Intelligence 41 (2013) 638–663

Contents lists available at ScienceDirect

Intelligence

Training working memory: Limits of transfer

Amber M. Sprenger [a], Sharona M. Atkins [a], Donald J. Bolger [c], J. Isaiah Harbison [b], Jared M. Novick [b], Jeffrey S. Chrabaszcz [a], Scott A. Weems [b], Vanessa Smith [a], Steven Bobb [a], Michael F. Bunting [b,*], Michael R. Dougherty [a,**]

[a] Department of Psychology, University of Maryland, USA
[b] Center for Advanced Study of Language, University of Maryland, USA
[c] Department of Human Development & Quantitative Methodology, University of Maryland, USA

ARTICLE INFO

Article history:
Received 17 October 2011
Received in revised form 12 July 2013
Accepted 15 July 2013
Available online xxxx

Keywords:
Working memory training
Fluid intelligence
Inhibition
Verbal abilities

ABSTRACT

In two experiments (totaling 253 adult participants), we examined the extent to which intensive working memory training led to improvements on untrained measures of cognitive ability. Although participants showed improvement on the trained task and on tasks that either shared task characteristics or stimuli, we found no evidence that training led to general improvements in working memory. Using Bayes Factor analysis, we show that the data generally support the hypothesis that working memory training was ineffective at improving general cognitive ability. This conclusion held even after controlling for a number of individual differences, including need for cognition, beliefs in the malleability of intelligence, and age.

1. Introduction

Working-memory (WM) processes, which support the purposeful, active maintenance of goals and information, are among the most important and widely studied components of human cognition, and for good reason. WM processes have been implicated in a variety of cognitive processes, such as visual and auditory attention, language learning and comprehension, problem solving, and fluid intelligence (see Conway, Jarrold, Kane, Miyake, & Towse, 2008). Simply put, WM is important for everyday activities, and poor WM is often associated with poor performance inside as well as outside the laboratory (Bull, Espy, & Wiebe, 2008; Gathercole, Alloway, Willis, & Adam, 2006; Geary, Hoard, Byrd-Craven, Nugent, & Numtee, 2007). In light of its importance, it is unsurprising that there is extensive interest in developing procedures to enhance WM. Improving WM even by a small amount could have enormous practical implications across a wide variety of contexts, ranging from educational to mental health contexts.

The very notion that WM in adults is changeable stands in stark contrast to the traditional view that most cognitive abilities (as opposed to acquired skills) reflect a *stable* individual trait (Neisser et al., 1996). Implicit in this view is the notion that cognitive abilities are fixed by early adulthood and immutable to positive change thereafter. Indeed, some studies have even suggested that WM has a strong genetic component (Friedman et al., 2008). Genetics aside, recent research challenges the traditional view that fluid cognitive abilities lack the capability to improve (Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Jaeggi, Buschkuel, Jonides, & Perrig, 2008), and that the neural systems underlying WM processes remain plastic throughout the lifespan and can be enhanced through intensive cognitive training (Klingberg et al., 2005; Mahncke, Connor, et al., 2006; but see Owen et al., 2010, for a contrasting view). Several studies have purported that targeted training of WM abilities leads to both behavioral (Chein & Morrison, 2010; Jaeggi et al., 2008; Klingberg et al., 2005;

* Correspondence to: M.F. Bunting, Center for Advanced Study of Language, University of Maryland, 7005 52nd Ave., College Park, MD 20742, USA.
** Correspondence to: M.R. Dougherty, Department of Psychology, University of Maryland, College Park, MD 20742, USA.
E-mail addresses: mbunting@casl.umd.edu (M.F. Bunting), mdougher@umd.edu (M.R. Dougherty).

0160-2896/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.intell.2013.07.013

Mahncke, Connor, et al., 2006; Thorell, Lindqvist, Bergman, Bohlin, & Klingberg, 2009) and neurophysiological changes (McNab et al., 2009; Olesen, Westerberg, & Klingberg, 2004; Westerberg & Klingberg, 2007).

The evidence supporting the efficacy of WM training is enticing but not perfectly robust. Jaeggi et al. (2008) reported transfer from n-back training to a matrix reasoning test that is presumed to tap general fluid intelligence, but not to verbal WM, as measured by the reading span task (see also Jaeggi et al., 2010, 2011). By contrast, Owen et al. (2010) examined whether a variant of popular training tasks would improve cognitive performance, and concluded that training on these tasks did not transfer to other, untrained tasks. Although the Owen et al. study calls into question the validity of cognitive training altogether, the results of Jaeggi et al. (2008) suggest that transfer from one particular type of training, using dual-task n-back, may yield fairly narrow transfer effects. Indeed, a recent review of the cognitive training literature by Klingberg (2010) suggests that transfer effects in cognitive training studies are frequently narrow in scope. That is, training-related transfer effects typically are limited to improvements on one or a couple of transfer tasks, rather than a broad spectrum of tasks. At the same time, most of the studies reviewed by Klingberg (2010) used narrowly defined training regimens consisting of one or a few tasks. For example, in studies by Jaeggi et al. (2008, 2011), participants trained only on an adaptive version of the n-back task. Likewise, in Klingberg, Forssberg, and Westerberg (2002), Klingberg et al. (2005) and also Olesen et al. (2004), participants trained on only three different tasks. Moreover, the bulk of the studies reviewed by Klingberg (2010) showed improvements only on tasks that were closely related to the trained abilities with respect to processing demands.

These studies suggest two important properties of cognitive training: First, cognitive training may be process-specific (cf. Dahlin, Neely, Larsson, Bäckman, & Nyberg, 2008); and second, narrow training yields narrow transfer. The implication of these two assertions is that the breadth of transfer effects should reflect the breadth of training. Training a narrowly defined set of cognitive processes should yield improvement on transfer tasks only to the extent that the transfer tasks share the same underlying cognitive processes with the training tasks. We refer to this as *process-specific* transfer. We prefer to use the terms process-specific and non-process-specific transfer as opposed to the terms 'near' and 'far' transfer, which appear elsewhere in the training literature and are unclear in our opinion. Far-transfer typically refers to improved performance on an assessment task that is ostensibly quite different from the training task(s) completed during the intervention regimen. However, one obviously expects transfer only when the underlying cognitive processes (and possibly the neuroanatomical systems that support them) are common across training and transfer tasks. Thus, the term 'far' may be a misnomer in view of the shared processes.

The possibility that training related effects might lead to generalizable transfer is both exciting and provocative, yet as discussed above the available evidence is hotly debated. If cognitive training can yield broad improvements in cognitive ability, beyond the trained tasks, it could be of enormous benefit for domains such as education and cognitive and neural remediation. However, some researchers have expressed skepticism that cognitive training works. For example, Shipstead, Redick, and Engle (2012) (see also Redick et al., 2013; Melby-Lervåg & Hulme, 2012) have argued that the majority of the empirical studies purported to show benefits of cognitive training were fundamentally flawed in ways that prevent drawing straightforward interpretations, for example, by lacking a proper control condition or a failure to keep both the participants and the experimenters blind to condition. In addition, several recent studies that have included so-called active control conditions have failed to demonstrate any advantage of cognitive training. Redick et al. (2013; see also Chooi and Thompson (2012) and Thompson et al. (2013), for example, failed to replicate findings reported by Jaeggi et al. (2008) using n-back training. Finally, using meta-analytic techniques, Melby-Lervåg and Hulme (2012) concluded that there was no evidence that WM training was effective at improving reasoning, intelligence, or Stroop performance. Yet, one shortcoming of these studies is the reliance on traditional null hypothesis significance testing (NHST) methodology. Obviously, claims made about the ineffectiveness of training imply that the null hypothesis is true, or approximately so. NHST methods are not well suited for quantifying the degree to which the data support the null, versus the alterative.

The present paper addresses some of the shortcomings in prior studies. First, rather than focusing on a single training task, we evaluated the impact of training on a battery of training tasks. Our goal was to test the hypothesis that broad training yields broad transfer. In Experiment 1, participants trained on eight different cognitive tasks, and in Experiment 2, we manipulated the process-specificity versus -generality of the training by manipulating the number and type of training tasks.

Second, in both of the studies reported herein, we included proper control conditions. In Experiment 1, we utilized a double-blind no-contact control where both the experimenter and the participant were blind to group-assignment: The experimenter did not know which participants had been assigned to training versus control, and participants were not informed about the nature of the comparison condition (or that one even existed, for that matter). In Experiment 2, we included an active control condition in which participants trained on tasks that resembled some of the training tasks, but which did not require much effortful processing beyond sustained attention.

Third, we utilized Bayesian methods to evaluate the strength of the evidence for and against the null hypothesis. Given that much of the debate regarding WM focuses on whether the existing data support the claim that training is effective or not, it is particularly important to evaluate the hypothesis that cognitive abilities are *invariant* to WM training. Indeed, the relevant question in our mind is the degree to which the evidence actually supports the hypothesis that WM training works (the alternative hypothesis) versus that it does not work (the null hypothesis). In what follows, we present data that are, by and large, consistent with the hypothesis that WM training, as implemented in our experiments, does not improve cognitive abilities unless the assessments share task or stimulus characteristics with the trained task.

2. Experiment 1

The purpose of Experiment 1 was to address two potential implications of cognitive training, within the context of

evaluating whether the WM training effect extends beyond the characteristics of the trained task. One issue is transfer effects. Specifically, we aimed to test the extent to which training on a battery of cognitive tasks that targeted a similar set of WM and attentional-control processes would transfer to untrained cognitive assessments that draw on this same set of processes. Our use of a battery of training tasks is in contrast to other studies, which typically have used a single training task (e.g. Jaeggi et al., 2008). This allowed us to examine whether training could in theory lead to broad transfer effects. Our assumption about the training task battery is that the training tasks share some underlying mechanistic commonality with each other, as well as the assessment tasks to which we expect training to transfer (see a priori hypotheses below). Multiple compatible theoretical positions on the unity of executive functions and the suggestion for some common mechanism across different working memory executive functions informs our assumption (e.g., Duncan, Johnson, Swales, & Freer, 1997; Engle, Tuholski, Laughlin, & Conway, 1999; Friedman et al., 2008; Miyake, Friedman, Emerson, Witzki, & Howerter, 2000).

The second issue relates to the persistence of training-induced improvements over time. Most studies illustrating gains in cognitive ability, including the current study, require participants to engage in many hours of training over several weeks (e.g., 20 h over 4 to 6 weeks). Given the duration of time required for observing training effects in prior studies, long-lasting effects are ideal. Therefore, we included two evaluation sessions occurring after the conclusion of the training: one at 1 week following training (post-test) and a second 3 months following the completion of training (follow-up). Although prior studies have shown process-specific training, relatively few studies have targeted a broad spectrum of cognitive abilities during training. Moreover, although there has been a fair amount of focus on transfer effects in the literature, there have been relatively few studies examining the persistence of such effects over time.

To begin addressing these issues, we asked two questions. First, does intensive cognitive training lead to improvements on untrained measures of cognitive ability and, if so, what is the extent of the transfer effects? Second, assuming positive transfer, do training-induced improvements persist over extended periods of time without additional training?

One way to address the first question is to define, a priori, a set of cognitive assessments that share cognitive processes with the training tasks and a set of cognitive assessments that are hypothesized to draw on different processes. At the broadest level, one can differentiate between fluid abilities and crystallized abilities (Cattell, 1971). Fluid abilities are individuals' ability to think and act quickly and solve novel problems. Fluid abilities are considered independent from education. In contrast, crystallized abilities stem from learning and knowledge. Crystallized abilities are reflected in tests of knowledge, general information, and the use of vocabulary. Examination of the literature illustrates that fluid and crystallized abilities are distinct, yet closely related. For example, it is common for measures of fluid ability to predict performance on tasks that reflect crystallized abilities, as is illustrated by the correlation between tests such as the SAT and measures of WM capacity (Engle et al., 1999). Our working assumption is that the relationship between fluid and crystallized abilities manifests because fluid abilities facilitate the acquisition of crystallized abilities, so that individuals with greater fluid ability tend to learn and acquire crystallized abilities *over time* at a greater rate. Thus, we hypothesize that training on a set of cognitive tasks should yield transfer only to cognitive tasks that share those fluid abilities tapped during training. No transfer is expected for measures of crystallized abilities because these are a long-term consequence of one's fluid ability.

Although recent studies have shown transfer to untrained tasks, few studies have assessed the transfer of improvements across a wide range of both fluid and crystallized abilities (Lövdén, Bäckman, Lindenberger, Schaefer, & Schmiedek, 2010). Our theoretical position that training should be process-specific implies that any positive transfer to non-process-specific crystallized abilities would be evidence of a placebo effect. That is, if training leads to improvement on both fluid and crystallized abilities, it would suggest that improvements across all of the ability measures were plausibly due to factors unrelated to the training per se, and possibly due to achievement motivation or mere practice effects.

The second question addressed by our study is whether training induced-improvements, if they exist, persist over extended periods of time without additional training. Practical considerations motivate this question. Most prior studies showing positive transfer required participants to engage in extensive training over weeks or even months; few studies have examined what occurs after training discontinues (but see von Bastian & Oberauer, 2013). Potential users of cognitive training methodologies (e.g., students, elderly, and patients with WM deficits) might be more willing to commit to the training regimen if they know that the benefits persist in the absence of continued training and are evident months after the training ends.

In Experiment 1, participants trained for 20 h over six weeks on a battery of performance-adaptive cognitive training tasks designed to enhance WM functioning. We administered a battery of cognitive assessments, including measures of fluid and crystallized abilities (the verbal reasoning and verbal skills' tasks, respectively, as listed in Table 1), prior to and immediately following 20-hours of cognitive training, and again three-

Table 1
Complete assessment battery (*italics* indicate assessment data that will be reported elsewhere).

Construct	Assessment	Duration (min)
Verbal working memory	Listening span	15
	Operation span	15
Spatial working memory	Rotation span	10
	Symmetry span	15
Inhibition	Stroop[a]	8
	Antisaccade	10
Verbal reasoning	ETS deciphering languages' test (RL4)	8
	ETS inference test (RL3)	6
Verbal skills	AFOQT reading comprehension	9
	AFOQT verbal analogies test	5
Ambiguity resolution	*Cloze task*	*30*
	Syntactic ambiguity resolution (3 tasks)	*45*
Metaphor comprehension	*Metaphor comprehension*	*15*
	Metaphor priming	*15*

[a] Note: Due to a computer error, Stroop data is missing for 35 participants, 17 from the training group and 18 from the control group.

months following cessation of training. We hypothesized that cognitive training would yield gains in cognitive ability, relative to a no-contact control group, and that these gains would transfer to untrained but distinct measures of cognitive ability. More specifically, we hypothesize that transfer effects would be limited to process-oriented, or fluid, measures of cognitive ability and that measures of crystallized abilities would yield little or no benefit. Previous research implies that crystallized abilities are acquired throughout one's lifetime (e.g., due to continued knowledge acquisition); therefore, training of fluid abilities should not confer benefits to crystallized abilities within the tested time frame. Considering the question of the persistence of training benefits, we hypothesized that cognitive training would show transfer effects immediately following training. Yet, the longevity of training-induced-improvements in WM following cessation of training remains unknown.

3. Method

3.1. Participants

One-hundred twenty-seven (N = 127) participants recruited from the University of Maryland, College Park community were randomly assigned to either the training (n = 70, 36 women, M_{age} = 22.97 years, range$_{age}$: 18–43 years) or control group (n = 57, 36 women, M_{age} = 23.05 years, range$_{age}$: 18– 36 years). The study adhered to a true double-blind pretest/posttest/follow-up experimental design, in which neither participants nor study moderators had knowledge of the condition to which participants were assigned. Participants in the control group had no knowledge of the existence of the training group, and vice-versa, as we recruited for several studies and across languages (Native English Speakers and Native speakers of Cantonese, Mandarin, Taiwanese and Korean). Even though we recruited people from diverse language backgrounds, there was no effect of language on any of the assessments or training; therefore, we collapsed analyses across language groups. All participants completed a battery of cognitive assessments prior to commencing training (pretest), approximately one-week after completing training (posttest), and three-months after completing training (follow-up). Participants were compensated for each training session that they completed, in addition to receiving a bonus at the end of the experiment. Participants in the training condition earned $500 for completing all three assessment periods plus all 20 training sessions. Participants in the control condition received $180 for completion of the three assessment periods.

