer regression-path times launched from the disambiguating region. Very little is known about how online and offline measures are related, and one might argue that it is unclear what specific process might be reflected in a correlation of these different dependent measures (regression-path time from the disambiguating region and comprehension-question accuracy). Nevertheless, one could construct the hypothesis that higher accuracy should be associated with longer regression-path times because, in order to effectively override an

initial misinterpretation, a reader will have to backtrack from the point of conflict to reanalyse the earlier evidence that was originally misunderstood, thus resulting in longer latencies. Similarly, at Assessment 1, shorter regression-path times might reflect little reanalysis in garden-path situations, thereby resulting in less accurate comprehension. Indeed, this was the Assessment 1 pattern for the 13 *n*-back responders, a correlation that was statistically reliable ($r = .64$; $p = .02$; see Figure 5A. This was also true for all subjects combined (not shown); $r = .45$; $p < .005$). However, at Assessment 2, the correlation weakens and fails to reach significance ($r = .33$; $p > .14$), essentially because, as a group, these individuals are getting more accurate and are taking less time (i.e., the data points on the scatter plot cluster around values associated with greater accuracy and shorter regression-path durations; see Figure 5A, Assessment 2). Nevertheless, the pattern reveals little about how these performance measures change *together* across assessments. Hence, we probed further by testing how *n*-back responders' accuracy *gains* are related to their regression-path *gains* (i.e., reduced durations) from Assessment 1 to Assessment 2. We were especially interested in examining responders who are already accurate at Assessment 1 and, consequently, have longer regression-path latencies.

Interestingly, inspection of Figure 5B reveals the following picture: a responder who exhibits large gains in accuracy shows small reductions if any in regression-path time; similarly, those responders who show large decreases in regression-path time to pass the disambiguating region show little gain in accuracy ($r = -.48$; $p = .06$). Finally, there are some *n*-back responders (approximately half) who demonstrate moderate gains on both measures. In sum, this pattern suggests that a subset of responders transfer their gains to improved accuracy, while others transfer theirs to shorter regression-path times (real-time revision immediately following entry into conflict regions) when they are responding to the items correctly.

We reasoned that how responders transfer training benefits might depend on their baseline sentence processing performance at Assessment 1. Namely, *n*-back responders who have relatively high accuracy before training might not have much room to improve in accuracy, as their offline garden-path recovery ability is already fairly good. These individuals, therefore, might instead transfer training gains to better on-the-fly revision, reflected in reduced regression-path durations associated with entry into the disambiguating region (e.g., "sparkled brightly"). Such an effect would suggest that, for those who respond well to the *n*-back task but already have a high accuracy rate (and thus little possibility of improving), training may confer earlier real-time syntactic conflict resolution. Said another way, individuals who start out with smaller lingering effects of ambiguity, reflected by greater accuracy to comprehension questions, may nevertheless demonstrate transfer, just in another way.

To explore this, we performed a median split on the 13 responders to separate them into "high accuracy" ($n = 6$; $x = 87.5\%$; range = 75–100%) and "low accuracy" ($n = 7$; $x = 40.48\%$; range = 8.3–67.7%) groups based on Assessment 1 performance. Indeed, responders who began with higher accuracy demonstrated significantly more reduction in regression-path time (1187 ms in terms of reading past the disambiguating region) than the responders who started with lower accuracy before training (141 ms; $F(1,10) = 8.58$, $p = .015$; see Figure 6A). Alongside the regression-path gains vs. accuracy gains analysis plotted in Figure 5B, it seems that while more than half of responders improved in both accuracy *and* reading time, the other half improved in one measure or the other; which benefit these responders achieve depends on the room he or she has to improve in accuracy. Indeed, the "low accuracy" group demonstrates greater gains in accuracy from Assessment 1 to Assessment 2 ($x_{gain} = 23.8\%$) compared to the "high accuracy" group ($x_{gain} = 4.2\%$), with the two lowest Assessment 1 performers exhibiting the greatest accuracy gains following training (58.3 and 33.4%). If one's garden-path recovery accuracy is low to start, then one demonstrates large gains in accuracy. If one is already accurate to start, then one demonstrates earlier reanalysis. In fact, how *much* of a reduction in regression-path time a responder exhibits can be predicted by an individuals' baseline accuracy on ambiguous items at Assessment 1 ($r = .62$; $p = .018$; see Figure 6B): those responders who start with high accuracy (between 75 and 100%) generally decrease their reading time the most (associated with regressions following entry into the disambiguating region), strikingly by about 1–2 seconds. These patterns suggest that, overall, the effects of *n*-back training are not acting separately on our two dependent measures, but that how the effects are transferred is determined, at least in part, by a person's baseline accuracy.

