The Effects of Dual n-back Training on the Components of Working Memory and Fluid Intelligence: An Individual Differences Approach

by

Andrew Clouter

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2013

Table of Contents

List of Tables

List of Figures

Abstract

Measures of working memory capacity are associated with performance on a variety of cognitive tasks, and reliably predict academic achievement, fluid intelligence, controlled attention, problem solving, the ability to follow directions, and a number of other cognitive abilities. A number of recent studies have provided evidence that training working memory can lead to improvements in fluid intelligence, working memory span, and performance on other untrained tasks. However, in addition to a number of mixed results, many of these studies suffer from design limitations. The aim of the present study was to experimentally investigate the effects of a dual n-back working memory training task on a variety of measures of fluid intelligence, reasoning, working memory span, and attentional control. The present study compared a training group with an active control group (a placebo group), using appropriate methods that overcame the limitations of previous studies. The dual n-back training group improved more than the active control group on some, but not all outcome measures. Differential improvement for the training group was observed on fluid intelligence, working memory capacity, and response times on conflict trials in the Stroop task. In addition, individual differences in pre-training fluid intelligence scores and initial performance on the training task explain some of the variance in outcome measure improvements. We discuss these results in the context of previous studies, and suggest that additional work is needed in order to further understand the variables responsible for transfer from training.

List of Abbreviations Used

ANOVA	Analysis of Variance
BOMAT	Bochumer Matrizentest
CFIT	Cattell's Culture Fair Intelligence Test
CI	Confidence Interval
CWMS	Categorization Working Memory Span Task
GES	Generalised Eta-Squared
GLMM	Generalised Linear Mixed Models
MAAS	Mindful Attention Awareness Scale
MHP	Monty Hall Problem
MMP	Mindfulness-based Meditation Practice
OSPAN	Operation Span
PE	Point Estimate
RAPM	Raven Advanced Progressive Matrices
RT	Response Time
SD	Standard Deviation
STM	Short-term Memory
WM	Working Memory

Acknowledgements

I would like to acknowledge the support of my thesis supervisory committee: Dr. Gail Eskes, Dr. Ray Klein, and Dr. Olav Krigolson. Dr. Eskes, in particular, has been invaluable in all stages, from the initial proposal of the study to the final submission. I would also like to acknowledge the assistance of Amber McConnell, Kerry Clifton, and Joshua Salmon, who were critical to the success of this project. Sources of funding included the Dalhousie University Faculty of Graduate Studies Graduate Student Fellowship, the RADIANT Graduate Student Scholarship, the ACOA Atlantic Innovation Fund, and the Dalhousie University Department of Psychiatry Summer Research Grant. Many thanks to all of the participants that volunteered their valuable time to take part in this experiment. Finally, I would like to thank Sarah Phelps for her unending understanding and support over the years.

Chapter 1: Introduction

Both short-term memory (STM) and working memory (WM) are involved in processing information that will be used in the short term. In modal models of memory, information is recorded in sensory memory and passes on to STM. Information that is passed on to STM is either rehearsed and retained, or else it is lost; both sensory memory and STM are limited in capacity and temporary in duration (Atkinson and Shiffrin, 1968, 1971). Atkinson and Shiffrin's (1971) model suggests that coding, rehearsal, organisation, and retrieval from STM requires active control processes; this model of STM argues for the importance of STM in cognitive tasks. However, STM tasks, such as digit span and word span tasks, are primarily passive storage and recall tasks; the tasks do not require processing or manipulation of information in addition to storage.

Baddeley and Hitch (1974, 2000; see also Baddeley, 2000; and Repovs and Baddeley, 2006) recognised this limitation of the STM model in explaining reading, reasoning, and other higher-order cognitive functions that require processing or manipulation of information. Baddeley and Hitch's reconceptualisation of the STM model treats STM as an active system for storage, maintenance, and retrieval of information for tasks in which both storage and processing are required. The WM model of Baddeley and Hitch is a multi-component system, with domain-specific subcomponents, including separate subcomponents for the storage and maintenance of verbal information and a similar subcomponent for visuospatial information. A third component, called the central executive, coordinates information, plans strategies, and is involved in reasoning, response inhibition, attentional control, and goal maintenance, among other cognitive operations. Baddeley and Hitch's WM system can store, maintain, retrieve, and update information while manipulating memory contents in order to complete tasks. Unsworth and Engle (2007) suggest that WM is also required to inhibit automatic response tendencies, to maintain information in the presence of distraction, and to retrieve relevant information in the presence of irrelevant information. WM, then, is not simply a system for maintaining information, but one that also keeps action plans, goals, and task-relevant information easily accessible; these same cognitive operations are thought to be involved in fluid intelligence (the ability to think and reason logically in novel situations), reading, and other higher-order functions.

Given the general nature of WM, researchers began to measure WM, predicting that individual differences in WM capacity should relate to performance on other tasks that measure higher-order cognitive functions. In order to appropriately measure WM capacity, given the role of WM in storage, maintenance, and retrieval and in information processing and manipulation, Daneman and Carpenter (1980) devised the reading span task. The reading span task required that participants read a series of sentences, storing the last word of each sentence for a memory test. The task required both the storage and processing components of WM. Reading span was measured as the number of final words recalled. Daneman and Carpenter found that the reading span measure of WM capacity correlated highly with multiple measures of reading comprehension. Other researchers have created similar WM span tasks, such as the operation span task (OSPAN; Turner and Engle, 1989) and the symmetry span task (Kane, Hambrick, Tuholski, Wilhelm, Payne, and Engle, 2004). Like the reading span task, these WM span tasks require that participants simultaneously process information and store information for later recall. The dual-task demand are a critical component of the WM span tasks, since WM involves storage of information and processing or manipulation of information (Miyake, 2001; see also Conway, Kane, Bunting, Hambrick, Wilhelm, and Engle, 2005). The dual-task demands imposed by these tasks agree with Baddeley and Hitch's (1974) model of WM, with its emphasis on both the storage and processing components of WM.

WM span tasks have been shown to be reasonably good predictors of performance on a variety of complex cognitive tasks. In a meta-analysis of 77 studies, Daneman and Merikle (1996) reported a correlation between WM span and reading comprehension of 0.41, indicating that WM span and reading comprehension are highly related; that is, WM span and reading comprehension are not wholly independent. Other studies have shown that WM span tasks are good predictors of a wide variety of cognitive abilities, including reasoning, fluid intelligence, academic achievement, spelling, the acquisition of logic, the ability to follow directions, controlled attention, response inhibition, and problem solving (see Baddeley, 2003; Bleckley, Durso, Crutchfield, Engle, and Khanna, 2003; Conway, Cowan, and Bunting, 2001; Conway, Kane, and Engle, 2003; De Neys and Verschueren, 2006; Engle, 2002; Engle, Tuholski, Laughlin, and Conway, 1999; Gathercole and

Pickering, 2000; Kane, Bleckley, Conway, and Engle, 2001; Kane and Engle, 2000; Kane and Engle, 2003; Kyllonen and Christal, 1990; Roberts, Hager, and Heron, 1994). These studies, and others, suggest that WM span is important in many higher-order cognitive tasks. Latent variable analyses (e.g., Kane, Hambrick, and Conway, 2005) suggest that WM span and fluid intelligence share up to 50% of their variance, indicating that WM span and fluid intelligence are not wholly independent of each other. Kane et al. (2004) also showed WM span to be associated with fluid intelligence, reporting that WM span accounted for up to 40% of the variance in fluid intelligence. Similar correlations between WM span and fluid intelligence were reported in Conway et al. (2002) and Engle et al. (1999).

Kane and Engle (2003) examined whether the executive component of WM was responsible for the correlation between WM span and higher-order cognitive operations in a series of experiments using the Stroop (1935) task, assuming that the processing components in the WM span tasks relate to the control of attention. In the Stroop task, participants respond to the colour of words that are presented on a computer screen. The words that are presented are colour words or word-like stimuli. When a colour word is presented, it may be presented in a colour that is congruent with the word (e.g, the word BLUE presented in a blue font colour) or incongruent with the word (e.g., the word RED presented in a blue font colour). Stroop interference is defined as the relative slowing of response times (RTs) and greater error rate on incongruent trials compared to neutral or congruent trials.

Noting that many studies that have attempted to link Stroop interference to fluid intelligence have used versions of the Stroop task, Kane and Engle's study varied the proportion of congruent and incongruent trials in each block. In contrast to the previously used blocked designs, in which incongruent trials are presented in sequence in a block, so that participants are reinforced for ignoring the word on each trial, the mixed-trials design used by Kane and Engle stressed the goal maintenance component of WM. Neglect of a task goal can occur when reinforcement for acting in accordance with a goal is not present, or when competing demands (e.g., task demands) require active processing, which may lead to temporary neglect of the goal. In their fourth experiment, Kane and Engle presented participants with a Stroop task in which 80% of the trials were congruent, making the incongruent trials relatively rare. The authors found that participants with high WM spans (measured using the OSPAN) had lower RT interference (measured as the difference in RTs between incongruent trials and congruent trials) than participants with low WM spans. In addition, participants with low WM spans also committed more errors on the 80% congruent Stroop task than participants with high WM spans. These results suggest that higher WM span is associated with goal maintenance in the Stroop task.

Conway, et al. (2001) investigated the importance of WM on the cocktail party phenomenon, which is also thought to be related to attentional control. The cocktail party phenomenon (Moray, 1959) refers to situations in which participants must attend to one stream of audio information in a noisy environment that contains multiple audio streams. It has been shown that highly relevant information (such as a participant's name) can capture attention, even if it is presented in a to-be-ignored audio stream (Cherry, 1953; Moray, 1959). However, not all participants are able to detect their name in the to-be-ignored stream. Conway et al. measured the WM span of participants using the OSPAN, and found that only 20% of participants with high WM spans detected their name in the to-be-ignored stream, while 65% of participants with low WM spans detected their name. The results suggest that participants with lower WM spans have greater difficulty ignoring a highly relevant lure than participants with higher WM spans.

Kane et al. (2001) and Roberts et al. (1994) investigated the role of WM in the antisaccade task, in which performance is thought to be a measure of attentional control. The antisaccade task requires that participants move their eyes in the opposite direction from a visually-presented cue. Roberts et al. demonstrated that performance on the antisaccade task breaks down when the WM of participants is taxed with a secondary task while completing the antisaccade task. Kane et al. showed that performance in the prosaccade condition did not differ between participants with low and high WM spans, participants with high WM spans made fewer eye movements towards the cue in the antisaccade condition, in addition to faster antisaccade recoveries. The results suggest participants with higher WM spans are better able to resolve the conflict between the task goal (to look away from the stimulus) and the automatic tendency (to look in the direction of the stimulus). The Stroop, cocktail party, and antisaccade tasks are attention tasks, and so it could be argued that higher-order cognitive processes are not involved in the differences in performance between the span groups on the tasks. However, if the central executive component of WM is involved in attentional control, which is assumed in the aforementioned models of WM, particularly when prepotent responses must be overcome, then WM span and performance on attentional control tasks would be related (see Hutchison, 2007; Kane and Engle, 2003; Logan and Zbrodoff, 1979).

WM span has also been related to classical reasoning problems. Stanovich and West (2000) note that, while participants' responses to reasoning problems are often in error, some participants do give responses in line with the normative solution to reasoning problems (identifying the best decision to make). Stanovich and West's studies on individual differences have suggested that participants who tend to give the correct response in reasoning tasks (e.g., the conjunction fallacy, in which it is assumed that specific possibilities are more likely than a more general possibility (Tversky and Kahneman, 1983) and the Wason selection task (Wason, 1966), a deductive reasoning task) have larger WM or cognitive capacity. Stanovich and West argue that correct normative reasoning requires the inhibition of the heuristic system, such as the representativeness heuristic (in which something that is judged to be more representative is incorrectly thought to be more likely) in the conjunction fallacy. De Neys and Verschueren (2006) argue that inhibiting the heuristic system requires the use of WM resources, suggesting that participants with higher WM spans may be more likely to inhibit prepotent heuristics and engage normative reasoning.

De Neys and Verschueren (2006) studied the association between WM capacity and normative reasoning in the Monty Hall Problem (MHP). The MHP is a well-known probability-based problem-solving dilemma based on the dilemma presented by Monty Hall, the host of the popular game show "Let's Make a Deal." In the game show, the contestant had to guess which one of three doors contained a prize. Once the contestant made a choice, Monty Hall opened one of the two unchosen doors to reveal that it did not contain the prize. Contestants were then given the opportunity to switch to the remaining unchosen door or stick with their initial choice. Most people believe that the probability of winning is 50% whether they switch to the unchosen door or stick with their initial choice, and most people choose to stick with their initial choice. However, switching to the remaining unchosen door doubles the probability of winning. One way to consider the problem is to recognise that when you make your first choice (one of the three doors), the probability that you have chosen the winning door is 1/3. If you stick with that choice when given the option to stick or switch, you will win 1/3 of the time. However, given that the probability of having chosen the winning door in your initial choice was 1/3, then the probability of having not chosen the winning door was 2/3. Switching to the remaining unchosen door, then, would allow you to win 2/3 of the time. More simply, if you switch on every trial, you will switch to a losing door only 1/3 of the time, but switch to the winning door 2/3 of the time. Switching, then, is the optimal strategy.

In studies of college students, the majority of students fail to give the correct response to the MHP, with switching rates varying from 9% to 65% (e.g., Burns and Wieth, 2004; Friedman, 1998; Granberg and Brown, 1995; Herbranson and Schroeder, 2010, Krauss and Wang, 2003; Tubau and Alonso, 2003). A number of heuristics have been studied as possible candidates for the consistency of incorrect responses (see De Neys and Verschueren, 2006), including the number-of-cases heuristic (Falk, 1992; Shimojo and Ichikawa, 1989) and the stick-with-your-pick heuristic (Geiger, 1997; Gilovich, Medvec, and Chen, 1995).

In modified versions of the MHP, De Neys and Verschueren reported that participants who made the correct response in the MHP had higher WM spans (measured with the GOSPAN task) than participants who failed to make correct responses, and that, when concurrently performing a secondary task designed to engage WM resources, correct responding on the MHP decreased. While these findings suggest that WM capacity may be associated with correct responses on the MHP, the modified tasks used by De Neys and Verschueren were significantly easier than typical presentations of the MHP. The modified problem used a "ball-and-cup" task, wherein the participant was presented with a problem in which a boy chose one of three upturned cups randomly. The girl in the problem received the remaining two cups. Under one of the cups was a winning marble. The girl looked under her two cups, and then revealed one cup in which the marble was not under. Both the boy and the girl in the problem are left with one cup each. The question posed to the participants in the experiment was to determine what the boy should do in order to have the most chance of winning the game; the options were to stick with the initially chosen cup, switch to the girl's cup, or decide that it does not matter – the chances are even (see De Neys and Verschueren, 2006, Appendix). The use of easier versions of the MHP by De Neys and Verschueren was justified, as in their first experiment only 5% of participants gave the correct answer to the problem. In order to examine the potentially detrimental effects of introducing a secondary working memory task, the MHP was modified to make it much easier so that performance would not be floored.

Herbranson and Schroeder (2010) investigated the MHP in pigeons and in humans, and found that the pigeons adjusted behaviour to the optimal strategy, while humans did not. Given that the birds were not given verbal task instructions, Herbranson and Schroeder modified the MHP presented to the human participants by eliminating the cover story that was used by De Neys and Verschueren, noting that Burns and Wieth (2004) showed that versions of the MHP with different cover stories, but the same underlying mathematical structure, were not optimally solved equally. With stripped-down task instructions, and 200 trials of the MHP presented over four blocks, Herbranson and Schroeder showed that human participants were no different from chance in selecting the optimal strategy in the first or last blocks of trials.

Kane et al. (2004) examined verbal and visuospatial measures of WM span, including the OSPAN and the symmetry span task, noting that the measures shared 70-85% of their variance, and were very highly correlated, suggesting that the abilities measured by the

OSPAN and symmetry span tasks are not independent. While this result suggests that these WM span tasks primarily measure a domain-general component of WM, others have argued that these WM span tasks are complex, and likely include domain-general and domain-specific components (Engle et al., 1999; Miyake and Shah, 1999) – a view that agrees with the Baddeley and Hitch model of WM. For example, Shah and Miyake (1996) presented results that suggested that reading span was correlated with reading comprehension, but not with measures of spatial ability, while spatial span measures were correlated with spatial ability but not reading comprehension. While the authors did not deny the existence of a domain-general WM component, they did argue that, at least for verbal and spatial WM measures, domain-specific components appear to be present.

The many studies that indicate a strong relationship between WM span and higher-order cognitive tasks provide the logic behind many of the recent studies that attempt to improve performance on cognitive tasks by "training" WM. If WM and performance on higher-order cognitive tasks are highly related, then a program of training that increases WM span should lead to improvements in tests of these higher-order processes. Of course, the latent variable studies and the correlational studies suggest that 50% or more of the variance in fluid intelligence and other cognitive tasks is not accounted for by variance in WM span, leaving plenty of room for improvements in WM span that do not transfer to improvements on other cognitive tasks, and also improvements on cognitive tasks that do not have improvements in WM span as the cause. Other researchers have proposed that WM and fluid intelligence share common capacity constraints: the number of items in WM, and the number of inter-relationships among them (Halford, Cowan, and Andrews, 2007); it may be possible to relax these capacity constraints (improve capacity) with WM training. The common capacity is assumed to be due to attentional control or central executive processes. The rationale behind some training studies (e.g., Jaeggi, Buschkuehl, Jonides, and Perrig, 2008; Jaeggi, Studer-Luethi, Buschkuehl, Su, Jonides, and Perrig, 2010; see also Buschkuehl and Jaeggi, 2010; Jaeggi, Buschkuehl, Etienne, Ozdoba, Perrig, and Nirkko, 2007; Jaeggi, Seewer, Nirkko, Eckstein, Schroth, and Groner, et al., 2003) is that if participants train in one domain (e.g., WM), the benefits from training should be apparent in domains that share capacities or processes (e.g, fluid intelligence), if the WM training task engages the overlapping capacities or processes. Buschkuehl and Jaeggi (2010) suggest that, in order to maximise the process overlap between the training task and the outcome measures, the training task should be complex, and should train multiple processes. In addition, Buschkuehl and Jaeggi propose that a training task should: (1) be a task in which it is difficult to develop a strategy, in order to minimise strategy development and focus on training core WM processes; and (2) be adaptive in difficulty so that the challenge level of the training task matches the skill level of each participant, thus providing a relatively similar high level of training demand to all participants. The training task should also be different enough from the outcome measures to avoid practice effects; the training task should not train participants on the outcome measures themselves.

The dual n-back task has been used as WM training task, as it appears to meet the above criteria. In a typical n-back task (Kirchner, 1958; Mackworth, 1959; Moore and Ross, 1963), a list of items is presented, one item at a time, and participants must report whether each item matches the item that had been presented n-back in the stream. For example, in a two-back task, participants must maintain the two most recently presented items, and indicate whether the current item matches the first of the two items that are maintained in memory. Then they must update the memory set with each new item by dropping the least recent item and maintaining the most recent item. The dual n-back task simultaneously presents verbal and visuospatial n-back tasks with two separate streams. The dual n-back task engages multiple processes involved in WM, including the inhibition of irrelevant items, managing multiple tasks simultaneously, binding items with their temporal context, and continuously updating and maintaining multiple items in memory. It is difficult to develop a strategy to improve performance on both the verbal and visuospatial components of the dual n-back task, and studies that have used the dual n-back task as the WM training task used an adaptive version of the task so that the difficulty level (the n-level) of the task increases as participants perform well. The adaptive nature of the dual n-back task allows for a match between the challenge level provided by the training task and the skill level of the participant, even as participants improve with successive training.

Shipstead, Redick, and Engle (2010) discussed the requirements for experimental designs of training studies in order to claim improvements in performance on outcome measures, and noted in their discussion that the adaptive dual n-back task appeared to be a promising WM training task. In some studies (e.g., Jaeggi, Buschkuehl, Perrig, and Meier, 2010; Jaeggi et al., 2010; Redick et al., 2012), performance on the dual n-back task was correlated with performance on tests of fluid intelligence. Jaeggi et al. (2008) were the first to report that WM training using the dual n-back task led to improved scores on tests of fluid intelligence. In their study, healthy young participants completed 20 blocks of trials, with 20 + n trials in each block, beginning with n = 1. Different groups in their series of experiments received either 8, 12, 17, or 19 days of training. Compared to a no-contact control group, Jaeggi et al. reported an increase in fluid intelligence in the training group. However, as Moody (2009) and Redick et al. (2012) point out, the results from Jaeggi et al. are difficult to interpret for a variety of reasons. For example, reported data were collapsed across different tests of fluid intelligence, including the Raven Advanced Progressive Matrices (RAPM) and the Bochumer Matrizentest (BOMAT), each of which were administered with different time limits. Moreover, the reported data were collapsed across different groups: some control participants completed an active control task, while others did not. Jaeggi et al. (2010) followed up their first study, with participants assigned to either a dual n-back group, a visuospatial single n-back group, or a no-contact control group. Results suggested that both single n-back and dual n-back training groups improved scores on tests of fluid intelligence (RAPM and BOMAT) compared to the no-contact control group. However, comparisons of training groups to a no-contact control group, and the use of matrix reasoning tests as the only measure of fluid intelligence, limit the usefulness of the results. As noted in Buschkuehl and Jaeggi (2010), Shipstead et al. (2010), and Redick et al. (2012), the most compelling results would come from studies that include an active control group, as opposed to a no-contact control group, in order to decrease the chances that observed effects are a result of Hawthorne or placebo effects.