3.2. Assessments

The assessment battery consisted of 16 cognitive tasks (see Table 1). For the purposes of this report, we consider only the measures of WM, inhibition, verbal reasoning, and verbal skills; data from the other tasks have been reported elsewhere (Novick, Hussey, Teubner-Rhodes, Harbison, & Bunting, in press). The tasks were computer administered and fully counterbalanced for version and order of completion. Based on our a priori assumptions about the unity of working memory executive functions and the kind of complex tasks that rely on these processes, we expected transfer to the WM, inhibition and verbal reasoning tasks, but not verbal skills' tasks, because the passages were short, affirmative, and unambiguous, as described below, and comprehending these sorts of materials typically does not heavily task WM. As we know from Engle and Conway (1998); also cf. Daneman and Merikle (1996) the relationship between WM and verbal comprehension is robust for unskilled readers or for skilled readers when the materials are not "short, simple, active, affirmative, declarative sentences" (p. 88) or when the sentences do not follow perfectly logically from one to the next.

3.2.1. Verbal working memory

3.2.1.1. Automated listening span task

3.2.1.1.1. Description. This is a quintessential dual-task WM span task and is based on a task described by Daneman and Carpenter (1980). It combines listening comprehension with a test of short-term memory for symbols to emulate the simultaneous processing and storage demands that are the hallmark of WM. The task was automatized for computerized delivery according to the procedure described by Unsworth, Heitz, Schrock, and Engle (2005). It consisted of three parts: (1) a brief symbol memory practice task, (2) a sentence listening and comprehension task, and (3) the sentence listening and symbol memory tasks combined (the listening-symbol span task).

For Part 1, the symbol span practice, the objective was to remember symbol strings for immediate serial recall. The symbol strings consisted of two non-repeating symbols from a pool of nine symbols ("Wingdings" font: 🕭, 🕮, 🕯, ☺, ✂, 💣, 🏱, 🖉, ☠). Following a 500-ms fixation point at trial onset, each symbol was visually presented serially at the rate of one symbol per second. Recall was cued immediately following the last symbol. The recall screen consisted of a 3 × 3 grid of nine possible symbols with instructions at the top of the screen to recall both symbols in the order presented. A check appeared to the left of each symbol as it was selected, and symbols that were selected appeared in a row at the bottom of the screen. Participants could click a button marked 'Blank' to mark the serial position of symbols they could not recall. A button marked 'Clear' could be pressed to clear all of the symbols and begin recall again. After recalling a particular sequence, participants pressed a button marked 'Next' to begin the next trial. A total of 3 symbols strings were presented for recall.

For Part 2, the sentence listening and comprehension task, the objective was to listen to nine- to fourteen-word sentences and indicate whether the sentence was semantically plausible ("The squirrel stored some nuts in the tree in preparation for a long winter") or implausible ("On their first visit to ketchup, they took a formal tour"). Sentences were played one at a time on a computer, and participants listened with headphones. As soon as the sentence stopped, buttons labeled "True" and "False" appeared on the computer screen with the question, "Did the sentence make sense?" Participants used the mouse to click the button corresponding to their answer. The task ended when 12 sentence judgments were made.

Finally, for Part 3, the sentence listening and symbol memory span tasks were combined. A 500-ms fixation point ("+") appeared at trial onset followed by a semantically plausible or implausible sentence. Participants listened to the sentence and answered true or false, as in Part 2. A to-be-remembered symbol then appeared on the screen for 1 s.

Sentences and symbols were presented in this manner until the set size was reached. As in Part 1, symbol recall was cued immediately following presentation of the last symbol in the set. The task had 12 sentence-and-symbol sets, three of each set size from three to six. Three practice trials (set size = two) preceded the real test.

3.2.1.1.2. Scoring. The listening span score was a percentage of correctly recognized symbols in the correct serial position.

3.2.1.2. Automated operation-letter span task

3.2.1.2.1. Description. This dual-task is structurally identical to listening span, but with different processing and storage components (see Turner & Engle, 1989; Unsworth et al., 2005). A mental arithmetic task (computing the solution to (8 − 2) + 7 = ?) served as the processing activity in place of the listening task. The memoranda were consonant letters (D, F H, J, L, P, R, T, Z) rather than symbols, as in the listening span task. The task specifications and procedure were otherwise identical to those for the automated listening symbol-span task.

3.2.1.2.2. Scoring. The operation span score was a percentage of correctly recognized letters in the correct serial position.

3.2.2. Spatial working memory

The symmetry span and rotation span tasks were dual-tasks, each with unique processing and storage components. They were automated for computer administration according to the procedure described by Kane, Conway, Hambrick, and Engle (2007).

3.2.2.1. Symmetry span

3.2.2.1.1. Description. Participants had to remember the spatial location of a set of red blocks presented serially for 650 ms in a 4-by-4 grid. Between presentations of blocks, participants saw an image constructed in an 8-by-8 grid and had to judge whether the image was symmetric or asymmetric. Set sizes ranged from two to five blocks and were presented in random order. Participants completed 3 sets of each set size, for a total of 12 trials.

3.2.2.1.2. Scoring. The symmetry span score was the number of correctly recognized blocks in the correct serial position.

3.2.2.2. Rotation span

3.2.2.2.1. Description. The rotation span task was adapted from Kane et al. (2004). Participants had to remember a series of short or long arrows originating at the center of the screen and pointing in any of 8 directions. Between presentations of the arrows, they made a keyboard response to indicate whether the orientation of a single letter presented on the screen was normal or mirror-reversed. The letter stimuli were normal and mirror images of capital G, F and R and were rotated at 0°, 45°, 90°, 135°, 180°, 225°, 270° or 315°, and participants had to rotate the letter mentally to respond correctly. The sequence of events included (1) the presentation of a short or long arrow for 1000 ms, (2) the participant's keyboard response, (3) a blank screen for 500 ms. Sequences repeated until the presentation of a recall cue appeared in place of the blank screen. The recall cue consisted of the image of two circles of arrows, one long and one short, with each arrow originating from the center and pointing in one of the 8 possible angles of rotation. Participants used the mouse to click on the image of the arrows in the recall cue that matched the length and direction of the arrows presented in the set. Set sizes ranged from two to five rotated letter-arrow displays per trial. Participants completed three sets of each set size, for a total of twelve trials.

3.2.2.2.2. Scoring. The rotation span score was a percentage of correctly recognized arrows in the correct serial position.

3.2.3. Inhibition

3.2.3.1. Stroop

3.2.3.1.1. Description. Participants made a keyboard response to indicate the font color (green, blue, red or yellow) of a word or character string presented on the screen. The words were *Blue*, *Green*, *Yellow* and *Red*, so the font color and the word were sometimes congruent (i.e., they matched) but also were sometimes incongruent (i.e., they did not match). The character string, which served as a baseline comparison to congruent and incongruent trial types, consisted of three to six asterisks. The stimulus remained on the screen until the participant's response. A 750 ms fixation was presented between the character series. Participants completed 12 practice trials, including 8 congruent and 4 baseline trials. They then completed 192 test trials, including 24 baseline, 24 incongruent, and 144 congruent trials, presented randomly.

3.2.3.1.2. Scoring. Mean accuracy and mean reaction time on accurate trials were collected for congruent, incongruent and baseline trials.

3.2.3.2. Antisaccade

3.2.3.2.1. Description. The antisaccade task was adapted from Kane, Bleckley, Conway, and Engle (2001). This test used a black background with text presented in white or cyan 12 point bold Courier New font. The test was comprised of the following blocks presented in this order: response-mapping practice block, prosaccade practice block, antisaccade practice block, antisaccade test block, and finally prosaccade test block. The blocks are described in turn.

The test began with a response-mapping practice block of 15 trials. Each trial began with a blank (black) screen for 400 ms. A fixation signal (three cyan asterisks: "***") then appeared in the center of the screen for a duration that varied unpredictably between 200 ms and 1800 ms. There were three trials (one with each target letter) for each fixation duration (200 ms, 600 ms, 1000 ms, 1400 ms, and 1800 ms). After the fixation signal disappeared, the screen went blank for 10 ms. Following this brief interval, a white attractor signal (an equal sign: "=") appeared in the middle of the screen. The attractor signal was displayed for 100 ms, replaced with a blank screen for 50 ms, shown again for 100 ms, and then replaced again with a blank screen for 50 ms. After the last of these blank screens, the target letter (a white "B," "P," or "R") was displayed in the center of the screen for 100 ms. This was then replaced with a mask (a white "H") in the center of the screen for 50 ms, and then a second mask (a white "8") that remained on the screen until the subject responded by pressing one of the buttons on the response box. The next trial then began (with a 400 ms blank screen).

The prosaccade practice block was similar to the response mapping practice block except for two differences. First, there

were 30 trials instead of 15 trials. Second, the attractor signal, target letter, and masks appeared unpredictably to either the right or left of the center of the screen, about two-fifths of the distance from the center of the screen to the edge of the screen. The target letter and masks would always appear in the same location as the attractor signal in each trial. The fixation signal remained in the middle of the screen. There was one trial for each combination of the three target letters, five fixation durations, and two sides of the screen that the attractor signal appeared on, to make the total of 30 trials.

The antisaccade practice block was similar to the prosaccade practice block except that the target letter and masks would appear on the opposite side of the screen from the attractor signal. There were 30 trials. The antisaccade test block was the same as the antisaccade practice block, but with twice as many trials. Finally, the prosaccade test block was the same as the prosaccade practice block, but with twice as many trials.

3.2.3.2.2. Scoring. Mean accuracy and mean reaction time on correct trials were collected for the prosaccade and antisaccade blocks.

3.2.4. Verbal reasoning

3.2.4.1. Deciphering languages' test

3.2.4.1.1. Description. This is a multiple-choice test of logical reasoning ability from the Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, & Haarmann, 1979) and requires that participants adapt to a new language system. Each test item consisted of a symbol(s), syllable(s) or word(s), and participants had to choose the translation from five possible translations. For each different artificial languages, three expressions in English and their translation into the language are given. From these mappings the participant had to figure out logically which syllable or symbol in the language is equivalent to which English word. There were three languages to decipher and 12 problems (3 each for the first two languages and 6 for the third language). Participants had eight minutes to complete the test.

3.2.4.1.2. Scoring. The score was the number of questions answered correctly.

3.2.4.2. Logical reasoning inference test

3.2.4.2.1. Description. This is a multiple-choice test of inference and logical reasoning ability from the Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1979) that relies on existing verbal knowledge. The stimuli were ten complete statements, each of which was one to two sentences in length. Each statement provided a few facts pertaining to a specific subject. Each statement was followed by five potential conclusions. Only one of the conclusions could be definitively supported by the information provided in the opening statement. The remaining four statements each require additional information or inferences in order to be valid conclusions. For each statement, the task was to decide which of the five given conclusions was valid based only on the information provided in the statement.

Participants were given the statements one at a time. Statements appeared simultaneously with the appropriate answer selections. The participants had 6 min to complete 14 problems, with a maximum of 1 min per problem. If the participant did not complete a problem within the allotted minute, the computer automatically moved forward to the next problem while recording an error.

3.2.4.2.2. Scoring. The score was the number of questions answered correctly.

3.2.5. Verbal skills

3.2.5.1. Reading comprehension

3.2.5.1.1. Description. This task measures English reading comprehension ability. The materials were adapted from a practice exam for the Air Force Officers Qualifying Test (AFOQT; Officer Candidates Test, 2005). In this multiple choice test, participants selected the appropriate word or phrase that best completed a short paragraph. Because the passages were short, affirmative, and unambiguous, we did not expect performance on this comprehension task to correlate with working memory (cf. Engle & Conway, 1998). They had 10 min to complete 20 items.

An example item is as follows: *We had to acknowledge that he was still at the peak of his maturity, although he had been at precisely that point for as long as we could remember. Old age seemed as alien to his being as callow youth. There was about him: (a) an inexplicable perpetuity, (b) a childish frivolity, (c) an intense desire to live, (d) an apparent change with time, or (e) a quality of old age.* The correct answer is (a).

3.2.5.1.2. Scoring. The score was the number of questions answered correctly.

3.2.5.2. Verbal analogies' test

3.2.5.2.1. Description. Stimuli were 18 incomplete analogies accompanied by 5 answer choices. Only one answer choice logically completed the analogy. The materials were taken from a practice exam for the AFOQT and administered according to the instructions for the practice exam (Officer Candidates Test, 2005). Participants had 5 min to complete all 18 analogies by typing in the number (1–5) corresponding to the correct answer choice. Analogies appeared one at a time, with a new analogy appearing after the participants made their selection for the previous analogy.

3.2.5.2.2. Scoring. The score was the number of questions answered correctly.

3.2.6. Training tasks

The training-group completed 20 one-hour supervised training sessions over 3–6 consecutive weeks ($M_{weeks} = 4.89$; range = 3–7 sessions per week). The training battery consisted of eight tasks designed to train executive WM functions, including a letter N-back task, auditory letter running-span, block-span, letter-number-sequencing, and tasks developed by Posit Science (match-it, sound-replay, listen-and-do, and jewel-diver) for their brain-fitness software packages (Brain Fitness Program, Version 2.1; Insight, Version 1.1). Four tasks were administered per session, each for a maximum of 15 min. All tasks adapted in difficulty to the level of participants' performance and repeated 10 times. The task order was the same for all participants.

3.2.7. Training tasks 1–4

3.2.7.1. Block span

3.2.7.1.1. Description. The block span training required participants to remember the serial order in which a sequence of black blocks appear in a 4×4 grid, where each trial is characterized by a set of 1 to J such sequences, and where each sequence consists of 2 to K blocks ($1 \leq J \leq 5$ and $2 \leq K \leq 4$). Each block within a sequence flashed for 1 s, one at a time, in one of the cells within the 4-by-4 grid. When the grid flashed for 1 s, it was an indication to the participant that one sequence was ending and another sequence would begin following a 1-s delay. After the final sequence within a set, participants were prompted to indicate (via mouse click) the spatial location (in serial order) of each block within the first sequence of the set, then spatial location (in serial order) of each block within the second sequence of the set, and so forth for all sequences within the set. This procedure was then repeated for the next set of sequences for the duration of the task.

The values of J and K and, thus, the difficulty of the task adapted automatically to the participant's performance. As a participant's performance improved, J, K, or both J and K increased as well. The difficulty level of block span progressed according to the following algorithm:

(a) If the participant correctly remembered the location and serial position of three consecutive sequences within a level of J and K, the value of K increased by 1, unless $K = 4$, in which case J was increased by 1 and K reduced to 2.
(b) If the participant correctly remembered the location and serial position of two of the three sequences within a level of J and K, the values of K and J were left unchanged.
(c) If the participant incorrectly remembered the location and serial position of two of the three sequences within a level of J and K, the value of K was reduced by 1, unless $K = 2$, in which case J was decreased by 1, and K was set to 4.

3.2.7.1.2. Scoring. Block span was scored by counting the number of blocks recalled in the correct serial order and spatial position.

3.2.7.2. Letter-number sequencing

3.2.7.2.1. Description. Participants saw a sequence of letters and numbers presented one at a time in random order on the screen, for 500 ms. Following each sequence of letters and numbers, participants used the keyboard to enter all the numbers and then all the letters that appeared in the set. They had to enter the number in forward numerical order and the letters in forward alphabetical order.

