Psychological Research (2020) 84:1635–1654
https://doi.org/10.1007/s00426-019-01172-8

ORIGINAL ARTICLE

The role of proactive interference in working memory training and transfer

Thomas S. Redick[1] · Elizabeth A. Wiemers[1] · Randall W. Engle[2]

Received: 20 August 2018 / Accepted: 25 March 2019 / Published online: 5 April 2019

Abstract

Recent work on working memory training has produced conflicting results regarding the degree and generality of transfer to other cognitive processes. However, few studies have investigated possible mechanisms underlying transfer. The current study was designed to test the role of proactive interference in working memory training and transfer. Eighty-six young adults participated in a pretest–posttest design, with ten training sessions in between. In the two working memory training conditions, subjects performed an operation span task, with one condition requiring recall of letters on every trial (operation-letters), whereas the other condition alternated between letters, digits, and words as the to-be-remembered items across trials (operation-mix). These groups were compared to an active-control group (visual-search). Working memory, verbal fluency, and reading comprehension measures were administered in pretest and posttest sessions. All groups significantly increased their performance over the ten training sessions. There was evidence of strategy-specific benefits on transfer, such that transfer to working memory measures was higher for the operation-letters group on tasks specifically involving letters, and no differential transfer to working memory tests without letters, to verbal fluency, or to reading comprehension. The results indicate that proactive interference does not appear to play a causal role in determining transfer from working memory training, and instead a strategy account based on stimulus content provides a more parsimonious explanation for the pattern of training and transfer.

Introduction

Working memory (WM) has several definitions (Cowan, 2016), but many conceptualize WM as the temporary storage and manipulation of information to fulfill current information-processing demands. Given WM's central role in numerous aspects of cognition, and the findings that better scores on various WM measures are associated with superior reasoning, math, language comprehension, attention, and multitasking abilities, psychologists began testing whether increasing an individual's WM via an intervention would lead to concomitant improvements in these related outcomes. Early WM training studies had subjects repeatedly practice memory span (e.g., Klingberg, Forssberg, & Westerberg, 2002) and *n*-back (e.g., Jaeggi, Buschkuehl, Jonides, & Perrig, 2008) tasks for multiple sessions, with the goal of improving the trainees' WM abilities. These studies concluded that not only had the subjects' WM levels increased after training, but they also argued that the training had transferred to unpracticed matrix reasoning tests. Many subsequent WM training studies have been conducted using a similar design—assign subjects to either a training or a control group, have the subjects complete multiple sessions of WM tasks across different days, and assess whether the training and control groups have differential amounts of pretest-to-posttest change on a variety of cognitive, clinical, and academic outcomes (Melby-Lervåg, Redick, & Hulme, 2016). Often, researchers have made a distinction between near and far transfer (Barnett & Ceci, 2002). Near transfer includes gains on tasks that are similar in content or methodology to the training materials, whereas far transfer would be shown via improvements on outcomes that are dissimilar to the training procedures. Reviews and meta-analyses differ in their conclusion about the efficacy of WM training, with some arguing that WM training produces far transfer (e.g., Au et al., 2015), and some arguing it does not (e.g., Melby-Lervåg et al., 2016).

✉ Thomas S. Redick
tredick@purdue.edu

[1] Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907, USA

[2] Georgia Institute of Technology, Atlanta, USA

Many WM training studies use training programs mainly because they have been used in previous WM studies—*n*-back or memory span tasks are prevalent. Here, we sought to understand why WM training might be effective in producing particular patterns of near and far transfer. Before describing the current study, we briefly review the literature arguing for the importance of proactive interference (PI), or forgetting caused by prior learning, in linking WM and outcomes such as reasoning and reading comprehension.

PI and variation in WM

Individual differences in WM have been interpreted in various ways, with some researchers arguing that they reflect variation in amount of resources (Engle, Cantor, & Carullo, 1992), ability to control attention (Kane, Bleckley, Conway, & Engle, 2001), and ability to deal with PI (May, Hasher, & Kane, 1999). The latter view, that one's susceptibility to PI determines an individual's WM functioning, is critical for the current study.

Lustig, May, and Hasher (2001) tested subjects on versions of reading span, a complex span task in which subjects alternate between reading a sentence and encoding an item for subsequent serial recall (Daneman & Carpenter, 1980). One between-subjects PI manipulation was a typical version of reading span and a modified version that presented 90s of spatial ability filler tasks (e.g., hidden figures, card rotation) between the end of the recall period on one trial and the presentation of the subsequent list. The results showed that adding the filler breaks between trials acted as a means to release the subjects from PI from previous trials, and increased reading span scores relative to the condition without the breaks (replicating May et al., 1999). Critically, not only did the manipulation intended to reduce PI significantly increase reading span scores, they also produced reduced and nonsignficant correlations with a measure of subjects' ability to recall information from a separate reading task, relative to the reading span conditions in which PI was higher. Blalock and McCabe (2011) observed similar results with a between-trial filler task (viz., symbol comparison) on a nonverbal complex span task—when PI was reduced, the correlation with Raven progressive matrices was significantly reduced, compared to the version without the filler tasks.

Bunting (2006) administered multiple versions of operation span, which is similar to reading span but instead presents math operations as the processing task between each to-be-remembered item (Turner & Engle, 1989). The standard version presented math operations and words on every trial; two versions created to reduce PI either manipulated the to-be-remembered items within a trial (switching from words to digits halfway through the list) or after every third trial (after three consecutive trials of remembering lists of words, switching to three consecutive trials of remembering lists of digits, and so on). The conditions were inspired by Young and Supa (1941), who showed that such manipulations produced release from PI when the category of to-be-remembered stimuli changed. Bunting found that operation span scores were higher, and correlations with Raven lower, in the conditions with varying memoranda compared to the standard operation span version with repeated presentation of words as the to-be-remembered stimuli. Emery, Hale, and Myerson (2008) replicated Bunting's results with young adults (although not with older adults). To summarize, previous research examining manipulations of PI within complex span tasks, and the concomitant effect on the WM tasks' relationship to higher-order cognition, suggests that the presence of PI may be a critical factor for WM training interventions.

PI and WM training

A few WM training studies have examined the idea that improving one's ability to counteract PI is the critical change that would produce far transfer after WM training. Oelhafen et al. (2013) tested young adults using dual *n*-back training (Jaeggi et al., 2008), in which subjects decide if the current item matches the one presented *n* items back. In the dual version, subjects make decisions about the location of visually presented squares and the identity of aurally presented letters. The training task was adaptive, meaning that as subjects reached certain levels of accuracy on the visual and auditory tasks, the level of *n* would increase (or decrease). In the high-PI version, a large proportion of correct 'no' responses, in which the current item did not match the one presented *n* back, were lure trials—it had recently been presented but not in the *n*-back position. Thus, during a 3-back block of trials, lure trials could show stimuli that had been presented on trial $n-2$. In the low-PI version, no lure trials were presented. Previous research has shown that performance on *n*-back tasks with and without lures is quite different (Szmalec, Verbruggen, Vandierendonck, & Kemps, 2011), with the PI experienced on lure trials important for the relationship between *n*-back and reasoning (Burgess, Gray, Conway, & Braver, 2011; Gray, Chabris, & Braver, 2003). Oelhafen et al. predicted subjects completing 14 sessions of *n*-back training with lures would produce greater transfer relative to *n*-back training without lures and a passive-control group. They found that subjects in the lure-training group showed less *n*-back training task improvement than the no-lure-training group, but no differences between the training groups in transfer to a non-adaptive dual *n*-back task. In addition, even when compared to a passive-control group, neither training group showed transfer to reading span (WM) or the

Bochumer Matrizentest (reasoning). Thus, these results do not support the idea that the presence of PI is the critical aspect for transfer after *n*-back training.

Similarly, Hussey et al. (2017) compared three *n*-back training groups that finished 16 training sessions: adaptive *n*-back training with lures (high-PI), adaptive *n*-back training without lures (low-PI), and 3-back training without lures. Compared to the low-PI group, the high-PI group produced significant transfer to specific conditions of recognition memory and verb generation tasks, but no transfer to a color-Stroop task. Most germane to the current work, the training groups did not differ in transfer to comprehension accuracy of sentences, although there were significant differences in favor of the high-PI group in the online processing of the sentences (eyetracking and/or reading time differences).