Borella et al. (2010) investigated the potential of transfer from a verbal WM training paradigm (the Categorization Working Memory Span task, CWMS; see Borella, Carretti, and De Beni, 2008; De Beni, Borella, Carretti, Marigo, and Nava, 2008) to measures of STM, visuospatial WM, inhibition, and fluid intelligence, in a sample of 40 healthy older adults (aged 65-75 years). The visuospatial WM task was the Dot Matrix task (Miyake, Friedman, Rettinger, Shah, and Hegarty, 2001), a standard dual-task measure of WM capacity. STM was measured with the Forward Digit Span and Backward Digit Span tasks (De Beni et al., 2008). Fluid intelligence was measured using Cattell's Culture Fair Intelligence Test (Cattell, 1971; Institute for Personality and Ability Testing, 1973), and response inhibition with a Stroop task. The training group completed three 60 minute training sessions using the CWMS, while the active control group completed three engaging activities, both over a two-week period. Results indicated significant group x session interactions for all of the outcome measures, including the CFIT measure of fluid intelligence and the Dot Matrix measure of WM capacity. These results suggested that verbal WM training can transfer to performance improvements on tasks that are associated with WM, and that the benefits from the training exceed any benefits from participating in a control task, at least among older adults.

However, Jaeggi et al. (2011) were not able to demonstrate similar transfer in a study involving elementary and middle school children. In their study, 62 children participated in either the training group, completing an adaptive single (spatial) n-back task, or an active control group that involved general knowledge practice. Group x session interactions did not reach significance in this study involving children. However, Jaeggi et al. report that, among children in the training group, those with the highest performance on the training task (children with above-median improvement on the task) improved on fluid intelligence by more than participants with lower performance on the training task. Moreover, the group x session interaction was significant between participants with higher performance on the training task and the active control group, suggesting that the amount of improvement on the training task may be an important variable mediating transfer from training.

Seider et al. (2010) investigated whether five weeks of training on a dual n-back task led to improvements in measures of cognition and other skills in a study involving both young and older adults, comparing results from the dual n-back group to an active control group that trained general knowledge. Among the younger adults (results for the older adults were not presented in the technical report, as the study was ongoing), a significant group x session interaction was reported for the operation span measure of WM capacity. Redick et al. (2012), using a similar training task, and participants from three different universities in order to broaden the sample of participants, were not able to replicate the findings of Seider et al. Moreover, Redick et al. were unable to demonstrate significant group x session interactions for any of the outcome measures used in their study, including composite scores for measures of spatial fluid intelligence, verbal fluid intelligence, and WM capacity.

Dahlin, Neely, Larsson, Bäckman, and Nyberg (2008) reported that transfer from WM training tasks may be mediated by the striatum. The extent to which the training tasks and the outcome measures engage the striatum may explain the divergent results from the WM training studies. Dahlin et al. demonstrated transfer from an updating WM training task to other measures of WM (letter memory and n-back tasks where n was greater than 3, due to performance ceiling being reached in the 1-back and 2-back tasks) with five weeks of training.

Since WM is closely related to a wide variety of cognitive processes, and is a central component of cognition, studies focussing on techniques for training and improving WM are theoretically justified. In the current study, we followed the advice of Buschkuehl and Jaeggi (2010) and Shipstead et al. (2010) by including an active control group and multiple measures of cognitive processes thought to be associated with WM. WM span has been correlated with a wide variety of cognitive tasks, including fluid intelligence, reasoning, and attentional control. In the current study, we measured fluid intelligence using Cattell's Culture Fair Intelligence Test (CFIT; Cattell, 1971; Institute for Personality and Ability Testing, 1973): a complete, two-form test of fluid intelligence with four subtests; normative reasoning with the MHP; WM span with both the OSPAN and the symmetry span; and attentional control with the 80% congruent Stroop task. In addition, analysis of individual differences based on performance on both the CFIT and the OSPAN was performed in order to determine whether improvements in the outcome measures were specific to participants with higher or lower fluid intelligence or OSPAN scores. The rationale for including this variety of outcome measures was that, although individual differences in WM span have been shown to be correlated with performance on these tasks, suggesting that WM is involved in each, it may be that the training task generalises to only some, but not all, of the tasks that measure related cognitive processes. The active control group completed the same task as the training group, except that the n-level in the active control task was set to n = 1. Instead of increasing their n-level over training sessions, participants in the active control group were provided with both accuracy and RT feedback, and worked to improve their RT and accuracy on match trials over the training sessions. The low WM load in the active control task ensured that WM was not overly engaged during training, while the use of an active control group controlled for participants' expectations and potential placebo-type effects.

The purpose of the present study was to examine the effects of training on the dual n-back task on a variety of measures that are related to WM capacity, and to address questions such as: Does training on the dual n-back task lead to improvements on tests of fluid intelligence, reasoning, WM span, and attentional control? Are improvements in the outcome measures related to performance on the training task? Will participants with lower WM span or lower fluid intelligence improve more on the outcome measures than participants with higher WM spans or fluid intelligence scores? We hypothesise that (1) compared to the active control group, the training group will show improved scores on the CFIT, the MHP, WM span as measured by the OSPAN, WM span as measured by the symmetry span, and attentional control in the Stroop task, based on the reported associations between these tasks in the literature; (2) the improvements in the outcome measures will be related to performance on the training task, based on results presented by Jaeggi et al., 2011; and (3) participants with lower WM spans and scores on the CFIT in the pre-training session will improve on the outcome measures more than participants with higher WM span scores and scores on the CFIT.

Chapter 2: Method

2.1 Participants

Fifty healthy young adults with no history of head injury, psychiatric, or neurological illness began the study. Half of the participants (25) participated in part 2 of the study (see Procedure, below); the data from these participants have been combined with the the data from the 25 participants that completed part 1 of the study. All participants had normal or corrected-to-normal vision. Seven participants (six in the active control group and one in the training group) withdrew from the study, primarily due to time constraints and the demands of their classes, and did not complete the study. The data from these participants were excluded from all analysis. The data from seven other participants (three in the active control group and four in the training group) were also excluded from the final analysis (see Data Screening, below). Data from 36 participants (4 male, 32 female, mean age = 20.39, *sd* = 2.54) were retained for the final analysis. Of these, 18 participants were in the active control group (mean number of training sessions completed = 15.2, *sd* = 0.55, min = 14, max = 16.3), and 18 in the training group (mean number of training sessions completed = 14.3, *sd* = 1.61, min = 10, max = 15.1). Participants estimated the number of hours per week that they used a computer and the number of hours per week spent playing digital games in the six months prior to the study. The number of hours spent using a computer per week did not differ between the active control group (mean = 32.50, *sd* = 11.61) and the training group (mean = 33.92, *sd*

= 10.43), nor did the number of hours spent playing digital games per week (active control group mean = 1.78, *sd* = 3.66; training group mean = 3.22, *sd* = 7.45). All participants were fluent in English, with 33 participants speaking English as their first language, and three participants having used English as their primary language in school. Three participants were left-handed (one in the active control group and two in the training group); the remaining 33 participants were right-handed. Participants were granted a choice between course credit and a small stipend, or a combination, for participating in the study.

2.2 Materials and Tasks

2.2.1 Apparatus

In the pre-training and post-training phases, participants sat in a well-lit testing room with the experimenter. The operation span, symmetry span, and Stroop tasks were programmed using PsychoPy (version 1.71.01; Peirce, 2007, 2009) and presented on a 24 inch iMac computer running Mac OS-X 10.7. Participants used the computer keyboard to respond to the operation span mathematical problems, symmetry span symmetry judgements, and Stroop trials.

The training tasks were loaded onto participants' personal laptop computers. The training tasks were modified from the free open-source n-back package Brain Workshop (version

4.8.4), available at http://brainworkshop.sourceforge.net. The program was modified so that the program window was 912 x 684 pixels. Participants used their laptop keyboards to respond to audio and location matches throughout the training sessions. The computer recorded all responses, and participants brought their laptop to the post-training session so that the experimenters could obtain the training data.

2.2.2 Cattell's Culture Fair Intelligence Test (fluid intelligence)

The CFIT Scale 3 (Cattell, 1971; Institute for Personality and Ability Testing, 1973) was administered in accordance with the standardised testing procedures that accompany the tests. The complete test includes Form A (the short form) and Form B. All participants completed Form A in the pre-training session and Form B in the post-training session, as scoring norms are provided for Form A when administered alone, but Form B only when it is administered after Form A.

The pen-and-paper test is composed of four timed subtests. The time allowed for each subtest varies between 2.5 and 4 minutes. Participants stopped working on each subtest immediately when time expired, and participants were not permitted to return to work on previous subtests at any time. In the first subtest (Series), participants had to choose one from six alternatives that would complete a progressive series of figures or shapes. In the second subtest (Classifications), participants had to choose two figures or shapes from five alternatives that differed from the remaining three. In the third subtest (Matrices), participants had to complete a matrix that contained between four and nine boxes by choosing the best box from a set of six alternatives. In the fourth subtest (Conditions), participants were presented with a figure that consisted of shapes, lines, and a single dot. Participants had to choose the one from the five alternatives where the dot could be placed in the same relationship to the other figures. The total CFIT score was the sum of all correct answers across the four subtests; the scores for each subtest were also used in the final analysis.

2.2.3 Monty Hall Problem (normative reasoning)

The MHP has been adapted from the dilemma presented by Monty Hall, the host of the popular game show "Let's Make a Deal." The MHP used in this study was modified from the task used by Herbranson and Schroeder (2010). Participants were told, "We will play a probability-based problem-solving task together. In each game, there will be three index cards on the table, numbered 1, 2, and 3. One of the cards is worth 100 points, and the other two cards are worth 0 points. I know which of the three cards is worth 100 points in each trial; the card worth 100 points has been determined randomly in advance, and I have the answer sheet with me so that we can track the total amount of points that you win. After each trial, I will tell you whether you have won or lost, and will keep track of your total score. The goal is to accumulate as many points as possible... I will ask you to choose one card. Then, I will reveal a card that you have not chosen, and that is not worth 100 points. You then have the opportunity to choose one of the remaining two cards. I will then reveal the card that is worth 100 points and tell you whether you have won or not." Figure 1 depicts a sample of the MHP. Participants were not informed that they were completing a MHP task, and terms that are frequently associated with the MHP (e.g., "stick" and "switch") were avoided.

Following the task instructions, participants were allowed three practice trials. The experimental trials began with the presentation of the three cards, and the instruction that the participant had to make their choice. The experimenter then revealed that one of the remaining two (unchosen) cards was not the winning card, and removed it from the game. With two cards remaining, participants were instructed to make their choice again. The experimenter then revealed the winning card, and informed the participant whether they had won and their total accumulated points. The chances of winning the MHP trials depended on the participant's second choice: if the participant's second choice was the same as their first choice, then the chances of winning were 1/3. However, if the participant's second choice was the other remaining card, then the chances of winning were 2/3. Pseudorandomly intermixed with the 30 MHP trials were 20 trials designed to counter learning by reinforcement; all participants received the trials in the same order. It is possible to learn the optimal game strategy by normative reasoning, or by reinforcement learning. In fact, Herbranson and Schroeder (2010) suggest that pigeons adopt the optimal strategy in a version of the MHP quickly, whereas humans do not, primarily as a result of reinforcement learning. In order to ensure that learning by reinforcement was not the primary driver of improvement on the MHP in this study, given that the purpose of this task was to determine the effects of working memory training on the normative reasoning aspect of the MHP, the 20 trials that were intermixed with the 30 MHP trials were similar to, but not the same as, the MHP trials. The intermixing of non-MHP trials was intended to counter the potential of learning the optimal strategy in the MHP trials by reinforcement learning alone, without recourse to normative reasoning. In ten of these trials, participants were instructed that they would choose one of the three cards, and then the winning card would be revealed immediately, with no opportunity for a second choice. The chances of correctly choosing the winning card in these trials was 1/3. In ten other trials, participants were instructed to choose two of the three cards, and that the winning card would be revealed after their choices were made, with no opportunity for a second choice. The chances of winning on these trials was 2/3. The dependent variable of interest was the proportion of trials in which the participant chose the remaining card that they had not chosen initially (indicating an understanding of the best strategy to use to win points), whether participants won or not, on the MHP trials.

2.2.4 Working Memory Tasks

Many working memory span tasks include a memory set to engage working memory storage and retrieval in addition to a demanding processing task designed to engage the processing functions of working memory (see Conway et al., 2005). The inclusion of a processing task defines "complex" working memory tasks; "simple" span tasks typically only involve memory storage and retrieval without engaging the processing functions of working memory. In the OSPAN task, the processing component was verbal, and required that participants solve mathematical problems that they read aloud; the memory component was a set of words. In the symmetry span task, the processing component was spatial, as participants had to examine 8 x 8 matrices and decide whether each matrix was symmetrical across the vertical axis; the memory component was a set of spatial locations within a 4 x 4 matrix.

2.2.4.1 Operation Span (executive working memory, verbal task)

The OSPAN task (Engle et al., 1999; modified from Kane et al., 2004**)** required that participants solve mathematical problems out loud, while trying to remember words. Each trial began with a math problem presented in the form of a question, in white text (arial font, 100% opacity, 1.2 degrees of visual angle, presented at the origin), on a black computer screen. For example, participants might have seen, "is (7 x 2) - 1 = 14?" Participants were required to read the equation aloud exactly as it was presented, and then solve the problem. Participants were instructed to say "yes" out loud and press the "y" key on the keyboard if the equation was correct, or state "no" out loud and press the "n" key on the keyboard if the equation was not correct. Half of the equations were correct. Immediately after pressing the response key, a word appeared on the screen (white text, arial font, 100% opacity, 5% screen height (1.2 inches on the testing computer screen), presented at the origin); participants were instructed to read the word aloud. The word remained on the screen for 1 second. After a 0.5 second delay, participants were presented with another trial and another word to remember, or a recall instruction. When presented with the recall instruction, participants were instructed to recall, out loud, and in the correct serial order, the words that were presented. Figure 2 depicts the OSPAN task. It was stressed in the task instructions and in the practice trials that the order of the recalled words was critical. All participants were instructed that if a situation arose in which the participant could not remember all of the words, to insert a statement to that effect in the serial position of that word during recall. For example, if the set of words was "arm", "tree", and "cake", and participants could only recall "arm" and "cake", than participants should respond with "arm", "can't remember", "cake", rather than "arm", "cake". Each block of trials varied in length from two to five equations, each followed by words to be recalled, with three blocks of each length presented in a pseudorandom order, for a total of 42 trials in 12 blocks, with two additional practice blocks excluded from the final analysis. The operation span score was taken as the proportion of correctly recalled words of the 42 words that were presented. Scores for each set size (2, 3, 4, and 5) were also used in the analysis. Since baseline performance was close to ceiling on the smaller set sizes, the proportion of correctly recalled words in the sets of size 4 and 5 (a total of 27 trials) was also used in the analysis. Different sets of word lists and equations were used in the pre-training and post-training phases, and the order of presentation was counterbalanced.

2.2.4.2 Symmetry Span (executive working memory, spatial task)

The symmetry span task (Kane et al., 2004**)** required that participants make symmetry judgments while trying to remember the locations of red-coloured squares. Each trial began with an 8 x 8 matrix (6cm x 6cm, 100% opacity, presented at the origin) presented on a grey computer screen. The 8 x 8 matrix contained a number of white (unfilled) and black (filled) squares. Participants were instructed to determine whether the presented matrix was symmetrical across its vertical axis, so that the left and right halves of the matrix were mirror images of each other. The matrices were symmetrical half of the time. Participants were instructed to press the "y" key on the keyboard to indicate that the matrix was symmetrical, or the "n" key on the keyboard to indicate that the matrix was not symmetrical. Immediately following the symmetry judgment, participants were presented with a 4 x 4 matrix on the computer screen (3cm x 3cm, 100% opacity, presented at the origin), with one red (filled) square and 15 white (unfilled) squares. Participants were instructed to remember the location of the red square for future recall. The 4 x 4 matrix was displayed for 1 second, and was followed by a 0.5 second delay before the next trial began, or the recall instruction was presented. When presented with the recall instruction, participants were instructed to use the provided answer sheet, which contained five unfilled 4 x 4 matrices for each block of trials, to recall the locations of the red squares, in the order that they were presented. Figure 3 depicts the symmetry span task. As with the operation span task, the importance of the serial order of the locations was stressed, and in the task instructions and practice trials, participants were informed that they could skip some answers, if it meant that correctly recalled locations could be placed in the proper order. Each block of trials ranged in length from two to five locations, with three blocks of each length presented in a pseudorandom order. The two practice blocks were excluded from the final analysis, leaving 12 blocks, with a total of 42 trials, retained for the analysis. The symmetry span score was taken as the proportion of correctly recalled locations of the 42 total locations that were presented. Scores for each set size (2, 3, 4, and 5) were also used in the analysis, as was the proportion of correctly recalled locations in the 27 trials that made up the blocks of length 4 and 5.

2.2.5 Stroop Task

The Stroop task (Stroop, 1935) required that participants identify the colour in which a word was presented on the computer screen, while ignoring the word itself. The mostly-congruent Stroop task used in the present experiment follows the task used by Kane and Engle (2003). Each trial began with a blank (grey) computer screen, lasting 0.75 seconds, followed by a white fixation cross (100% opacity, 1.2 degrees of visual angle, presented at the origin), lasting 1 second. The fixation cross was replaced with one of three colour words ("red", "green", or "blue", presented in uppercase, arial font, 100% opacity, 2.4 degrees of visual angle, presented at the origin), which were presented in a pseudorandom order. The words were presented in either a red, green, or blue font colour; colours used the RGB colour palette, with each colour set at 100% of the required colour and 0% for the remaining two colours (e.g., red was 100% red, 0% green and 0% blue on the RGB palette). Figure 4 depicts the Stroop task. Participants responded to the font colour of the word by pressing one of three arrows keys (left, down, and right) that were clearly marked to represent the three colours. All participants were given clear instructions and ten practice trials prior to the experimental trials. The stimulus list contained 180 words, 80% of which were trials in which the colour of the word matched the font colour (congruent trials; 144 total trials, with 48 trials for each word). In 20% of trials, the presented word did not match the font colour in which the word was presented (incongruent trials; 36 total trials, with 12 incongruent trials for each word, with half of the trials for each word being presented in each of the two incongruent font colours). In order to obtain enough responses for each condition, while keeping the length of the task manageable, neutral trials were not used (following Kane and Engle, 2003). Response times for correct responses and accuracy scores were calculated for the congruent trials and incongruent trials. Accuracy scores were transformed to log-odds error for analysis of errors on congruent and incongruent trials. The Stroop interference measures were calculated as the difference between the response times on correct responses, and the difference between the errors, on the incongruent and congruent trial types.

2.2.6 Training Tasks

The training tasks were modified dual n-back tasks. The dual n-back tasks (Kirchner, 1958; Mackworth, 1959; Moore and Ross, 1963) presented two sets of stimuli that participants continually monitored. Participants responded whenever one of the stimuli matched the stimulus that was presented n items ago in the set. In the training group, the "n" level varied. In a dual one-back task, participants must maintain the last letter (presented auditorally) and the last spatially presented location. For each subsequent stimuli presentation, participants must update the memory set by remembering the newly presented items and dropping the previous items. Both the training group and the active control group completed 20 blocks of trials during each training session. All participants were instructed to complete five training sessions each week over three weeks, for 15 total training sessions.

2.2.6.1 Training Group Task

For the training group, each daily session began with an n level of one. Participants were presented with $20 + n^2$ trials in each of the 20 blocks per training session (e.g., 21 trials for $n = 1$). Participants were instructed to complete the training sessions in a quiet, comfortable room. The training sessions began with a main screen that displayed the number and proportion of correct and incorrect trials for both the verbal and spatial tasks for the previous block of trials, the average of the proportion correct over both tasks for each block of trials completed in that training session, and the overall average n-level achieved during the training session. The main screen also informed participants the n-level of the next block of trials (Figure 5). Participants began each block of trials by pressing the spacebar. The game screen contained a 3 x 3 matrix presented in a tic-tac-toe-like design, without an outer border. The middle square contained a cross, and no visual stimuli appeared in the middle square. Each trial began with a 2.5-second delay, followed by the simultaneous presentation of a letter spoken aloud through the computer speakers or headphones and a blue square appearing in one of the eight matrix locations. The location stimuli remained for 0.5 seconds, and each trial was followed by a 2.5-second inter-trial interval, during which participants made responses. On each trial, participants were instructed to press the "A" key on the keyboard to indicate when the audio stimulus (the letter) matched the letter that was presented n-items ago, and the "L" key on the keyboard to indicate when the location of the blue square matched the location of the square n-items ago. Figure 6 shows a sample sequence of trials in a dual 1-back task. The n-level of the auditory and visual stimuli were always matched. No trial-by-trial feedback was given. Following each block of trials, participants were returned to the main screen, which contained feedback for the previously completed block and all of the blocks completed during that training session. Letters and locations were randomly drawn from a list of eight possible letters and locations. The random chance of a match for each stimulus presentation was 12.5%, which was increased by an additional 12.5% chance for each stimulus, for a 25% chance of each stimulus presentation being a match. There was an inherent 12.5% chance that any stimulus presentation was a n+1 foil, and a 12.5% chance that an item would be an n-1 foil (on blocks with $n > 1$), for a 25% total foil chance.