Letter stimuli were drawn from the English alphabet (uppercase A through Z), and number stimuli consisted of Arabic numerals 1–9. Each letter-number sequence always contained from a minimum of one to maximum of four letters and numbers. Within a set, any given letter or number appeared only once.

One way in which the difficulty level of this task was manipulated was the presentation of multiple letter-number sequences in a to-be-remembered set. As a warning to participants that they would have to remember multiple sets of sequences, the set size was shown at the top of the screen. An asterisk (*) separated each sequence in a set during presentation. At recall in a multiple-sequence condition, participants recalled each sequence in order; therefore, they entered the numbers from the first sequence in forward numerical order followed by the letters from that sequence in forward alphabetical order before recalling the numbers and letters from the next sequence.

This task also adapted automatically to the participant's performance, beginning with a set size of one sequence that contained two characters, 1 letter, and 1 number. The difficulty level increased by increasing the number of characters in each sequence of a set, if each sequence in the set was already at the maximum number of characters, set size was increased by one sequence to a maximum of five sequences per set. When the number of sequences per set was increased by one, the number of characters per set was always reset to two. The difficulty level decreased by reducing the number of characters in the sequence to a minimum of 1 letter and 1 number, if the sequence already had just two characters, the number of sequences per set was decreased by one sequence to a minimum of one sequence. When the number of sequences per set was reduced by one, the number of characters per set was always reset to a maximum level of six.

The difficulty level of this task was reevaluated every four trials according to this algorithm:

(a) If the participant correctly reproduced the sequence of letters and numbers on four consecutive trials, the difficulty level increased by one level.
(b) If the participant correctly reproduced the sequence of letters and numbers on three of four consecutive trials, the difficulty level remained the same. But, if they were 100% accurate on three of four consecutive sequences three times in a row, the difficulty level increased by one level.
(c) If the participant correctly reproduced the sequence of letters and numbers on two of four consecutive trials, the difficulty level remained the same. But, if they were 100% accurate on two of four consecutive sequences three times in a row, the difficulty level decreased by one level.
(d) If the participant failed to recall any letter-number sequences on four consecutive trials correctly, the difficulty level decreased by one level.

3.2.7.2.2. Scoring. Letter-number-sequencing is scored based on correct recollection of the serial reordering of the characters.

3.2.7.3. N-back

3.2.7.3.1. Description. Participants saw a sequence of letter stimuli one at a time and had to indicate by key-press when the current stimulus matched one presented *n* items prior in the sequence. A sequence contained 25 items, of which 5 were targets (i.e., they matched an item n back), 0 or 5 were lures (i.e., they matched an item close to n back) and the rest were non-targets (i.e., letters that had last occurred more than 10 letters prior).

There were three levels of lures. The easiest level (level 0) consisted of no lures. At the next difficulty level (level 1) lures appeared in position $n + 1$. In the most difficult lure

level (level 2) lures appeared both in position n + 1 and n − 1. Participant performance on each sequence was used to determine whether and how the task difficulty should adapt on the subsequent sequence of 25.

The difficulty level of this task adapted to the participant's performance upon completion of a sequence. When performance on the previous sequence was 85% accuracy or above, the difficulty level of the next sequence increased by one. When performance on the previous sequence was less than or equal to 65% accuracy, the difficulty level of the next sequence decreased by one. Otherwise, the difficulty level remained the same on the next sequence.

When difficulty level had to be increased or decreased by one, lure level was either increased or decreased, respectively. However, if the difficulty needed to be increased and the lure level was less than 2, the lure level would increase (from 0 to 1 or 1 to 2). If lure level was already at the maximum, n would increase by one (n could range from 1 to 8) and the lure level would reset to zero. Similarly, when the task needed to be made easier and the lure level was greater than 0, the lure level would be decreased by one level. If the lure level was already at the minimum, n would decrease by one and the lure level would be reset to two. All participants started at n = 2, lure level = 0.

3.2.7.3.2. Scoring. Accuracy of response and reaction times was computed for each level of *n* and a mean level of *n* was computed for the session overall.

3.2.7.4. Running memory span task

3.2.7.4.1. Description. This task is based upon a task developed by Pollack, Johnson, and Knaff (1959) and more recently updated by Cohen and Heath (1990) and Bunting, Cowan, and Saults (2006). When used as a measurement tool, it measures the immediate serial order recall of the last *n* letters in auditorily-presented strings of 12–20 pseudo-random letters, where n is typically a constant from 5 to 7. Each letter string had 10 unique letters from a pool of 12 consonants (C, F, H, J, L, N, P, R, T, V, X, Z); letters could repeat up to three times per string but not within a window of six letters. A further restriction on the randomization of letters was that no two letters occurred in adjacent forward alphabetical order (e.g., P, R was permitted). The letters were digitally recorded in a male voice and compressed to play within 250 ms each, without a change in fundamental frequency. The letters were computer delivered and played in serial order over noise-canceling headphones at the rate of one letter per 500, 750, or 1000 ms. A mouse click initiated each trial, and a total of 20 trials were presented.

The recall screen depicted all 12 possible letters in a 3 × 4 grid of small rectangular buttons. Beneath the 12 letter buttons was a larger rectangular button labeled "Blank," and to the right of the letter buttons were larger buttons labeled "Clear All" and "Enter". Recall was prompted immediately upon the completion of the presentation of the last letter in a string. Participants had to recall the last *n* letters beginning with the *nth* letter from the end of the list, where *n* was a number from 2 to 9. They chose the memoranda in serial order, using a mouse to select the letters that corresponded to their memory of the list. They were instructed to click "Blank" to demarcate the serial position of letters they could not remember, and they clicked "Clear All" when they wanted to start over. As letters were selected, the buttons turned from white to yellow, and the letter appeared in a horizontal list to the right of the letter buttons.

At the beginning of each training session, initial $n = 2$ and initial rate = 500 ms, and *n* and rate were evaluated for increase or decrease after every run (i.e., every four successive trials). If accuracy was 100% on the previous run, the participant would advance by one level of difficulty. If accuracy was 25% or less, the participant would regress by one level of difficulty. Other accuracy scores did not automatically result in a change in the level of difficulty. However, if there was no change in level of difficulty for three consecutive runs, a change in the level of difficulty was forced. If on the previous three runs accuracy was 50% or less, the participant would regress by one level of difficulty. Otherwise, the participant would advance by one level of difficulty.

Level of difficulty was determined by n (2–9) and presentation rate (1000, 750, and 500 ms). From the initial starting values, the next two lesser levels of difficulty were (1) less difficult: n = 2, rate = 750 ms, and (2) even less difficult: n = 2, rate = 1000 ms. Or, the next two greater levels of difficulty were (1) more difficult: n = 3, rate = 1000 ms, and (2) even more difficult: n = 3, rate = 750 ms. Importantly, while this task was adaptive within training session, participants always started each session at the same level of difficulty (as opposed to resuming where they left off) due to experimenter error.

3.2.7.4.2. Scoring. Participants received one point per item recalled in the correct serial position. Points were summed by trial, and the critical score, running memory span, is the mean proportion correct across trials.

3.2.8. Training tasks 5–8

Posit Science contributed the four executive function 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-spatial tracking of multiple objects), "Match-It" (targeting auditory and visual-spatial memory), "Sound-Replay" (targeting sequential ordering of information in auditory WM), and "Listen-and-Do" (targeting auditory working-memory span). We describe each task briefly below.

3.2.8.1. Jewel diver

3.2.8.1.1. Description. Participants viewed a display consisting of *n* objects, with a subset of them identified as targets and then were hidden by an object. Participants were required to keep track of only the objects hiding the targets, while all the objects moved randomly throughout the display. The number of objects, speed of motion, unity of motion, size of display and number of obstruction were manipulated.

3.2.8.1.2. Scoring. Points were awarded for correct detection of the target.

3.2.8.2. Match-it

3.2.8.2.1. Description. Participants saw an array of 'cards' presented in a grid and tried to match up pairs of identical cards by clicking two per trial to reveal their stimuli. The target stimulus on each card was an auditory representation of a phoneme. The size of the array, the degree of speech processing and the phonemic similarity were manipulated.

3.2.8.2.2. Scoring. Points were awarded for correct matching of phoneme pairs.

3.2.8.3. Listen-and-do

3.2.8.3.1. Description. Participants were presented with a visual array of several objects, and were given auditory instructions about how to rearrange them. After hearing the instruction, participants were required to reorganize the objects on the screen by dragging them to the correct location, and in the correct serial order. Participants carried out a series of audio instructions by manipulating objects and characters around graphic interface. As the task difficulty increases, the series of instructions becomes longer and more variables are introduced into the graphic interface, leaving the participant with more commands to hold in memory and more possible wrong choices.

3.2.8.3.2. Scoring. Points were awarded for correctly following the instructions.

3.2.8.4. Sound replay

3.2.8.4.1. Description. Participants heard a sequence of phonemes and then clicked on a visual depiction of each phoneme in the same order in which they were played. This exercise required the listener to sequence a set of syllables; the syllables were played auditorily and listeners had to click on written representations of those syllables in the proper sequence. When a participant successfully completed enough exercises for the program to progress to the next level, the sequences became longer and individual tokens repeated, which made the sequences more challenging to remember. The size of the array, the degree of speech processing and the phonemic similarity were manipulated.

3.2.8.4.2. Scoring. Points were awarded for correct repetition of the sequences.

4. Results and discussion

Of the 127 participants to start the study, 115 (59 in the training condition and 55 in the control condition) completed post-test assessments and 93 completed the follow-up assessment. Data from a small number of tasks was lost due to experimenter error or computer error. In particular, a programming error resulted in the loss of pre-test Stroop data from 38 participants.

4.1. Analysis of working memory training

There were no language differences, so we collapsed across language groups. The training data were examined to investigate participants' improvement on the trained tasks. Fig. 1 plots the learning curves for the eight training tasks. These training curves show that participants' performance on the training tasks improved substantially. Paired-sample t-tests comparing performance on the average of the first two training sessions for each task with the average of the last two training sessions revealed significant training effects for all training tasks (block span: $t(58) = -9.02$, $p < 0.001$; letter-number sequencing: $t(58) = -13.86$, $p < 0.001$; n-back: $t(58) = -10.78$, $p < 0.001$; running span: $t(58) = -17.77$, $p < 0.001$; jewel diver: $t(58) = -12.62$, $p < 0.001$; sound replay: $t(58) = -33.00$, $p < 0.001$; match-it: $t(58) = -105.68$, $p < 0.001$; listen-and-do: $t(58) = -31.16$, $p < 0.001$).

As the participants in the training group improved on all the training tasks, we next investigated the transfer of the training improvement to the non-trained cognitive assessments. In addition to conducting analyses using traditional NHST, we also conducted statistical tests using Bayes factors (BF). There are many reasons to prefer Bayesian tests to standard null hypothesis tests (e.g., Wagenmakers, 2007), but the principal one for our purposes is that it allows one to quantify the strength of the evidence for the null compared to the alternative. As stated in the Introduction section, this is particularly important in the context of the debate over whether WM training is effective at improving cognitive skills. Interpretation of the BF is straightforward. The BF expresses the odds in favor of the null hypothesis compared to the alternative hypothesis: a BF < 1.0 indicates evidence in favor of the alternative hypothesis and a BF > 1.0 indicates evidence in favor of the null hypothesis. BFs were computed using the web-applet on Dr. Jeffrey Rouder's Website (http://pcl.missouri.edu/bayesfactor) using the Jeffrey–Zellner–Siow prior (Rouder, Speckman, Sun, Morey, & Iverson, 2009). The BF statistics were based on the usual *t*. As a point of reference, a BF = .33 (1/3) corresponds to positive support for the alternative hypothesis whereas a BF < .05 (1/20) corresponds to strong evidence. Conversely, BF > 1 corresponds to evidence for the null hypothesis, with values greater than 3 denoting strong support. BFs between 0.33 and 3 generally are interpreted as providing weak evidence, and values close to 1 are essentially uninformative. Because the hypothesis of 'no difference' is a valid (and important) hypothesis to test, our statistical conclusions are based on interpretations of the BF, though traditional NHST results are reported for completeness.

4.2. Analysis of transfer to assessments of working memory capacity

We examined the individual cognitive tasks using Analysis of Covariance (ANCOVA) testing for post-test differences between conditions controlling for pre-test performance. These analyses were performed separately for posttest and the 3-month follow-up. To ease interpretation of the BFs, we present t-tests on the least-squared adjusted means, with variance due to pretest factored out. This provides for a straightforward test of whether the training and control groups differed at posttest (and follow-up) when participants are equated on pre-test performance. Raw means and standard deviations are presented in Appendix A.

The results of these analyses are presented in Table 2. The second to last column provides the results using traditional NHST, and the final column provides the BF. Although there are four total significant differences according to NHST at post-test, only operation span and deciphering symbols yielded positive evidence (BF < 0.33); symmetry span and listening span showed only weak evidence for WM training effectiveness (.33 < BF < 1.0). The remaining statistical analyses all favor the invariance hypothesis. This same general pattern of findings held for analyses based on the follow-up assessments: Participants show positive training benefits for operation span and weak evidence for symmetry span and rotation span. While we expected there to be little effect of

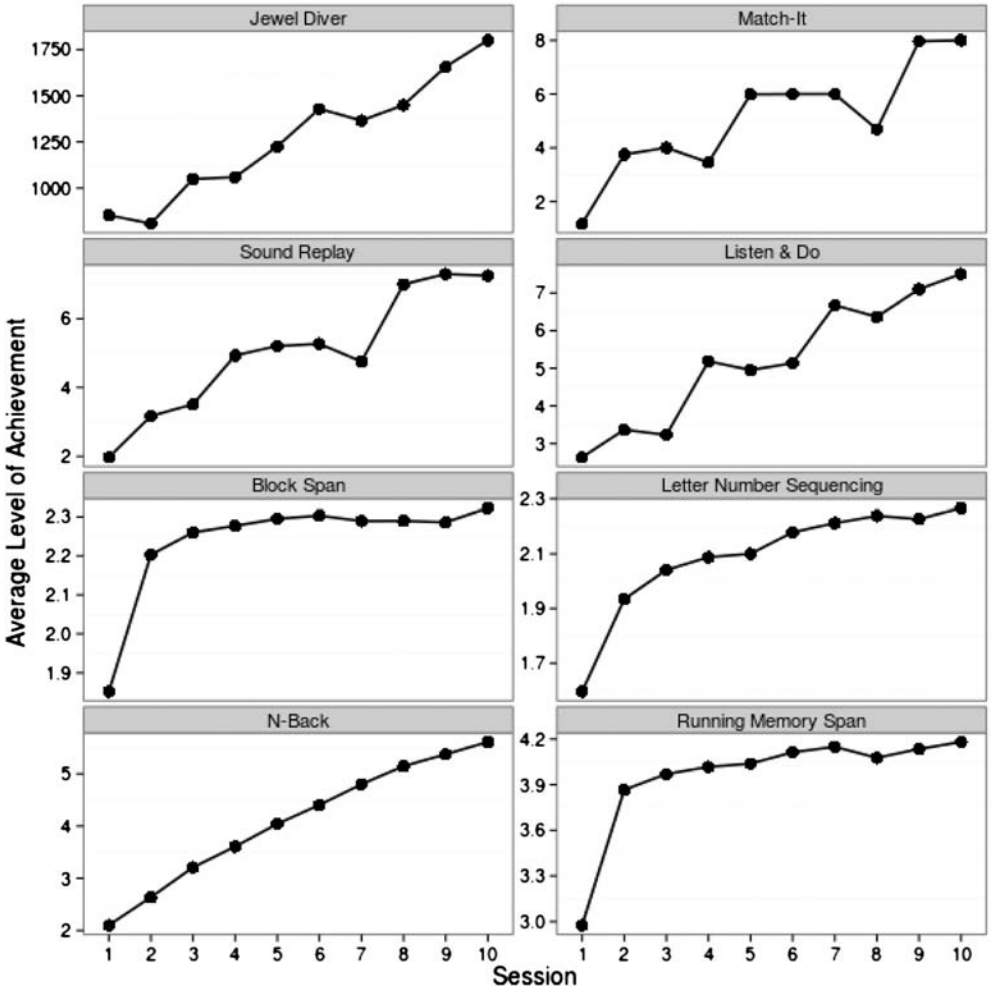

Fig. 1. Training curves for Study 1: The charts plot the average performance on all eight of the working memory training tasks. As these training curves indicate, participants' performance on the training tasks improved substantially. The Y-axis varies as an indication of performance in each task: block-span and letter-number-sequencing performance are depicted as number of items correctly identified. N-back performance is depicted as the mean number of *n* back reached by participants for that session. Running-span performance is depicted as the highest span achieved. Jewel diver performance is depicted as the points achieved in two hundred units. Match-it, sound replay and listen-and-do performance are depicted as the stage achieved weighted by the temporal speech processing level. Note: Although all participants in the training group completed 20 1-hour training sessions, each individual training task was used only 10 times and never more than once per session.

training on crystallized abilities, we hypothesized that there would be sizable effects of training on many of the fluid abilities. This was clearly not the case. There were no training benefits for performance on Stroop, anti-saccade, and the training effects for the verbal reasoning tasks (deciphering languages and verbal inference) were inconsistent across task and measurement time points. The only consistent findings across post-test and follow-up are for operation span and symmetry span. Although the effects of training on operation span and symmetry span represent reasonably strong evidence for the hypothesis that training worked, it is somewhat puzzling that the effects did not extend consistently to the other measures of WM. Why might this be the case?