Two studies (Loosli et al., 2016; Persson & Reuter-Lorenz, 2011) administered WM training via a non-adaptive *n*-back task and a recent-probes task.[1] PI within the *n*-back was manipulated as described above; PI in the recent-probes task was also manipulated, which is described elsewhere (Jonides & Nee, 2006). Despite differences in the training materials (3-back with words vs. 2-back with images) and samples (young vs. older adults), both studies found the same pattern of results. When examining the high-PI training groups, there was no evidence for transfer to the trained memory tasks, untrained memory tasks (verb generation, paired-associates), or far transfer tasks (Stroop, test of nonverbal intelligence, digit-symbol substitution test), relative to either an active-control group (Persson & Reuter-Lorenz, 2011) or the low-PI training group (Loosli et al., 2016).

Finally, Bomyea, Stein, and Lang (2015) is particularly relevant for the current study. Based on Bunting (2006), they randomly assigned women with post-traumatic stress disorder to training groups that completed eight sessions of a non-adaptive reading span task. In the high-PI condition ($n = 13$), all of the to-be-remembered items were words; in the low-PI condition ($n = 8$), the to-be-remembered items alternated between words and digits every three trials (as in Bunting, 2006). Bomyea et al. (2015) focused on clinically relevant outcomes, and operation span was the only cognitive assessment included in pretest/posttest sessions. They reported significantly greater operation span posttest scores for the high-PI group compared to the low-PI group, along with other reductions in various clinical symptoms. Although promising, some caution should be noted when interpreting Bomyea et al. (2015)—the samples were small, only women with post-traumatic stress disorder were recruited, there was no control group that did not train on WM, the training was non-adaptive, and there was no assessment of cognition other than a verbal complex span task that was very similar to training.

Current study

The current work was based on the research discussed above, while also adhering to 'best practices' advocated in the WM training literature (Redick et al., 2015). In all five training studies just reviewed, there were a total of 14 training and control groups—with the exception of Hussey et al. (2017), the sample sizes for each group in the other four studies were $n \leq 16$. Underpowered studies are common in the WM training literature (Bogg & Lasecki, 2015; Melby-Lervåg et al., 2016). Therefore, we enrolled more subjects in each of our three groups, with a comparable final sample size to Hussey et al. (2017). Our training sessions adapted in difficulty based on the subject's performance. Also, we used an active-control condition from previous studies (Foster et al., 2017; Harrison et al., 2013; Redick et al., 2013), in an effort to distinguish any possible training effects from improvements due to placebo and/or expectancy effects. The visual-search control condition was motivated by extensive work showing individual differences in WM, as measured by complex span tasks, are unrelated to visual-search performance (Kane, Poole, Tuholski, & Engle, 2006). In addition, by having a separate control group, we can contrast the training effects for both the high- and low-PI conditions (in contrast to Bomyea et al., 2015; Hussey et al., 2017; Loosli et al., 2016).

Finally, we administered numerous verbal and nonverbal measures of WM, and two separate far transfer domains that might be susceptible to training PI—verbal fluency and reading comprehension. Recent research suggests that PI manipulations affect performance on change detection WM tasks, and their corresponding relationships with complex span tasks (Shipstead & Engle, 2013). In addition, as seen in the previously discussed WM training studies, the PI present on lure trials within *n*-back tasks is critical to the relationship between *n*-back performance and higher-order cognition (Burgess et al., 2011; Gray et al., 2003). Verbal fluency and reading comprehension are frequently used as outcome measures in WM training (Schwaighofer, Fischer, & Buhner, 2015), and previous work suggests that one's ability to counteract PI may underlie the relationship between WM and both verbal fluency (Rosen & Engle, 1997) and reading comprehension (Borella, Carretti, & Pelegrina, 2010; see also Hussey et al., 2017).

We created training versions of operation span that we predicted would have varying amounts of PI. Based on Bomyea et al. (2015) and Hussey et al. (2017), and the idea

[1] Persson and Reuter-Lorenz (2008) was retracted due to undetected programming issues in the training groups (see Persson & Reuter-Lorenz, 2011). When correcting the programming problem, Persson and Reuter-Lorenz (2011) were unable to replicate the positive transfer results published in their 2008 article.

that increased resistance to PI is fundamental to far transfer, one prediction is that the high-PI training condition will produce more transfer, in particular on fluency and reading comprehension. However, we also considered other possible patterns of transfer results that would be consistent with alternative accounts. It is possible that because subjects in the high-PI condition practice repeatedly with the same kind of to-be-remembered stimuli (letters), they will be more likely than the low-PI condition to develop task- and stimulus-specific mnemonic strategies that are beneficial for transfer, as long as the outcome measure affords the use of the strategy (Dunning & Holmes, 2014; Sprenger et al., 2013). Thus, a strategy account would predict that what we are labeling as the high-PI condition would be expected to produce transfer to tasks that are similar methodologically and/or use the same type of mnemonic materials (viz., letters), but not produce far transfer to tasks that are dissimilar (change detection, fluency, reading comprehension). This pattern of results would be consistent with Harrison et al. (2013), who observed complex span training benefits on other memory span tasks but not to change detection, fluid intelligence, or multitasking transfer tasks. Finally, because subjects in the low-PI condition train with different kinds of to-be-remembered stimuli, they may be less likely to develop specific strategies and instead engage in what has been labeled 'core' (Morrison & Chein, 2011) or 'process-based' (Jolles & Crone, 2012) WM training. The idea is that exposure to variability in tasks and memoranda in WM training is one critical factor in producing broad transfer. This training-variability account would predict that the low-PI condition would actually produce the most transfer, not because the PI is reduced, but because the trainee is exposed to a variety of training materials. In particular, the training-variability account predicts that the low-PI group is most likely to lead to improvements on structurally dissimilar outcomes (far transfer); notably, this prediction is counter to the transfer results seen in Bomyea et al. (2015) and Hussey et al. (2017).

Method

Subjects

Subjects were recruited from Purdue University and the surrounding community. Inclusion criteria included: (a) age 18–30 years old; (b) normal or corrected-to-normal vision; (c) native-English speaker; and (d) availability to come to the laboratory for 12 separate sessions within a 3-week period. Other than attempting to reach $n = 30$ subjects per group, our criterion for stopping data collection was to recruit as many subjects as possible given the amount of funding budgeted for subject compensation and the time constraints of the summer and fall semester academic calendar (Simmons, Nelson, & Simonsohn, 2011). Ninety-two subjects completed at least 1 session, but 6 subjects did not complete all 12 sessions. Two subjects quit after the pretest session before beginning training, 2 additional subjects quit after completing the first training session, another subject quit after the eighth training session (a family emergency required international travel), and a final subject quit after the ninth training session (no reason provided). Therefore, complete training and transfer data were available for 86 subjects, and the demographic information for each group is shown in Table 1.[2]

Table 1 Demographic information for each group

Group	Age	Gender (M/F)	Days to complete study (min/max)
Operation-letters	20.53 (1.17)	13/17	18.47 (15/23)
Operation-mix	20.41 (2.41)	12/15	19.33 (17/22)
Visual-search	20.59 (2.43)	11/18	18.62 (15/22)

Days to complete study calculated as number of days between pretest and posttest session (weekends included although subjects did not have option of participating on weekends)

Transfer tasks

Pretest and posttest tasks were administered in the order listed below. Pretest sessions additionally administered demographics (age, gender, handedness, native-English) before the cognitive battery, and posttest sessions additionally administered a survey about enjoyment, strategies, and expectations after the cognitive battery. Dependent variables (DVs) from all cognitive tasks are scored such that higher values reflect better performance (except fluency repetitions). 3-back, fluency, and reading comprehension tasks used alternate versions at pretest and posttest, counterbalanced across subjects.

Running span letter

Letters were presented one at a time onscreen for 300 ms (200 ms ISI), and subjects were instructed before the presentation of each list that they were to recall in order the last n letters upon completion of the list ($n = 3$–9). There were two

[2] For one subject in the operation-letters group, there was a power outage halfway through the subject's 8th training session. For a second subject in the operation-letters group, there was an internet network outage during the subject's 3rd training session. In order to retain these subjects' data for the training analyses, we used the highest level the subjects had obtained during the session to that point. Therefore, it would be more accurate to say these subjects completed 9.5 and 9.75 training sessions, respectively.

trials at each n, with each set of two trials presenting $n+1$, or $n+2$ letters. At recall, subjects were presented with the possible 12 letters, and clicked the order in which the last n items were presented. We used a partial-credit scoring method, with any letters recalled in the correct serial position given 1 point.

Running span matrix

This task was identical to running span letter, with the exception that the to-be-remembered stimuli were locations of red squares within a 4 × 4 grid.