Scoring was calculated as the number of correct responses divided by the sum of the numbers of correct responses, incorrect responses, and missed responses, with both verbal and spatial tasks summed for the average block score. For the training group, if the average score on a block of trials was 80% or greater, than the n-level of the subsequent blocks was increased by one. If participants scored less than 50% on three blocks of trials, then the n-level was decreased by one for subsequent blocks. Participants were provided with a scoring sheet, and were instructed to fill in the date, start time, finish time, and the average score and n-level for each of the 20 blocks of trials in each training session. The goal was to try to improve the overall average n-level over the training sessions. Participants returned the scoring sheet to the experimenters, and the experimenters collected the training data from each participant's laptop computer.

2.2.6.2 Active Control Group Task

The active control group completed a task modified from the same program used by the training group. Unlike the training group, where the n-level of each block adapted to the performance of the participants, the active control group was always completing a dual 1-back task. The program was further modified to display the reaction time for each response. Following each block of trials, participants were returned to the main screen that displayed both accuracy and the average response time for correct responses (Figure 7). Participants in the active control group were given the goal to improve their combination of accuracy and reaction time over each block and session. Given that responses were withheld when no matches were present, the active control task resembled a dual 1-back speeded go/no-go task. The active control task was otherwise equivalent to the training group task, and data were collected from participants' laptop computers in the post-training session.

2.3 Procedure

The study was conducted in two parts (Part 1 and Part 2), each with three phases: the pre-training phase, the training phase, and the post-training phase. In Part 1, in the pre-training phase, participants were randomly assigned to either the active control group or the training group. Following informed consent, participants provided demographic information and completed a screening questionnaire. No participants were excluded from participating in the study. Outcome measures were administered by a second experimenter, blind to the group assignment, in the same order for all participants. Participants completed the CFIT, the Stroop, the MHP, the OSPAN, and the symmetry span tasks in that order. Following completion of the tasks, the training program was loaded onto participants' laptops, and participants were instructed on how to complete the training. Each participant's expectations of whether the training task would likely result in an improvement in performance on each of the outcome measures were also measured. During the training phase, participants were instructed to complete 20 blocks of training trials during each training session, and complete five training sessions per week for three weeks. Three weeks after the pre-training session, participants returned for the post-training session in which participants were asked to record their expectations of whether they thought that the training would have any impact on their outcome measure performance, and then the same outcome measures were administered following the same procedure. Participants were debriefed and thanked upon completion.

In Part 2 of the study, participants were randomly assigned to one of four groups: an active control group with mindfulness-based meditation practice (MMP) , an active control group with an audiobook control for the MMP, a training group with MMP, and a training group with an audiobook control for MMP. The pre-training session in Part 2 of the study differed from Part 1 of the study in that participants watched a 10-minute video containing instructions on how to do MMP, or a control video (a non-emotionally-arousing TED talk). The post-training session in Part 2 of the study differed from the post-training session in Part 1 of the study in the debriefing of the participants. The training phase of Part 2 of the study differed from that in Part 1 of the study in that participants in the MMP groups were instructed to listen to a 10-minute audio file designed to guide the participant through MMP, or a 10-minute clip of a public-domain audiobook (Siddhartha, by Herman Hesse). In both the pre-training and post-training sessions, participants in all groups completed two standard questionnaire measures of mindfulness: the Mindful Attention Awareness Scale (MAAS; Brown & Ryan, 2003) and the Neff Self-Compassion Scale (Neff, 2003). Analyses indicated that the MMP had no effect on scores on either of the two scales, and that performance on the training tasks and outcome measures did not differ between the MMP groups and the audiobook control groups, and did not differ between groups in Part 2 and the groups in Part 1. Therefore, the data from Part 1 and Part 2 of the study were combined in the final analyses.

2.4 Data Screening

One participant used the answer sheet for the symmetry span task incorrectly, and the data were excluded from the analysis of the symmetry span; this participant was in the training group. Data from five participants who were familiar with the MHP were excluded from the analysis of the MHP; four of these participants were in the active control group, and one in the training group. Data from six participants who completed fewer than ten of the required 15 training sessions were excluded from the analysis. One participant in the training group was unable to provide data from the training phase due to problems with their personal laptop computer, and their data were excluded from the final analysis. The active control group and training group had $N = 18$ each. For analysis of the symmetry span, the training group had $N = 17$. For analysis of the MHP, the active control group had $N = 14$, and the training group had $N = 17$.

2.5 Statistical Procedure

Point estimates and confidence intervals for each effect of interest were obtained by a posteriori bootstrapping (100,000 iterations) from generalised linear mixed models (GLMMs; Pinheiro and Bates, 2000). GLMMs require fewer assumptions than analysis of variance, while providing greater statistical power and the possibility of multiple random effects. In the present analysis "participant" is defined as a random effect, as interparticipant variability is not an effect of interest. GLMMs account for the interparticipant variance, enhancing the power of the models. Prior to the computation of the bootstrapped confidence intervals, the estimates for the variance of the grand mean of each effect and its covariance with the other effects of interest were zeroed. This approach yields confidence intervals for cells in repeated-measures designs that are inferentially useful for within- and between-groups comparisons (cf. Masson and Loftus, 2003).

While the point estimates and confidence intervals provided evidence of changes in performance on the pre-training and post-training sessions, analysis of variance (ANOVA) were used to test for the presence of training group x session interactions. ANOVAs were used to test for training group x session and training group x split group x session interactions in the individual differences analysis.

Analysis was coded in R (Version 3.0.0; R Development Core Team, 2013) using the functions in the ez package (https://github.com/mike-lawrence/ez) for analysis and the ggplot2 package (Wickham, 2009, 2012).

Chapter 3: Results

3.1 Pre-training Individual Differences on Outcome Measures

For brevity, only significant results are mentioned in the text. Descriptive statistics and results from all significance tests appear in the appropriate tables.

Analysis of individual differences was conducted by dividing participants into groups based on their pre-training performance on the CFIT and the operation span task, in a manner similar to that used by many research groups (e.g., Bleckley et al., 2003; Colzato, Spape, Pannebakker, and Hommel, 2007; Conway et al., 2001; Kane and Engle, 2003; Kane, Poole, Tuholski, and Engle, 2006; Poole and Kane, 2009; Redick and Engle, 2006; Redick, Gay, Calvo, and Engle, 2011; Unsworth and Engle, 2007). Median split groups based on pre-training CFIT scores and OSPAN scores were used as between-subjects factors in two sets of ANOVAs, with pre-training scores on the remaining outcome measures as the dependent variables[1]. Significance testing results are shown in Table 1, with descriptive data shown in Table 2. Of the ANOVAs on the CFIT split groups, there were no significant results. In contrast, the median split groups by OSPAN differed significantly in performance on the symmetry span task, $F(1, 33) = 6.818$, $p = 0.013$; the difference on the average score of set sizes 4 and 5 on the symmetry span was marginal,

1 Identical analyses were performed by dividing participants into groups based on the top third and bottom third of performance on the CFIT and OSPAN tasks. Results were similar to the median split, except where reported in the text.

$F(1, 33) = 4.085$, $p = 0.051$. In both cases, participants with higher OSPAN scores had higher scores on the symmetry span task, which is expected if both tasks are measures of working memory span. In the top/bottom thirds split of the OSPAN scores, performance on the matrices subtest of the CFIT differed significantly by group, $F(1, 22) = 4.568$, $p = 0.044$, with participants in the high span group having higher scores on the matrices subtest.

3.2 Training Performance

As shown in Figures 8 and 9, both the training group and the active control group improved on the respective training tasks over the course of the training.

There were substantial individual differences in performance on both tasks, as shown in Figures 10 and 11. Across the 18 participants in the dual n-back training group, participants ranged from a maximum daily average n-level of 2.20 to 4.40. Linear regressions on each participant's daily average n-level revealed slopes ranging from −0.073 to 0.190, which was used to proxy improvement on the training task over the training sessions, and intercepts ranging from 1.53 to 2.50, which was used to proxy "day zero" (initial) performance in the n-back training task. All participants but one improved their daily average n-level over the training sessions, with improvements ranging from −0.94 to 2.15, with the mean improvement between the first and last training sessions = 1.02, $sd = 0.72$. Across the 18 participants in the active control group, minimum daily average RTs for correct responses ranged from 343ms to 782ms. All participants improved their RT over the course of the training, with improvements ranging from 137ms to 868ms. The mean improvement in RTs for correct responses was 471ms, *sd* = 159ms. Accuracy remained unchanged throughout the active control task, ranging from 89.0% to 93.3% (mean = 91.2%, *sd* = 1.1) over the sessions. Accuracy averaged 92.0% on the first day of training, and 89.5% on the last day of training.

3.3 Individual Differences in Culture Fair Intelligence Test and Operation Span on Training Task Performance

To test for individual differences in performance on the training task based on performance on the CFIT and operation span in the pre-training session, a series of ANOVAs with CFIT median split group as the between-subjects factor, and with OSPAN median split group as the between-subjects factor, were performed.

The top/bottom thirds split of CFIT scores revealed a significant main effect of CFIT split group on the n-back slope, $F(1, 10) = 6.251$, $p = 0.031$, with participants with higher CFIT scores in the pre-training session also having a higher slope on the n-back training task (mean = 0.098, *sd* = 0.056) than participants with lower CFIT scores (mean = 0.022, *sd* = 0.049).

3.4 Changes in Performance on Outcome Measures Over Test Sessions

The design to evaluate the primary hypotheses of changes in performance over the pre-training versus post-training testing sessions were a set of 2 x 2 designs, with training group (active control, training) as between-subjects factors and testing session (pre-training, post-training) as within-subjects factors.

G*Power (version 3.1, Faul, Erdfelder, Lang, and Buchner, 2007) was used to conduct a post hoc power analysis. The power analysis indicated sufficient power to detect group x session interactions in the outcome measure data. The power to detect large ($f = 0.40$) and medium ($f = 0.25$) effect sizes was over 0.99, given the sample size.

3.4.1 Culture Fair Intelligence Test

Descriptive statistics for all of the outcome measures are presented in Table 3. Point estimates (*PE*) represent the best estimation of the variables of interest, and are based on the GLMMs described earlier. Confidence intervals (*CI*) represent a range of values: 95% *CIs* reflect the fact that 95% of the time, the actual value of the variable that we are measuring is within the interval, so we can be confident that our *CI* contains the actual value of the variable that we are interested in. *PEs* and *CIs* are provided for the difference scores (post-training score minus pre-training score).

Results from significance testing to compare pre-training and post-training performance are provided in Table 4. Of the five CFIT measures, point estimates and 95% confidence intervals suggest improvements in the outcome measures for two of the measures for the active control group, and four measures for the training group. As shown in Table 3 and Figure 12, the training group improved on the CFIT classifications subtest, *PE* = 1.94, 95% *CI* = [0.99, 2.90], the CFIT matrices subtest, *PE* = 0.56, 95% *CI* = [0.09, 1.02], the CFIT conditions/topology subtest, *PE* = 1.56, 95% *CI* = [1.04, 2.07], and the overall CFIT score, *PE* = 4.33, 95% *CI* = [2.93, 5.74]. The active control group also improved on the CFIT conditions/topology subtest, *PE* = 1.00, 95% *CI* = [0.60, 1.41], and the overall CFIT score, *PE* = 2.00, 95% *CI* = [0.25, 3.73] (Table 3).

The 2 x 2 ANOVAs (Table 4) revealed significant main effect of training group for the CFIT classifications subtest, $F(1, 34) = 4.489$, $p = 0.042$, with the training group scoring slightly higher than the active control group. While the main effects of session were significant for three of the reasoning measures, the training group x session interactions approached significance only for the overall CFIT score, $F(1, 34) = 3.948$, $p = 0.055$, generalised eta-squared (*ges*) = 0.028, with the training group improving by more than 100% more on the overall CFIT score than the active control group.

3.4.2 Monty Hall Problem

Neither the active control group nor the training group improved performance on the MHP (see Table 3). In both the pre-training session and post-training sessions for both groups, the proportion of trials in which participants switched was at or below chance (50%).

3.4.3 Working Memory Tasks

While the OSPAN and symmetry span tasks are often used as measures of domain-general executive WM capacity, our analysis of the data from both the active control and training groups showed clear ceiling effects in the smaller set sizes on these tasks, indicating that including the smaller set sizes in the pre-training and post-training comparisons may make improvements on the overall OSPAN and symmetry span scores harder to detect. The high end of the 95% *CIs* for the OSPAN set size of two averaged 0.95 in the pre-training session and 1.0 in the post-training session, and averaged 0.81 and 0.96 in the pre-training and post-training sessions, respectively, for set sizes of three. For this reason, the average OSPAN score for set sizes of 4 and 5 was also used as a dependant variable to determine effects of training.

As shown in Table 3, while the active control group showed an improvement on the overall OSPAN score, $PE = 0.08$, 95% $CI = [0.03, 0.12]$, only the training group showed an improvement on both the overall OSPAN score, $PE = 0.15$, 95% $CI = [0.10, 0.19]$, and the average score for sets of 4 and 5, $PE = 0.18$, 95% $CI = [0.12, 0.24]$. Figure 13 shows the improvements in OSPAN and symmetry span scores for both groups.

ANOVAs indicated significant training group x session interactions for both the overall OSPAN score, $F(1, 34) = 4.734$, $p = 0.037$, and the average score for set sizes 4 and 5, $F(1, 34) = 5.547$, $p = 0.024$ (Table 4). Compared to the active control group, the training group improved significantly on the OSPAN measures of WM capacity, with the training group improving by 285% more than the active control group on set sizes of 4 and 5.

Unlike the OSPAN, both the active control group and the training group improved similarly on the symmetry span task, when measured as the overall score and as the average of set sizes of 4 and 5 (see Table 3). Scores for the active control group improved on the overall task, $PE = 0.08$, 95% $CI = [0.03, 0.14]$, and on the average of set sizes of 4 and 5, $PE = 0.12$, 95% $CI = [0.04, 0.20]$. Scores for the training group improved by similar amounts, $PE = 0.08$, 95% $CI = [0.04, 0.13]$ for the overall score, and $PE = 0.10$, 95% $CI = [0.04, 0.16]$ for the average of set sizes 4 and 5.

The training group x session interactions for the symmetry span task (Table 4) were not significant.

3.4.4 Stroop Task

Table 3 and Figure 14 show that the Stroop RT interference effect (calculated as the difference between RTs on incongruent and congruent trials) for the 80% congruent Stroop task was significant in both groups in both the pre-training and post-training session, as the 95% *CI* of the interference effect did not include zero. The RT interference effect decreased significantly in the active control group, *PE* = −84ms, 95% *CI* = [-126, -43], but the 95% *CIs* for the decrease in the RT interference effect for the training group crossed zero, *PE* = −47ms, 95% *CI* = [-98, 3]. The Stroop error interference effect (calculated as the difference in errors on incongruent and congruent trials, see Figure 15) was also significant in both groups in both the pre-training and post-training sessions, as the 95% *CI* of the error interference effect did not include zero, but the change in the error interference was not significant for either group.

Training group x session interactions (Table 4) were non-significant for both RT and error interference measures.

In order to examine the Stroop effects in more detail, we analysed the RTs and errors for both congruent and incongruent Stroop trials separately. As shown in Table 3, while the active control group improved RTs on both incongruent (*PE* = −132ms, 95% *CI* = [-160, -105]) and congruent (*PE* = −47, 95% *CI* = [-57, -37]) trials, the active control group also increased errors (measured as log-odds error) on both trial types: *PE* = 0.45, 95% *CI* = [0.04, 0.87] for incongruent trials and *PE* = 0.91, 95% *CI* = [0.50, 1.32] for congruent trials. The training group showed no such speed-accuracy tradeoffs, with RT improvements on both incongruent (*PE* = −103ms, 95% *CI* = [-132, -73]) and congruent trials (*PE* = −58ms, 95% *CI* = [-70, -46]), with no increases in errors.

The training group x session interaction (Table 4) for errors on congruent trials was significant, $F(1, 34) = 6.607$, $p = 0.015$.

3.5 Individual Differences in Changes in Performance Over Test Sessions

The design to test for individual differences in changes in performance over the testing sessions was a series of 2 x 2 x 2 designs, with training group (active control, training) and quantile split group based on performance (high, low) as between-subjects factors, and testing session (pre-training, post-training) as a within-subjects factor. As in the earlier analysis of individual differences, two separate quantile splits were used in the final analysis: participants were split into low and high groups based on a median split of scores on the CFIT in the pre-training session, and participants were also split into the high and low groups in the same way using the scores on the OSPAN task in the pre-training session[2]. Results of the significance testing are shown in Table 5 for the CFIT split groups and Table 6 for the OSPAN split groups. Descriptive statistics are presented in Tables 7 and 8 for the CFIT split groups and OSPAN split groups, respectively.

3.5.1 Culture Fair Intelligence Test Split Groups on CFIT Performance

When the groups are split by the median overall CFIT score, the training group x session interaction for the overall CFIT score is significant, $F(1, 32) = 5.399$, $p = 0.027$ (Table 5),

2 As before, identical analyses were performed by dividing participants into groups based on the top third and bottom third of performance on the CFIT and OSPAN tasks.

with the training group improving on the CFIT more than the active control group (see Table 3). The training group x split group x session interaction was also significant, with $F(1, 32) = 9.598$, $p = 0.004$ (Table 5; see Figure 16). Data for the three-way interaction (Table 7) reveals that, while participants with low CFIT scores improve on the CFIT whether they are in the active control group or the training group, participants with high CFIT scores in the training group improve by more on the CFIT (mean = 4.78, $sd = 2.95$) than participants with high CFIT scores in the active control group (mean = -0.67, $sd = 3.81$), $F(1, 16) = 11.502$, $p = 0.004$.

The three-way training group x split group x session interaction was also significant for the CFIT Series subtest, $F(1, 32) = 10.563$, $p = 0.003$ (Table 5). Participants with high CFIT scores improved by more on the Series subtest if they were in the training group than the active control group (see Table 7), $F(1, 16) = 5.714$, $p = 0.029$. The three-way interaction was also significant for the CFIT Classifications subtest, $F(1, 32) = 6.709$, $p = 0.014$ (Table 5), with the same pattern of improvements as the Series subtest (see Table 7), $F(1, 16) = 6.443$, $p = 0.022$.

3.5.2 Culture Fair Intelligence Test Split Groups on Working Memory Task Performance

The training group x session interactions for the OSPAN measures were significant in the earlier analysis of changes in performance over testing sessions (Table 4), and remain significant when accounting for the variance in the OSPAN scores that is explained by the CFIT split groups (Table 5). Participants in the training group improved by more than participants in the active control group on the OSPAN score and the average OSPAN score for set sizes 4 and 5 (see Tables 3 and 7). No other effects were significant.

3.5.3 Culture Fair Intelligence Test Split Groups on Stroop Task Performance

The training group x session interaction for errors on congruent Stroop trials from the earlier analysis of changes in performance over testing sessions (Table 4) remained significant when split by median CFIT scores (Table 5). The three-way training group x split group x session interaction was significant for Stroop RT interference, $F(1, 32) = 4.212$, $p = 0.048$ (Table 5; see Figure 17). Table 7 shows that participants in the high CFIT split group improved (decreased) RT interference if they were in the training group (mean = -88ms, *sd* = 108) or the active control group (mean = −57ms, *sd* = 89.9). In contrast, participants in the low CFIT split group improved Stroop RT interference more if they were in the active control group (mean = −111ms, *sd* = 92.5) than the training group (mean = -6ms, *sd* = 106.8), $F(1, 16) = 4.979$, $p = 0.040$.

3.5.4 Operation Span Split Groups on CFIT Performance

The training group x session interaction for the overall CFIT score that did not reach significance in the earlier analysis of changes in performance over testing sessions (Table

4) remained non-significant (Table 6) when participants were split by median scores on the OSPAN task (high and low spans). However, the training group x session interaction was significant when split by top/bottom thirds, $F(1, 20) = 6.066$, $p = 0.023$. When we account for the variance in the overall CFIT score resulting from the OSPAN split groups, the training group improved by more than the active control group on the CFIT, mean = 4.31, *sd* = 2.84 for the training group, versus mean = 0.73, *sd* = 4.29 for the active control group.

The training group x session interaction was not significant in the top/bottom thirds OSPAN split for the CFIT Matrices subtest, $F(1, 20) = 3.252$, $p = 0.086$, with the training group improving by more than the active control group (mean = 0.69, *sd* = 0.95 for the training group versus mean = -0.55, *sd* = 2.11 for the active control group). The three-way training group x split group x session interaction was also not significant, $F(1, 20) = 3.647$, $p = 0.071$. In the training group, participants with high spans improved on the Matrices subtest (mean = 1.20, *sd* = 1.30) by more than low spans (mean = 0.38, *sd* = 0.52), while in the active control group, low spans improved by more than high spans on the Matrices subtest (mean = 0.50, *sd* = 1.00 for low spans, versus mean = 1.14, *sd* = 2.41 for high spans). While low spans improved on the Matrices subtest regardless of the training task, high span participants improved on the Matrices subtest only if they were in the training group.

3.5.5 Operation Span Split Groups on Working Memory Task Performance

The significant training group x session interactions for the OSPAN and average OSPAN score on sets of 4 and 5 that were reported in the earlier analysis of changes in performance over testing sessions (Tables 3 and 4) remained significant when participants were split by median OSPAN scores (Table 6).

3.5.6 Operation Span Split Groups on Stroop Task Performance

The significant training group x session interaction that was observed in the earlier analysis of changes in performance over testing sessions (Tables 3 and 4) for RTs on congruent Stroop trials remained significant (Table 6).