One possible explanation for this mixed pattern of results it that transfer effects were limited to tasks that shared characteristics or stimuli with the training tasks themselves. For instance, the operation-span task required participants to remember sequences of letters, as did several of our training tasks (LNS, n-back, and running span). Symmetry-span required participants to remember the spatial location of a dot flashed in a 4×4 grid, which is similar to the block-span training task. Given that the transfer tasks shared these important characteristics, we cannot rule out the possibility that training gains were stimulus or task specific, rather than reflective of a general increase in cognitive ability. We return to this issue in the General discussion section.

5. Experiment 2

The results of Experiment 1 show some promise for WM-training, yet we could not rule out the possibility that

Table 2
Least-squared adjusted means for post-test and follow-up scores in Study 1. Reported t-tests are based on the adjusted means after controlling for variance due to pretest and variance due to the pretest × condition interaction.

Assessment	Control		Training		Group comparisons		
	N	*M*	*N*	*M*	*t*	*p*	*BF*
Posttest							
Working memory							
Operation span	55	40.11	58	47.35	3.60	<.01	0.019
Listening span	52	35.71	57	39.19	2.30	<.05	0.58
Symmetry span	55	29.97	59	32.73	2.32	<.05	0.56
Rotation span	49	27.59	55	28.86	1.16	ns	3.52
Inhibition							
Stroop effect	35	−335.01	39	−299.80	1.23	ns	2.82
Antisaccade	53	−120.11	59	−106.86	0.45	ns	6.22
Verbal reasoning							
Deciphering languages	55	.719	58	.812	2.59	<.05	0.31
Inference	54	6.82	59	6.74	0.21	ns	6.73
Verbal abilities							
Reading comprehension	54	9.49	59	9.02	0.34	ns	6.51
Verbal analogies	47	14.68	59	15.85	1.36	ns	6.29
3-Month follow-up							
Working memory							
Operation span	47	37.62	49	46.07	3.62	<.01	0.019
Listening span	46	35.74	48	39.08	1.83	ns	1.34
Symmetry span	47	30.06	49	34.51	3.37	<.01	0.39
Rotation span	41	27.40	49	30.33	2.11	<.05	0.80
Inhibition							
Stroop effect	29	−361.63	29	−311.10	1.60	ns	1.62
Antisaccade	44	−79.48	46	−58.64	0.84	ns	4.45
Verbal reasoning							
Deciphering languages	46	.756	46	.790	1.21	ns	3.16
Inference	47	7.09	49	6.96	0.40	ns	5.92
Verbal abilities							
Reading comprehension	47	10.11	49	9.67	0.82	ns	4.65
Verbal analogies	47	15.68	49	15.47	0.78	ns	4.79

BF = Bayes factors.

the observed transfer effects were due to shared task characteristics between the training and transfer tasks. At the most general level, however, the results of Experiment 1 argue against the idea of process-general transfer, while potentially showing evidence for process-specific transfer. In Experiment 2, we more directly tested the process-specific hypothesis by designing training regimens that targeted two abilities: Inhibition/memory updating and spatial WM. The overall experimental design included four conditions: A placebo control condition in which participants practiced two tasks that were structurally similar to two of the training tasks but which lacked any demands on WM, and three 'training' conditions. In one condition (inhibition), participants played two tasks designed to target inhibition and memory updating; in a second condition (spatial), participants played two tasks designed to target spatial WM; and in the third condition (combo), participants played all four tasks. The inclusion of the combo condition was premised on the assumption that broad transfer could only be achieved by broad training, while the other two training conditions allowed us to test the process-specific hypothesis. Importantly, our battery of pre- and post-test assessments consisted of both tasks that shared process with the training tasks as well as tasks that were essentially assessment forms of at least one training task from each of the training conditions. This design allowed us to examine the degree to which transfer was limited to specifically trained tasks, versus generalized to tasks that shared process but had relatively little overlap in task characteristics.[1]

In addition to measuring cognitive abilities both pre and post training, we also collected data in three experiments at post-test only. These three experiments were designed to test the hypothesis that WM training might improve cognitive resiliency. That is, we hypothesized that if WM training improves cognitive ability, it might help offset the negative effects of divided attention and proactive interference, and improve learning rates.

6. Method

6.1. Participants

Two-hundred sixty-four participants (171 females), ages 22–50, enrolled in the study and completed the pre-training assessments. Of these, 138 participants (94 females) completed the pre-training assessment, the training regimen, and the post-training assessment (see Table 3 for more descriptive

[1] Experiment 2 was not designed to differentiate between the two explanations put forth in the discussion of Experiment 1 for why operation span but not listening span showed transfer effects. Indeed, the task analysis of Experiment 1 was not conducted until after we had finished data collection on Experiment 2.

Table 3
Comparison of participants who dropped out of the study with participants who completed the study.

	Dropouts (n = 126)	Completers (n = 138)	Significance test	Bayes factor (scaled JZS)
Sex (% F)	0.67	0.61	$X^2(2) = 1.15$, $p > 0.05$	–
Age	34.24 (9.62)	35.51 (9.14)	$t(257) = -1.08$, $p > 0.05$	5.80
Education	4.23 (1.49)	4.72 (1.29)	$t(258) = -2.82$, $p < 0.05$	0.22
Need for cog.	24.90 (19.05)	27.10 (17.54)	$t(258) = -0.97$, $p > 0.05$	6.48
Malleability IQ	23.13 (7.15)	24.40 (6.22)	$t(257) = -1.52$, $p > 0.05$	3.33
Reading span	54.22 (14.82)	56.68 (13.55)	$t(260) = -1.40$, $p > 0.05$	3.97
Shapebuilder	1265.70 (480.8)	1391.7 (550.6)	$t(250) = -1.93$, $p > 0.05$	1.66
Ravens # correct	8.84 (3.42)	9.89 (3.09)	$t(261) = -2.62$, $p < 0.05$	0.38
N-back hit-FA	−0.04 (0.36)	0.05 (0.30)	$t(259) = -2.32$, $p < 0.05$	0.77
ANT executive	97.70 (45.38)	92.48 (39.44)	$t(261) = 1.00$, $p > 0.05$	6.33
ANT alerting	43.68 (25.04)	39.88 (24.47)	$t(261) = 1.25$, $p > 0.05$	4.82
ANT orienting	40.87 (26.77)	38.09 (20.89)	$t(261) = 0.94$, $p > 0.05$	6.70
Task switch cost	258.8 (285.7)	275.1 (280.6)	$t(258) = -0.46$, $p > 0.05$	9.31
Nelson # correct	22.17 (8.07)	23.92 (6.84)	$t(20) = -1.89$, $p > 0.05$	1.82
Nelson time	35,867 (19,328)	37,070 (17,440)	$t(260) = -0.53$, $p > 0.05$	8.96

statistics). Participants were recruited from the Washington, DC community via listservs and paper fliers. Participants were compensated $30 for completing the pre-test and $70 for completing the post-test.

6.2. Design & procedure

Participants were randomly assigned to one of 4 possible conditions: a placebo control training group; an interference training group; a visuo-spatial training group, and a combination training group (which performed both the interference and visuo-spatial WM training tasks). Training took place remotely by logging onto a training website. Similar to Experiment 1 this study maintained a true double-blind pretest/posttest experimental design, in which neither participants nor study moderators had knowledge of the condition to which participants were assigned. All participants were given a brief practice session on all of the training tasks (including the placebo control tasks) during pre-test, to familiarize them with each of the tasks. Only after logging into the training website did participants know which tasks they were required to complete as part of their training. Participants first came to campus for a 3-hour pre-assessment session, in which they completed 7 assessment tasks, which measured (broadly) WM capacity, ability to resist interference, cognitive flexibility, and reading comprehension. Then participants completed an average of 14.03 h of cognitive (or active control) training, which was delivered via internet. Participants completed the training at home on their own computers, for 26 min/day for 28–35 days. Compliance with the training regime was monitored online and participants were sent e-mail reminders if they fell behind schedule. Following the completion of the training, participants returned to campus to complete post-training assessments. The post-training assessments included the same 7 tasks that were completed at pre-test, and an additional 3 tasks that measured ability to learn foreign language vocabulary under full and divided attention conditions; resistance to proactive interference; and Artificial Grammar Learning ability.

In addition to the above measures, participants completed a standard demographic questionnaire that included measures of age, sex, education level, handedness, and medication use (self report). Education was coded as: 1 = less than high school degree; 2 = high school degree; 3 = some college; 4 = bachelor's degree; 5 = some graduate school; 6 = master's degree; 7 = PhD, MD, JD. Participants were also given the option to provide a saliva sample (to be used for genotyping) and to complete a battery of measures related to physical fitness and exercise habits. These measures were taken during pre-test. Of the 264 participants, 89 participants were selected to be included in an fMRI study looking at neural correlates of WM-training. Forty-four participants completed both the pre-test and post-test fMRI sessions. These fMRI sessions took place in separate testing sessions after the behavioral assessments were administered. Behavioral data from these 44 participants is included in the analyses presented below. However, analysis of the genetic, fitness measures, and fMRI data is ongoing and will be reported elsewhere.

6.3. Assessment tasks

6.3.1. Automated Reading Span (Daneman & Carpenter, 1980; Turner & Engle, 1989)

Participants were asked to memorize letters while reading sentences in order to recall them at a later time. The Automated Reading Span task allowed participants to complete the task independently at their own pace. Participants were shown 3–7 sentence/letter items, and then asked to recall the letters. There were 75 possible items to remember. Reading span is composed of two sections, which provide a practice session before participants proceed with the actual experiment. The practice sessions were divided into three sections. During the first practice session, participants saw a simple letter span. This round of the practice session consisted of seeing letters that appear on the screen one at a time, and then recalling them in the same order. Participants were asked to click a box next to the correct letters during recall. The second part of the practice session consisted of practicing the sentence portion of the experiment. Participants were shown 15 sentences and asked to determine if they made sense or not by selecting “true” or “false”. The final practice session included both letters and sentences, in which participants were shown a sentence and asked to verify whether it made sense, and then they saw a letter to be recalled. During the experimental blocks, the letters

appeared on the screen for 800 ms. After the letters disappeared, the participants were asked to recall the letters in the same order as they appeared. After the recall portion was completed, participants were given feedback about their performance and told how many letters they recalled correctly. Participants were asked to keep an accuracy level of at least 85%, which was presented in red in the upper right hand corner of the screen during recall. Participants' absolute scores were equal to the number correct when the entire trial was recalled correctly. Participants' total scores were equal to the number of correctly recalled letters.

6.3.2. Raven's Advanced Progressive Matrices (Raven, Raven, & Court, 1998)

Participants viewed eight black and white figures arranged in a 3 × 3 grid with one figure missing. Participants chose the image that best completed the pattern from eight possible choices. Participants completed either the odd or even problems at pre-test, and they then completed the complementary set of problems at post-test. Participants were given 10 min to complete as many problems as possible. The dependent variable was the number of correctly solved problems.

6.3.3. Shapebuilder

Shapebuilder is a visuo-spatial WM task in which participants were asked to remember the order and spatial position that a series of colored shapes were presented (Atkins et al., submitted for publication). Participants viewed a 4 × 4 grid of connected squares. Then, participants observed a sequence of between 2 and 4 colored (red, blue, yellow, or green) shapes (circles, triangles, squares, or diamonds) appearing one at a time in one of the 16 possible grid locations. Participants were asked to remember the location of each item, the shape of each item, the color of each item, and the order in which items appeared. After the final shape of a trial was presented, participants were asked to recreate the sequence by clicking on the correct colored shape and dragging it to the appropriate location. Participants completed 26 trials, of which 6 had 2 stimuli per trial, 9 had 3 stimuli per trial, and 11 had 4 stimuli per trial. The Shapebuilder task varies in difficulty in two ways: (1) trial length increased from 2 to-be-remembered shapes to 4; and (2) within each set of trials of a given trial length, the trials varied the number of distinct dimensions. At the easiest level, items were all the same shape or color, but appeared in different locations. At the most difficult level, items were all different colors and shapes, and appeared in different locations. Participants received immediate feedback about the accuracy of each item; the Shapebuilder task displayed the points awarded for each item immediately after the participant released the mouse button.

The dependent variable on this task was participants' final score, which was calculated as follows. Participants only received points for items that were placed in the correct location and the correct order. Participants received 15 points for getting the first item of a trial correct (correct location, color and shape) and received increasingly more points for each additional correct item: an additional 30 points for getting the second item correct after getting the first item correct, an additional 60 points for getting the third item correct after getting the first two items correct, and an additional 120 points for getting the fourth item correct after getting the first three items correct. If participants missed an item in the sequence (either entirely or partially—by forgetting one or more features), the scoring started over such that they received 15 points for the next item that was completely correct, and then 30 points if the following item was correct, and 60 if the following item was correct. Participants received partial points for items that were partially correct. Participants received 5 points for any item placed in the correct location (and right order) with the correct color, but the incorrect shape. Participants received 10 points for any item placed in the correct location with the correct shape, but the incorrect color. The task duration was between 5 and 10 min.

6.3.4. N-back

The stimuli for the N-back task included upper- and lower-case letters from the English alphabet. Letters were presented one at a time in white, 22 point Courier New font on a black screen. Each letter was displayed for 500 ms followed by an interstimulus interval of 2000 ms, after which the next letter was displayed. Participants were instructed to respond by pressing the 1 key on the number keypad if the letter shown matched the one shown N letters ago or 2 if the letter did NOT match the one shown N letters ago. Participants were instructed to treat upper- and lower-case versions of the same letter as a match. Participants were shown several examples for different levels of N before beginning the task. In the task, participants completed 50 trials at each of three levels of N: N = 2, N = 4, and N = 6. Participants first completed all 50 N = 2 level trials, then completed the N = 4 trials, and completed the N = 6 trials last. The sequences were created so that each sequence has 9–11 target items, 16–17 lure items, and 32–34 distractors. The dependent variables examined for this task included D-Prime and C-Bias (for target and lure trials).