Change detection color

This task presented 4, 6, or 8 colored squares simultaneously for 250 ms, followed by a blank screen for 900 ms, and then a display of all of the squares in their same onscreen location for an unlimited duration. At test, one square was circled, and the subject was to respond 'yes'/'no' with a left-/right-mouse click if the circled square was the same/different color as in the initial display. Other than the circled square on non-match trials, all squares at test are presented in the same color as the initial display. Subjects completed 1 practice block of 6 trials with feedback before 2 blocks of 30 trials without feedback. The DV was k (Cowan et al., 2005).

Change detection orientation

This task is similar in structure to change detection color, with the following exceptions. In this task, subjects are presented with either 5 blue and 7 red, or 7 blue and 5 red, rectangles simultaneously. A cue presented 300 ms before the initial rectangle display instructed subjects to remember the location and orientation of either the blue or red rectangles. At test, one of the blue or red rectangles was cued, and subjects were to indicate whether or not its orientation was the same as its initial presentation. Again, k was used as the DV.

3-back letter

Letters were presented onscreen for 2500 ms, with an 800 ms ITI. Subjects responded to each letter using a left click if the letter was the same as the letter presented 3 letters back, or a right click if not. Stimulus lists were generated such that 25% of trials were targets (left) and 75% of trials were nontargets (right). In addition, given previous research showing the critical role of lures on n-back tasks in relation to PI (Gray et al., 2003), one-third of the nontargets were lures (80% were $n-1$ lures, 10% were $n-2$ lures, 10% were $n+1$ lures). To ensure that the percentage of lures was equivalent across subjects, two lists were created using the 20 letters in the stimulus pool. A 20-trial practice block (with feedback) proceeded the 40-trial experimental block (without feedback). The main DV was d', which was calculated based on hit rates to targets minus false alarms to nontargets using the loglinear correction for hit and false alarm rates (Stanislaw & Todorov, 1999). In addition, we conducted supplementary analyses on lure accuracy to focus on the effects of training upon PI resolution.

3-back matrix

The task was identical to 3-back letter, with the exception that the to-be-remembered stimuli were locations of red squares within a 4 × 4 grid (same as the running span matrix task). Thus, the stimulus pool was composed of 16 unique locations. Again, d' and lure accuracy were used as DVs.

Letter fluency

Subjects were instructed to type as many words as possible in 2 min that started with either the letter 'F' or 'S'. The main DV was the number of unique words generated using the criteria provided in Unsworth, Spillers, and Brewer (2011). In addition, we conducted supplementary analyses on word repetitions to focus on the effects of training upon PI resolution.

Category fluency

The task was identical to letter fluency, with the exception that subjects generated exemplars of specific categories ('animals' or 'supermarket'). The DVs were the same as letter fluency.

Reading comprehension (Nelson-Denny, 1993)

Reading comprehension was assessed with versions G/H of the Nelson-Denny test (counterbalanced across subjects as pretest and posttest), modified for the current study. In the paper-and-pencil version, each page presents 1 of the 6 reading passages and its associated 5 multiple-choice comprehension questions. In our computerized version, the passage also remained onscreen throughout the presentation of all 5 question-and-answer options, but only 1 question with A-E answer options was shown at a time (hereafter called problems). When the subject clicked to choose an answer, the next problem was presented. After 5 problems, a new passage and the first problem associated with that passage was shown, which was repeated until all 6 passages and 30 problems were shown. Following the instructions provided in the manual, subjects were provided 20 min to work on the test. The DV was the number of problems correctly answered. We did not convert the raw score to the scaled score (Chein & Morrison, 2010) because we did not administer the 1 passage in each version that had 8 associated problems. The reason for excluding this passage is that the 8-problem passage is longer and, thus, would have taken up the entire screen, leaving insufficient space to present each of the problems.

Survey

Upon completion of the posttest cognitive tasks, subjects were asked a variety of questions about their subjective experience during the study. Subjects were asked 'Did you find the 10 shorter practice sessions enjoyable?' and to rate their response (from 1 = not at all to 4 = very much so). Subjects were asked to 'Rate your level of effort during the 10 shorter practice sessions' (from 1 = low to 4 = high). Similar to Redick et al. (2013), we asked subjects to provide open-ended answers about strategies they used. Specifically, the strategy questions were "If a friend were to come into the lab to begin the study next week and take the same 10 shorter, practice sessions, what tips/tricks could you tell your friend that would increase their performance?" followed by "What strategy or strategies did you use to try to reach a higher level during the 10 shorter, practice sessions?". The strategy questions were followed with six questions asking subjects 'Do you feel that your participation in this experiment has changed your _______ ?' and to respond 'yes' or 'no'. Each question asked separately about intelligence, memory, attention, visual perception, language, and the way subjects carried out daily activities, similar to Redick et al. (2013).

Training tasks

All training tasks were computerized, with stimuli presented visually and subjects using either the mouse (operation-letters, operation-mix) or keyboard (visual-search) to make responses.

Adaptive operation-letters (high-PI)

The adaptive operation-letters training task (Fig. 1a) was very similar to the operation span training program in Harrison et al. (2013). Subjects were presented with math operations to solve, such as (9/3) − 1 = , for a maximum of 4000 ms, and then presented with a digit between 0 and 9, such as 4, for a maximum of 2000 ms, and asked to answer whether or not the provided answer was the correct solution or not. After the subject clicked the TRUE or FALSE box, a to-be-remembered letter was presented for 1000 ms. Each trial contained a variable number of operation-letter item presentations, determined by the subject's current level (see below). After the appropriate number of items occurred for that trial, the recall screen was presented and subjects were instructed to click on the letters they saw in serial order. There were 15 letters in the stimulus pool, and letters never repeated within a trial. Math operations were scored as correct if the subject selected the correct TRUE/FALSE response. Item recall was scored using a partial-credit method (Redick et al., 2012), such that letters recalled in the correct serial position in each trial were summed to create a total score for the trial.

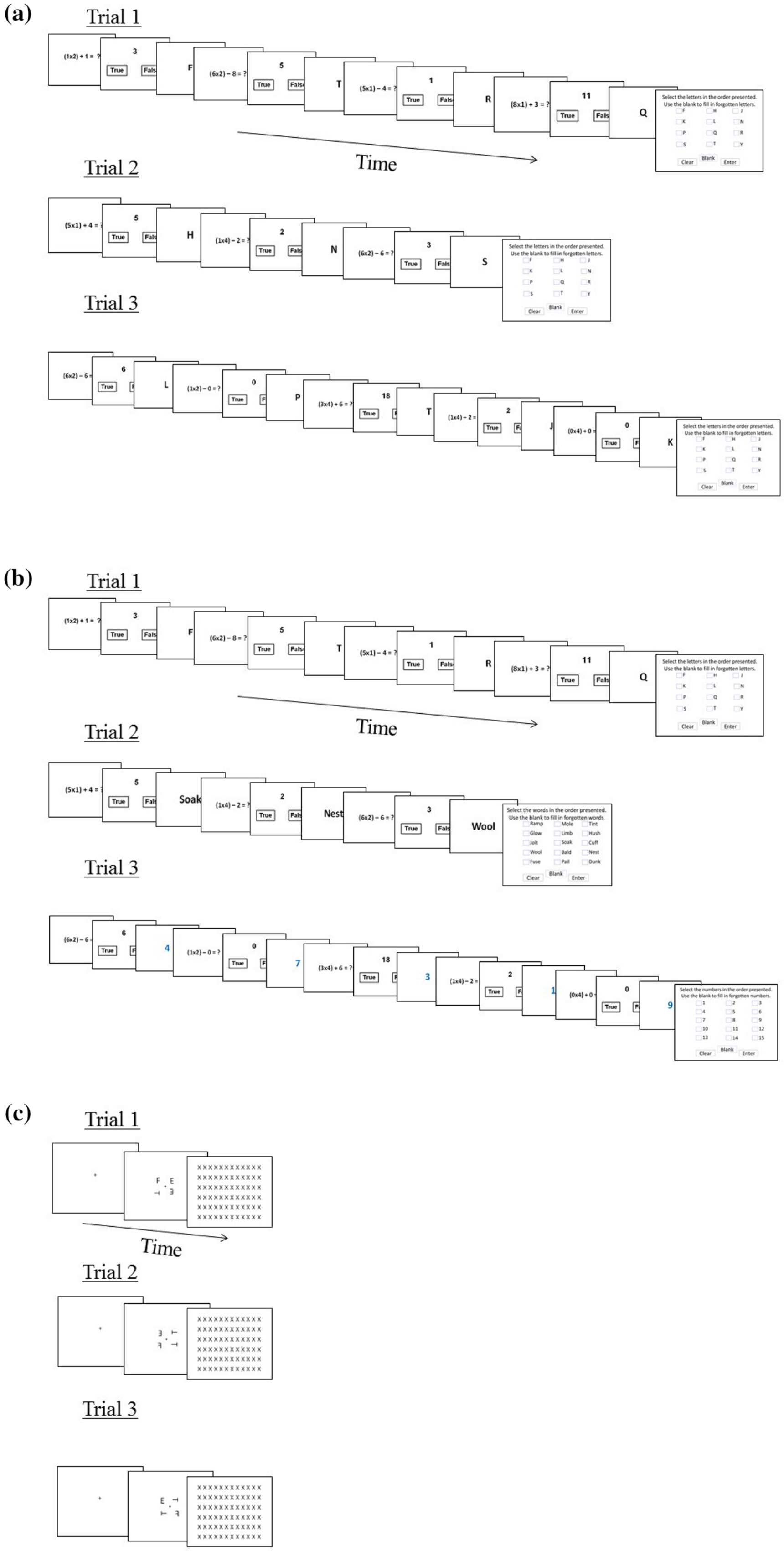
There were 8 blocks of trials in each training session, with 3 trials in each block. The number of operation-letter item presentations during a given block was determined according to the subject's current level (x). Randomly ordered within each block, one trial presented $x + 1$ items, another trial presented $x + 2$ items, and the remaining trial presented $x + 3$ items. For example, at level 1, subjects saw a 2-item, 3-item, and 4-item trial, in random order. At level 12, subjects saw a 13-item, 14-item, and 15-item trial, in random order.