3.6 Individual Differences of Changes in Performance on Outcome Measures in the Training Group Based on n-back Performance

The design to test for individual differences within the training group was a series of 2 x 2 designs, with quantile split (high, low) as between-subjects factors and testing session (pre-training, post-training) as within-subjects factors. Training group participants were split into the high and low groups based on a median split[3] of both the slope of the n-level over the training sessions and the intercept, which were calculated using regression equations over the number of training days completed by each participant.

3 Identical analyses were performed by dividing participants into groups based on the top third and bottom third of performance on the n-back slope and n-back intercept.

A series of ANOVAs was used to determine the effects of performance on the training task on improvements on the outcome measures. Results of the significance testing are shown in Table 9, with descriptive statistics provided in Table 10.

3.6.1 n-back Slope Split Groups on Changes in Performance in the Training Group

Split groups by the n-back slope revealed a split group x session interaction on the CFIT Matrices subtest, $F(1, 16) = 3.821$, $p = 0.068$ (Table 9) that did not reach significance. Participants with high n-back slopes improved by more than participants with low n-back slopes (mean = 1.00, *sd* = 1.00 for participants with high slopes, versus mean = 0.11, *sd* = 0.93 for participants with low slopes; see Table 10).

The median split of n-back slopes revealed a split group x session interaction for errors on congruent Stroop trials, $F(1, 16) = 3.810$, $p = 0.069$ (Table 9), which did not reach significance. As shown in Table 10, measured with log-odds error, participants with higher n-back slopes worsened (mean = 0.009, *sd* = 0.011), relative to participants with lower n-back slopes (mean = −0.005, *sd* = 0.020).

3.6.2 n-back Intercept Split Groups on CFIT Performance

Splitting the training group by n-back intercept revealed a significant split group x session interaction on the CFIT Conditions/Topology subtest, $F(1, 16) = 15.077$, $p = 0.001$ (Table 9, see Figure 18). As shown in Table 10, participants with a lower n-back intercept improved by more on the CFIT Conditions/Topology subtest than participants with a higher n-back intercept (mean = 2.33, *sd* = 0.71 for the low-intercept group, versus mean = 0.78, *sd* = 0.97 for the high-intercept group).

3.6.3 n-back Intercept Split Groups on Working Memory Task Performance

The split group x session interaction was significant on the two OSPAN measures of working memory capacity. For the median split of the n-back intercept, the split group x session interaction was significant for the overall OSPAN score, $F(1,16) = 19.345$, $p < 0.001$ (Table 9, see Figure 19). Participants with a lower n-back intercept improved by more on the OSPAN than participants with higher n-back intercepts, with mean = 0.22, *sd* = 0.08 for the low intercept group and mean = 0.08, *sd* = 0.06 for the high intercept group (Table 10). The same trend holds with the average score on set sizes of 4 and 5 on the OSPAN, $F(1, 16) = 10.558$, $p = 0.005$ (See Tables 9 and 10).

3.6.4 n-back Intercept Split Groups on Stroop Task Performance

The split group x session interaction was significant for RTs on congruent Stroop trials, $F(1, 16) = 4.596$, $p = 0.048$ (Table 9). Table 10 shows that participants with higher n-back intercepts improved by more on congruent Stroop trials than participants with a lower n-back intercept, with mean = −94ms, *sd* = 63 for the high intercept group and mean = −23ms, *sd* = 76 for the low intercept group. However, the split group x session interaction was also significant for errors on congruent Stroop trials, $F(1, 16) = 7.987$, $p = 0.012$ (Table 9). The high intercept group improved log-odds error on the congruent Stroop trials compared to the low intercept group, with mean = 0.012, *sd* = 0.010 for the high intercept group and mean = −0.007, *sd* = 0.018 for the low intercept group (Table 10). Thus, the improvement in RTs on congruent trials for the high intercept group was not a result of a speed-accuracy tradeoff. The same trend held for errors on incongruent Stroop trials, with the high intercept group improving log-odds error more than the low intercept group, however the split group x session interaction was marginal, $F(1, 16) = 3.200$, $p = 0.093$ (Table 9). The effect on incongruent Stroop trials was significant when the training group was split by top/bottom thirds on the n-back intercept, $F(1, 10) = 6.785$, $p = 0.026$, with the high intercept group improving log-odds error by more than the low intercept group (mean = 0.046, *sd* = 0.033 for the high intercept group, versus mean = −0.051, *sd* = 0.085 for the low intercept group).

3.7 Expectations

In the pre-training session, after being instructed on how to use the training program (after the administration of the outcome measures), participants in both groups completed a short questionnaire in order to gauge their expectations of whether the training task would affect their performance on the outcome measures. The same questionnaire was repeated in the post-training session, after the training had been completed, and prior to the administration of the outcome measure tasks. While the inclusion of an active control task might have reduced the likelihood of positive results occurring due to Hawthorne or placebo-like effects, these measures were included in order to measure the expectations and beliefs of whether the training task would improve performance on the training tasks. Participants were asked, "On a scale from 1 to 5, where 1 is "not at all" and 5 is "very much", where 3 is "unsure", do you think that the training will improve your performance on (each task)?"

Table 11 shows the *PEs* and 95% *CIs* for the expectations for each task for the active control group and the training group, for the questionnaires given in both the pre-training and post-training sessions. The only difference between the groups was for the symmetry span task, where participants in the training group expected their performance on the task to improve by more than participants in the active control group, $F(1, 31) = 16.39$, $p = 0.003$.

In some, but not all, cases, the expectations of participants matched the actual improvements. For the CFIT, neither the training group nor the active control group thought that their scores would improve, since the 95% *CI* for both groups included 3 (unsure). However, both groups improved their scores on the CFIT significantly. Based on the 95% *CIs*, both groups expected improved performance on the OSPAN and symmetry span tasks, which turned out to be the case. Contrary to the group difference in expectations, the training group did not improve by more than the active control group on the symmetry span task. Interestingly, both groups did not expect improvement on the MHP, which also turned out to be the case (in the post-training session, the 95% *CI* for the MHP for the training group included 3, but only just; the training group x session interaction was not significant). Finally, both groups expected improvement on the Stroop task. In reality, both groups improved on Stroop RTs, but only the training group improved on RTs without increasing errors (in the pre-training session, the 95% *CI* for the Stroop task includes 3 for the training group, but only just; the training group x session interaction was not significant).

Chapter 4: Discussion of Results

The purpose of this study was to investigate the effects of training on a dual n-back task on a variety of cognitive processes that have been related to WM capacity, including fluid intelligence, reasoning, WM capacity as measured using two tasks drawing upon different domains, and attentional control. In addition, we investigated whether individual differences in fluid intelligence and WM capacity were related to the degree of improvement in the outcome measures. Our hypotheses were that (1) compared to the active control group, the training group will show improved scores on the CFIT, the MHP, WM span as measured by the OSPAN, WM span as measured by the symmetry span, and attentional control in the Stroop task; (2) the improvements in the outcome measures will be related to performance on the training task; and (3) participants with lower WM spans and scores on the CFIT in the pre-training session will improve on the outcome measures more than participants with higher WM span scores and scores on the CFIT.

Our investigation revealed numerous interesting findings. Both the training and active control groups were able to improve performance on the training tasks over the training sessions. Also, both the training group and the active control group showed improvements on some of the outcome measures; the active control group and the training group had similar expectations as to whether performance would improve as a result of the training. However, these improvements differed by training group, with differential improvements between the groups for the operation span and some measures of the Stroop task, and two of the CFIT subtests, partially supporting our first hypothesis. Within the training group, individual differences in regression intercepts of the n-back training data explained performance improvements on a number of the outcome measures, partially supporting our second hypothesis. Finally, improvements also differed based on individual differences in fluid intelligence, but not by individual differences in WM span, partially supporting our third hypothesis.

Our data suggest that fluid intelligence, as measured by the CFIT, may improve as a result of the dual n-back training. While both the active control group and the training group improved on the CFIT score overall, the improvement in scores in the training group was almost twice that than in the active control group; the training group x session interaction did not quite reach significance (Tables 3 and 4), although the training group x session interaction was significant when we account split the groups by median scores on the CFIT in the pre-training session. In terms of the CFIT subtests, neither group improved significantly on the Series subtest, while both groups improved on the Conditions/Topology subtest. Only the training group improved on the Classifications and Matrices subtests. The improvement on the Matrices subtest agrees with improvements on matrix reasoning fluid intelligence tests in other training studies (e.g., Jaeggi et al., 2008, 2010, but see Redick et al., 2012 for a study that did not yield such improvements). As far as we are aware, this is the first study to examine performance on subtests of fluid intelligence. While both Borella, Carretti, Riboldi, and De Beni (2010) and Redick et al. (2012) included the CFIT as an outcome measure in their respective training studies, neither have reported results for the individual subtests. Given that the active control group improved on only the Conditions/Topology subtest, and the training group did as well, the improvement in the overall CFIT score for the active control group is largely driven by improvement on this subtest (in fact, 50% of the improvement on the overall CFIT score for the active control group resulted from improvement on the Conditions/Topology subtest). Therefore, it may be that the n-back task in general, even a non-adaptive version, may train WM processes that are common to performance on the Conditions/Topology subtest. Alternatively, the Conditions/Topology subtest may be more susceptible to test-retest improvements than the other subtests. However, even though we presented results for individual CFIT subtests, we do not argue for the legitimacy of scores on the individual subtests as a measure of fluid intelligence itself, much as we (and others) have argued against the use of matrix-only fluid intelligence tests.

However, the analysis of individual differences yielded other interesting results for the CFIT. When participants were split into high (above-median) and low (below-median) groups based on their CFIT score in the pre-training session, we found significant three-way training group x split group x session interactions for the CFIT Series subtest, the Classifications subtest, and overall CFIT score (Table 5). Among participants with above-median CFIT scores, only participants in the training group improved significantly on these two CFIT subtests, and on the overall CFIT test. For the overall CFIT score, the training group x session interaction was significant, with the training group improving by more than the active control group on the measure of fluid intelligence. The split group x session interaction was also significant; participants with below-median CFIT scores improved by more on the CFIT than did participants with above-median CFIT scores. Finally, the significant three-way interaction reveals that, among participants with higher CFIT scores, only participants in the training group demonstrated significant improvement on the test of fluid intelligence. Based on these results, it appears that participants with lower CFIT scores are likely to show improvements on the CFIT whether they undertake the dual n-back training or the active control training. Participants with higher CFIT scores show improvements on the CFIT only if they undertook the dual n-back training. These individual differences may explain some of the divergent results from a number of other training studies, and suggest that initial fluid intelligence scores may be an important variable that affects transfer from training. Sternberg (2008) noted that the participants in Jaeggi et al. (2008) and Jaeggi et al. (2010) were primarily healthy young college students, and are not a representative sample of the general population. Other studies (e.g., Borella et al., 2010) make use of older adults, which are also a selective sample. Redick et al. (2012) used a more diverse sample of healthy young adults by including participants from three different universities in their study. However, it is not clear whether their participants are not all of above-average intelligence. While most of the participants in our study were students from a single university, the analysis of individual differences in fluid intelligence on the outcome measures provides some evidence that dual n-back training may have differential impact based on initial fluid intelligence scores.

While the CFIT is one measure of fluid intelligence, and the CFIT subtests measures of reasoning in Series, Classifications, Matrices, and Conditions contexts, each with different loadings on fluid intelligence (Cattell, 1963), the MHP allows a measure of normative reasoning in the presence of faulty heuristics. In the present experiment, neither the active control group nor the training group chose the payoff-maximising strategy (to switch to the remaining, initially unchosen card) in either the pre-training or post-training sessions. In the pre-training session, both groups chose the optimal strategy at a rate less than chance (the 95% *CI* for the proportion of switches was less than, and did not include 50% for both groups; Table 3). These results agree with the results of other human studies of performance on the MHP, and demonstrate the presence of faulty heuristics that tend to favour not deviating from the original choice (e.g., Burns and Wieth, 2004; Friedman, 1998; Granberg and Brown, 1995; Herbranson and Schroeder, 2010, Krauss and Wang, 2003; Tubau and Alonso, 2003). In Herbranson and Schroeder, from whose version of the MHP we have derived the one used in our experiment, participants in a multiple-block study of performance on the MHP did not differ from chance in choosing the optimal strategy on either the first block (*PE* = 56.67%, 95% *CI* = [34.51, 78.83]) or the final block (*PE* = 65.67, 95% *CI* = [45.04, 86.29]). The small increase in the number of switches was not significant. Likewise, in our experiment, the use of the optimal strategy by both the active control group and the training group did not differ from chance in the post-training session, and any increases in the use of the optimal strategy across the testing sessions was not significant for either group.

The individual differences analysis did not reveal any differences in the use of the optimal strategy between participants with higher fluid intelligence or lower fluid intelligence. Similarly, individual differences in WM span did not reveal any differences in the use of the optimal strategy between participants with high WM spans and low WM spans, suggesting that individual differences in fluid intelligence and WM capacity are not related to performance on the MHP. These results contrast with those reported by De Neys and Verschueren (2006), who, in their first experiment, found that participants that gave the correct response on a version of the MHP had larger WM spans than participants that did not give the correct response. Based on their results, De Neys and Verschueren suggest an association between WM capacity and reasoning on the MHP. However, the MHP that was presented in their experiment differed in important ways from the MHP used in our experiment. In our experiment, participants made a choice of one of the three cards that were presented on the table. The experimenter removed one of the remaining two unchosen cards (that was not the winning card), from the game. Participants then made a second choice from the remaining two cards that were present on the table. Words that are commonly associated with the MHP were carefully avoided (such as "stick," "switch," and "door"), in order to avoid cueing participants as to the nature of the problem. De Neys and Verschueren presented the problem as a game show with three doors, and name the game show host was identified as Monty Hall (see De Neys and Verschueren, 2006, Appendix). De Neys and Verschueren present the problem by asking the question, "What should you do to have the most chance of winning the main prize? a. Stick with your first choice, door 1; b. Switch to the other door, door 2; or c. It does not matter. Chances are even" (De Neys and Verschueren, 2006, p. 131). This presentation of the MHP may cue astute participants that the chances of winning are not even between the options presented. The primary cues in the presentation include the question itself (What should you do to have the most chance of winning the main prize?), which implies that the chances of winning the main prize differ between the choices. Also, the inclusion of the "It does not matter. Chances are even" option may cue astute participants that something is amiss, and that the chances are not, in fact, even. Based in the strong associations between WM span and a variety of measures of higher-order cognitive processes, if we assume, perhaps reasonably, that the participants with higher WM span are more astute than participants with lower WM span, then the results presented by De Neys and Verschueren are not surprising, and do not reflect normative reasoning on the MHP, but rather reflect individual differences in the ability to pick up on the cues presented in the formulation of the MHP. As Burns and Wieth (2004) noted, Monty Hall-style problems with differing cover stories, but the same underlying mathematical structure, are not solved equally; the cover story can influence performance of participants undertaking the problem. In the present study, both the active control group and the training group improved OSPAN scores between the pre-training and post-training sessions (and the training group x session interaction was significant, such that the training group improved WM spans by more than the active control group), but neither group improved performance on the MHP, despite improved WM spans.

Moreover, the median split by WM span did not reveal any differences in adopting the optimal strategy on the MHP between participants with low and high WM spans. Finally the three-way training group x OSPAN split group x session interaction was not significant. We therefore suggest that the differences in the presentation of the MHP between our study and that of De Neys and Verschueren is primarily responsible for the divergent results.

As mentioned above, both the active control group and the training group improved scores on the OSPAN. However, when examining the OSPAN score for the larger set sizes of 4 and 5 items, only the training group showed any improvement in WM span. Moreover, the training group x session interaction was significant for both the overall OSPAN score and the score on set sizes of 4 and 5, with the training group improving WM significantly more than the active control group. Conversely, both the active control group and the training group improved scores on the symmetry span measure of WM capacity, whether measured as the overall score or as the score for the set sizes of 4 and 5. The training group x session interactions were not significant, and both groups improved by a similar amount on the symmetry span measures. Seider et al. (2010), and others (e.g., Bergman et al., 2011; Chein and Morrison, 2010; Jaeggi et al., 2010; Klingberg et al., 2005; Morrison and Chein, 2011; Shipstead et al., 2012) have noted improvements in untrained WM tasks, such as the OSPAN, as a result of WM training, suggesting that in some WM training studies, the WM training task effectively improves WM span, even when the span measure is different from the training task.

The different results between the OSPAN and symmetry span can be explained in a number of ways. The simplest explanation may be that the memory component of the symmetry span task, which involved remembering the location of a filled square in a 3 x 3 matrix, has much in common with the visuospatial component of the dual n-back task, which also involves remembering the location of a filled square in a 3 x 3 matrix. In both the symmetry span and n-back task cases, the 3 x 3 matrices were presented on a computer screen. Also, the dual n-back task involved a second task (the verbal component of the task), which, while it is not in the same domain as the visuospatial component of the task, allows participants to practice remembering visuospatial information under dual-task conditions. The dual n-back task, then, has much in common with the symmetry span task, which may explain why both the active control group and the training group improved performance on the symmetry span task. Unlike the symmetry span task, the verbal component of the dual n-back task has little in common with the OSPAN: the verbal component of the dual n-back task requires that participants remember letters that are read aloud through computer speakers, while the OSPAN task requires the memory of words that are read by the participants. A second explanation for the different results between the OSPAN and symmetry span tasks is that, while the two tasks measure a domain-general component of WM, the tasks differ in domain and also include domain-specific components. Kane et al. (2004) noted that the OSPAN and the symmetry span task shared between 70-85% of their variance; in the present study, we found that pre-training performance on the OSPAN task explains pre-training performance on the symmetry span task. That is, participants with higher OSPAN scores also had higher scores on the symmetry span task (Table 2), suggesting that the two tasks measure a common component of WM. As noted by Engle, et al. (1999) and Miyake and Shah (1999), WM span tasks likely include domain-specific components. Miyake (2001) reviewed evidence that supports the domain-specific components of WM, noting that even Daneman and Carpenter (1980), in the study that introduced the reading span task, suggested a domain-specific view of WM.

> Consider, for example, an individual who, for whatever reason, is highly efficient in reading but not in arithmetic processing. If this individual is performing the reading span task, then he or she needs to allocate only a small amount of resources to the processing component (i.e., reading), so a large amount of resources are available for the maintenance of target words. The same person, however, should have more difficulty performing the operation span task because a large amount of resources needs to be allocated for the processing requirement (i.e., arithmetic), thereby leaving only a small amount of resources available for the storage of target memory items. (Miyake, 2001, p. 166)

The different results between the OSPAN and symmetry span tasks may reflect the different domain-specific components of WM that are used in each task. Taken together with the potential for the dual n-back task to have trained the visuospatial domain in a way similar to the visuospatial domain measure of WM capacity, we can account for the different results between the two measures of WM capacity.

In both the pre-training and post-training sessions, the measures of Stroop RT interference and Stroop error interference were significant for both groups (Table 3).We found that the active control group significantly improved RTs on the 80% congruent Stroop task for both congruent and incongruent trials types. However, the active control group also significantly increased errors on both trial types, suggesting a speed-accuracy performance tradeoff. While the measure of Stroop RT interference (the difference in RT for incongruent trials and congruent trials) decreased significantly for the active control group, the improvement in RT interference was not without the cost of an increase in errors, thus we cannot conclude that the active control group improved Stroop RT interference in a meaningful way. Importantly, the training group significantly improved RTs on both the incongruent and congruent trial types, with no increase in the error rate for either trial type. The measure of Stroop RT interference did not decrease significantly, which is to be expected, since the RT interference measure is the difference between RTs on incongruent and congruent trials, and RTs decreased on both trial types. Given the nature of the 80% congruent Stroop task, and the relative rarity of the conflict (incongruent) trials in the task, the improvement in RTs on the conflict trials provides another measure of improvements in attentional control in addition to improvements in RT interference, as improvements in RTs on the common congruent (non-conflict) trials contribute to the standard interference measure. However, given that both groups improved RTs on the incongruent trials (albeit with a speed-accuracy tradeoff in the case of the active control group), the training group x session interaction did not reach significance (Table 4). Given the significant increases in errors for both congruent and incongruent trial types for the active control group, and the lack of significant increases in error for the training group, it is difficult to draw general conclusions regarding improvements on the Stroop task. Given that both groups improved RTs on both trial types, but only the training group improved RTs without the cost of an increase in errors, in particular on the rare conflict trials, we can tentatively conclude that the WM training led to some improvements on the Stroop task.

The analysis of individual differences in CFIT scores revealed a three-way training group x CFIT split group x session interaction for Stroop RT interference (Table 5). Among participants with a lower CFIT score, those in the active control group improved Stroop RT interference more than participants in the training group. However, among participants with higher CFIT scores, participants in the training group improved Stroop RT interference more than participants in the active control group (Table 7).