6.3.5. Task switching

Adapted from a paradigm developed by Rogers and Monsell (1995), participants were asked to make magnitude (lower/higher than 5) and parity (odd/even) judgments of target digits (1–9, excluding 5). The words *Magnitude* and *Low-High* cued the magnitude task and the words *Parity* and *Odd–Even* cued the parity task.

Each trial in a block began with a 500 ms fixation display. A cue was then presented centrally, replacing the fixation display. After a Cue-Target Interval (0, 100, 200, 400, or 800 ms), a target was presented. Cue and target remained visible until participants made a response, after which the screen was cleared for 500 ms. The next trial commenced immediately thereafter. The responses were made with the Z and slash keys on a QWERTY keyboard, with same-task categories assigned to different keys and category response assignments counterbalanced across participants. Reminders of the category-response assignments appeared in the bottom corners of the screen during the experiment. Participants were instructed to respond quickly and accurately. Participants completed one practice block with 62 trials and one main block with 200 trials. Cued trials were randomly selected from the full set of Cue × Target × CTI combinations. The dependent variables examined for this task included task switching cost, calculated as the difference between the mean

reaction time for trials in which the task switched versus the mean reaction time for trials in which the task and cue remained the same. We also examined cue-switching cost, calculated as the difference between the mean reaction time for trials in which the cue (but not task) switched versus the mean reaction time for trials in which the cue and task remained the same. Because the findings for the cue-switch costs were identical to those for task-switching, we report only the latter below.

6.3.6. Attention networks' task (ANT)

Participants were presented with a series of trials in which they viewed 5 symbols (arrows or straight lines), and for each display they were asked to determine whether the middle arrow was pointing right or left (Fan, McCandliss, Sommer, Raz, & Posner, 2002). Participants first viewed a fixation point, indicated by a + symbol, then saw one of four possible cues, and then saw the target display. The target display always appeared in one of two locations: either directly above or below the fixation point. Participants either saw no cue, a central cue (an asterisk appearing at the location of the fixation point), a double cue (an asterisk appearing at each of the two possible locations of the target), or a spatial cue (a single asterisk occurring at the same location that the target would ultimately occur). Further, there were three possible Flanker types. The central arrow either was shown with no arrows but rather straight lines on either side of it (neutral condition); with arrows pointing in the same direction as the target arrow (congruent condition); or with arrows pointing in the opposite direction as the target arrow (incongruent condition).

Participants were asked to focus their attention on the fixation point and then press the right arrow button on the keyboard if the central arrow pointed right, or to press the left arrow button on the keyboard if the central arrow pointed left. Participants were asked to respond as quickly and accurately as possible. Participants completed a practice block of 24 trials (with feedback) and three test blocks of 96 trials each with no feedback. The entire task took around 25 min to complete.

The alerting effect was calculated by subtracting the mean RT of the double-cue conditions from the mean RT of the no-cue conditions. When no cue is presented, attention tends to be spread across both possible cue locations. The double cue keeps attention spread in these two locations, but provides temporal information that the target will appear very soon. The orienting effect was calculated by subtracting the mean RT of the spatial cue conditions from the mean RT of the center cue. The executive control effect was calculated by subtracting the mean RT of all congruent flanking trials (across all cue types) from the mean RT of incongruent flanking trials.

6.3.7. Swahili divided attention task

This task, presented post-intervention, examined participants' ability to learn foreign language vocabulary, while performing a secondary task that divides attention during learning. Swahili words were presented in pairs along with their corresponding English word. The Swahili/English word pairs were selected from Nelson and Dunlosky's (1994) word norms. Participants were asked to remember these pairings while simultaneously performing a finger-tapping task, as a task of divided attention. After viewing 20 word-pairs, each presented for 7 s, participants were presented with each of the Swahili words and asked to recall the English definition that matched the Swahili word. Participants completed 3 learning/recall trials with the same 20 words, so that we could measure learning over time. Participants then completed 3 trials of 20 new words. One set of 3 trials was completed while performing the "hard" level of the divided attention task, and the other set of three trials was completed while performing the "easy" level of the divided attention task. Participants performed the divided attention task only during the encoding portion of the task, and not during the recall portion of the task.

For the divided attention task (based on Moscovitch, 1994; Sprenger et al., 2011), participants performed a concurrent finger-tapping task. Participants placed the four fingers of their right hand on the "j", "k", "l", and ";" keys in typing position (i.e., index finger in the "j" key, middle finger on the "k" key, ring finger on the "l" key, and littlest finger on the ";" key). Participants were required to press each key when they heard a tone associated with that key. Each key was associated with a different-pitched tone. The lowest pitch was associated with the "j" key, and the highest pitch was associated with the ";" key. In the easy divided attention condition, the tone sequence always began with the lowest-pitched tone and incremented sequentially to the highest-pitched tone. Thus participants pressed keys in order from index finger, to middle finger, to ring finger, and finally to littlest finger. The sequence then began again with the lowest-pitched tone and continued in the same cycle. In the difficult divided attention condition, the tone sequence was random. Thus, participants were required to pay more attention to the tones because there was no predictable pattern to the sequence of tones. Participants had 500 ms to respond before the next tone played. If participants did not respond in time, the trial was counted as incorrect and the next trial began.

The dependent variables examined for this task included mean percent correct recall for the 3 trials under the Easy DA condition and for the 3 trials under the Hard DA condition, as well as mean percent correct finger tapping trials for the same trials.

6.3.8. Proactive interference

The proactive interference (PI) task was modeled after Jacoby, Wahlheim, Rhodes, Daniels, and Rogers (2010)'s task. Participants learned 40 word pairs that were semantically related (i.e., ale-brew, sugar-candy) on a first list. After viewing all 40 word-pairs 3 times (each for 2 s), participants were presented with the first word and asked to recall the second word of the word pair. In a second trial, participants viewed 20 word pairs, which matched exactly the pairings from the first list (ale-brew). These were considered "facilitated" pairings, because they had been learned on 4 separate learning trials. Twenty word pairs from the first list were transformed to PI items by keeping the root word (the word that appeared on the left during the first trial) but switching the word paired with it (i.e. ale-brew became ale-beer). Then, participants learned 20 new word pairs (control items). After viewing these 60 items (20 facilitate, 20 PI, 20 control), participants were asked to recall the words paired with the root word. The dependent variables examined for this task included mean percent correct recall on List 1, and mean percent correct recall for Control, PI,

and Facilitate items on List 2. We also examined the mean number of false alarm recalls participants gave for PI items on List 2.

6.3.9. Artificial grammar task

The artificial grammar task is presented to participants during post-intervention cognitive testing to evaluate transfer of cognitive training to proficiency in this process. The task evaluates participants' ability to discern rules of grammar from one session where strings of letters are shown that adhere to a set of rules, by asking them to endorse new sequences of letters (that may or may not follow the rules presented in the first session) as grammatical or not. The grammar rules presented in the first session are complex enough that participants are unlikely to be aware of any strategy, and must instead rely on a "hunch" in determining if the rule applies with the novel letter strings in the second session. This task was developed by Knowlton and Squire (1994) to isolate implicit relative to explicit (or conscious) rule learning. This task manipulated Chunk Strength (the frequency with which sequences that appear at test also appear in training, thus resulting in exemplar-based similarity effects) and grammaticality. The dependent variable for this task was the percent of correct decisions at test.

6.3.10. Nelson–Denny reading test

Participants completed a computerized version of the Nelson Denny Reading Test (Brown, Fishco, & Hanna, 1993). Participants completed either form G or H at pre-test and the opposite version at post-test. Participants were presented with a series of reading selections and asked to answer multiple-choice questions about them. Participants were shown 7 reading selections and asked to answer a total of 38 comprehension questions, each with 5 answer choices. Each selection was presented on the screen and participants pressed the space bar to advance to the following page. Participants were instructed to read through the entire passage and then answer the questions following that passage by choosing the key corresponding to the correct answer. Participants were informed that they should pay careful attention while reading, because they would not be able to look back at the passages to review the material. Participants were given 15 min to complete the task. The dependent variables for Nelson Denny were reading time for a sample passage and the number of correctly answered comprehension questions.

6.3.11. Need for cognition (NFC)

This scale identifies an individual's "tendency to engage in and enjoy thinking," (Cacioppo & Petty, 1982), and is used in this study as a covariate to account for individual differences in motivations to engage in challenging cognitive activities. We hypothesized that those individuals who are motivated to engage in cognitive activities would be more likely to show transfer effects. Participants are asked to rate their agreement or disagreement with 18 different statements using a −4 to +4 Likert-type scale. Some examples of items to be rated include such ideas as to whether an individual enjoys coming up with new solutions to problems, prefers to "let things happen" rather than trying to understand why, prefers small projects to long-term ones, or whether he or she puts more thought into a task than minimally required. This scale was developed and validated by Cacioppo and Petty, with additional validation contributed by Osberg (1987).

6.3.12. Implicit theories of intelligence scale

This scale is a measure of the extent to which individuals believe that intelligence can be changed (Dweck, 2000). Beliefs assessed with this scale can affect motivation, which in turn can potentially affect the outcome of cognitive training (Mangels, Butterfield, Lamb, Good, & Dweck, 2006). For example, a person who believes that intelligence is malleable is going to perceive learning in a positive manner, maintaining perseverance towards learning goals, whereas one with a belief that intelligence is fixed may harbor a feeling of helplessness in relation to learning, possibly undermining success. This scale contains 8 statements to rate as for agreement or disagreement—such as "The effort you exert improves your intelligence." or "You are born with a fixed amount of intelligence."

6.4. Training tasks

6.4.1. Active placebo control

Participants in the active control condition completed two tasks each day, each for 13 min. The design of these tasks was based on the desire to have participants engage in a task that is minimally different from the actual training tasks in terms of their structural characteristics. These tasks are called RememberMe and FollowMe. The RememberMe task required participants to view a series of images and decide for each image whether they had seen that image previously in the sequence. Participants responded "yes" or "no" for each image. This task was basically a continuous recognition memory task, and similar in instructions but without WM-load demand, to the n-back training task (discussed below). The main difference between n-back and RememberMe is that participants did not have to match items on the nth trial but rather if the images had been presented previously in the sequence of pictures presented in the training session. Points were awarded for correct responses to 'targets' only, using the equation: $15 * (2\text{^}(M - 1))$, where M = the lessor of the number of correct responses from last incorrect response and 5.

For the FollowMe task, participants viewed a grid (as in the Memnosyne and Shapebuilder task), and were instructed to click on a square if a flower image appeared in one of the squares within the grid. On the majority of trials, a square would be highlighted with no flower appearing. Thus, the task required vigilance on the part of participants; they needed to pay attention and respond when a trial had a target image appear, and do nothing otherwise. Points were awarded for correct responses to 'targets' only, using the equation: $15 * (2\text{^}(M - 1))$, where M = the lesser of the number of correct responses from last incorrect response and 5.

6.4.2. Memory updating & interference training

The second group trained on the N-back task, similar to the n-back task in the first experiment, and a new task called "Floop" (a cross between the Flanker task and the Stroop task). Participants played each task for 13 min each day (for a total of 26 min of training/day). For the N-back task, participants viewed letter stimuli and decided (yes/no) whether the current stimulus matched the stimulus "n" trials before. Participants advanced through three lure conditions (no lures, lures at n + 2

and n − 2, and lures at n + 1 & n − 1). The total number of lures was the same as the number of targets. Participants advanced to the next lure or n level when they performed <=3 incorrect responses in the trial sequence, and fell back a step if they performed >=5 incorrect responses. When participants made 4 incorrect responses, they remained at the same difficulty level. Points were awarded for correct responses to 'targets' only, using the equation: 15 * (2^(M − 1)), where M = the lesser of N + 2 or the number of correct responses from last incorrect response.

For the Floop task, participants viewed a stimulus that consisted of 5 letters (e.g. CCXCC). Participants also heard an auditory presentation of a letter. If the letter presented auditorily matched the center letter in the visual display, participants were instructed to respond by clicking the "yes" button. If the auditory stimulus did not match the center letter of the visual stimulus, participants were asked to click the "no" button (participants also had the option of pressing the right arrow key for "yes" and the left arrow key for "no" if they preferred). There were several types of trials. First, the "flanker letters" (the 4 letters surrounding the center letter) could either match the center letter (congruent trials, i.e. AAAAA) or not match (incongruent—note that the 4 flanker letters are always the same letter, i.e. AABAA). Furthermore, the sound could match the center letter, the flanker letters, or neither. The visual stimulus appeared in random locations on the screen, rather than always appearing in the same location in the center of the screen.

The first time participants played the task, the perceptual threshold for participants was found. The stimulus duration remained fixed after this trial. The inter-trial interval (ITI) started at the users' response, and was always 500 ms. Auditory stimulus onset began 100–150 ms before the visual stimulus onset. Stimuli were presented in blocks of n = 50.

The Floop task became more difficult over time in several ways, based on the following criteria: If performance (percent of correct responses) was between 65% and 85%, the response window remained the same. If performance was below 65%, the response window was increased by 10%, and if performance was above 85%, the response window decreased by 10%. This process continued until either the participant had 3 (non-consecutive) blocks at the same level, at which time participants moved to the next level of difficulty on the task. The first difficulty level varied the distance of the visual stimulus (the number of character spaces between each character). The task began with 3 spaces between each letter, and incremented down in 0.5 steps to the level 0.5 between letters. The second level of difficulty was color pop-out. In 50% of trials, all letters were presented in black font. In 10% of trials, one of the flankers closest to the target was a different color from the rest of the letters. In 10% of trials, one of the flankers in the outside positions was a different color from the rest of the letters. In 30% of trials, the target was a different color compared to flankers. When this level began, the letters moved back to the widest distance and moved through the distances back to 0.5 again before moving to the next difficulty level. The third difficulty level was the similarity of the items in the stimulus set (all characters in black font again). Targets and flanker letters were similar to each other (i.e. O's and Q's, X's vs. K's, etc.). Again, this level reset the distance back to the widest distance and moved through the distances before moving to the next difficulty level. The fourth (and final) difficulty level was spatial cueing. Asterisks cued the location of the visual stimulus. On 30% of trials, the asterisk appeared in the location of a flanker and on 70% of trials the cue appeared in the position of the Target. Again, participants moved through the distance levels with the spatial cueing level. They were awarded points for correct responses according to the following equation: Base * (2^(M − 1)). The base was 10 for congruent trials and 15 for incongruent trials and M was the number of correct responses since the last incorrect response, but was capped at the current level number.

6.4.3. Visuo-spatial WM training

Participants in the third group trained on Shapebuilder and block span tasks (the latter was renamed "Memnosyne", but was similar to the block span training task from Experiment 1). For Memnosyne, participants viewed a 4 × 4 grid. Stimuli consisted of N-long sequences of squares lighting up on the grid at a specified inter-stimulus-interval (ISI). On each trial, the participant recalled the correct sequence by clicking the same squares, in order, on the grid. For each n, participants progressed through 3 different inter-stimulus-intervals: 750 ms, 500 ms, and 250 ms. The stimulus duration was fixed at 500 ms. Participants advanced when they performed three correct trials in a row. Participants moved back a level when they performed incorrectly for three trials in a row. Participants received 15 points for the first correct response, 30 points for the second, and 60 points for the third, with the doubling continuing for each additional sequential correct response.

The Shapebuilder training task was based on the Shapebuilder assessment except that it adjusted in difficulty adaptively based on participants' performance. The stimuli consisted of N multi-feature (color, shape) stimuli objects displayed on the grid at a specified ISI. The participant was asked to repeat the sequence by dragging the appropriate sequence of objects onto the grid at the same locations as the stimulus sequence. For each N, participants progressed through 3 ISI's, 750 ms, 500 ms, and 250 ms. The stimulus duration was set to 500 ms. The stimulus set started at n = 2 with ISI = 250. Participants moved up a level when they performed 3 correct trials in a row, and moved back a level when they performed 3 trials incorrectly in a row. Participants received 15 points for the first correct response (location, shape, and color); 30 points for the 2nd correct response, 60 points for the third correct response, and so forth with the doubling continuing for each sequential correct response. For partial corrects, participants received 0 points if the location was incorrect; 5 points if the location was correct; 5 points if the location and color were correct, and 10 points if the location and shape were correct.