All subjects began the first training session at level 1, but thereafter the subject's current level on a block adjusted based on their performance on both the math operations and letter recall from the previous block of trials. This included the first block of trials of sessions 2 through 10, in which the levels were determined by performance on the final block of the previous session. The current level was determined according to the following rules. If on the current block the subject correctly solved greater than 87.5% of the math operations and recalled more than 87.5% of the letters, the level on the next block increased by 1. If on the current block the subject correctly solved less than 75% of the math operations or recalled fewer than 75% of the letters, the level on the next block decreased by 1. Any other combination of math and letter performance resulted in the subsequent block having the same level as the one just completed. Feedback was provided after each trial and block, and the current level was indicated at the beginning of the block. In the first training session, before beginning the real training task, subjects practiced the letter recall task by itself, then the math operation task by itself, and then a block of level 0 of the training task.

Adaptive operation-mix (low-PI)

The adaptive operation-mix task was very similar to the operation-letters task (Fig. 1b). The only difference was that the category of to-be-remembered items alternated in a fixed order among letters, words, and digits, on each block. The first trial of every block presented letters for recall, the second trial of every block presented words, and the third trial of every block presented digits. Words were randomly sampled from a pool of 15, four-letter, monosyllabic words, each of which started with a unique letter. Digits were randomly sampled from 1 to 15, and were presented in blue to distinguish them from the math operation stimuli. The

Fig. 1 Example of three trials at level 2 for operation-letters (**a**), operation-mix (**b**), and visual-search (**c**)



subject's level adjusted each block based on the same formulas as the adaptive operation-letters condition, but recall averaged across all 3 categories of to-be-remembered items determined the level on the next block of trials.

Adaptive visual-search (active-control)

The adaptive visual-search task (Fig. 1c) was very similar to the program used in Harrison et al. (2013). Subjects were instructed to locate a target (normal F or mirror-reversed F) among heterogeneous distractors (normal E, mirror-reversed E, left-rotated T, or right-rotated T) on each trial. Subjects had to indicate the correct response using a compatibly mapped left- or right-index finger keyboard button press using labeled keys (*z* and, keys, respectively). During each training session, subjects completed 8 blocks of trials, and each block contained 24 trials, equally divided between the two target types. On each trial, a central fixation cross appeared for 500 ms, followed by the search display for 500 ms containing a variable number of stimuli depending on the current level (see below), and a 2500 ms perceptual mask (Xs in the location of each possible stimulus). Subjects could make their response during the mask presentation, and accuracy was emphasized over speed.

The current level determined the number of distractors presented on each trial. On level 1, 1 target was presented in a single row with 1 distractor (2 columns × 1 row). On level 2, 1 target was presented with 3 distractors in a 2 column × 2 row configuration. On level 3, 1 target was presented with 5 distractors in a 3 column × 2 row configuration. Subsequent higher levels increased the search display accordingly, adding a new row on even levels and a new column on odd levels.

The starting level for each block and session followed the same rules as the training tasks. The current level was determined according to the following rules. If on the current block the subject accurately responded on 87.5% or greater of the trials, the level on the next block increased by 1. If on the current block the subject accurately responded on 75% or lower of the trials, the level on the next block decreased by 1. Any other accuracy level resulted in the subsequent block having the same level as the one just completed. Feedback was provided after each block, and the current level was indicated at the beginning of the block. In the first training session, before beginning the real training task, subjects practiced a response-mapping block with lone targets.

Procedure

Subjects were assigned to one of the three groups at pretest. Unlike Redick et al. (2013), no attempt was made to match groups on pretest scores. Recruitment material and consent forms referred to the study as involving practice and learning, in contrast to other studies that have explicitly mentioned 'brain training' or similar terms in the study materials (Foroughi et al., 2016). Pretest and posttest sessions occurred in a room where up to three subjects could be tested simultaneously, and were ~ 75 min in duration. Subjects completed 10 training sessions individually in laboratory rooms, and each session was ~ 30 min in duration. Subjects could not complete more than 1 of the 12 sessions per day, and sessions were conducted on weekdays. Table 1 displays the average number of days between the pretest and posttest, which did not differ among groups, $F(2, 83) = 1.331$, $p = 0.270$, $\eta_p^2 = 0.031$.

Subjects were paid $20, $25, and $10 USD for completing the pretest, posttest, and each of the 10 training sessions, respectively. Additional bonus compensation per training session was calculated according to the following criteria: $0 if they ended on Level 1, Level 2 or 3; $1 on Levels 4 and 5; $2 on Levels 6 and 7; $3 on Levels 8 and 9; $4 on Levels 10 and 11; and $5 on Level 12. The bonuses were used to encourage subjects to constantly attempt to maximize their training.

Analyses

Training data for the operation-letters and operation-mix groups were analyzed using a factorial ANOVA, with group (2) as a between-subjects variable, and training session (10) as a within-subjects variable. Because the 'level' used as the DV from the visual-search group was markedly different from the DV in the WM training groups, the visual-search group's training performance was examined separately using a repeated-measures ANOVA with training session as the within-subjects variable. In addition, a repeated-measures ANOVA with memoranda type (3) and training session (10) was conducted on the training data for the operation-mix group only.

Transfer data were analyzed using two different methods. First, factorial ANOVAs were computed with group (3) as a between-subjects variable, and transfer session (2) as a within-subjects variable. Second, ANCOVA was used with group (3) as a between-subjects variable, posttest score as the DV, and pretest score as the covariate. In terms of providing evidence in favor of an effect of an intervention providing transfer, the critical test in the ANOVA model is the group × session interaction, and the group effect in the ANCOVA model. The debate about the appropriate use of ANOVA versus ANCOVA to assess differential change has a long history in psychology (Huck & McLean, 1975; Lord, 1967; Miller & Chapman, 2001; Wright, 2006). Although the ANOVA results are discussed and interpreted in the text, the ANOVA and ANCOVA results are both provided. Given there are two training groups, separate analyses comparing the operation-letters training and control groups, and the