In a series of experiments, Kane and Engle (2003) investigated the relationship between WM capacity and performance on the Stroop task. In their experiment 4, Kane and Engle presented participants with two versions of the Stroop task: one version in which 80% of the trials are congruent, making the incongruent (conflict) trials relatively rare, and one version in which 80% of the trials were incongruent. Participants completed both tasks, with half of the participants completing the 80% congruent task, followed by the 80%

incongruent task, and the other half of participants completing the tasks in the reverse order. In the present experiment, we closely followed the 80% congruent task used by Kane and Engle. While Kane and Engle noted that RT interference on the 80% congruent Stroop differed significantly between participants with low WM spans and participants with high WM spans (regardless of the order in which participants completed the two Stroop tasks), our analysis of individual differences did not yield the same results. In order to match the methods used by Kane and Engle, we also split participants into high and low WM span groups based on the top 25% and bottom 25% of scores on the OSPAN; however, the quartile split also failed to yield significant differences between groups. In fact, Kane and Engle report that, among participants who completed the 80% congruent Stroop task first, participants with higher WM spans were slower on congruent trials than participants with lower WM spans (643 ms versus 624 ms, respectively), while the reverse was true on incongruent trials (high spans = 834 ms, low spans = 856 ms). However, we have found the opposite pattern of results: high span participants were faster on congruent trials than low span participants (631 ms versus 643 ms), and slower on incongruent trials (881 ms versus 860 ms). In addition, we have found that the high span group committed more errors on both the congruent trials (0.016, versus 0.013 for the low span group) and incongruent trials (0.079, versus 0.062 for the low span group). Stroop error interference was larger in the high span group (0.062) than in the low span group (0.049). Unlike Kane and Engle, who found that high spans showed significantly lower Stroop RT and error interference on the 80% congruent Stroop task, our results suggest no differences between the groups on either measure of Stroop interference.

Considering that we closely followed the methods and procedure used by Kane and Engle, we are unsure why we were unable to replicate their results; however Kane and Engle had a substantially higher sample size (274 participants, 136 in each of the high and low WM span groups). Moreover, the results of our individual differences analysis, which suggest no differences between low and high WM span groups on the Stroop task, do not agree with the individual differences found on other tasks involving attentional control, such as the cocktail party phenomenon (Conway et al., 2001) and the antisaccade task (Kane et al., 2001; Roberts et al., 1994).

We also found that, among participants in the training group, individual differences in performance improvements on the dual n-back training task over the training sessions (measured by the regression slope of the n-level over the number of training sessions completed) did not explain improvements in any of the outcome measures. The median split by n-back slope explained improvements on the CFIT Matrices subtest and errors on congruent Stroop trials only marginally ($p = 0.068$ and $p = 0.069$, respectively; Table 9). Interestingly, individual differences in the regression intercept of the average n-level reached by participants, which can be thought of as the initial n-level on the dual n-back task, significantly explained improvements on a number of outcome measures. Individual differences in the n-back intercept explained performance improvements on the CFIT Conditions/Topology subtest, in which participants with a lower n-back intercept improved by more than participants with a higher n-back intercept (Tables 9 and 10), again suggesting the importance of individual differences on measures taken prior to training on the effectiveness of the training. The n-back intercept also explained improvements on both the overall OSPAN and the OSPAN score for set sizes of 4 and 5. Again, it is the participants with lower n-back intercepts who improve by more on the OSPAN than participants with higher n-back intercepts. It is interesting to note that individual differences in the n-back intercept did not explain performance improvements on the symmetry span task as well, perhaps for the reasons outlined earlier. Finally, individual differences in the n-back intercept explained RT and error improvements on congruent Stroop trials. In the case of RTs, participants with higher n-back intercepts improved more than participants with lower intercepts, but participants with lower intercepts improved errors more than participants with higher intercepts.

Jaeggi et al. (2011) split the training group based on performance improvements between the first and last training sessions, and reported that the children in their study with the highest improvements in the n-level on the training task showed more transfer to other outcome measures than the children that improved the least on the training task, suggesting that improvement on the training task may be critical in determining whether transfer to untrained tasks occurs. However, Redick et al. (2012) performed a similar analysis on a more varied group of healthy young adults, using the same dual n-back training task, and did not reach the same conclusion. In the present study, we failed to show a relation between improvement in the n-back training task and transfer to untrained tasks. Rather, individual differences in the n-back intercept explained some transfer effects. While the relationship between the n-back intercept and improvements on the outcome measures was somewhat unexpected, the lack of relationship between improvements on the n-back level and the outcome measures was not entirely unexpected in retrospect. The adaptive nature of the dual n-back training task ensures that each participant is always training WM at an n-level where, for that participant, the challenge level of the task (the n-level) matches the skill level of the participant. That is, all participants are completing the dual n-back training task at their personal maximum level; while the actual n-level achieved by each participant may differ, the relative difficulty of the task is similar across all participants. We suspect that any transfer observed in the training group is related to the match between the challenge level of the task and the skill level of the participants, and not related to the n-level that is reached by each participant. It is important to note that in our study, as in Jaeggi et al. (2011) and Redick et al. (2012), the sample sizes are small when splitting only the training group by the median. However, our multiple analyses of individual differences suggests that while the improvement in the n-level on the dual n-back task is not related to improvements on the outcome measures, other variables, including the pre-training CFIT scores and the dual n-back intercept, play a role in the transfer of training to untrained tasks.

Chapter 5: Conclusion

A number of recent studies of training working memory have provided mixed results. However, many of these studies have suffered from design limitations. In the present study, we addressed many of these design limitations, and present some evidence of transfer from training on the dual n-back task to fluid intelligence, working memory capacity, and conflict resolution. The mixed results from recent studies, including our own, suggest the presence of variables, such as initial scores on fluid intelligence tests, that may mediate the effectiveness of working memory training. In addition to replicating and extending the results of previous research, future research aiming to continue to identify and explore these variables, including important individual differences, is warranted.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–95). New York: Academic Press.

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American, 225*, 82–90.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423.

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4 (10)*, 829-839.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press.

Baddeley, A. D., & Hitch, G. J. (2000). Development of working memory: Should the Pascual-Leone and the Baddeley and Hitch models be merged? *Journal of Experimental Child Psychology, 77 (2)*, 128-137.

Bergman Nutley, S., Söderqvist, S., Bryde, S., Thorell, L. B., Humphreys, K., & Klingberg, T. (2011). Gains in fluid intelligence after training non−verbal reasoning in 4−year−old children: A controlled, randomized study. *Developmental Science, 14 (3)*, 591-601.

Bleckley, M. K., Durso, F. T., Crutchfield, J. M., Engle, R. W., & Khanna, M. M. (2003). Individual differences in working memory capacity predict visual attention allocation. *Psychonomic Bulletin & Review, 10 (4)*, 884-889.

Borella, E., Carretti, B., & De Beni, R. (2008). Working memory and inhibition across the adult life-span. *Acta Psychologica, 128*, 33–44.

Borella, E., Carretti, B., Riboldi, F., & De Beni, R. (2010). Working memory training in older adults: Evidence of transfer and maintenance effects. *Psychology and Aging, 25 (4)*, 767.

Brown, K.W. & Ryan, R.M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology, 84*, 822-848.

Burns, B. D., & Wieth, M. (2004). The collider principle in causal reasoning: Why the Monty Hall dilemma is so hard. *Journal of Experimental Psychology: General, 133 (3)*, 434.

Buschkuehl, M., & Jaeggi, S. M. (2010). Improving intelligence: A literature review. *Swiss Medical Weekly, 140 (19-20)*, 266-272.

Cattell, R. B., (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54 (1)*, 1-22.

Cattell, R. B., (1971). *Abilities: their structure, growth, and action*. Oxford, England: Houghton Mifflin.

Cattell, R. B., & Cattell, A. K. S. (1973). Measuring intelligence with the culture fair tests. *Institute for Personality and Ability Testing.*

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 (2)*, 193-199.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*, 975.

Colzato, L. S., Spapé, M. M., Pannebakker, M. M., & Hommel, B. (2007). Working memory and the attentional blink: Blink size is predicted by individual differences in operation span. *Psychonomic Bulletin & Review, 14 (6)*, 1051-1057.

Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review, 8 (2)*, 331-335.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30 (2)*, 163-183.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12 (5)*, 769-786.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7 (12)*, 547-552.

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 (4)*, 450-466.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3 (4)*, 422-433.

De Beni, R., Borella, E., Carretti, B., Marigo, C., & Nava, L. A. (2008). *BAC. Portfolio per la valutazione del benessere e delle abilita cognitive nell' eta adulta e avanzata* [The assesment of well-being and cognitive abilities in adulthood and aging]. Firenze, Italy: Giunti OS.

De Neys, W., & Verschueren, N. (2006). Working Memory Capacity and a Notorious Brain Teaser. *Experimental Psychology, 53 (2)*, 123-131.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11 (1)*, 19-23.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control.* 102-134.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128 (3)*, 309.

Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition, 77*, 197-213.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39 (2)*, 175-191.

Friedman, D. (1998). Monty Hall's three doors: Construction and deconstruction of a choice anomaly. *American Economic Review, 88*, 933-946.

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. *British Journal of Educational Psychology, 70 (2)*, 177-194.

Geiger, M. A. (1997). Educators' warnings about changing examination answers: Effects on students perceptions and performance. *College Student Journal, 31*, 429-432.

Gilovich, T., Medvec, V. H., & Chen, S. (1995). Commission, omission, and dissonance reduction: Coping with regret in the "Monty Hall" problem. *Personality and Social Psychology Bulletin, 21 (2)*, 185-190.

Granberg, D., & Brown, T. A. (1995). The Monty Hall dilemma. *Personality and Social Psychology Bulletin, 21*, 182-190.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences, 11 (6)*, 236-242.

Herbranson, W. T., & Schroeder, J. (2010). Are birds smarter than mathematicians? Pigeons (Columba livia) perform optimally on a version of the Monty Hall Dilemma. *Journal of Comparative Psychology, 124 (1)*, 1-13.

Hutchison, K. A. (2007). Attentional control and the relatedness proportion effect in semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33 (4)*, 645.

Jaeggi, S. M., Buschkuehl, M., Etienne, A., Ozdoba, C., Perrig, W. J., & Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cognitive, Affective, & Behavioral Neuroscience, 7 (2)*, 75-89.

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

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

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory, 18 (4)*, 394-412.

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

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010). 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., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130 (2)*, 169.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26 (2)*, 336.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General, 132 (1)*, 47.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *American Psychological Association, 131 (1)*, 66.

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, 133 (2)*, 189.

Kane, M. J., Poole, B. J., Tuholski, S. W., & Engle, R. W. (2006). Working memory capacity and the top-down control of visual search: Exploring the boundaries of "executive attention". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32 (4)*, 749.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology, 55 (4)*, 352.

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., ... & Westerberg, H. (2005). Computerized training of working memory in children with ADHD-A randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry, 44 (2)*, 177-186.

Krauss, S., & Wang, X. T. (2003). The psychology of the Monty Hall problem: Discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General, 132 (1)*, 3.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence, 14 (4)*, 389-433.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition, 7 (3)*, 166-174.

Mackworth, J. F. (1959). Paced memorizing in a continuous task. *Journal of Experimental Psychology, 58 (3)*, 206.

Masson, M. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 57 (3)*, 203.

Miyake, A. (2001). Individual differences in working memory: Introduction to the special section. *Journal of Experimental Psychology: General, 130 (2)*, 163.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent variable analysis. *Journal of Experimental Psychology: General, 130*, 621–640.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge, England: Cambridge University Press.

Moore, M. E., & Ross, B. M. (1963). Context effects in running memory. *Psychological Reports, 12 (2)*, 451-465.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology, 11 (1)*, 56-60.

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review, 18 (1)*, 46-60.

Neff, K. D. (2003). The development and validation of a scale to measure self compassion. *Self and Identity, 2*, 223–250

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods, 162 (1)*, 8-13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2-10.

Pinheiro, J. C., & Bates, D. M. (2000). *Linear mixed-effects models: basic concepts and examples* (pp. 3-56). Springer: New York.

Poole, B. J., & Kane, M. J. (2009). Working-memory capacity predicts the executive control of visual search among distractors: The influences of sustained and selective attention. *The Quarterly Journal of Experimental Psychology, 62 (7)*, 1430-1454.

R Development Core Team. (2012). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org

Redick, T. S., Calvo, A., Gay, C. E., & Engle, R. W. (2011). Working memory capacity and go/no-go task performance: Selective effects of updating, maintenance, and inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37 (2)*, 308.

Redick, T. S., & Engle, R. W. (2006). Working memory capacity and attention network test performance. *Applied Cognitive Psychology, 20 (5)*, 713-721.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2013). No Evidence of Intelligence Improvement After Working Memory Training: A Randomized, Placebo-Controlled Study. *Journal of Experimental Psychology: General, 142 (2)*, 359.

Repovš, G., & Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience, 139 (1)*, 5-21.

Roberts, R. J., Hager, L. D., & Heron, C. (1994). Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology: General, 123 (4)*, 374.

Seidler, R. D., Bernard, J. A., Buschkuehl, M., Jaeggi, S., Jonides, J., & Humfleet, J. (2010). Cognitive training as an intervention to improve driving ability in the older adult (Tech. Rep. No. M-CASTL 2010-01). Ann Arbor: University of Michigan.

Shimojo, S., & Ichikawa, S. (1989). Intuitive reasoning about probability: Theoretical and experimental analyses of the "problem of three prisoners". *Cognition, 32*, 1-24.

Sternberg, R. J. (2008). Increasing fluid intelligence is possible after all. *Proceedings of the National Academy of Sciences, 105 (19)*, 6791-6792.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General, 125*, 4-27.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2010). Does working memory training generalize? *Psychologica Belgica, 50*, 3-4.

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

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23 (5)*, 645-665.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

Tubau, E., & Alonso, D. (2003). Overcoming illusory inferences in a probabilistic counterintuitive problem: The role of explicit representations. *Memory & Cognition, 31*, 596-607.

Turner, M. K., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28 (2)*, 127-154.

Tversky, A., & Kahneman, D. (1983). Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90 (4)*, 293–315.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133 (6)*, 1038.

Wason, P. C. (1966). Reasoning. In Foss, B. M. *New horizons in psychology*. Harmondsworth: Penguin.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer Verlag.

Wickham, H. (2012). ggplot2: An implementation of the grammar of graphics [Computer software]. Retrieved from http://CRAN.R- project.org/package=ggplot2

Appendix A: Tables

Table 1

Descriptive Statistics for Individual Differences on Pre-training Tasks, Split by Operation Span

	Operation Span			
	Low OSPAN		High OSPAN	
Task	*Mean*	*SD*	*Mean*	*SD*
Culture Fair Intelligence Test				
CFIT Series (raw score)	7.67	1.53	7.44	1.20
CFIT Classifications (raw score)	7.17	1.20	6.61	1.97
CFIT Matrices (raw score)	5.89	1.13	6.28	1.41
CFIT Conditions/Topology (raw score)	6.06	1.30	6.33	0.97
CFIT Score (raw score)	26.78	2.88	26.67	4.24
Monty Hall Problem				
Monty Hall Problem Switches (%) [a]	41.11	17.62	36.67	14.09
Working Memory Tasks				
Operation Span Score (%)	*N/A*	*N/A*	*N/A*	*N/A*
Operation Span, Sets 4 & 5 (%)	*N/A*	*N/A*	*N/A*	*N/A*
Symmetry Span Score (%) [b]	56.63	12.53	67.52	12.12
Symmetry Span, Sets 4 & 5 (%) [b]	39.03	13.82	49.41	16.52
Stroop Task				
Stroop Incongruent RT (ms)	860	223	881	137
Stroop Congruent RT (ms)	643	116	631	79
Stroop RT Interference (ms)	217	127	250	93
Stroop Incongruent Errors (%)	6.17	6.57	7.87	6.12
Stroop Congruent Errors (%)	1.30	1.25	1.65	2.32
Stroop Error Interference (%)	4.87	6.51	6.22	5.20

Note. SD = standard deviation, OSPAN = Operation Span, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 18 in all groups except [a] N = 15 in the low operation span group and N = 16 in the high operation span group. [b] N = 17 in the high operation span group.

Table 2

Descriptive Statistics for Individual Differences on Pre-training Tasks, Split by CFIT

	CFIT			
	Low CFIT		High CFIT	
Task	*Mean*	*SD*	*Mean*	*SD*
Culture Fair Intelligence Test				
CFIT Series (raw score)	*N/A*	*N/A*	*N/A*	*N/A*
CFIT Classifications (raw score)	*N/A*	*N/A*	*N/A*	*N/A*
CFIT Matrices (raw score)	*N/A*	*N/A*	*N/A*	*N/A*
CFIT Conditions/Topology (raw score)	*N/A*	*N/A*	*N/A*	*N/A*
CFIT Score (raw score)	*N/A*	*N/A*	*N/A*	*N/A*
Monty Hall Problem				
Monty Hall Problem Switches (%) [a]	39.61	13.69	37.86	18.52
Working Memory Tasks				
Operation Span Score (%)	64.37	9.97	63.44	14.63
Operation Span, Sets 4 & 5 (%)	45.56	12.21	44.17	21.69
Symmetry Span Score (%) [b]	61.45	13.43	62.36	13.63
Symmetry Span, Sets 4 & 5 (%) [b]	44.46	16.65	43.70	15.56
Stroop Task				
Stroop Incongruent RT (ms)	881	211	860	155
Stroop Congruent RT (ms)	638	113	636	85
Stroop RT Interference (ms)	243	119	224	105
Stroop Incongruent Errors (%)	7.87	6.41	6.17	6.29
Stroop Congruent Errors (%)	1.92	2.25	1.03	1.23
Stroop Error Interference (%)	5.95	6.01	5.14	5.82

Note. SD = standard deviation, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 18 in all groups except [a] N = 17 in the low CFIT group and N = 14 in the high CFIT group. [b] N = 17 in the low CFIT group.

Table 3

Significance Testing Results for Individual Differences on Pre-training Tasks, Split by CFIT and Operation Span

Task	CFIT Median Split			Operation Span Median Split		
	F	*p*	*ges*	*F*	*p*	*ges*
Culture Fair Intelligence Test						
CFIT Series	*N/A*	*N/A*	*N/A*	0.234	0.631	0.007
CFIT Classifications	*N/A*	*N/A*	*N/A*	1.040	0.315	0.030
CFIT Matrices	*N/A*	*N/A*	*N/A*	0.836	0.367	0.024
CFIT Conditions/Topology	*N/A*	*N/A*	*N/A*	0.525	0.474	0.015
CFIT Score	*N/A*	*N/A*	*N/A*	0.008	0.927	0.000
Monty Hall Problem						
Monty Hall Problem Switches [a]	0.092	0.764	0.003	0.605	0.443	0.020
Working Memory Tasks						
Operation Span Score	0.049	0.826	0.001	*N/A*	*N/A*	*N/A*
Operation Span, Sets 4 & 5	0.056	0.814	0.002	*N/A*	*N/A*	*N/A*
Symmetry Span Score [b]	0.040	0.843	0.001	6.818	0.013	0.171 *
Symmetry Span, Sets 4 & 5 [b]	0.019	0.890	0.001	4.085	0.051	0.110 †
Stroop Task						
Stroop Incongruent RT	0.121	0.730	0.004	0.118	0.733	0.003
Stroop Congruent RT	0.003	0.960	0.000	0.130	0.720	0.004
Stroop RT Interference	0.282	0.599	0.008	0.802	0.377	0.023
Stroop Incongruent Errors	0.643	0.428	0.019	0.643	0.428	0.019
Stroop Congruent Errors	2.120	0.155	0.059	0.308	0.582	0.009
Stroop Error Interference	0.171	0.682	0.005	0.475	0.496	0.014

Note. ges = generalized eta-squared, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 18 in all groups, except [a] N = 17 in the low CFIT group, N = 14 in the high CFIT group, and N = 15 in the low operation span group, N = 16 in the high operation span group. [b] N = 17 in the low CFIT group and N = 17 in the high operation span group. * = significant at $p < 0.05$, † = $p < 0.10$.

Table 4

Point Estimates and 95% Confidence Intervals for the Outcome Measures as a Function of Training Group and Session

	Active Control Group				Training Group			
	Pre-training	Post-training	Difference		Pre-training	Post-training	Difference	
Task	Point Estimate 95% CI	Point Estimate 95% CI	Point Estimate 95% CI		Point Estimate 95% CI	Point Estimate 95% CI	Point Estimate 95% CI	
			Culture Fair Intelligence Test					
CFIT Series (raw score)	7.61 [6.90, 8.32]	7.94 [7.23, 8.66]	0.33 [-0.18, 0.84]		7.50 [6.84, 8.16]	7.78 [7.11, 8.44]	0.28 [-0.29, 0.85]	
CFIT Classifications (raw score)	6.78 [6.12, 7.44]	7.67 [7.00, 8.33]	0.89 [-0.04, 1.82]		7.00 [6.30, 7.70]	8.94 [8.25, 9.65]	1.94 [0.99, 2.90]	*
CFIT Matrices (raw score)	6.44 [5.89, 7.00]	6.22 [5.67, 6.77]	-0.22 [-1.01, 0.56]		5.72 [5.11, 6.33]	6.28 [5.66, 6.89]	0.56 [0.09, 1.02]	*
CFIT Conditions/Topology (raw score)	6.00 [5.37, 6.62]	7.00 [6.37, 7.63]	1.00 [0.60, 1.41]	*	6.39 [5.88, 6.89]	7.94 [7.44, 8.45]	1.56 [1.04, 2.07]	*
CFIT Score (raw score)	26.83 [25.22, 28.46]	28.83 [27.23, 30.45]	2.00 [0.25, 3.73]	*	26.61 [25.04, 28.17]	30.94 [29.37, 32.51]	4.33 [2.93, 5.74]	*
			Monty Hall Problem					
Monty Hall Problem Switches (%) [a]	38 [28, 48]	46 [36, 56]	8 [-3, 19]		40 [31, 48]	41 [32, 50]	1 [-8, 11]	
			Working Memory Tasks					
Operation Span Score (%)	67 [62, 71]	75 [70, 79]	8 [3, 12]	*	61 [56, 66]	76 [71, 81]	15 [10, 19]	*
Operation Span, Sets 4 & 5 (%)	48 [40, 56]	52 [44, 60]	5 [-4, 14]		42 [34, 50]	60 [52, 68]	18 [12, 24]	*
Symmetry Span Score (%) [b]	60 [54, 66]	68 [62, 74]	8 [3, 14]	*	64 [56, 72]	72 [65, 80]	8 [4, 13]	*
Symmetry Span, Sets 4 & 5 (%) [b]	42 [34, 50]	54 [46, 62]	12 [4, 20]	*	46 [37, 56]	57 [47, 66]	10 [4, 16]	*
			Stroop Task					
Stroop Incongruent RT (ms)	857 [778, 936]	724 [645, 804]	-132 [-160, -105]	*	876 [809, 943]	773 [706, 840]	-103 [-132, -73]	*
Stroop Congruent RT (ms)	618 [567, 669]	571 [520, 622]	-47 [-57, -37]	*	655 [616, 694]	597 [558, 636]	-58 [-70, -46]	*
Stroop RT Interference (ms)	242 [195, 289]	158 [111, 205]	-84 [-126, -43]	*	225 [182, 268]	178 [135, 221]	-47 [-98, 3]	
Stroop Incongruent Errors [c]	-2.84 [-3.29, -2.40]	-2.39 [-2.80, -1.98]	0.45 [0.04, 0.87]	*	-2.67 [-3.10, -2.25]	-2.54 [-2.96, -2.13]	0.13 [-0.28, 0.54]	
Stroop Congruent Errors [c]	-4.63 [-5.17, -4.10]	-3.73 [-4.19, -3.26]	0.91 [0.50, 1.32]	*	-4.38 [-4.85, -3.91]	-4.23 [-4.68, -3.77]	0.16 [-0.26, 0.57]	
Stroop Error Interference (%)	5 [3, 8]	7 [4, 9]	1 [-1, 4]		6 [3, 9]	7 [4, 9]	1 [-2, 4]	

Note. CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 18 in all groups except [a] N = 14 in the active control group and N = 17 in the training group. [b] N = 17 in the training group. [c] log-odds error. * = significant within-group difference.