6.4.4. Combination training

The fourth group trained on N-back, Floop, Shapebuilder, and Memnosyne (block span) each day, playing each task for 6.5 min/day.

7. Results and discussion

Of the 264 participants who completed pretest, only 138 completed both the required training and the post-test assessments. Therefore, as a first step, we examined whether participants who failed to complete the study differed from those who completed. As shown in Table 3, completers and

dropouts differed only in terms of mean education level, number correct on Raven's progressive matrices, and n-back performance. Those who completed the study had slightly higher education, and performed slightly better on Raven's and n-back. Other than these three variables, the BFs all supported the hypothesis of no-difference between groups. Although completers performed significantly better on Shapebuilder, the BF for Shapebuilder slightly favored the null hypothesis. Nevertheless, inspection of the mean scores for the cognitive measures indicates that across all of the measures, those who completed the study performed slightly better than those who dropped out.

7.1. Pre-test correlations

The main goal of Study 2 was to examine the degree to which WM training transferred to cognitive tasks that tapped cognitive abilities targeted by the training tasks, but which differ in terms of task and stimulus specific characteristics. The first step in this analysis involves examining how the training tasks themselves correlate with non-trained abilities. For this analysis, we focus on the subset of assessment tasks that were used for training: Shapebuilder assessment, n-back assessment, and ANT-assessment. Table 4 presents the zero-order correlations among all of the assessment tasks. Several findings are noteworthy. First, n-back correlated moderately highly with Shapebuilder, Ravens, and reading-span, thereby suggesting that n-back shares processes with these other tasks. Given this pattern of correlations, it is reasonable to assume that, if n-back training improves the shared process, then training on n-back should transfer to Shapebuilder, ravens, and reading span. Likewise, Shapebuilder correlated highly with ravens, and reasonably high with n-back, reading-span, ANT and Nelson–Denny, again suggesting that Shapebuilder shares processes with these other tasks. Given this pattern of correlations, it is reasonable to assume that if training with Shapebuilder improves the processes shared by these tasks, then training on Shapebuilder should transfer to Raven's, n-back, reading-span, ANT and Nelson–Denny.

7.2. Training and transfer effects

Fig. 2 depicts the training curves for the various conditions. As is clear, participants improved considerably on the training tasks. But did WM training lead to improvements in untrained tasks? The answer to this question is no, despite the fact that participants improved considerably on the training tasks. Table 5 presents the results of the ANCOVA analyses using pre-test as the covariate. A full-reporting of the means is provided in Appendix B. As should be clear from Table 5, the only tasks that showed effects of WM training were those that were either identical or highly similar to the training tasks. For example, comparing the control group to the interference training condition revealed that people who trained on n-back and floop, showed significant effects on the assessment version of n-back and on the executive component of the ANT—two tasks that were either identical (n-back) or highly similar to the trained tasks (ANT). All other transfer results failed to reach significance and had BFs favoring the null hypothesis.

Likewise, comparing the control group to the spatial WM training condition revealed that the only task that participants showed significant effects on was Shapebuilder. This particular result is striking since the analysis of pre-test data revealed that Shapebuilder was well correlated with a variety of the other assessments, including Ravens ($r = .55$), n-back ($r = 0.45$), and reading-span ($r = .34$), yet none of these tasks even showed trends suggestive of transfer, let alone convincing support for the alternative hypothesis.

Finally, a comparison of the combination training condition with the control group revealed that participants in the combination group improved only on n-back, ANT, and Shapebuilder, and nothing else. Again, variants of these three tasks were included among the training tasks for this condition. Although, there was an effect of training on reading span, it was in the direction opposite of what was predicted: Participants in the combination group actually performed worse at post-test on reading span relative to the control group. In sum, the above data indicate that WM training was largely ineffective at improving performance on the pre/post assessment battery beyond performance on the trained tasks.

7.3. Post-session experiments

We had initially hypothesized that if WM-training was effective, then it could be used to improve people's resiliency to the effects of stress and divided attention. Here, we describe the results of three experiments that were carried out at post-test to

Table 4
Zero-order correlations for pre-test cognitive assessments, self-report, and demographic variables.

	1	2	3	4	5	6	7	8	9	10	11	12
1. Age	1											
2. Education	0.25	1										
3. NFC	0.09	0.24	1									
4. Beliefs IQ	−0.01	0.11	−0.05	1								
5. Shape-builder	−0.26	0.16	0.19	0.10	1							
6. N-back hit-FA	−0.05	0.22	0.23	0.08	0.45	1						
7. Reading span	−0.01	0.14	0.24	0.02	0.34	0.35	1					
8. Ravens	−0.06	0.35	0.36	0.18	0.55	0.52	0.35	1				
9. ANT exec	0.17	−0.04	−0.10	−0.02	−0.19	−0.08	0.05	−0.12	1			
10. ANT alert	−0.02	0.01	0.02	−0.06	−0.03	0.02	−0.04	−0.02	0.03	1		
11. ANT orient	0.02	0.07	0.03	0.05	0.01	0.13	−0.02	0.07	0.04	0.13	1	
12. Task switch	−0.04	−0.06	−0.12	−0.10	−0.13	−0.13	−0.13	−0.11	0.13	0.12	0.07	1
13. Nelson #corr	0.16	0.23	0.28	0.12	0.23	0.24	0.26	0.32	−0.04	0.00	−0.01	−0.19

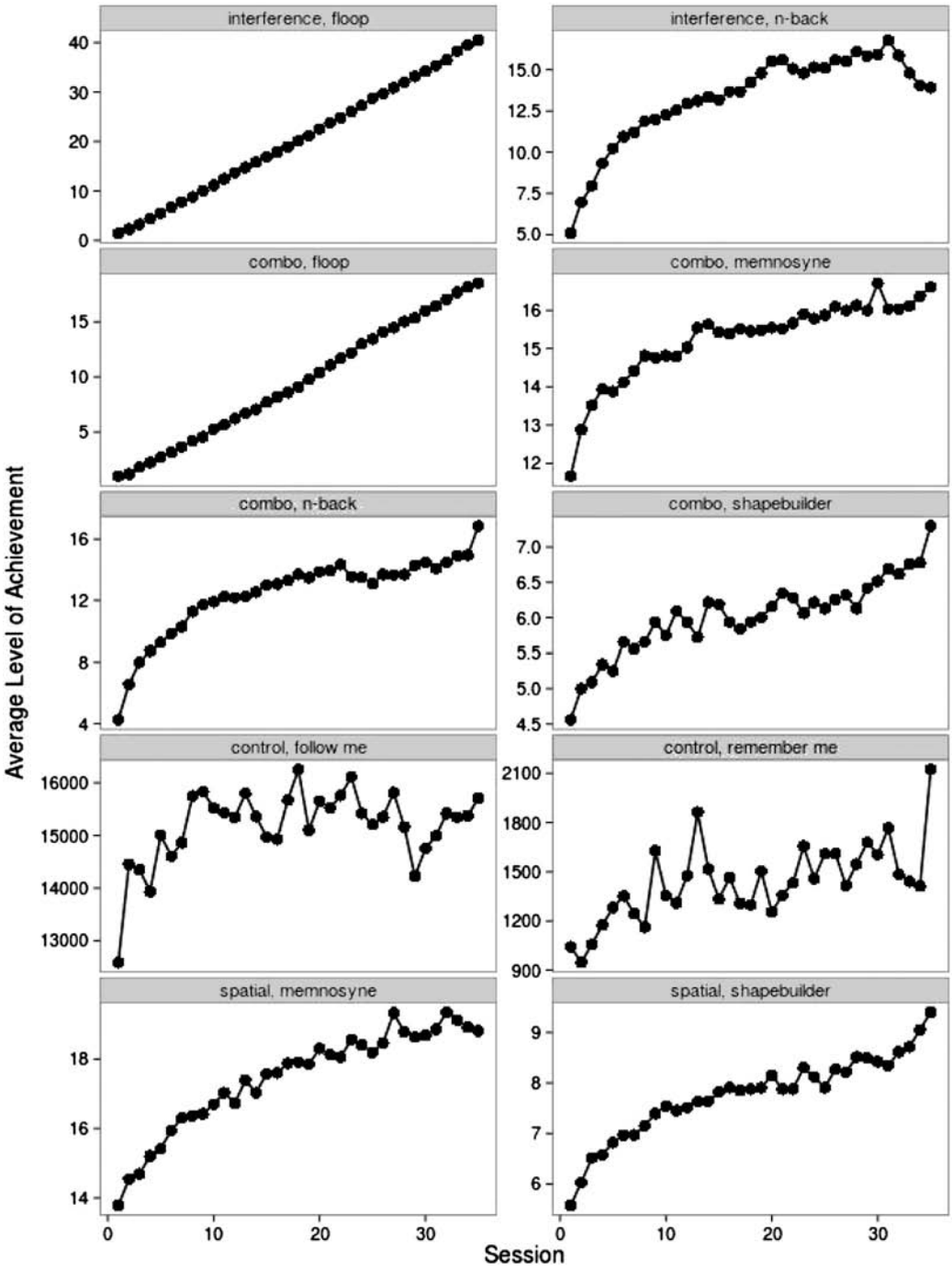

Fig. 2. Training curves for Study 2.

test the hypothesis that WM training improves participants' ability to effectively handle increased cognitive loads and proactive interference. These included the Swahili–English paired associate task, an experiment examining resistance to proactive interference, and an artificial-grammar task.

7.3.1. Effect of divided attention on Swahili–English learning

This study involved a multi-trial learning paradigm in which participants studied a list of Swahili–English word pairs on each of three learning trials. After each learning trial, participants were provided with a Swahili word, and asked to recall the English translation. This task was completed once under full attention and once under divided attention, as described in the procedure section above. Means and standard deviations are presented in Table 6.

Statistical analyses were conducted controlling for mean tapping accuracy on the finger tapping task. Not surprisingly, participants' recall improved across trials ($F(2, 132) = 50.55$,

Table 5
Least-squared adjusted means for post-test scores Study 2. Reported t-tests are based on the adjusted means after controlling for variance due to pretest and variance due to the pretest × condition interaction.

Assessment	*N*	*M*	*N*	*M*	*t*	*p*	*BF*
		Control		Interference			
N-back hits-FA	37	.19	34	.36	t(69) = 2.63	<.05	0.26[a]
Shapebuilder	35	1538.40	32	1402.62	t(65) = 1.79	ns	1.27
ANT-exec	37	76.61	34	61.61	t(69) = 3.54	<.01	0.02[a]
ANT-orient	37	35.60	34	26.29	t(69) = 2.49	<.05	0.35[a]
ANT-alert	37	45.15	34	43.92	t(69) = 0.25	ns	5.39[a]
Reading-span	37	58.42	34	54.86	t(69) = 1.71	ns	1.47
Task-switching (task)	37	134.19	34	102.52	t(68) = 1.06	ns	3.31
Raven's	37	10.09	34	9.60	T(69) = 0.86	ns	3.95
Nelson–Denny P (correct)	37	26.16	34	26.47	t(69) = 0.32	ns	5.29
Nelson–Denny RT	37	31,363.91	34	27,296.31	t(69) = 1.15	ns	3.02
		Control		Spatial WM			
N-back hits-FA	37	0.20	33	0.18	t(68) = 0.49	ns	4.93
Shapebuilder	35	1571.35	32	2408.57	t(65) = 7.60	<0.01	<0.0001[a]
ANT-exec	37	74.20	33	77.26	t(68) = 0.71	ns	4.36
ANT-orient	37	35.47	33	34.60	t(68) = 0.22	ns	5.38
ANT-alert	37	44.10	33	46.76	t(68) = 0.53	ns	4.84
Reading-span	37	59.84	33	58.08	t(68) = 0.91	ns	3.76
Task-switching (task)	37	137.199	33	146.41	t(68) = 0.32	ns	5.25
Raven's	37	10.03	33	10.51	t(68) = 0.76	ns	4.22
Nelson–Denny P (correct)	37	26.09	33	26.71	t(68) = 0.67	ns	4.48
Nelson–Denny RT	37	31,285.43	33	31,007.07	t(68) = 0.08	ns	5.49
		Control		Combination			
N-back hits-FA	37	0.21	33	0.40	t(68) = 3.36	<0.01	0.04[a]
Shapebuilder	35	1481.51	30	2055.90	t(63) = 5.66	<0.01	<0.0001[a]
ANT-exec	37	77.21	34	67.72	t(69) = 1.85	ns	1.17[a]
ANT-orient	37	37.48	34	29.08	t(66) = 2.29	<0.05	0.53[a]
ANT-alert	37	43.79	34	39.90	t(69) = 0.70	ns	4.40[a]
Reading-span	37	59.19	34	54.39	t(69) = 2.36	<0.05	0.46[b]
Task-switching (task)	37	143.40	34	109.52	t(69) = 1.33	ns	2.45
Raven's	37	9.56	34	9.66	t(69) = 0.17	ns	5.43
Nelson–Denny P (correct)	37	25.93	34	25.17	t(69) = 0.68	ns	4.45
Nelson–Denny RT	37	30,410.4	34	32,232.8	t(69) = 0.51	ns	4.88

[a] Assessment task was similar or identical to training tasks.
[b] Effect in wrong direction, combination condition worse than control.

$p < 0.01$), but there was no main effect of divided attention (tapping) on recall accuracy ($F(1, 133) = 0.64$), which held regardless of whether tapping was included as a covariate. More importantly, there was neither an effect of training condition, nor were there interactions between training condition and any of the other independent variables. This pattern of results held even when controlling for cognitive ability (which was defined as the composite of pretest Shapebuilder, Ravens, and Reading Span) and NFC. Thus, consistent with the findings from the pre-post assessments, training had no impact on participants' ability to learn Swahili–English word pairs.

Table 6
Mean percent correct (standard deviations) per learning trial for the full attention (Easy) and divided attention (Hard) conditions in the Swahili paired associates task.

	Control	Interference	Spatial WM	Combo
Full attention				
Trial 1	0.20 (0.20)	0.22 (0.18)	0.11 (0.09)	0.16 (0.15)
Trial 2	0.44 (0.26)	0.51 (0.30)	0.34 (0.20)	0.38 (0.24)
Trial 3	0.62 (0.29)	0.67 (0.26)	0.56 (0.25)	0.57 (0.28)
Divided attention				
Trial 1	0.14 (0.16)	0.14 (0.16)	0.16 (0.20)	0.17 (0.18)
Trial 2	0.37 (0.32)	0.37 (0.27)	0.39 (030)	0.40 (0.30)
Trial 3	0.50 (0.34)	0.54 (0.30)	0.56 (0.34)	0.58 (0.33)

Table 7
Mean percent correct for cued recall for the proactive interference task by item type: List 1 recall, List 2 proactive interference items, List 2 facilitation items, and List 2 control items.