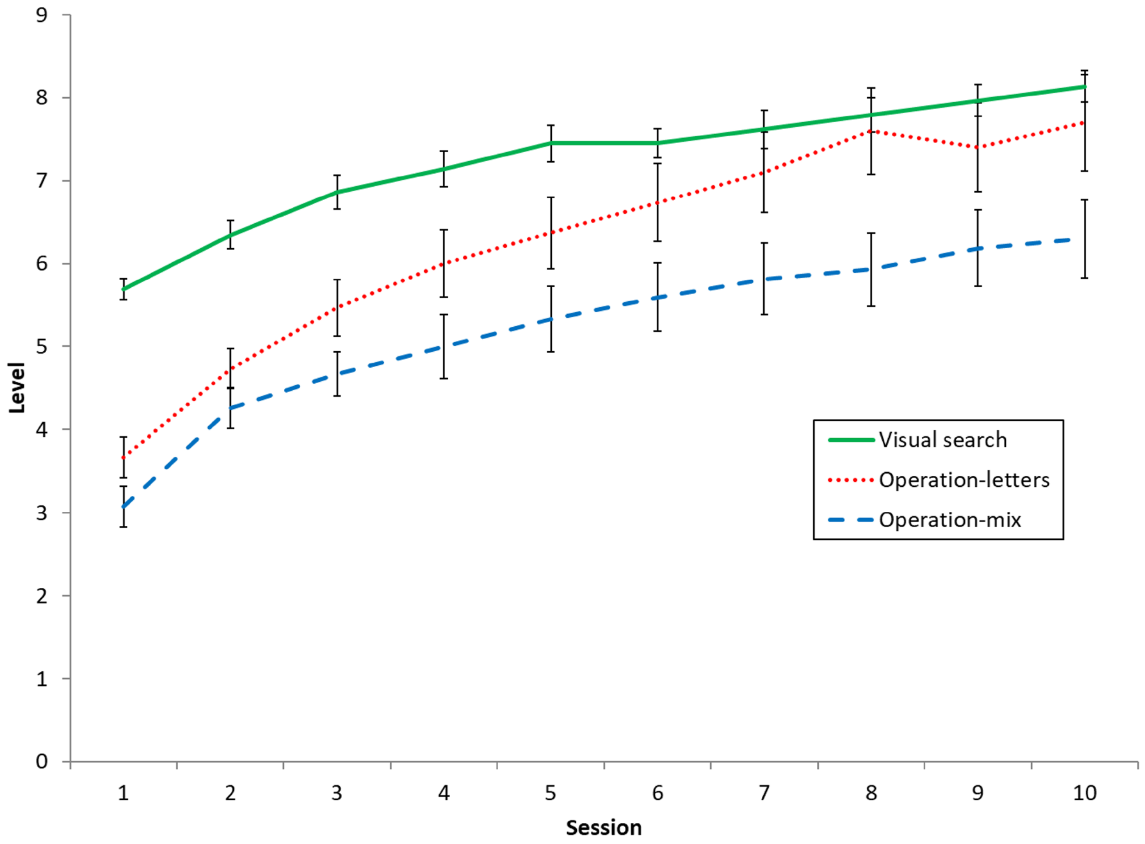


Fig. 2 Mean highest level obtained per training session as a function of group. Error bars represent ± 1 standard error of the mean

operation-mix training and control groups, were also conducted. In addition, for each group, paired-samples *t* tests were used to determine whether pretest-to-posttest changes were significantly different from 0. Partial-eta squared (η_p^2) and Hedge's g_{av} (Lakens, 2013) are provided as indices of effect size. Proportion data from the surveys regarding self-reported strategy use and perceived improvements were analyzed using Chi-square tests. All analyses were two-tailed and conducted using $\alpha = 0.05$.

Results

Training data

Figure 2 shows the average level achieved during each of the 10 training sessions for each of the three groups. As can be seen, all three groups improved over the course of the training sessions. In addition, although there was substantial variation within each group in terms of the amount of training improvement over the course of the 10 sessions, the majority of subjects achieved a higher level during session 10 versus session 1 (operation-letters: 27 of 30 subjects; operation-mix: 25 of 27 subjects; visual-search: 29 of 29 subjects). Finally, the operation-letters group appeared to outperform the operation-mix group across the 10 sessions, in contrast to the hypothesis that the mixed memoranda of the operation-mix group would lead to less PI and better performance than the operation-letters group.

These impressions were confirmed with the ANOVA results. First, for each group, the main effect of session was significant, operation-letters: $F(9, 261) = 32.353$, $p < 0.001$, $\eta_p^2 = 0.527$; operation-mix: $F(9, 234) = 27.062$, $p < 0.001$, $\eta_p^2 = 0.510$; visual-search: $F(9, 252) = 38.635$, $p < 0.001$, $\eta_p^2 = 0.580$. When comparing the operation-letters and operation-mix training performance, there was a significant main effect of group, $F(1, 55) = 4.261$, $p = 0.044$, $\eta_p^2 = 0.072$, and a significant main effect of session, $F(9, 495) = 57.447$, $p < 0.001$, $\eta_p^2 = 0.511$, although the group × session interaction was not significant, $F(9, 495) = 1.434$, $p = 0.170$, $\eta_p^2 = 0.025$.

An additional analysis focused on the training data from the operation-mix group, investigating recall performance separately for each of the to-be-remembered item categories (letters, words, digits).[3] Figure 3 displays the mean number

[3] The operation-mix analysis does not include one subject who completed the entire first training session but the file did not save cor-

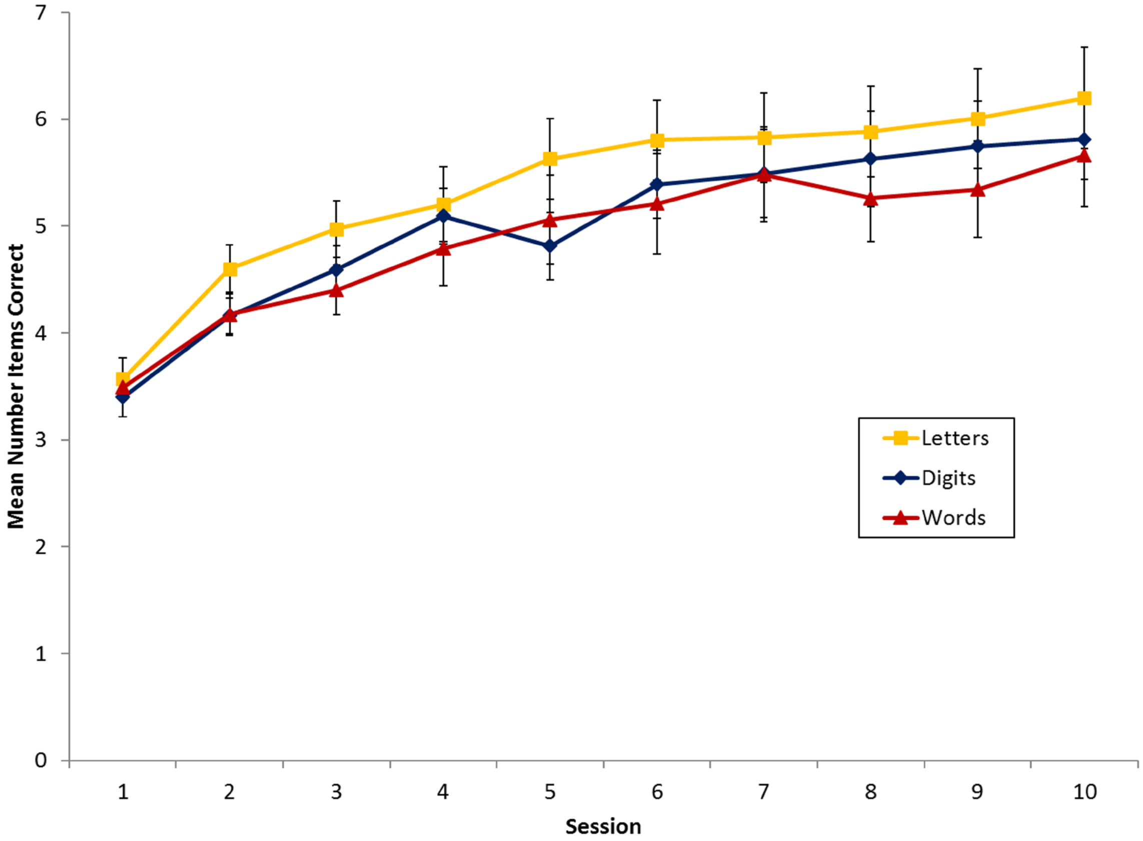


Fig. 3 Mean items per trial recalled correctly from individual memoranda categories for operation-mix group. Error bars represent ± 1 standard error of the mean

of items recalled correctly on each of the 8 trials in each of the 10 training sessions for each memoranda type. As can be seen, performance improved from session 1 to session 10, and although memory performance did not appear to differ as a function of item type in early sessions, letters tended to be recalled better than the words or digits with more practice, which is consistent with previous serial memory span tasks (Crannell & Parrish, 1957). The ANOVA results supported this interpretation, with significant main effects of memoranda type, $F(2, 50) = 23.875$, $p < 0.001$, $\eta_p^2 = 0.488$, and session, $F(9, 225) = 21.933$, $p < 0.001$, $\eta_p^2 = 0.467$, which were both qualified by a significant memoranda type × session interaction: $F(18, 450) = 1.640$, $p = 0.047$, $\eta_p^2 = 0.062$.