Table 4b

Mean Error Rate on Stroop Trials as a Function of Training Group and Session

	Active Control Group			Training Group		
Task	Pre-training	Post-training	Difference	Pre-training	Post-training	Difference
Stroop Incongruent Errors (%)	6.48%	9.72%	3.24%	7.56%	8.49%	0.93%
Stroop Congruent Errors (%)	1.30%	3.14%	1.84%	1.65%	1.92%	0.27%

Table 5

Significance Testing Results of 2 (Group) x 2 (Session) ANOVAs for the Outcome Measures

Task	Group			Session			Group x Session		
	F	*p*	*ges*	*F*	*p*	*ges*	*F*	*p*	*ges*
Culture Fair Intelligence Test									
CFIT Series	0.088	0.768	0.002	2.303	0.138	0.010	0.019	0.891	0.000
CFIT Classifications	4.489	0.042	0.061 *	15.477	0.000	0.188 *	2.148	0.152	0.031
CFIT Matrices	0.883	0.354	0.017	0.450	0.507	0.004 †	2.450	0.127	0.023
CFIT Conditions/Topology	2.989	0.093	0.068 †	54.835	0.000	0.212 *	2.591	0.117	0.013
CFIT Score	0.846	0.364	0.018	29.085	0.000	0.174 *	3.948	0.055	0.028 †
Monty Hall Problem									
Monty Hall Problem Switches [a]	0.067	0.797	0.002	1.284	0.266	0.013	0.723	0.402	0.008
Working Memory Tasks									
Operation Span Score	0.451	0.506	0.011	47.531	0.000	0.217 *	4.734	0.037	0.027 *
Operation Span, Sets 4 & 5	0.049	0.825	0.001	16.043	0.000	0.102 *	5.547	0.024	0.038 *
Symmetry Span Score [b]	0.888	0.353	0.023	18.604	0.000	0.075 *	0.004	0.952	0.000
Symmetry Span, Sets 4 & 5 [b]	0.385	0.539	0.010	17.161	0.000	0.081 *	0.070	0.793	0.000
Stroop Task									
Stroop Incongruent RT	0.358	0.553	0.009	26.554	0.000	0.108 *	0.260	0.613	0.001
Stroop Congruent RT	0.819	0.372	0.020	14.316	0.001	0.058 *	0.240	0.627	0.001
Stroop RT Interference	0.004	0.947	0.000	14.553	0.001	0.102 *	1.151	0.291	0.009
Stroop Incongruent Errors	0.002	0.967	0.000	3.400	0.074	0.026 †	1.049	0.313	0.008
Stroop Congruent Errors	0.498	0.485	0.012	11.890	0.002	0.063 *	6.607	0.015	0.036 *
Stroop Error Interference	0.050	0.825	0.001	0.896	0.350	0.008	0.117	0.734	0.001

Note. ges = generalized eta-squared, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 18 in all groups except [a] N = 14 in the active control group and N = 17 in the training group. [b] N = 17 in the training group. * = significant at $p < 0.05$, † = $p < 0.10$.

Table 6

Descriptive Statistics for Individual Differences in CFIT on Outcome Measures (Session Difference Scores)

	Active Control Group				**Training Group**			
	Low CFIT		High CFIT		Low CFIT		High CFIT	
Task	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Culture Fair Intelligence Test								
CFIT Series (raw score)	1.00	1.12	-0.33	0.71	-0.22	1.20	0.78	1.20
CFIT Classifications (raw score)	2.11	0.93	-0.33	2.45	1.44	2.07	2.44	2.19
CFIT Matrices (raw score)	0.44	1.01	-0.89	2.26	0.67	1.00	0.44	1.13
CFIT Conditions/Topology (raw score)	1.11	0.93	0.89	0.93	2.00	0.87	1.11	1.27
CFIT Score (raw score)	4.67	1.22	-0.67	3.81	3.89	3.41	4.78	2.95
Monty Hall Problem								
Monty Hall Problem Switches (%) [a]	8.33	24.30	7.22	21.23	0.00	17.87	2.92	23.80
Working Memory Tasks								
Operation Span Score (%)	5.09	11.79	10.39	6.37	14.94	8.13	14.81	12.25
Operation Span, Sets 4 & 5 (%)	-1.54	23.65	10.90	13.75	20.93	11.57	15.12	15.50
Symmetry Span Score (%) [b]	4.04	12.76	12.81	11.38	7.83	12.92	8.52	7.38
Symmetry Span, Sets 4 & 5 (%) [b]	4.07	12.17	19.44	20.37	10.10	13.40	10.56	13.34
Stroop Task								
Stroop Incongruent RT (ms)	-12.93	13.29	-12.99	11.31	-3.87	14.46	-17.37	13.58
Stroop Congruent RT (ms)	-1.80	11.47	-7.31	4.24	-3.26	7.81	-8.55	6.98
Stroop RT Interference (ms)	-11.13	9.25	-5.68	9.00	-0.61	10.69	-8.82	10.82
Stroop Incongruent Errors (%)	2.78	5.01	3.70	8.10	-0.31	9.47	2.16	3.34
Stroop Congruent Errors (%)	1.84	1.79	1.84	2.21	-0.08	1.56	0.61	1.87
Stroop Error Interference (%)	0.94	5.09	1.86	7.71	-0.23	9.34	1.55	2.48

Note. SD = standard deviation, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 9 in each group, except [a] N = 8 in the active control group, low CFIT, N = 6 in the active control group, high CFIT, N = 8 in the training group, high CFIT. [b] N = 8 in the training group, low CFIT.

Table 7

Descriptive Statistics for Individual Differences in Operation Span on Outcome Measures (Session Difference Scores)

	Active Control Group				Training Group			
	Low OSPAN		High OSPAN		Low OSPAN		High OSPAN	
Task	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Culture Fair Intelligence Test								
CFIT Series (raw score)	0.25	1.28	0.40	1.07	-0.20	1.23	0.88	1.13
CFIT Classifications (raw score)	1.00	1.07	0.80	2.86	1.80	1.87	2.13	2.53
CFIT Matrices (raw score)	0.38	1.19	-0.70	2.16	0.40	0.84	0.75	1.28
CFIT Conditions/Topology (raw score)	0.75	0.89	1.20	0.92	1.60	1.35	1.50	0.93
CFIT Score (raw score)	2.38	2.13	1.70	4.97	3.60	2.72	5.25	3.54
Monty Hall Problem								
Monty Hall Problem Switches (%) [a]	2.67	24.54	10.74	21.72	-4.00	15.06	9.05	25.22
Working Memory Tasks								
Operation Span Score (%)	13.78	7.73	2.90	8.29	18.64	10.85	10.17	7.10
Operation Span, Sets 4 & 5 (%)	12.99	18.21	-1.97	19.35	19.50	13.95	16.18	13.85
Symmetry Span Score (%) [b]	13.51	12.43	4.36	11.69	8.71	12.29	7.46	6.35
Symmetry Span, Sets 4 & 5 (%) [b]	14.17	13.73	9.83	21.52	9.08	15.35	12.14	9.32
Stroop Task								
Stroop Incongruent RT (ms)	-7.81	11.05	-17.07	11.55	-11.38	17.76	-9.67	12.64
Stroop Congruent RT (ms)	-2.76	10.97	-5.99	7.03	-5.66	7.74	-6.21	8.14
Stroop RT Interference (ms)	-5.05	9.97	-11.09	8.22	-5.72	13.34	-3.46	8.70
Stroop Incongruent Errors (%)	2.78	6.80	3.61	6.68	1.11	5.27	0.69	9.12
Stroop Congruent Errors (%)	1.90	2.32	1.79	1.73	0.21	1.63	0.34	1.92
Stroop Error Interference (%)	0.88	6.59	1.82	6.48	0.90	5.79	0.35	8.10

Note. SD = standard deviation, OSPAN = Operation Span, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 8 in the active control group, low OSPAN, N = 10 in the active control group, high OSPAN, n = 10 in the training group, low OSPAN, and N = 8 in the training group, high OSPAN, except [a] N = 5 in the active control group, low OSPAN, N = 9 in the active control group, high OSPAN, N = 7 in the training group, high OSPAN. [b] N = 7 in the training group, high OSPAN.

Table 8

Significance Testing Results for Individual Differences in CFIT on Outcome Measures

	Median Split by CFIT								
	Training Group x Session			Split Group x Session			Training Group x Split Group x Session		
Task	*F*	*p*	*ges*	*F*	*p*	*ges*	*F*	*p*	*ges*
				Culture Fair Intelligence Test					
CFIT Series	0.024	0.878	0.001	0.216	0.646	0.007	10.563	0.003	0.248 *
CFIT Classifications	2.520	0.122	0.073	1.180	0.286	0.036	6.709	0.014	0.173 *
CFIT Matrices	2.587	0.118	0.075	2.587	0.118	0.075	1.320	0.259	0.040
CFIT Conditions/Topology	2.721	0.109	0.078	2.721	0.109	0.078	0.980	0.330	0.030
CFIT Score	5.399	0.027	0.144 *	4.897	0.034	0.133 *	9.598	0.004	0.231 *
				Monty Hall Problem					
Monty Hall Problem Switches [a]	0.684	0.416	0.025	0.020	0.889	0.001	0.064	0.802	0.002
				Working Memory Tasks					
Operation Span Score	4.633	0.039	0.126 *	0.608	0.441	0.019	0.667	0.420	0.020
Operation Span, Sets 4 & 5	5.714	0.023	0.152 *	0.353	0.557	0.011	2.667	0.112	0.077
Symmetry Span Score [b]	0.010	0.923	0.000	1.615	0.213	0.050	1.119	0.298	0.035
Symmetry Span, Sets 4 & 5 [b]	0.103	0.750	0.003	2.495	0.124	0.074	2.095	0.158	0.063
				Stroop Task					
Stroop Incongruent RT	0.281	0.600	0.009	2.369	0.134	0.069	2.326	0.137	0.068
Stroop Congruent RT	0.255	0.617	0.008	4.049	0.053	0.112 †	0.002	0.969	0.000
Stroop RT Interference	1.232	0.275	0.037	0.172	0.681	0.005	4.212	0.048	0.116 *
Stroop Incongruent Errors	1.008	0.323	0.031	0.542	0.467	0.017	0.112	0.740	0.003
Stroop Congruent Errors	6.337	0.017	0.165 *	0.305	0.584	0.009	0.305	0.584	0.009
Stroop Error Interference	0.111	0.741	0.003	0.369	0.548	0.011	0.037	0.849	0.001

Note. ges = generalized eta-squared, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 9 in each group, except [a] N = 8 in the active control group, low CFIT, N = 6 in the active control group, high CFIT, N = 8 in the training group, high CFIT. [b] N = 8 in the training group, low CFIT. * = significant at $p < 0.05$, † = $p < 0.10$.

Table 9

Significance Testing Results for Individual Differences in Operation Span on Outcome Measures

	Median Split by Operation Span								
	Training Group x Session			Split Group x Session			Training Group x Split Group x Session		
Task	*F*	*p*	*ges*	*F*	*p*	*ges*	*F*	*p*	*ges*
				Culture Fair Intelligence Test					
CFIT Series	0.001	0.975	0.000	2.405	0.131	0.070	1.371	0.250	0.041
CFIT Classifications	2.031	0.164	0.060	0.007	0.934	0.000	0.124	0.727	0.004
CFIT Matrices	2.214	0.147	0.065	0.535	0.470	0.016	2.067	0.160	0.061
CFIT Conditions/Topology	2.649	0.113	0.076	0.245	0.624	0.008	0.606	0.442	0.019
CFIT Score	3.975	0.055	0.110 †	0.166	0.687	0.005	0.942	0.339	0.029
				Monty Hall Problem					
Monty Hall Problem Switches [a]	0.257	0.616	0.009	1.942	0.175	0.067	0.100	0.754	0.004
				Working Memory Tasks					
Operation Span Score	4.271	0.047	0.118 *	10.873	0.002	0.254 *	0.170	0.683	0.005
Operation Span, Sets 4 & 5	4.925	0.034	0.133 *	2.704	0.110	0.078	1.096	0.303	0.033
Symmetry Span Score [b]	0.068	0.796	0.002	1.940	0.174	0.059	1.056	0.312	0.033
Symmetry Span, Sets 4 & 5 [b]	0.076	0.784	0.002	0.020	0.889	0.001	0.446	0.509	0.014
				Stroop Task					
Stroop Incongruent RT	0.174	0.680	0.005	0.675	0.418	0.021	1.423	0.242	0.043
Stroop Congruent RT	0.304	0.585	0.009	0.443	0.510	0.014	0.221	0.641	0.007
Stroop RT Interference	1.003	0.324	0.030	0.295	0.591	0.009	1.425	0.241	0.043
Stroop Incongruent Errors	0.958	0.335	0.029	0.008	0.930	0.000	0.071	0.791	0.002
Stroop Congruent Errors	6.134	0.019	0.161 *	0.001	0.978	0.000	0.036	0.850	0.001
Stroop Error Interference	0.103	0.750	0.003	0.007	0.933	0.000	0.110	0.743	0.003

Note. ges = generalized eta-squared, OSPAN = Operation Span, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 8 in the active control group, low OSPAN, N = 10 in the active control group, high OSPAN, n = 10 in the training group, low OSPAN, and N = 8 in the training group, high OSPAN, except [a] N = 5 in the active control group, low OSPAN, N = 9 in the active control group, high OSPAN, N = 7 in the training group, high OSPAN. [b] N = 7 in the training group, high OSPAN. * = significant at $p < 0.05$, † = $p < 0.10$.

Table 10

Descriptive Statistics for Individual Differences in the Training Group, Split by n-back Slope and Intercept (Session Difference Scores)

	Median Split by Slope				**Median Split by Intercept**			
	Low Slope		High Slope		Low Intercept		High Intercept	
Task	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
	Culture Fair Intelligence Test							
CFIT Series (raw score)	0.00	1.12	0.56	1.42	-0.11	1.27	0.67	1.22
CFIT Classifications (raw score)	1.44	2.24	2.44	2.01	1.67	2.35	2.22	1.99
CFIT Matrices (raw score)	0.11	0.93	1.00	1.00	0.33	1.12	0.78	0.97
CFIT Conditions/Topology (raw score)	1.67	1.32	1.44	1.01	2.33	0.71	0.78	0.97
CFIT Score (raw score)	3.22	3.42	5.44	2.51	4.22	3.73	4.44	2.60
	Monty Hall Problem							
Monty Hall Problem Switches (%) [a]	-0.74	15.16	3.75	25.72	2.59	15.88	0.00	25.39
	Working Memory Tasks							
Operation Span Score (%)	17.16	8.79	12.59	11.28	22.13	7.84	7.62	6.03
Operation Span, Sets 4 & 5 (%)	21.98	10.28	14.07	15.89	26.36	5.89	9.69	14.22
Symmetry Span Score (%) [b]	6.44	10.61	9.75	9.83	12.20	8.91	4.63	10.05
Symmetry Span, Sets 4 & 5 (%) [b]	9.06	12.46	11.48	14.01	10.52	12.65	10.19	13.96
	Stroop Task							
Stroop Incongruent RT (ms)	-7.86	12.69	-13.38	17.83	-8.08	14.94	-13.16	16.08
Stroop Congruent RT (ms)	-4.12	8.23	-7.70	7.12	-2.38	7.65	-9.44	6.25
Stroop RT Interference (ms)	-3.74	10.99	-5.69	12.09	-5.71	11.18	-3.72	11.91
Stroop Incongruent Errors (%)	-1.54	8.91	3.40	3.34	-1.85	8.78	3.70	3.11
Stroop Congruent Errors (%)	-0.46	1.95	1.00	1.10	-0.69	1.79	1.23	0.96
Stroop Error Interference (%)	-1.08	8.83	2.40	3.22	-1.16	8.80	2.48	3.20

Note. SD = standard deviation, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 9 in each group, except [a] N = 8 in the active control group, low CFIT, N = 6 in the active control group, high CFIT, N = 8 in the training group, high CFIT. [b] N = 8 in the training group, low CFIT.

Table 11

Significance Testing Results for Individual Differences in the Training Group, Split by n-back Slope and Intercept

	Median Split by Slope			**Median Split by Intercept**		
	Split Group x Session			Split Group x Session		
Task	*F*	*p*	*ges*	*F*	*p*	*ges*
	Culture Fair Intelligence Test					
CFIT Series	0.847	0.371	0.050	1.750	0.204	0.099
CFIT Classifications	0.994	0.334	0.058	0.294	0.595	0.018
CFIT Matrices	3.821	0.068	0.193 †	0.810	0.381	0.048
CFIT Conditions/Topology	0.160	0.694	0.010	15.077	0.001	0.485 *
CFIT Score	2.473	0.135	0.134	0.021	0.885	0.001
	Monty Hall Problem					
Monty Hall Problem Switches [a]	0.198	0.663	0.013	0.065	0.802	0.004
	Working Memory Tasks					
Operation Span Score	0.918	0.352	0.054	19.345	0.000	0.547 *
Operation Span, Sets 4 & 5	1.570	0.228	0.089	10.558	0.005	0.398 *
Symmetry Span Score [b]	0.447	0.514	0.029	2.673	0.123	0.151
Symmetry Span, Sets 4 & 5 [b]	0.140	0.714	0.009	0.003	0.959	0.000
	Stroop Task					
Stroop Incongruent RT	0.573	0.460	0.035	0.481	0.498	0.029
Stroop Congruent RT	0.975	0.338	0.057	4.596	0.048	0.223 *
Stroop RT Interference	0.127	0.726	0.008	0.133	0.721	0.008
Stroop Incongruent Errors	2.427	0.139	0.132	3.200	0.093	0.167 †
Stroop Congruent Errors	3.810	0.069	0.192 †	7.987	0.012	0.333 *
Stroop Error Interference	1.235	0.283	0.072	1.359	0.261	0.078

Note. ges = generalized eta-squared, CFIT = Cattell's Culture Fair Intelligence Test, RT = response time. N = 9 in each group except [a] N = 8 in high slope group and N = 8 in high intercept group. [b] N = 8 in low slope group and N = 8 in high intercept group. * = significant at $p < 0.05$, † = $p < 0.10$.

Table 12

Point Estimates and 95% Confidence Intervals for Expectations for Improvement as a Function of Training Group and Session

	Active Control Group						Training Group					
	Pre-training			Post-training			Pre-training			Post-training		
Task	Estimate	95% CI		Estimate	95% CI		Estimate	95% CI		Estimate	95% CI	
CFIT	3.12	[2.65, 3.58]		3.35	[2.89, 3.81]		3.25	[2.78, 3.72]		2.75	[2.28, 3.22]	
Operation Span	4.24	[3.84, 4.63]	*	3.59	[3.20, 3.98]	*	4.44	[4.03, 4.84]	*	4.38	[3.97, 4.78]	*
Symmetry Span	4.12	[3.77, 4.46]	*	3.76	[3.42, 4.11]	*	4.75	[4.39, 5.11]	*	4.69	[4.33, 5.04]	*
Monty Hall Problem	1.82	[1.30, 2.35]	†	1.71	[1.18, 2.23]	†	2.19	[1.64, 2.73]	†	2.50	[1.96, 3.04]	
Stroop	3.59	[3.06, 4.12]	*	3.76	[3.24, 4.29]	*	3.44	[2.89, 3.98]		3.63	[3.08, 4.16]	*

Note. CFIT = Cattell's Culture Fair Intelligence Test. Scores range from 1 (“not at all”) to 5 (“very much”), with 3 being “unsure”. * = significantly above 3, † = significantly below 3.