	Control	Interference	Spatial WM	Combo
List 1 (percent correct)	0.79 (0.21)	0.79 (0.20)	0.82 (0.19)	0.81 (0.18)
List 2 PI items (percent correct)	0.63 (0.26)	0.63 (0.22)	0.69 (0.21)	0.60 (0.24)
List 2 Facilitate items (percent correct)	0.87 (0.16)	0.88 (0.14)	0.86 (0.16)	0.87 (0.11)
List 2 Control items (percent correct)	0.52 (0.20)	0.58 (0.21)	0.60 (0.19)	0.58 (0.22)
Interference intrusion errors	3.30 (2.68)	2.64 (2.53)	2.78 (2.71)	4.58 (3.27)

7.3.2. Proactive interference task

We analyzed four dependent measures drawn from the proactive interference task, based on word type: List 1 learning accuracy (L1), accuracy on List 2 proactive interference items (L2PI), accuracy on List 2 facilitation items (L2F), and accuracy on List 2 control items (L2C). Means and standard deviations are presented in Table 7. Controlling for pre-test cognitive ability, there was no effect of training condition on any of the four measures, all p's > 0.20. Thus, consistent with the verbal learning data from the Swahili experiment, training did not improve learning in this task. However, pre-test cognitive abilities were a significant predictor for all four dependent measures (all F's (1,128) > 9.12, p's < .01). Additionally, inclusion of NFC as a second co-variate did not change these conclusions, though NFC was a significant predictor of the four dependent measures (all F's(1,135) > 4.88, p's < .05).

7.3.3. Artificial grammar learning task

For the AGL task, the measure of interest was percent correct on judgments of whether a sequence was grammatical or not at the testing phase. Means and standard deviations are presented in Table 8. Again using pre-test cognitive ability as a covariate, we found a main effect of Chunk Strength on the number of sequences endorsed as grammatical, $F(1, 123) = 67.23$, $p < 0.01$. There was also a main effect of whether items were grammatical, $F(1, 123) = 92.15$, $p < 0.01$, and there was an interaction between chunk strength and grammaticality, $F(1, 133) = 99.97$, $p < 0.01$. There was neither a main effect of training condition, nor were there any interactions between training condition and chunk strength or grammaticality, p's > 0.20. The inclusion of NFC as a covariate did not alter any of the statistical conclusions, though NFC was a significant predictor of performance overall, $F(1, 130) = 9.48$, $p < .01$. In contrast, cognitive ability was not a significant predictor, $F(1, 123) = 2.40$, $p = 0.12$. Thus, as with both the verbal learning (Swahili) and proactive interference experiments, training did not lead to significant improvements.

7.3.4. Beliefs about malleability of intelligence

Thus far, our analyses have ignored the potential moderating roll of beliefs about the malleability of intelligence and other cognitive abilities. Do the results change when controlling for beliefs about malleability? The answer to this question is no. We had speculated a priori that individual differences in beliefs about the malleability of intelligence would be a significant predictor of training gains, such that those who believe more strongly that intelligence is malleable would be most likely to show gains on the training and transfer tasks. However, this was not the case, as inclusion of beliefs about the malleability of intelligence as a covariate did not change any of the statistical conclusions regarding the effectiveness of WM training, nor did beliefs about malleability significantly predict performance on any of the tasks.

Table 8
Mean percent correct (and standard deviations) for the artificial grammar task by chunk strength (high versus low chunk strength) and grammaticality. If the item was "grammatical" it reflects the proportion of people who respond "yes, it is grammatical". For non-grammatical items, it reflects the proportion of people who say "no, it was not grammatical." Hits for grammatical items and correct rejections for non-grammatical ones.

	Control	Interference	Spatial WM	Combo
High chunk strength				
Grammatical	0.78 (0.18)	0.71 (0.17)	0.72 (0.20)	0.71 (0.23)
Non-grammatical	0.33 (0.16)	0.39 (0.21)	0.30 (0.17)	0.37 (0.21)
Low chunk strength				
Grammatical	0.73 (0.21)	0.68 (0.27)	0.75 (0.24)	0.75 (0.22)
Non-grammatical	0.62 (0.23)	0.66 (0.24)	0.58 (0.19)	0.61 (0.25)

8. General discussion

The overarching goals of the studies presented in this paper were to (a) evaluate the degree to which the empirical data support the hypothesis that WM training works (i.e., people improve on the tasks themselves *AND* the training transfers in some meaningful way), (b) characterize the nature of transfer. In two studies, we illustrated that the extent of transfer effects observed in our studies may be limited to tasks that shared specific characteristics of the training tasks.

In Experiment 1, the training battery consisted of eight cognitive tasks, three of which required short-term memory of serially presented letters (n-back, running-span, and LNS) and two of which required spatial memory for items in a grid (block-span and match-it). Among all of the transfer tasks in Experiment 1, the strongest evidence of process-specific transfer was for the ones in which the BF was consistently less than 1.0 in both post-test and follow-up, operation span and symmetry span. This appears to provide support for process-specific transfer, yet the evidence is not completely unambiguous since the effects did not extend to the other measures of WM. Why would training lead to consistent transfer on one measure of verbal WM (operation span) but not a second one (listening span), and why would we find consistent evidence of training for symmetry span (a measure of spatial WM) but not rotation span (a second measure of spatial WM)?

Considering these results from Experiment 1 and taking an optimistic view of WM training, one could attribute the results to process-specific transfer. This is a perfectly plausible conclusion given the results of Experiment 1 and assuming that the training and transfer tasks do not all engage the same processes but that some do. This view would be consistent with recent findings suggesting that transfer effects are dependent upon process-specific overlap between the training and assessment measures. von Bastian and Oberauer (2013) found specific transfer related to shared sub-functions of WM. Novick et al. (in press) likewise reported process-specific transfer from practicing information recharacterization in the n-back task to performing information recharacterization in a garden-path ambiguity resolution task. Further research and task analysis are needed to determine what processes are required for perceiving and remembering items in the training and transfer tasks.

Conversely, considering these results from Experiment 1 and taking a more pessimistic view of WM training, one could also argue that the transfer results are due to task-specific transfer and not process-specific transfer. The basis of this argument stems from the observation that the training and transfer tasks share characteristics (e.g., memoranda, operations, and other task demands). Arguably, operation span and symmetry span, which are the only two tasks to show positive evidence of transfer by the BF analyses, overlap with the training tasks in

important ways. For example, operation span required participants to remember a sequence of serially presented letters, as did 3 of the 8 training tasks (n-back, running span, and LNS); symmetry span required participants to remember the spatial location of serially presented stimuli, as did one of the training tasks, memnosyne, and a second training task, match-it, required memory for grid locations. Listening span required serial memory for wingding (non-letter) characters and failed to show consistent evidence of transfer effects. Taken together, this suggests that the observed transfer effects on operation span and symmetry span likely reflect task- or stimulus-specific transfer, rather than process-specific. It is impossible to conclude from Experiment 1 alone whether the observed transfer effect is due to the commonality of the underlying processes (and thus process-specific transfer) or to the commonality of task or stimulus-specific effects (and thus task/stimulus-specific transfer).

Two limitations of Experiment 1 were the lack of an active control, and that the fact that participants trained on a large battery of cognitive tasks. These two limitations make it difficult to interpret (a) the degree to which the limited transfer for the training condition was due to expectancies (Hawthorne effects), and (b) the degree to which training is specific to the set of processes targeted by training. In particular, because the battery of training tasks was so diverse, it is difficult to test the process-specific hypothesis cleanly. Experiment 2 addressed these limitations by including an active control condition plus three different forms of cognitive training: One focused on inhibition and memory updating, one focused on spatial WM, and one condition that combined inhibition and memory updating with spatial WM training. Importantly, Experiment 2 largely confirmed the interpretation of Experiment 1. In particular, we observed large and reliable pre-post changes for the trained tasks (with BF firmly indicating evidence that training led to task-specific improvements) but no evidence that the training gains extended beyond the trained tasks. For example, participants who trained with N-back and Floop improved on n-back and ANT, but not spatial WM or any other assessment; participants who trained on Shapebuilder and Memnosyne improved on pre/post assessments of Shapebuilder, but not on n-back, ANT, or any other task; and participants who trained on all four training tasks improved on n-back, ANT, and Shapebuilder, but nothing else. This pattern of data clearly supports the hypothesis that the effects of training are task specific.

Given that our data failed to reveal convincing evidence in favor of the hypothesis that WM training yields improvements on non-trained tasks, how do our results square with the existing literature? To be sure, there is still much debate over the validity of many of the previously reported examples of successful transfer effects. On the one hand, researchers from several different labs maintain that there is solid evidence for the effectiveness of WM training (e.g., Jaeggi et al., 2011; Kundu, Sutterer, Emrich, & Postle, 2013[2]). On the other hand, some efforts to replicate key findings have failed or produced limited and inconsistent results. Thompson et al. (2013) reported a failure to find transfer from WM training, and while they could not offer much explanation for why, they do call for new work to "discern the factors across studies that are associated with success or failure in having WM training improve fundamental faculties of the human mind as measured by improved performance on a range of untrained tasks…" (p. 13). Engle and colleagues have questioned the validity of some empirical studies of WM training on methodological grounds (Redick et al., 2013; Shipstead et al., 2012). Still others have raised questions about some of the analysis techniques used to support claims of WM training effectiveness (Tidwell, Dougherty, Chrasbaszcz, Thomas, & Mendoza, submitted for publication).

Importantly, in assessing claims that WM works (or does not), it is instructive to evaluate the degree to which the available evidence supports the hypothesis that WM training works, versus the hypothesis that it does not work. Doing so calls for a Bayesian approach to data analysis. Obviously, the results of our experiments and Bayesian analysis argue against the hypothesis that WM training is effective at improving anything more broadly than performance on the trained task or tasks that share common characteristics, such as stimuli or basic operations. Nevertheless, there are some important caveats to this claim. For one, both experiments reported herein utilized normally functioning adults, whereas many prior studies showing successful transfer have used children, a population with potentially greater plasticity, or elderly adults. There is reason to believe that training may be less effective in normally functioning adults compared to developing children or mildly impaired elderly. Second, it is entirely possible that the tasks used in our training studies were poorly designed and simply did not yield training effects due to inadequacies in the design features. Although it is impossible to entirely rule out this explanation, we find it unlikely for several reasons. First, in terms of Experiment 1, the majority of the training tasks had high face validity with many of the measures of fluid ability used as pre-/post-tests. For example, prior work has established that n-back, running span, multiple-object tracking, and block span are well correlated with standard measures of complex span. Second, as illustrated in the pre-test data from Experiment 2, it is clear that the training tasks used in that experiment were quite well correlated with several of the transfer measures. As one example, Shapebuilder and n-back assessment were correlated $r = .45$, yet training on Shapebuilder did not lead to improvements on the n-back assessment and training on n-back did not lead to improvements on Shapebuilder. Third, the construction of our training tasks included features believed to be necessary for successful WM training. For example, the tasks induced a high-cognitive load and adapted to individual participants' ability. For most of the training tasks, the difficulty of the task increased as a function of the participant's skill level. Moreover, there was clear evidence in the data that participants did indeed show significant training gains from pre to post tests, as evidenced by the pre/post assessments of Shapebuilder, n-back, and ANT. Thus, given our findings and the caveats pointed out above, we believe

[2] In our view, Kundu et al. (2013) evidence of transfer has to be viewed with some circumspect. They report significant gains for n-back training and non-significant gains for control training, but no interaction. They report the effect as evidence that training transferred, but we argue that it is just not there without the interaction.

that whether WM-training works is indeed a valid and important question in need of being addressed.

Moving beyond the results of the present studies, there is a good theoretical reason to believe that general cognitive functioning can be improved through training. However, we speculate that the effects could be limited to special populations (e.g., children, Melby-Lervåg & Hulme, 2012) and the gains more modest than evidenced by the existing empirical studies (Melby-Lervåg & Hulme, 2012). Exploration of the potential moderating role of individual differences, such as motivation, physical fitness, and genetic markers, may yield promising avenues for future research. Nevertheless, we suspect that meaningful transfer to general cognitive functioning is likely going to require much more training than typically used in most laboratory studies, including ours. Indeed, we suspect that the effect sizes of existing WM training studies are probably over-estimated in published results, for two reasons. First, unsuccessful training studies are likely to go without being published (publication bias). Second, and more importantly, most laboratory-based training studies have utilized relatively small samples. Although not widely recognized, measures of effect sizes tend to be over-estimated by small-samples (see Ioannidis, 2008; Yu, Sprenger, Thomas, & Dougherty, in press). Given that most training studies consist of relatively small samples, it is likely that the reported effect sizes are themselves overestimated.

8.1. A Bayesian analysis of WM training effectiveness

To our knowledge, our study is the first to include a Bayesian analysis of working memory training, which we view as particularly well suited for evaluating its effectiveness. For example, we suspect that at least some of the existing studies reporting positive transfer of WM training will fail the Bayesian "sniff test." Indeed, even for studies that have faithfully observed statistically significant effects of training it is instructive to evaluate these findings in light of one's subjective prior probabilities. For illustrative purposes, suppose a pessimist adopts prior odds of 10:1 *against* the effectiveness of WM training, citing the plethora of historical evidence that cognitive abilities are stable. In contrast, suppose an optimist adopts a prior odds of 1:10 in *favor* of the effectiveness of WM training. How might these two individuals change their beliefs in light of the available evidence?

Chein and Morrison (2010, Table 2) report significant one-tailed t-tests on the gain scores for both Stroop ($t(40) = 1.80$) and reading comprehension ($t(38) = 1.80$). The corresponding BFs = 1.06 and BF = 1.067, respectively, using the JZS prior. These BFs are interpreted as providing equivalent support for the null and the alternative—that is, the BF indicates that the data are equally supportive of both the alternative and null hypotheses. The t-tests for fluid IQ ($t(40) = 0.24$) and reasoning ($t(40) = 1.39$) were both non-significant, and have corresponding BFs of 4.37 and 1.92 in favor of the null hypothesis. The average BF across all four tasks is 2.10 in favor of the null. Turning to the experiments reported above, across all measures of *fluid* abilities in Experiment 1, the average BF at post-test is 2.59 in *favor of the null*, and this includes operation span and symmetry span which arguably reflects stimulus specific training effects. Similarly, the average BF of the untrained assessment tasks in Experiment 2 across all three training groups is 4.18, again in favor of the null. Multiplying these BFs with the priors gives us the posterior odds ratios. For the pessimist, the posterior odds *against* the effectiveness of WM is over 227:1 ($10 * 2.10 * 2.59 * 4.18$). This corresponds to a posterior probability $p(\text{null is true}|\text{data}) = 227 / 228 = 0.996$. But, even for the optimist, the posterior odds favors the null at a ratio of 2.27:1 ($0.1 * 2.10 * 2.59 * 4.18 = 2.27$), with a posterior probability $p(\text{null is true}|\text{data}) = 2.27 / 3.27 = 0.694$. In other words, based on the result of Chein and Morrison (2010) and the experiments reported herein, even the optimist should express some skepticism in the hypothesis that WM-training is effective.[3]

8.2. Summary

While our experiments and analyses offer a rather pessimistic view of the prospects that WM training is efficacious, much more work is needed to fully evaluate whether or not WM training does indeed work, and if so, the boundary conditions of the effects. To this end, we advocate for studies that challenge participants to engage in training for a much longer duration. We also advocate for both large-scale multisite studies, which is likely necessary to obtain a large sample size in a reasonable period of time, and the use of Bayesian methods that will allow researchers to evaluate the degree to which the evidence supports the null versus the alternative hypotheses.

Notes & acknowledgements

The work presented in this paper was supported by the University of Maryland Center for Advanced Study of Language with funding from the United States Government, and the Office of Naval Research (ONR) grant N000141010605 awarded to MRD, DJB, MFB, JMN, and JIH. First authorship is shared equally between Amber Sprenger and Sharona Atkins. The authors thank David Alexander, Stephanie Berger, Carrie Clarady, Ryan Corbett, Erika Hussey, Daniel Levitas, Alexei Smaliy, and Susan

[3] Chein and Morrison (2010) used adaptive versions of verbal complex working memory span (i.e., memory for letters plus lexical decision) and spatial complex working memory span (i.e., memory for locations plus symmetry decisions) as their training tasks. We excluded their cognitive assessments measure of temporary memory (which were also the verbal complex working memory test and spatial complex working memory test) from our calculation of the BFs from Chein and Morrison because the temporary memory measures were identical to the training tasks. Obviously, taking the average BF across the assessment tasks is an oversimplification, as there may be good reasons to expect some tasks to show evidence of transfer and others not. However, at least for the Chein and Morrison (2010) study, there is no a priori theory for which of the pre/post tasks should show transfer. The measures of crystallized abilities (AFOQT verbal analogies and AFOQT reading comprehension) were omitted when computing the mean BF for Experiment 1, as we did not expect transfer to these two measures. Finally, it is important to acknowledge that there are many more studies that could be folded into a Bayesian analysis, some of which show significant transfer effects, and others that do not. Our goal was to merely illustrate the usefulness of Bayesian analysis for helping to shed light on the question of whether WM training works.