Transfer data

Table 2 displays the pretest and posttest means for each cognitive outcome as a function of group, along with the paired-samples *t* test results each group. One-way ANOVAs with group as the between-subjects factor were conducted on all 9 pretest DVs in Table 2. The results indicated no differences among the three groups before training (3-back spatial $p = 0.099$; all other p's > 0.618). Inspection of Table 2 reveals pretest-to-posttest improvements in several cognitive outcomes, particularly the running span letter, running span matrix, 3-back letter, and letter fluency tasks. However, only possibly the running span letter (Fig. 4a) and 3-back letter (Fig. 5a) tasks exhibit any evidence of differential improvement from pretest to posttest as a function of group, which is the critical aspect of assessing whether training produced transfer or not. There was no evidence for far transfer to reading comprehension, as pretest and posttest performance was virtually identical for all three groups. In addition, there was no evidence for transfer to any of the transfer tasks that did not involve serial-order memory for letter stimuli. The full ANOVA results (F values, p values, and η_p^2 for each main effect and interaction) for each outcome are provided in Table 3.

For running span letter, all three groups showed evidence of improvement from pretest to posttest via paired-samples *t* tests (Table 2). The marginal group × session interaction

Footnote 3 (continued)

rectly, resulting in a loss of the necessary recall data from the last trial involving digits as the to-be-remembered items.

Table 2 Transfer session performance as a function of group

Outcome	Operation-letters ($n=30$)			Operation-mix ($n=27$)			Visual-search ($n=29$)		
	Pre	Post	g_{av}	Pre	Post	g_{av}	Pre	Post	g_{av}
Running span letter	32.73 (10.53)	41.13 (12.39)	0.72*	32.22 (12.52)	35.59 (13.26)	0.26†	31.69 (8.23)	36.03 (10.36)	0.46*
Running span matrix	32.10 (9.25)	36.10 (10.86)	0.39*	30.89 (9.04)	34.15 (9.50)	0.35*	31.76 (10.29)	37.79 (11.92)	0.54*
Change detect color	4.40 (0.96)	4.22 (0.93)	−0.19	4.33 (0.70)	4.37 (0.87)	0.06	4.44 (0.93)	4.53 (0.85)	0.09
Change detect orient	2.88 (1.16)	2.58 (1.21)	−0.24	2.76 (0.97)	2.49 (1.13)	−0.25	2.75 (1.13)	2.29 (1.20)	−0.39*
3-back letter	1.28 (0.68)	1.70 (0.72)	0.59*	1.08 (0.92)	1.40 (0.92)	0.35*	1.19 (0.84)	1.36 (0.66)	0.23
3-back matrix	1.77 (0.89)	1.82 (0.74)	0.06	1.44 (0.85)	1.66 (0.94)	0.24	1.25 (1.02)	1.43 (0.92)	0.18
Letter fluency	26.06 (6.55)	30.50 (7.97)	0.60*	27.19 (4.90)	27.98 (5.29)	0.15	26.59 (6.60)	29.33 (6.48)	0.41
Category fluency	35.28 (7.01)	36.51 (7.06)	0.17	33.47 (7.54)	31.47 (8.17)	−0.25	33.68 (8.45)	32.59 (6.37)	−0.15
Nelson-Denny	26.10 (3.03)	26.17 (2.39)	0.03	25.37 (2.96)	25.93 (2.06)	0.22	25.86 (3.26)	25.34 (4.24)	−0.14

g_{av} pretest-to-posttest effect size within each group

*$p<0.05$; †$p=0.054$

(Table 3) reflects the seemingly larger improvement in the operation-letter group, whereas the operation-mix group and visual-search group gains did not differ (Fig. 4a). These impressions were consistent with two follow-up ANOVAs (Table 4), separately comparing the operation-letters and visual-search groups, and the operation-mix and visual-search groups. Note that none of the significance tests reached $p<0.05$, indicating relatively weak evidence for transfer. For running span matrix, a different pattern emerged—although all three groups showed significant improvement from pretest to posttest (Table 2), there were no differences in the amount of transfer among the three groups (Fig. 4b, Tables 3, 4).

For 3-back letter, both training groups, but not the control group, exhibited a significant pretest-to-posttest improvement (Table 2). However, the group × session interaction did not approach significance (Table 3), nor did either follow-up ANOVA comparing each training group separately versus the control group (Table 4).[4] In addition, there was no evidence of transfer to 3-back matrix (Tables 2, 3, 4). Finally, supplementary analyses focusing on lure accuracy indicated no significant effects for 3-back letter or 3-back matrix (Tables 3, 4).

There was no evidence for transfer to any other tasks. Specifically, neither the operation-letters nor the operation-mix group exhibited significant pretest-to-posttest improvements on either change detection task (color, orientation). Although the within-group pretest-to-posttest improvement on letter fluency was significant only for the operation-letters group, neither the overall group × session interaction (Table 3) nor the separate operation-letters versus visual-search control group × session interaction approached significance (Table 4).[5] In both letter and category fluency, supplementary analyses using repetitions as the DV showed no evidence of transfer (Tables 3, 4). Finally, the results for reading comprehension were clear—there was no effect of training (Tables 2, 3, 4).

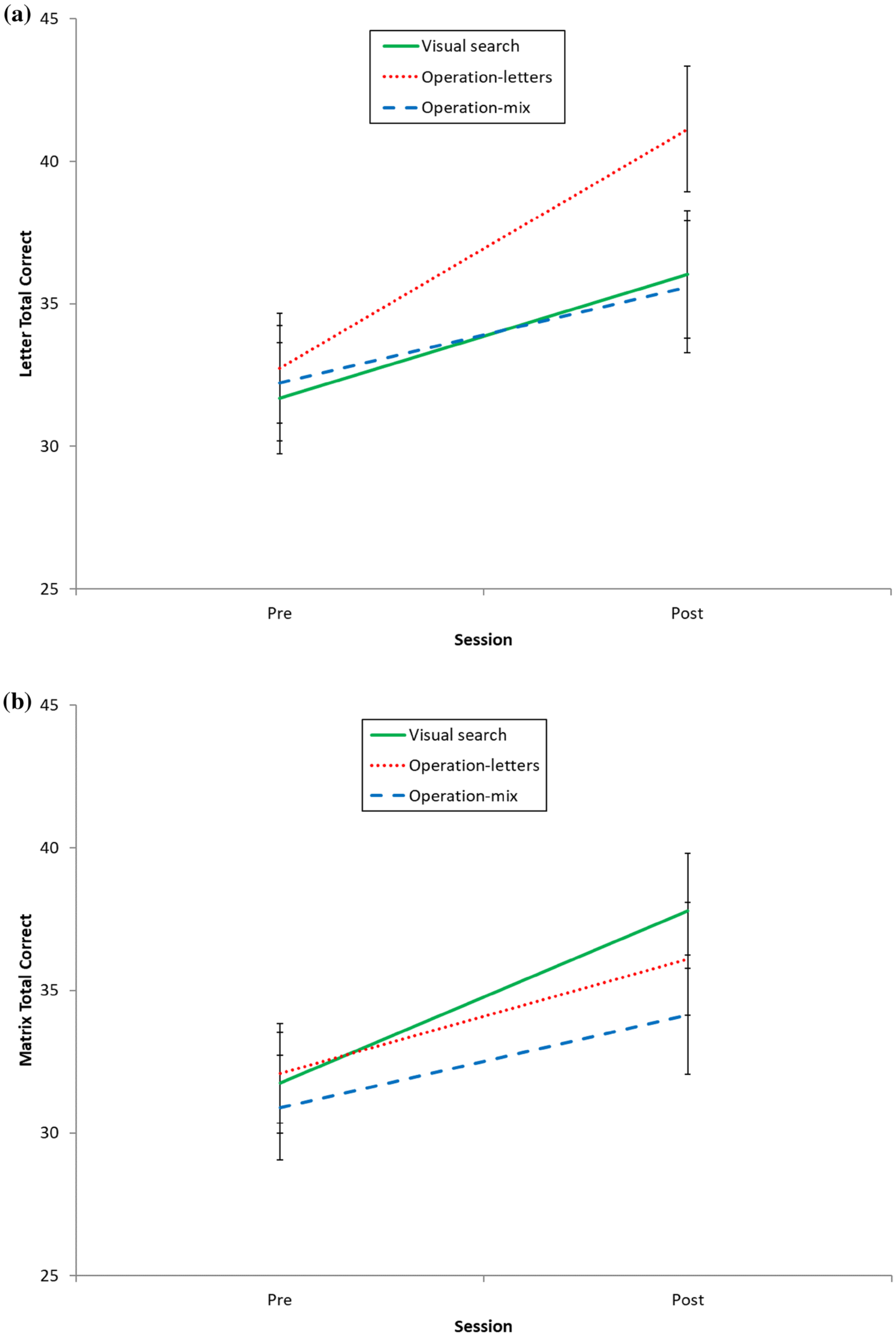
Survey data

Ratings of enjoyment, effort, and proportions of subjects who endorsed improvements in the cognitive abilities queried are presented in Table 5. Notably, the three groups did not differ in their ratings of enjoyment or effort during the 10 training sessions. The only self-perceived improvement outcome that differed as a function of group was for visual perception, where subjects in the visual-search control group were more likely than either training group to rate their visual perception to have changed ($p=0.042$). These results

[4] The letter 3-back result is one finding with a slight divergence between the ANOVA and ANCOVA results. While the follow-up ANCOVA comparing the operation-letters and visual-search groups for 3-back letter was marginally significant, the 2×2 follow-up ANOVA group x session interaction was not significant (Table 4).