Appendix B: Figures

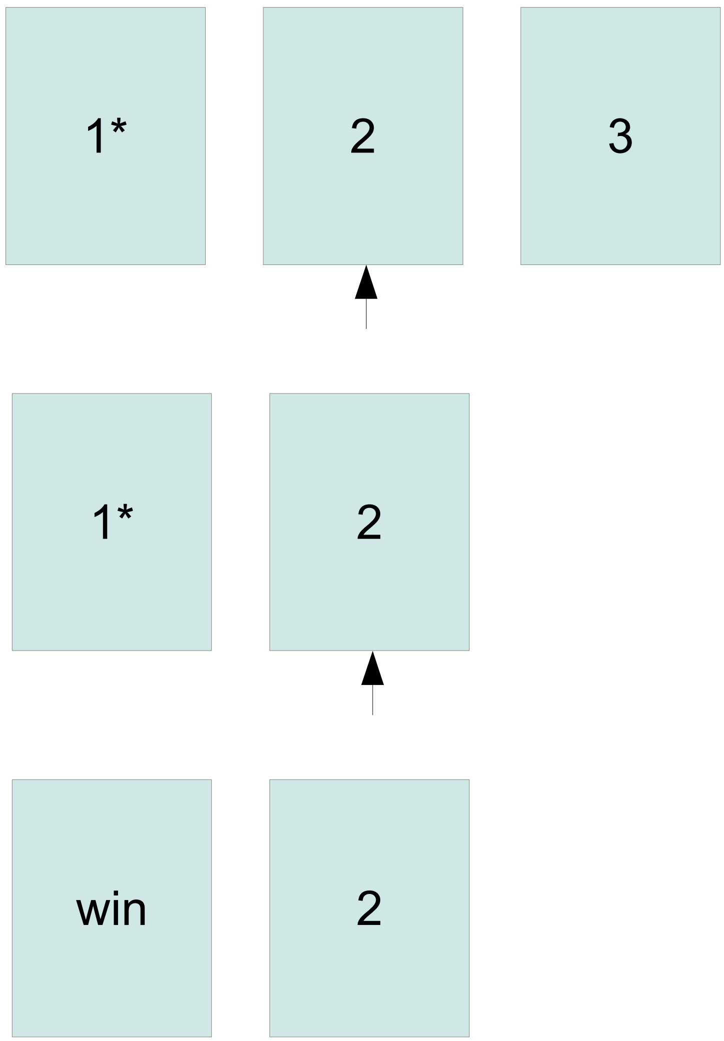


Figure 1. Depiction of the Monty Hall Problem. Participants first choose one from three cards, marked with an arrow. The winning card has been randomly determined in advance (marked with *). Following the choice, a remaining card that is not the winner is removed from the game. The participant chooses a card again. The winning card is then revealed.

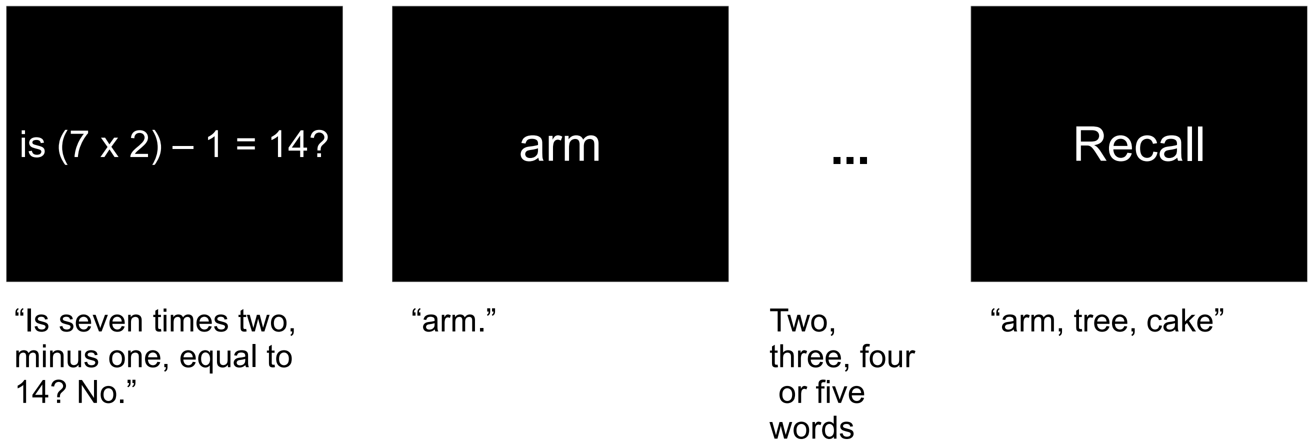


Figure 2. Depiction of the Operation Span task. Participants read a math equation that is presented on the screen out loud. After responding as to whether the equation is correct or not, participants are presented with a word that they have to remember, which they read out loud. After two, three, four, or five words, participants are presented with a recall instruction and respond verbally.

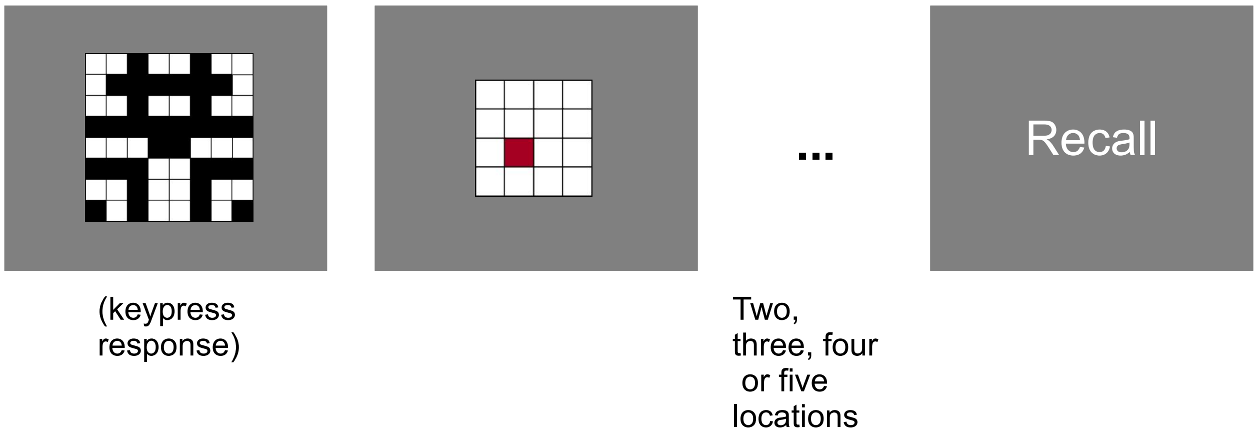


Figure 3. Depiction of the symmetry span task. Participants are presented with an 8 x 8 matrix on the computer screen. Participants respond as to whether the matrix is symmetrical across its vertical axis. Following each response, a 4 x 4 matrix is presented: participants are to remember to location of the filled square. After two, three, four, or five locations, participants are presented with a recall instruction, and fill in an answer sheet with the sequence of locations.

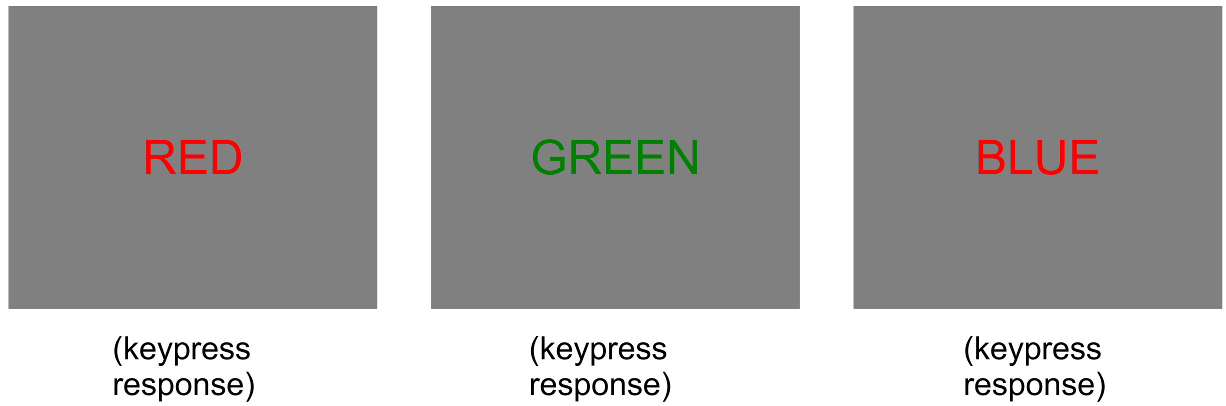


Figure 4. Depiction of the Stroop task. Participants are presented with a colour word on the computer screen. Participants are to ignore the identity of the word and respond to the colour in which the word is presented. The colour of the word and the identity of the word were congruent on 80% of trials.

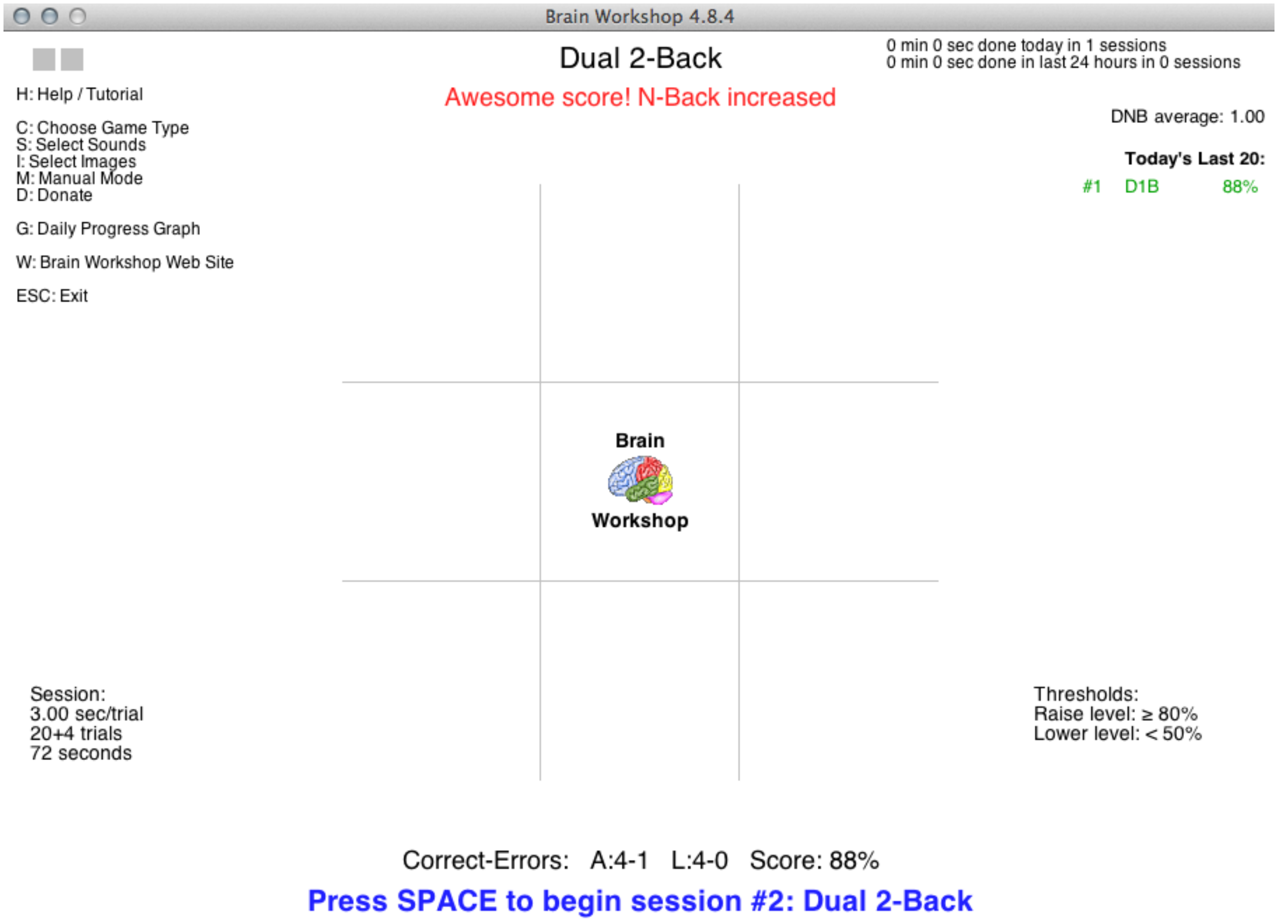
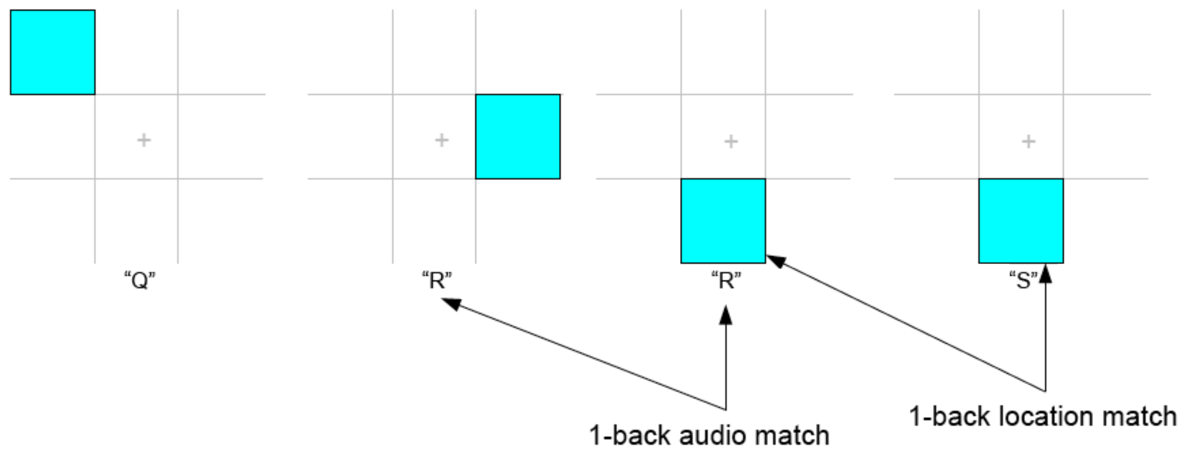


Figure 5. Screen capture of the main screen for the dual n-back task. The main screen provides detailed feedback on the previous block of trials (bottom of screen), summary feedback for up to 20 training blocks completed during the training session (right of screen), and the average n-level achieved during the current training session (right, top). The current n-level is displayed at the top and bottom centre of the main screen. DNB = dual n-back, D1B = dual 1-back, A = audio stimuli, L = letter stimuli.



(Press A key) *(Press L key)*

Figure 6. Depiction of the sequence of trials in the dual 1-back training task. In each trial, a letter is presented through the computer speakers and a blue square is presented in a 3 x 3 matrix on the computer screen simultaneously. In each participants are required to respond when the letter matches the letter that was presented 1 trial ago (1-back) and whenever the square is presented in the same location as it was 1 trial ago.

Figure 7. Screen capture of the main screen of the dual 1-back active control task. The main screen provides detailed feedback on the previous block of trials (bottom of screen), summary feedback for up to 20 training blocks completed during the training session (right of screen), and the average RT achieved during the current training session (right, top). A = audio stimuli, L = letter stimuli, RT = response time.

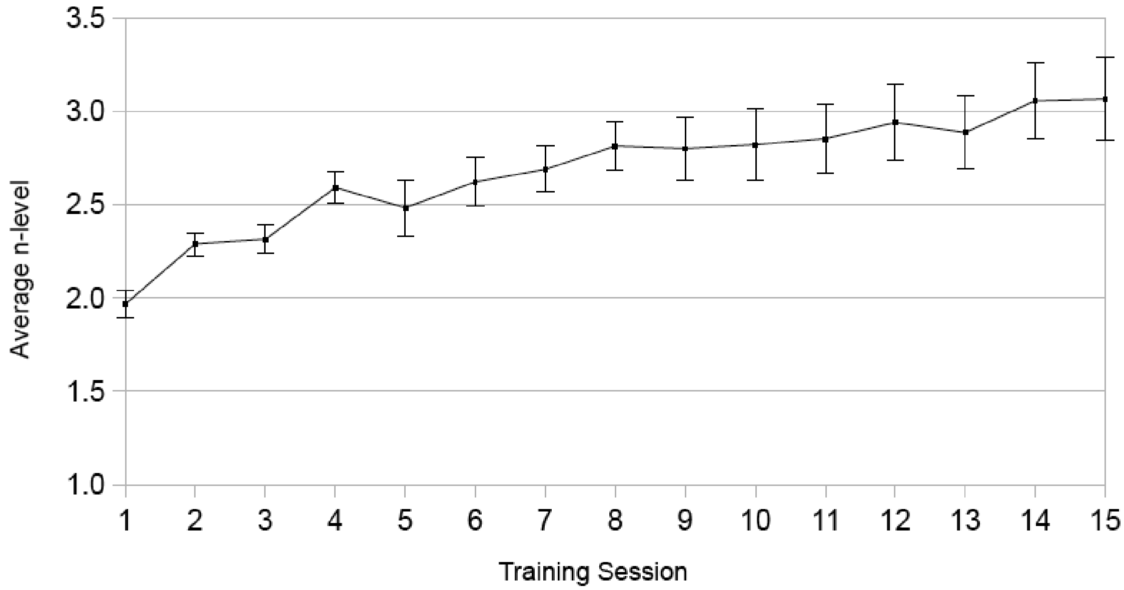


Figure 8. Performance on the training task. Average n-level reached for each training session, training group. Error bars represent +/- 1 standard error of the mean.

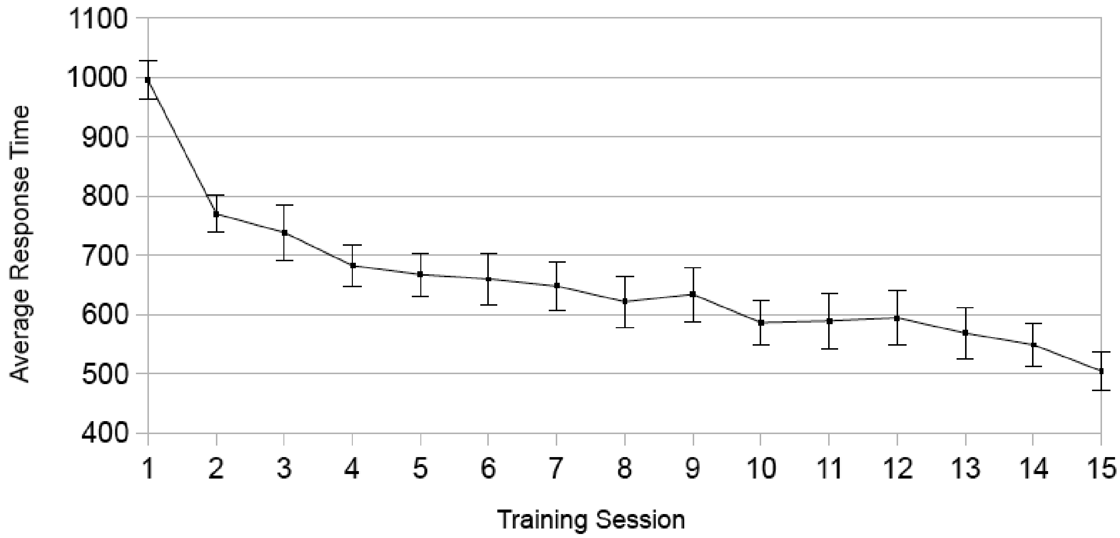


Figure 9. Performance on the active control task. Average response time (ms) for each training session, active control group. Error bars represent +/- 1 standard error of the mean.

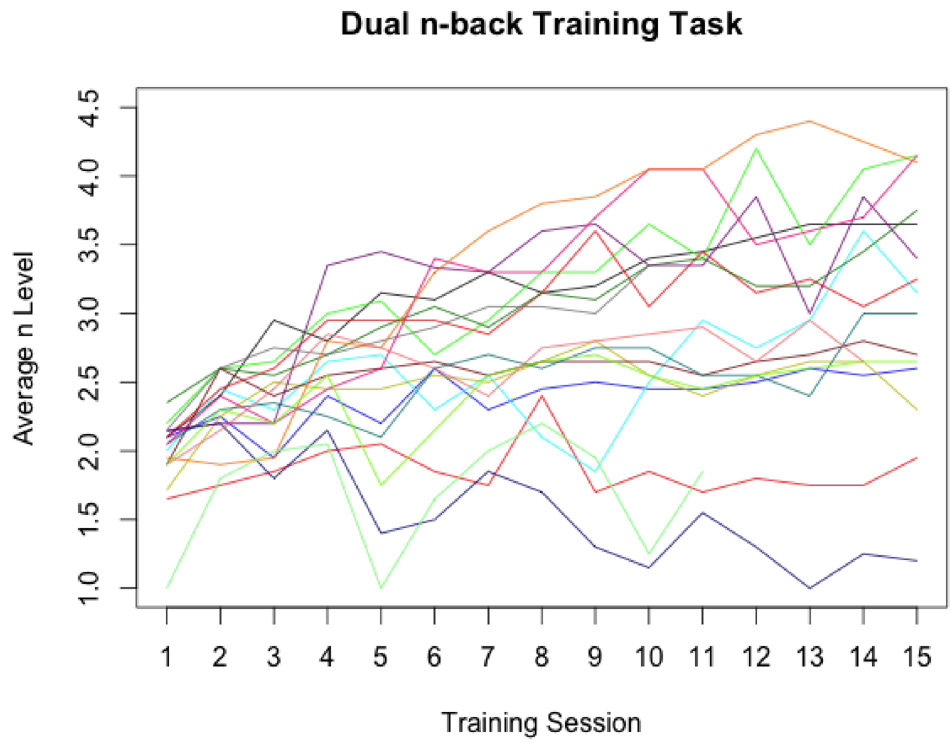


Figure 10. Individual performance on the training task. Average n-level reached for each training session for participants in the training group.