Teubner-Rhodes for their assistance with collecting and scoring data. Additionally, we thank Barbara Forsyth for her valuable feedback in designing and implementing the first experiment. Correspondence should be addressed to Michael Dougherty, Department of Psychology, University of Maryland, College Park, MD 20742, mdougher@umd.edu, or Michael Bunting, Center for Advanced Study of Language, University of Maryland, 7005 52nd Ave., College Park, MD 20742, mbunting@casl.umd.edu.

Appendix A

Means and standard deviations for pre- and post-test for both the training and control conditions for Experiment 1. Note that there was a programming error for the Stroop task that affected 39 participants' data at pretest.

Assessment	Training			Control		
	N	M	Std	N	M	Std
Pre-training						
Working memory						
Operation span	59	35.59	17.20	55	37.16	15.42
Listening span	58	36.21	11.55	53	35.43	10.05
Symmetry span	59	29.73	7.34	55	29.18	7.93
Rotation span	59	27.24	7.95	53	25.75	8.71
Inhibition						
Stroop (RT)	41	−398.28	326.72	35	−369.76	223.82
Stroop (% correct congruent)	41	99.19	0.98	35	98.97	1.1
Stroop (% correct incongruent)	41	91.67	15.59	35	91.07	9.595
Antisaccade (RT)	59	−175.19	183.71	53	−214.49	384.91
Antisaccade (% correct prosaccade)	59	90.48	10.03	53	87.86	13.88
Antisaccade (% correct antisaccade)	59	55.96	16.96	53	51.73	16.35
Verbal reasoning						
Deciphering languages	58	8.66	2.68	55	7.96	2.89
Inference	59	6.54	2.31	55	5.76	2.47
Verbal abilities						
Reading comprehension	59	9.24	3.82	55	8.60	3.57
Verbal analogies	59	15.80	5.38	55	14.93	5.50
Working memory						
Operation span	58	47.09	13.59	55	40.55	16.78
Listening span	58	39.03	11.63	54	35.17	10.42
Symmetry span	59	32.88	7.39	55	29.71	9.75
Rotation span	55	29.47	7.98	52	26.77	8.96
Inhibition						
Stroop (RT)	57	−294.08	142.00	55	−322.25	156.37
Stroop (% correct congruent)	57	98.11	2.32	55	98.64	1.30
Stroop (% correct incongruent)	57	88.08	12.39	55	90.00	13.74
Antisaccade (RT)	57	−95.74	174.45	55	−125.44	200.31
Antisaccade (% correct prosaccade)	57	88.89	13.03	55	88.52	12.77
Antisaccade (% correct antisaccade)	57	59.33	19.27	55	59.76	16.35
Verbal reasoning						
Deciphering languages	59	9.81	2.45	55	8.47	2.60
Inference	59	6.88	1.71	54	6.69	2.27
Verbal abilities						
Reading comprehension	59	9.25	3.67	54	9.19	3.95
Verbal analogies	59	15.86	4.01	55	14.62	5.14
Post-training						
Working memory						
Operation span	49	46.63	14.00	47	36.85	17.80
Listening span	48	39.54	9.59	46	35.09	12.12
Symmetry span	49	34.78	5.88	47	29.60	9.79
Rotation span	49	31.37	8.20	44	26.34	10.32
Inhibition						
Stroop (RT)	46	−306.56	153.38	46	−325.08	147.46
Stroop (% correct congruent)	46	98.31	2.32	46	98.79	1.39
Stroop (% correct incongruent)	46	92.30	8.10	46	91.30	11.91
Antisaccade (RT)	46	−54.26	123.84	46	−82.89	159.53
Antisaccade (% correct prosaccade)	46	90.11	11.88	46	86.63	16.09
Antisaccade (% correct antisaccade)	46	62.36	17.65	46	60.80	17.57
Verbal reasoning						
Deciphering languages	46	9.70	2.29	46	8.87	2.62
Inference	49	7.14	2.00	47	6.89	2.14
Verbal abilities						
Reading comprehension	49	9.88	4.02	47	9.89	4.35
Verbal analogies	49	15.94	5.54	47	15.21	5.96

Appendix B

Means and standard deviations for pre and post testing sessions for Experiment 2.

Assessment	Pre-test			Post-test		
	N	*M*	*SD*	*N*	*M*	*SD*
Control condition						
N-back hits-FA	37	0.10	0.28	37	0.21	0.30
Shapebuilder	35	1440.71	592.22	37	1598.51	517.72
ANT-exec	37	88.38	30.20	37	74.76	23.91
ANT-orient	37	39.44	20.88	37	36.70	20.11
ANT-alert	37	33.32	19.07	37	42.75	21.49
Reading-span	37	55.86	14.95	37	58.84	12.28
Task-switching (task)	37	235.17	177.49	37	130.65	126.57
Raven's	37	10.27	3.42	37	9.95	3.42
Nelson–Denny correct	37	24.46	6.97	37	26.27	5.68
Nelson–Denny RT	37	31,636.76	10,629.63	37	29,897.30	12,819.27
Need for cognition	37	27.97	19.08	–	–	–
Beliefs in malleability of IQ	37	23.70	6.24	–	–	–
Interference training						
N-back hits-FA	34	−0.01	0.34	34	0.31	0.44
Shapebuilder	32	1390.00	640.58	32	1368.44	546.67
ANT-exec	34	96.57	49.05	34	63.43	25.79
ANT-orient	34	35.54	18.97	34	26.15	14.93
ANT-alert	34	46.72	28.54	34	47.01	22.29
Reading-span	34	54.38	14.24	34	54.41	11.97
Task-switching (task)	34	258.19	260.79	33	105.47	141.20
Raven's	34	9.59	3.23	34	9.76	3.68
Nelson–Denny correct	34	24.09	7.45	34	26.35	5.95
Nelson–Denny RT	34	41296.79	22666.57	34	27,716.88	15,426.67
Need for cognition	34	26.03	17.30	–	–	–
Beliefs in malleability of IQ	34	24.06	5.45	–	–	–
Spatial WM training						
N-back hits-FA	33	0.05	0.30	33	0.16	0.26
Shapebuilder	32	1496.41	475.38	32	2416.72	642.98
ANT-exec	33	85.91	32.32	33	76.66	21.45
ANT-orient	33	35.01	22.21	33	33.64	19.32
ANT-alert	33	40.97	23.08	33	47.82	21.50
Reading-span	33	59.52	11.88	33	59.42	10.87
Task-switching (task)	33	283.28	314.09	33	152.20	140.85
Raven's	33	10.52	2.83	33	10.61	3.58
Nelson–Denny correct	33	23.82	6.50	33	26.48	5.70
Nelson–Denny RT	33	40,924.33	16,052.01	33	32,486.88	14,804.32
Need for cognition	32	27.31	13.63	–	–	–
Beliefs in malleability of IQ	32	23.91	5.87	–	–	–
Combination training						
N-back hits-FA	33	0.06	0.31	34	0.39	0.23
Shapebuilder	30	1224.67	450.44	32	1951.72	706.28
ANT-exec	34	99.25	43.79	34	69.92	28.44
ANT-orient	34	42.17	21.51	34	29.29	14.66
ANT-alert	34	39.13	25.72	34	40.88	26.88
Reading-span	34	57.12	12.83	34	54.94	13.65
Task-switching (task)	34	327.43	351.86	34	115.95	99.44
Raven's	34	9.18	2.75	34	9.24	3.34
Nelson–Denny correct	34	23.26	6.64	34	24.76	6.83
Nelson–Denny RT	34	35,016.15	17,470.80	34	32,484.24	17,088.16
Need for cognition	33	27.03	19.94	–	–	–
Beliefs in malleability of IQ	32	26.06	7.22	–	–	–

References

Atkins, S. M., Sprenger, A. M., Colflesh, G. J. H., Briner, T. L., Buchanan, J. B., Chavis, S. E., et al. (submitted for publication). *Measuring working memory is all fun and games: A four-dimensional spatial game predicts cognitive task performance.* (submitted for publication).

Brain Fitness Program (Version 2.1) [Computer software]. San Francisco, CA: Posit Science, 2009.

Brown, J. I., Fishco, V. V., & Hanna, G. (1993). *Nelson–Denny reading test: Manual for scoring and interpretation, forms G and H.* Itaskca, IL: Riverside.

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology, 33*, 205–228.

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

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116.
Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston, MA: Houghton Mifflin.
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.
Chooi, W. T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*, 531–542.
Cohen, R. L., & Heath, M. (1990). The development of serial short-term memory and the articulatory loop hypothesis. *Intelligence*, *14*, 151–171.
Conway, A. R. A., Jarrold, C., Kane, M. J., Miyake, A., & Towse, J. N. (Eds.). (2008). *Variation in working memory*. New York, NY: Oxford University Press.
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.
Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.
Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433.
Duncan, J., Johnson, R., Swales, M., & Freer, C. (1997). Frontal lobe deficits after head injury: Unity and diversity of function. *Cognitive Neuropsychology*, *14*, 713–741.
Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development.* Philadelphia, PA: Psychology Press.
Ekstrom, R. B., French, J. W., & Haarmann, H. H. (1979). Cognitive factors: Their identification and replication. *Multivariate Behavioral Research Monographs*, *79*, 7–56.
Engle, R. W., & Conway, A. R. A. (1998). Working memory and comprehension. In R. H. Logie, & K. J. Gilhooly (Eds.), *Working memory and thinking* (pp. 67–92). East Sussex, UK: Psychology Press.
Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology. General*, *128*, 309–331.
Fan, J., McCandliss, B. D., Sommer, T., Raz, M., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *3*(14), 340–347.
Friedman, N. P., Miyake, A., Young, S. E., Defries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, *137*, 201–225.
Gathercole, S. E., Alloway, T. P., Willis, C., & Adam, A. M. (2006). Working memory in children with reading disabilities. *Journal of Experimental Child Psychology*, *93*, 265–281.
Geary, D. C., Hoard, M. K., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development*, *78*, 1343–1359.
Insight (Version 1.1) [Computer software]. San Francisco, CA: Posit Science, 2008.
Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648.
Jacoby, L. L., Wahlheim, C. N., Rhodes, M. G., Daniels, K. A., & Rogers, C. S. (2010). Learning to diminish the effects of proactive interference: Reducing false memory for young and older adults. *Memory & Cognition*, *38*(6), 820–829.
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 of the United States of America*, *108*, 10081–10086.
Jaeggi, S. M., Buschkuel, M., Jondies, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 6829-6833.
Jaeggi, S. M., Studer, B., Buschkuehl, M., Su, Y. -F., Jonides, J., & Perrig, W. J. (2010). On the relationship between N-back performance and matrix reasoning—Implications for training and transfer. *Intelligence*, *38*(6), 625–635.
Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working memory capacity. *Journal of Experimental Psychology. General*, *130*, 169–183.
Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21–48). New York, NY: Oxford University Press.
Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *153*, 128–217.
Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Science*, *14*, 317–324.
Klingberg, T., Fernell, E., Olesen, P., Johnson, M., Gustafsson, P., Dahlström, K., et al. (2005). Computerized training of working memory in children with ADHD—A randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*, 177–186.
Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology*, *24*, 781–791.
Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 79–91.
Kundu, B., Sutterer, D. W., Emrich, S. M., & Postle, B. R. (2013). Strengthened effective connectivity underlies transfer of working memory training to tests of short-term memory and attention. *Journal of Neuroscience*, *33*, 8705–8715.
Lövdén, M., Bäckman, L., Lindenberger, U., Schaefer, S., & Schmiedek, F. (2010). A theoretical framework for the study of adult cognitive plasticity. *Psychological Bulletin*, *136*, 659–676.
Mahncke, H. W., Connor, B. B., Appelman, J., Ahsanuddin, O. N., Hardy, J. L., Wood, R. A., et al. (2006). Memory enhancement in healthy older adults using a brain plasticity-based training program: A randomized, controlled study. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 12523–12528.
Mangels, J. A., Butterfield, B., Lamb, J., Good, C., & Dweck, C. S. (2006). Why do beliefs about intelligence influence learning success? A social cognitive neuroscience model. *Social Cognitive and Affective Neuroscience*, *1*(2), 75–86.
McNab, F., Varrone, A., Farde, L., Jucaite, A., Bystritsky, P., Forssberg, H., et al. (2009). Changes in cortical dopamine D1 receptor binding associated with cognitive training. *Science*, *323*, 800–802.
Melby-Lervåg, M., & Hulme, C. (2012). Is working memory training effective? A meta-analytic review. *Developmental Psychology*. http://dx.doi.org/10.1037/a0028228 (Advance online publication).
Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex "front lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100.
Moscovitch, M. (1994). Cognitive resources and dual-task interference effects at retrieval in normal people: The role of the frontal lobes and medial temporal cortex. *Neuropsychology*, *8*(4), 524.
Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101.
Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory*, *2*(3), 325–335.
Novick, J., Hussey, E., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. (2013). Clearing the garden path: Improving sentence processing through executive control training. *Language & Cognitive Processes*. http://dx.doi.org/10.1080/01690965.2012.758297 (in press).
Officer Candidates Test (2005). *Master the office candidate tests* (7th ed.) Lawrenceville, NJ: Peterson's.
Olesen, P. J., Westerberg, H., & Klingberg, T. (2004). Increased prefrontal and parietal activity after training of working memory. *Nature Neuroscience*, *7*, 75–79.
Osberg, T. M. (1987). The convergent and discriminant validity of the need for cognition scale. *Journal of Personality Assessment*, *51*(3), 441–450.
Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., et al. (2010). Putting brain training to the test. *Nature*, *465*, 775–778.
Pollack, I., Johnson, I. B., & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology*, *57*, 137–146.
Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual section 4: Advanced progressive matrices.* Oxford: Oxford Psychologists Press.
Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., et al. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology. General*, *142*(2), 359–379
Rogers, R. D., & Monsell, S. (1995). The costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*, 207–231.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.
Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective. *Psychological Bulletin*, *138*, 628–654.
Sprenger, A. M., Dougherty, M. R., Atkins, S. M., Franco-Watkins, A. M., Thomas, R. P., Lange, N., et al. (2011). Implications of cognitive load for hypothesis generation and probability judgment. *Frontiers in Psychology*, *2*, 1–15.
Thompson, T. W., Waskom, M. L., Garel, K. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., et al. (2013). Failure of working memory training to enhance cognition or intelligence. *PLoS One*, *8*(5), e63614. http://dx.doi.org/10.1371/journal.pone.0063614.

Thorell, L. B., Lindqvist, S., Bergman, S., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science*, *12*, 106–113.

Tidwell, J., Dougherty, M. R., Chrasbaszcz, J., Thomas, R. P., & Mendoza, J. (submitted for publication). *What counts as evidence for working-memory training.* (submitted for publication).

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127–154.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language*, *69*, 36–58.

Wagenmakers, E. -J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Westerberg, H., & Klingberg, T. (2007). Changes in cortical activity after training of working memory: A single-subject analysis. *Physiology & Behavior*, *92*, 186–192.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2013). When decision heuristics and science collide. *Psychonomic Bulletin and Review* (in press).