[5] Category fluency is the other outcome in which the ANOVA and ANCOVA approaches differ. Although the interaction term from the ANOVA model was not significant, the group effect in the ANCOVA model was significant (Table 3). There are two reasons we do not think real transfer occurred for category fluency, in spite of the significant ANCOVA group effect. First, the within-group paired-samples t tests was not significant for either the operation-letters group, $t(29)=0.833$, $p=0.412$, $g_{av}=0.173$; operation-mix group, $t(26)=-1.446$, $p=0.160$, $g_{av}=-0.251$; or visual-search group, $t(28)=-0.663$, $p=0.513$, $g_{av}=-0.145$. These non-significant changes mean neither training group produced a change that was significantly different from 0. Second, the significant effect was likely driven by the non-significant increase from pretest to posttest for the operation-letters group, and the non-significant decrease from pretest to posttest for the operation-mix and visual-search groups (Table 2), a pattern that might produce statistical significance but is inconsistent with training producing transfer (see Redick, 2015, for further discussion).

Fig. 4 Running letter span (**a**) and running matrix span (**b**) transfer as a function of group. Error bars represent ± 1 standard error of the mean



suggest that the visual-search condition served as a believable control for the training groups, while also suggesting that any transfer that was observed on the cognitive tasks occurred separately from expectancy or placebo effects. Because the survey was the same as used in Redick et al. (2013, cf. Table 5), where we also used a visual-search control group, we note that the relative pattern of self-perceived improvements was very similar across both studies. The visual-search control groups in both experiments were most likely to self-report improvements in attention and visual perception and least likely to self-report improvements in daily activities and language.

The open-ended strategy questions were coded for responses indicating use of (a) overt rehearsal (examples included "Saying the letters out loud so I could hear them and see them" and "During the shorter practice sessions I

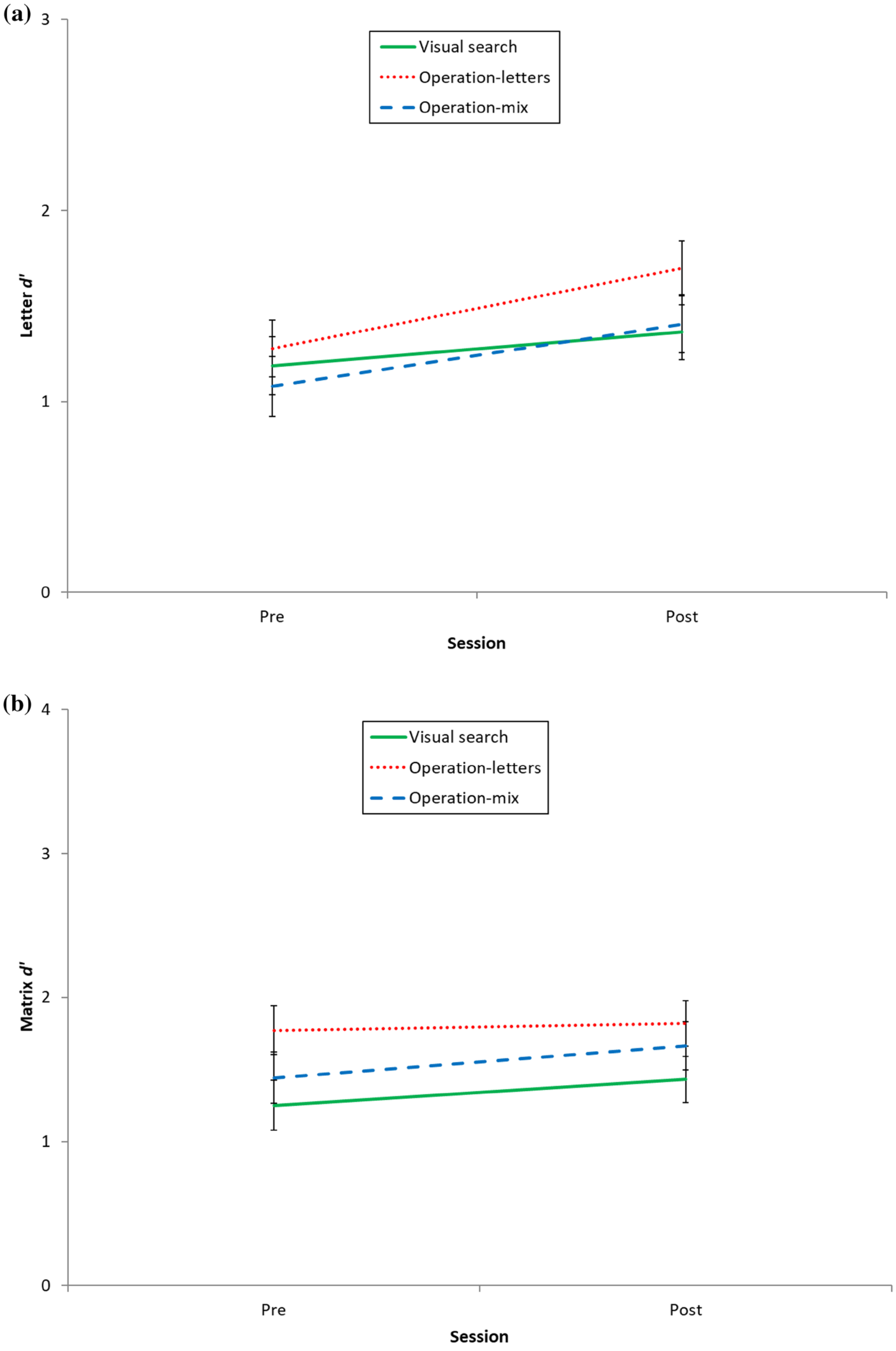


Fig. 5 3-back letter (**a**) and 3-back matrix (**b**) transfer as a function of group. Error bars represent ± 1 standard error of the mean

would say the letters out loud to help remember") and/or (b) associations between letter stimuli and words (examples included "Assign each letter a word or a name to make it into a sentence" and "To come up with words to remember the letters"). For overt rehearsal, 50% of the operation-letters group and 41% of the operation-mix group reported using the strategy, which did not differ between groups, $\chi^2(1) = 0.491$, $p = 0.483$. For the strategy of creating words from letter stimuli, 37% of the operation-letters group and 15% of the operation-mix group reported using this strategy, which marginally differed between groups, $\chi^2(1) = 3.499$, $p = 0.061$.

Table 3 Full statistics for the 3 (group) × 2 (session) factorial ANOVAs and group ANCOVAs

Task	ANOVA									ANCOVA	
	Group			Session			Group by session			Group	
	F	*p*	n_{p}^2	*F*	*p*	n_{p}^2	*F*	*p*	n_{p}^2	*F*	*p*
Running span letter	0.831	0.439	0.020	33.590	0.001	0.288	2.802	0.066	0.063	3.066	0.052
Running span matrix	0.423	0.656	0.010	26.030	0.001	0.239	0.905	0.409	0.021	1.041	0.358
Change detect color	0.450	0.639	0.011	0.030	0.863	0.000	0.718	0.491	0.017	1.017	0.366
Change detect orient	0.326	0.723	0.008	9.049	0.003	0.098	0.283	0.754	0.007	0.436	0.648
3-back letter	1.075	0.346	0.025	13.838	0.001	0.143	0.776	0.464	0.018	1.506	0.228
3-back matrix	2.328	0.104	0.053	3.189	0.078	0.037	0.393	0.676	0.009	0.229	0.796
Letter fluency	0.134	0.874	0.003	9.835	0.002	0.106	1.536	0.221	0.036	1.328	0.271
Category fluency	2.451	0.092	0.056	0.501	0.481	0.006	1.223	0.300	0.029	3.366	0.039
Nelson-Denny	0.346	0.709	0.008	0.012	0.912	0.000	0.954	0.389	0.022	0.882	0.418
3-back letter lure	0.362	0.697	0.009	2.506	0.117	0.029	0.610	0.546	0.014	0.427	0.654
3-back matrix lure	0.741	0.480	0.018	2.295	0.134	0.027	0.219	0.804	0.005	0.658	0.520
Letter fluency rep.	3.308	0.041	0.074	0.479	0.491	0.006	1.009	0.369	0.024	0.100	0.905
Category fluency rep.	0.735	0.483	0.017	1.351	0.249	0.016	0.365	0.695	0.009	0.460	0.633