Nevertheless, it is important to reiterate that these correlational analyses reflect an uncertain relationship concerning what specific process is captured in the shared variance between regression-path latencies and accuracy to a comprehension question that follows the target sentence. Although we offered a hypothesis earlier in this appendix about how these online and offline measures might be expected to relate, such notions are speculative, and future research designed to address this relationship should be conducted to further our understanding.

In sum, the method by which successful *n*-back trainees transferred gains to garden-path recovery largely depended on their comprehension accuracy for ambiguous sentences at Assessment 1; those who were already relatively accurate demonstrated greater improvements in regression-path times, whereas those with low Assessment 1 accuracy demonstrated greater improvements in accuracy (Figures 5 and 6). This

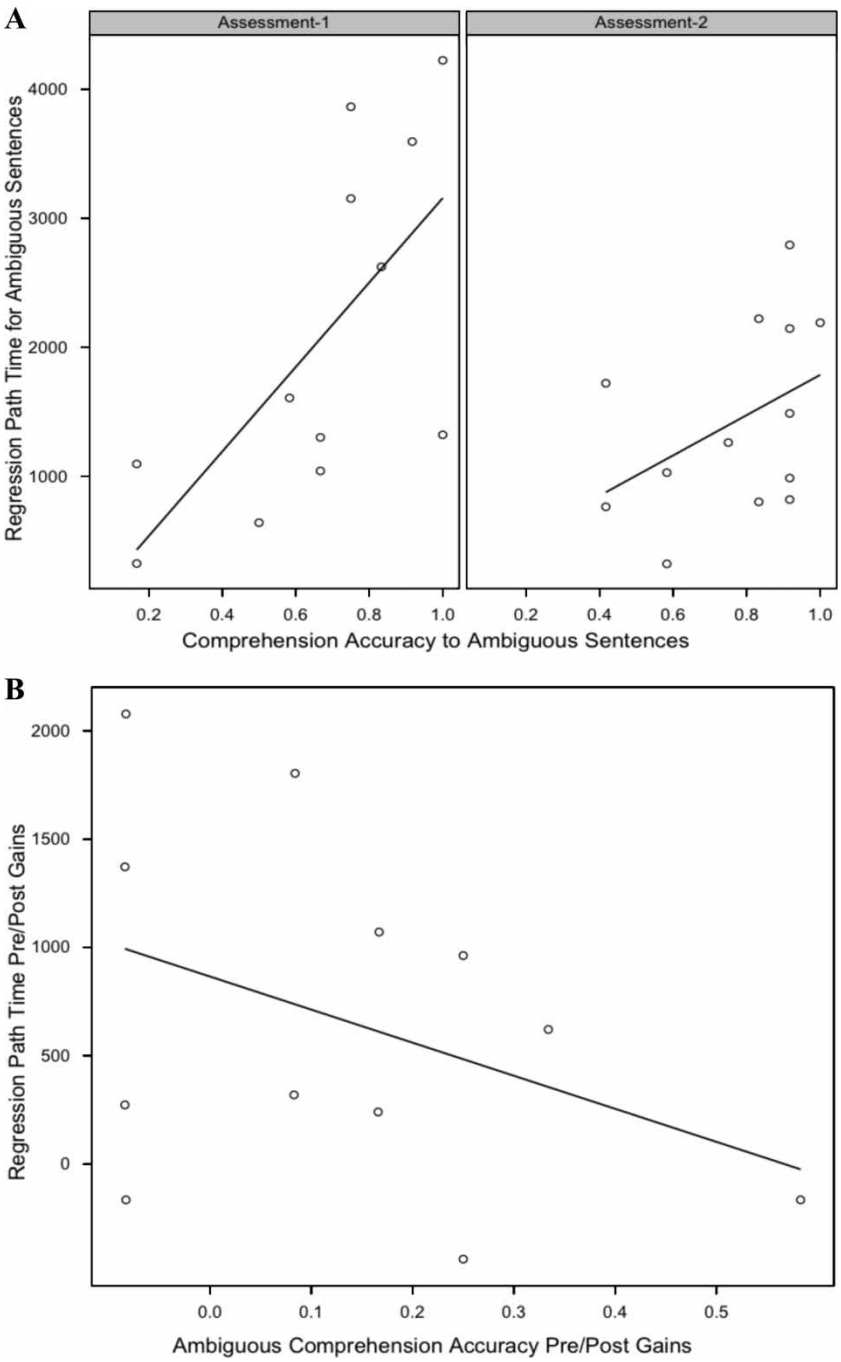

Figure 5. Responders' regression-path times from the disambiguating region as a function of accuracy to comprehension questions. (A) Accuracy to comprehension questions at Assessment 1 reliably predicts individuals' regression-path time immediately following entry into the disambiguating region at Assessment 1. This accuracy/regression-path time relation is weakened at Assessment 2 (see text).
Notes: The Assessment 1 correlation has one fewer data point (12) than the Assessment 2 correlation (13) because one responder did not contribute any regression-path data from the final region at pretest, but did at posttest. Importantly, removing this subject from regression-path analyses does not change any of the data patterns discussed in the paper. (B) The gains from Assessment 1 to Assessment 2 in accuracy to questions following ambiguous sentences are negatively correlated with improvements in regression-path time from the disambiguating region of ambiguous sentences from Assessment 1 to Assessment 2 (see text, Appendix B).