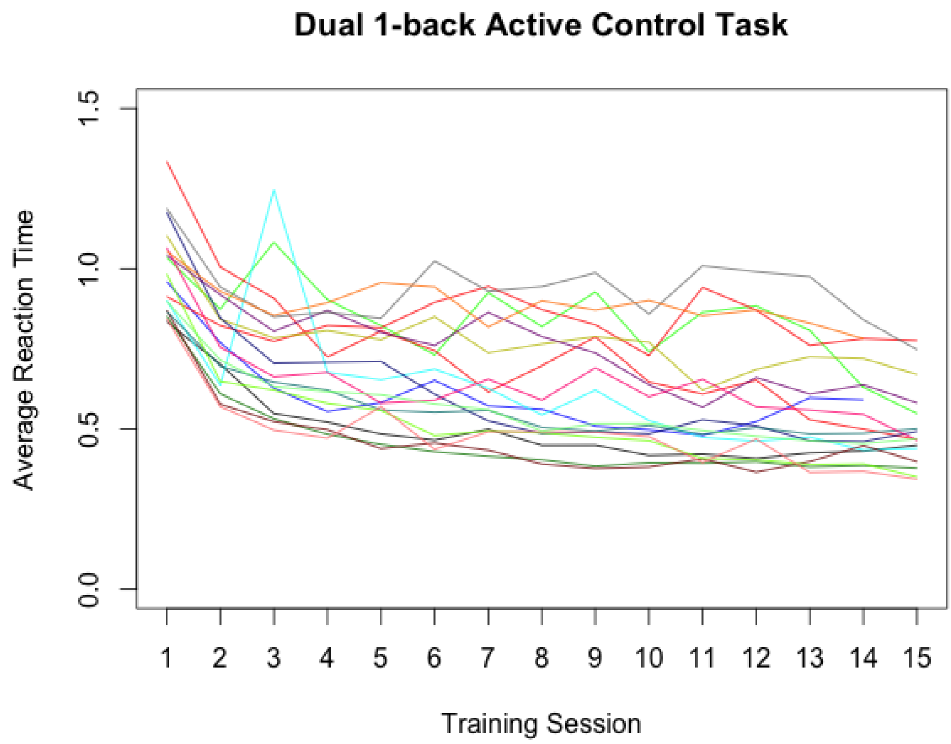


Figure 11. Individual performance on the active control task. Average response time (seconds) for each training session for participants in the active control group.

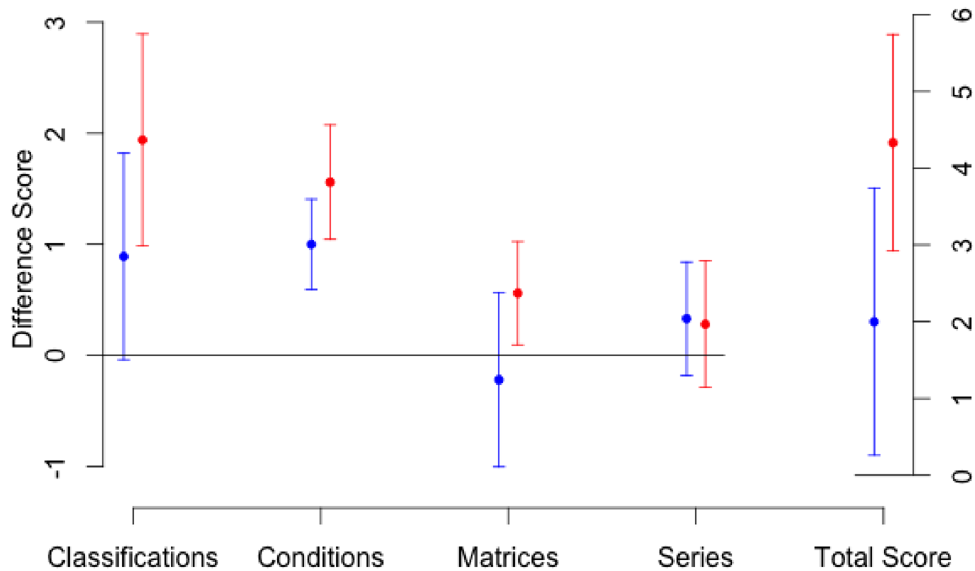


Figure 12. Point estimates for post-training minus pre-training differences for the four CFIT subtests (left axis) and the total CFIT score (right axis) for the training group (red) and the active control group (blue). Error bars represent within-groups 95% confidence intervals.

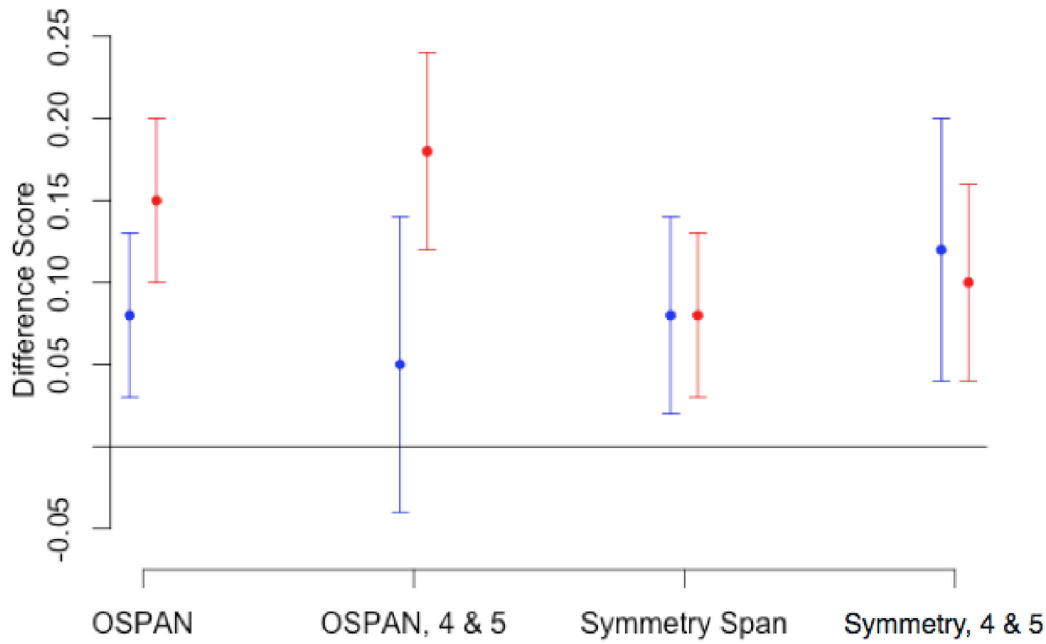


Figure 13. Point estimates (proportion score) for post-training minus pre-training differences for the working memory measures for the training group (red) and the active control group (blue). Error bars represent within-groups 95% confidence intervals. OSPAN = operation span. “4 & 5” = average of scores on set sizes of 4 and 5 for the task.

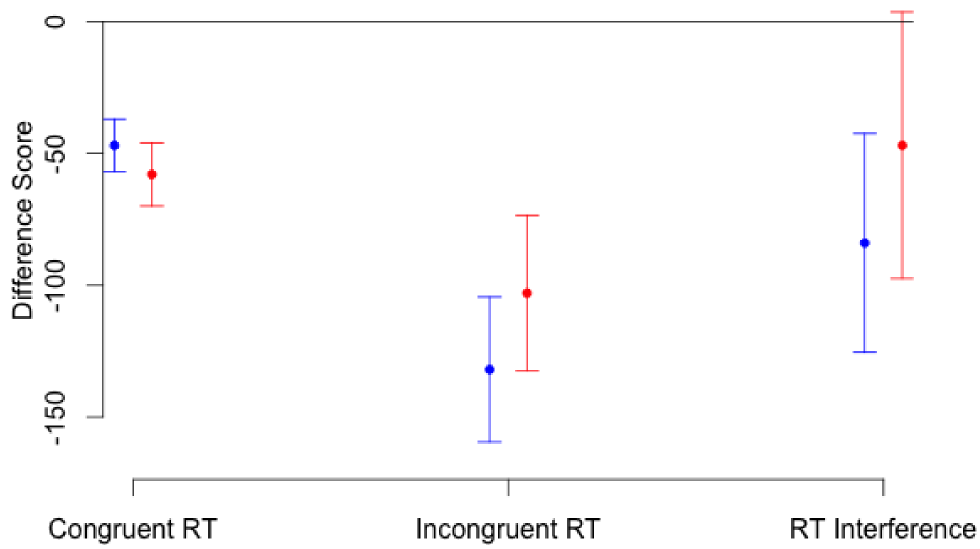


Figure 14. Point estimates (ms) for pre-training minus post-training differences for the Stroop RT measures for the training group (red) and the active control group (blue). Error bars represent within-groups 95% confidence intervals. RT = response time.

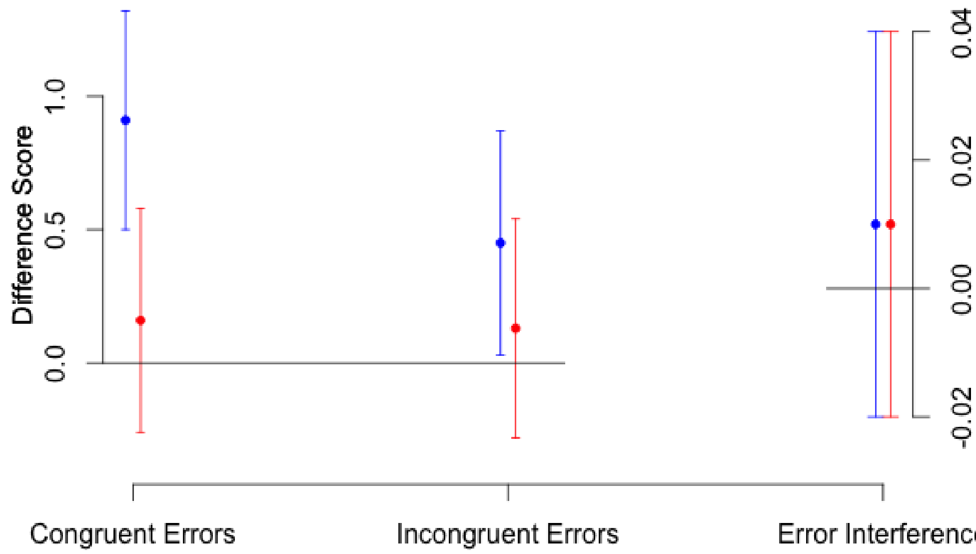


Figure 15. Point estimates for pre-training minus post-training differences for log-odds error for congruent and incongruent Stroop trials (left axis) and Stroop error interference (error rate; right axis) for the training group (red) and the active control group (blue). Error bars represent within-groups 95% confidence intervals.

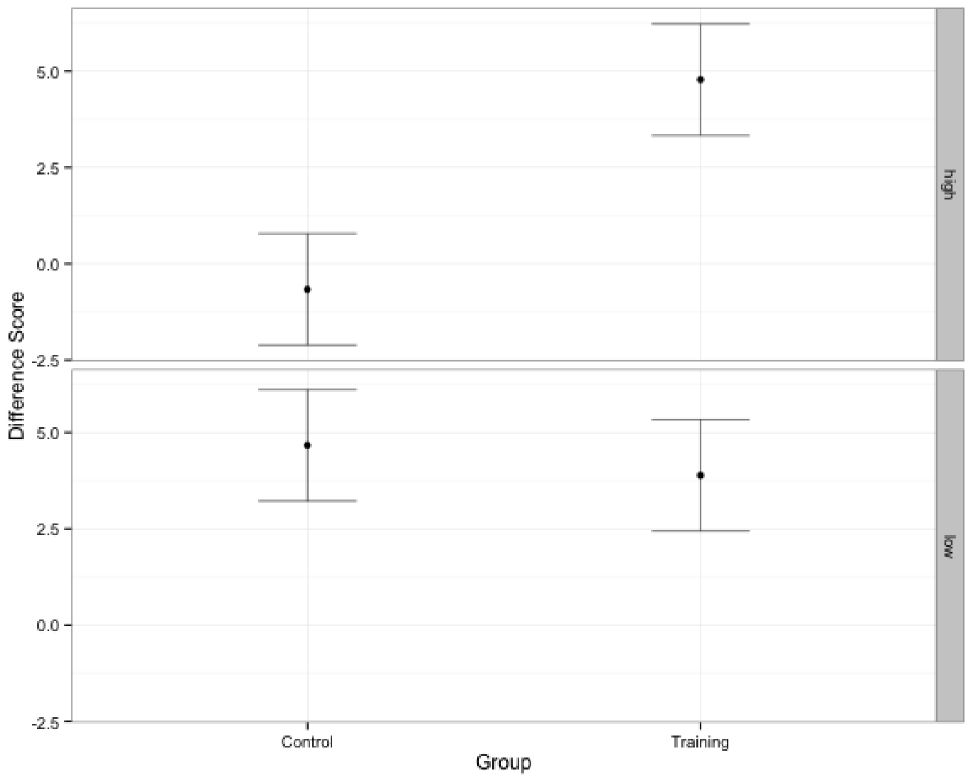


Figure 16. Difference scores (post-training minus pre-training) for the total CFIT score, with the active control and training groups (horizontal axis) divided into groups with above-median pre-training CFIT scores (high, upper portion) and below-median pre-training CFIT scores (low, lower portion). Error bars represent Fisher's Least Significant Difference.

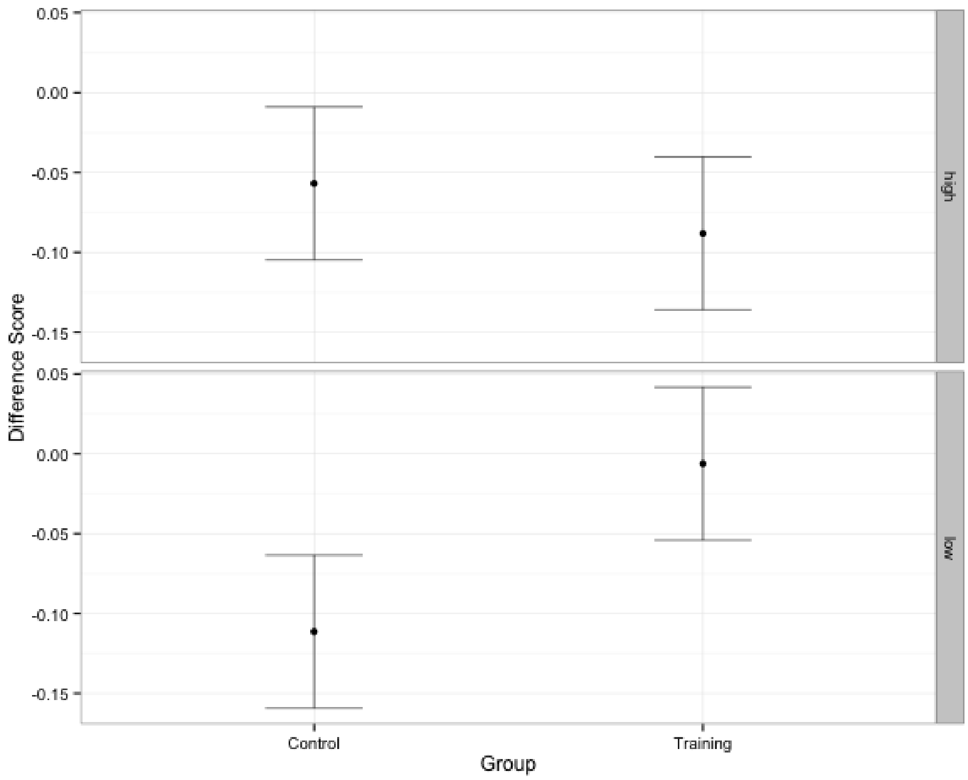


Figure 17. Difference scores (post-training minus pre-training) for Stroop response time interference (seconds), with the active control and training groups (horizontal axis) divided into groups with above-median pre-training CFIT scores (high, upper portion) and below-median pre-training CFIT scores (low, lower portion). Error bars represent Fisher's Least Significant Difference.

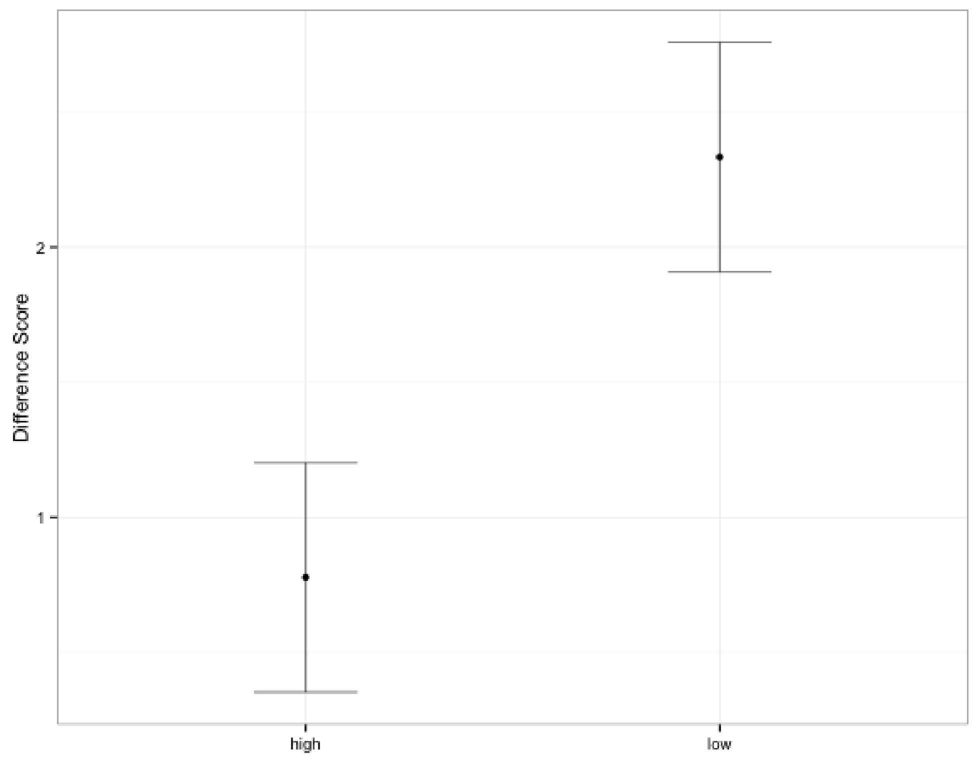


Figure 18. Difference scores (post-training minus pre-training) for total CFIT scores, with the training group divided into groups with above-median intercept on the dual n-back training task (high) and below-median intercept on the dual n-back training task (low). Error bars represent Fisher's Least Significant Difference.

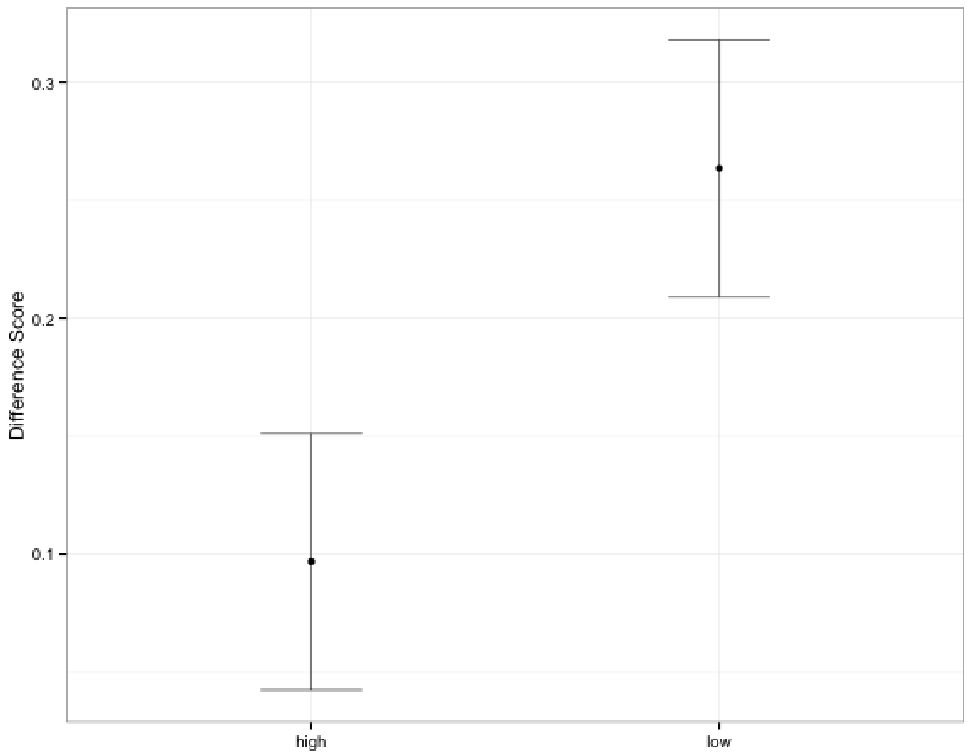


Figure 19. Difference scores (proportion correct, post-training minus pre-training) for the operation span, with the training group divided into groups with above-median intercept on the dual n-back training task (high) and below-median intercept on the dual n-back training task (low). Error bars represent Fisher's Least Significant Difference.