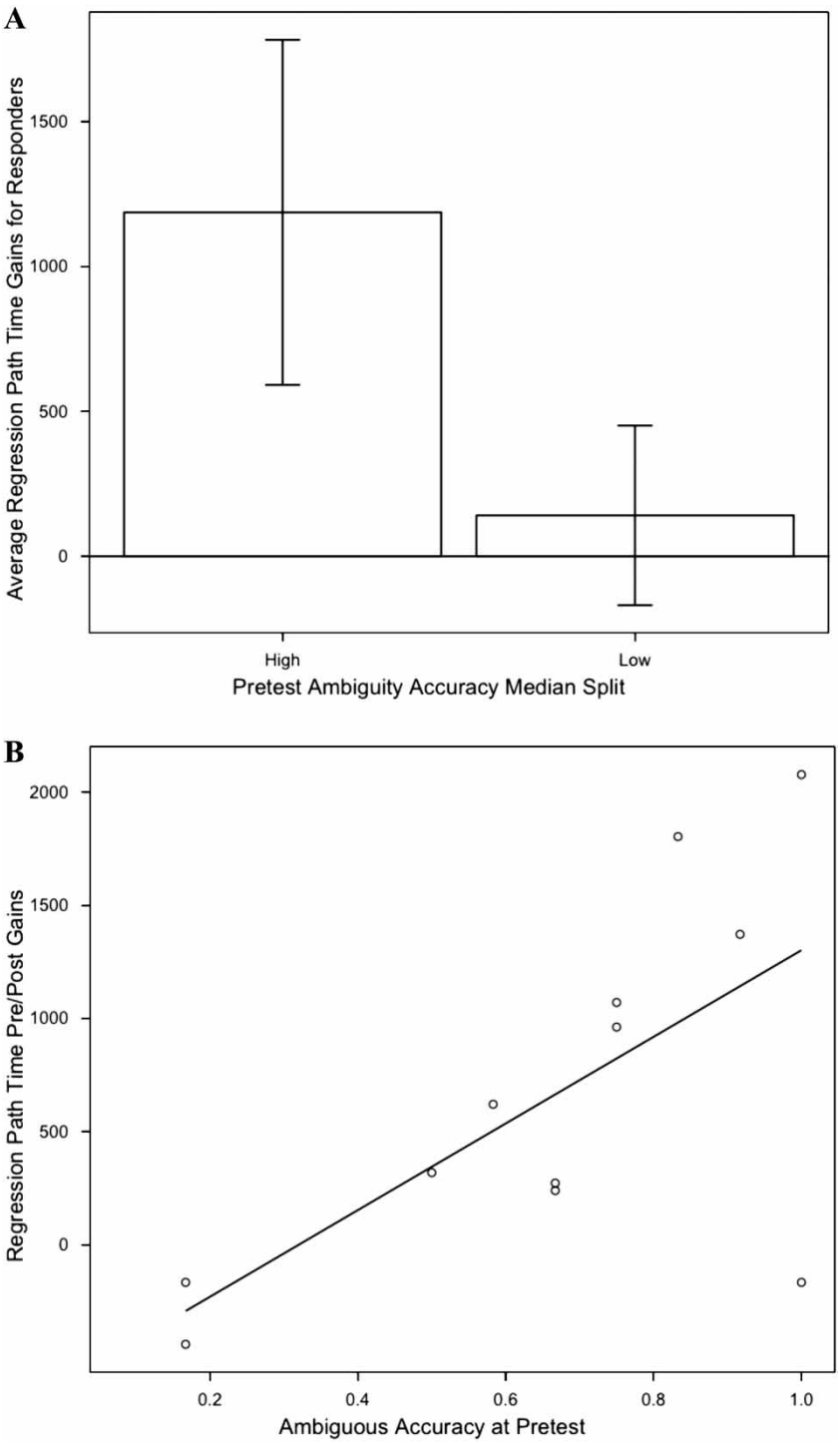

Figure 6. *N*-back responders' accuracy at Assessment 1 accounts for their regression-path-time reductions. A median split of responders in terms of accuracy at Assessment 1 revealed two clear patterns of decreases in regression-path durations. (A) Responders with the greatest accuracy at Assessment 1 ($n = 6$; $x = 0.875$) showed larger improvements in regression-path time (a decrease of 1187 ms) than those with low initial accuracy ($n = 7$; $x = 0.40$), who demonstrated reliably lower gains in regression-path time (a decrease of 141 ms). Error bars reflect ± 1 SEM. (B) Moreover, Assessment 1 accuracy significantly predicts total decreases in regression-path time from Assessment 1 to Assessment 2 on a subject-by-subject basis.

suggests that, regardless of their pre-existing conflict-resolution abilities, all successful trainees transfer improvements in cognitive control to garden-path recovery, but that the type of transfer experienced may be influenced by an individual's starting point. It is probably not surprising that those who are already good at syntactic ambiguity resolution, as indexed by high Assessment 1 comprehension accuracy, demonstrate smaller gains in accuracy than those with more room to improve. However, it is quite striking that such high-performing individuals nevertheless improve in their real-time reading behaviour, exhibiting shorter regression-path durations that may index more efficient revision processes selectively after encountering disambiguating information. Thus, both individuals with high and low cognitive control and conflict-resolution abilities may benefit from EF training, albeit differently. It is also worth reiterating the following point: as there are no accuracy differences at Assessment 1 among responders, non-responders and controls (i.e., even some non-responders and untrained controls have high accuracy), non-responder and control subjects still do not enjoy any reading-time transfer; only the *n*-back responders do.