For ANOVA, group main effect and group × session interaction, numerator *df* = 2; for session main effect, numerator *df* = 1; and for all effects, denominator *df* = 83. For ANCOVA, group main effect, numerator *df* = 2, denominator *df* = 82

rep. repetitions

Table 4 Statistics for the 2 (group) × 2 (session) factorial ANOVAs and group ANCOVAs

Task	Operation-letters					Operation-mix				
	Group by session			Group		Group by session			Group	
	F	*p*	n_{p}^2	*F*	*p*	*F*	*p*	n_{p}^2	*F*	*p*
Running span letter	3.329	0.073	0.055	3.708	0.059	0.166	0.685	0.003	0.135	0.715
Running span matrix	0.832	0.366	0.014	0.806	0.373	1.656	0.204	0.030	2.002	0.163
Change detect color	1.200	0.278	0.021	2.019	0.161	0.019	0.891	0.000	0.310	0.580
Change detect orient	0.354	0.554	0.006	0.705	0.405	0.457	0.502	0.008	0.457	0.457
3-back letter	1.430	0.237	0.024	3.312	0.074	0.569	0.454	0.010	0.355	0.554
3-back matrix	0.421	0.519	0.007	0.342	0.561	0.033	0.857	0.001	0.362	0.550
Letter fluency	0.621	0.434	0.011	0.533	0.468	0.837	0.364	0.015	0.773	0.383
Category fluency	1.105	0.297	0.019	4.265	0.044	0.176	0.677	0.003	0.342	0.561
Nelson-Denny	0.513	0.477	0.009	0.815	0.371	1.724	0.195	0.031	1.347	0.251
3-back letter lure	0.101	0.751	0.002	0.650	0.423	0.563	0.456	0.010	0.014	0.905
3-back matrix lure	0.151	0.699	0.003	0.020	0.888	0.473	0.494	0.009	0.099	0.754
Letter fluency rep.	1.146	0.289	0.020	0.205	0.653	1.233	0.272	0.022	0.102	0.751
Category fluency rep.	0.523	0.473	0.009	0.101	0.752	0.493	0.486	0.009	1.127	0.293

For ANOVA, group × session interaction, numerator *df* = 1, denominator *df* = 57/54 (operation-letters/operation-mix). For ANCOVA, group main effect, numerator *df* = 1, denominator *df* = 56/53 (operation-letters/operation-mix)

rep. repetitions

Discussion

The current study adds to a burgeoning literature examining the potential efficacy of WM training to improve cognitive abilities, as measured by transfer to untrained tasks. However, in contrast to many studies in the literature that atheoretically compare a WM training group against a control group (research question: does it work?), we manipulated the training content across two groups in an effort to investigate potential mechanisms underlying WM training and transfer (research question: if it works, why does it work?). Specifically, we compared two WM training programs that varied only in the to-be-remembered content,

Table 5 Posttest survey data

Topic	Operation-letters	Operation-mix	Visual-search	F/χ^2	p
Enjoyment	2.67 (0.88)	2.56 (0.80)	2.76 (0.58)	0.492	0.613
Effort	3.43 (0.63)	3.41 (0.50)	3.31 (0.54)	0.390	0.678
Intelligence	33%	26%	24%	0.697	0.706
Memory	67%	67%	62%	0.179	0.914
Attention	50%	48%	62%	1.314	0.518
Visual perception	37%	37%	66%	6.354	0.042
Language	3%	0%	3%	0.938	0.626
Daily activities	27%	11%	17%	2.324	0.313

Enjoyment and effort values represent mean (SD) on scale of 1–4; All other values represent proportion of subjects in each group that replied yes to the question "Do you feel that your participation in this study has changed your _______ ?". For one-way ANOVA, numerator $df=2$, denominator $df=83$. For Chi-square test, $df=2$

with one group (operation-letters) repeatedly practicing memory for letters and another group (operation-mix) practicing memory for letters, words, and digits in a predictable pattern. The use of different to-be-remembered materials embedded within the same overall training program (operation span) allowed us to compare predictions primarily about the role of PI in WM training and transfer, but also secondary predictions involving strategy and training variability. Next, we interpret the results in relation to these predictions.

Training and transfer: evidence for strategy account

Looking first at the training results, we observed that the operation-letters group outperformed the operation-mix group over the course of the 10 training sessions. This pattern of results is inconsistent with the PI account, but the results are consistent with a strategy account. Because subjects in the operation-letters repeatedly recalled from a fixed pool of stimuli across all memory trials, they should have experienced more PI than the operation-mix group, who had PI release built into the task design by predictably alternating among memory categories of letters, words, and digits (e.g., Bunting, 2006). However, instead of an impediment to performance, it appears that consistently remembering the same stimuli across training sessions was beneficial to performance. Perhaps this finding should not be surprising—Ericsson and colleagues' work with SF and DD in the 1970s and 1980s (recently reviewed and extended in Yoon, Ericsson, & Donatelli, 2018) demonstrated that subjects can develop very effective stimulus-specific strategies on verbal memory span tasks but still show typical levels of performance on unpracticed verbal memory span tasks (Ericsson, Chase, & Faloon, 1980).

The operation-letters group ($M=7.70$, SD $=3.18$) on average attained a marginally higher level in session 10 than the operation-mix group ($M=6.30$, SD $=2.46$), $t(55)=1.850$, $p=0.070$. This pattern is consistent with the operation-letters subjects developing letter-specific strategies and using them across sessions, whereas the intermixed nature of the stimuli for the operation-mix condition may have prevented or made less obvious the potential benefit of developing a letter-based mnemonic strategy. Limited data in support of this idea were obtained in a posttest survey, in which we asked subjects to provide strategies they used during the training sessions. While 37% of operation-letters subjects explicitly listed a strategy of associating letters with words to aid memory, only 15% of operation-mix subjects provided this strategy, even though both groups had practice remembering letters in the context of the operation span training. While such retrospective reports of strategy use should be interpreted cautiously, they are consistent with differential strategy usage between the two highly similar training conditions.

The transfer data also provided evidence for the strategy account versus the PI and training-variability accounts. Recall that relatively weak evidence for transfer was observed for the operation-letters group only versus the active-control group (running letter span and, based on ANCOVA, possibly 3-back letter). The PI account predicts greater transfer for the operation-letters group than the operation-mix group, but the transfer should be broad, across tasks that require resolution of PI. The lack of far transfer is inconsistent with the PI account. The training-variability account predicts greater transfer for the operation-mix group than for the operation-letters group, because the change in to-be-remembered content limits stimulus-specific strategies in favor of training 'core' WM processes (Lustig, Shah, Seidler, & Reuter-Lorenz, 2009; Morrison & Chein, 2011). In addition, the training-variability account predicts broad transfer. The results were inconsistent with both predictions of the training-variability account. The combination of training and transfer results is thus most consistent with the strategy account, in which subjects develop stimulus- or task-specific strategies to improve performance on the repeatedly practiced training task, but transfer is narrow and only observed on tasks in which the learned strategies can also be utilized (Dunning & Holmes, 2014; Ericsson et al., 1980; Soveri et al., 2017; Sprenger et al., 2013). In any case, the very limited transfer observed does not bode well for more general cognitive improvements following WM training.

Relation to previous research

We followed recommended procedures in the training literature (Redick et al., 2015), specifically using: (a) larger sample sizes than typical—a recent WM training meta-analysis (Au et al., 2015) reported that the average group sample size is just under 20 subjects; (b) an active-control instead of a passive-control group, to help control for placebo and expectancy effects; (c) using multiple measures of each outcome (except reading comprehension), to provide more conclusive evidence in favor of or against transfer than when relying on a single measure; and (d) pretest and posttest session durations of approximately 75 min, to mitigate the potential criticism that transfer effects were not observed because subjects were fatigued. Studies with passive controls have an elevated likelihood of observing statistically significant transfer compared to studies that employ active-control groups. However, we did not include a passive-control group here because, as discussed elsewhere (Melby-Lervåg et al., 2016; Redick et al., 2015), the limitations of passive-control groups are so severe as to render them minimally effective as a meaningful comparison condition.

The lack of a WM training group by session interaction in the training data is consistent with the results of previous studies that attempted to manipulate high- and low-PI WM training procedures (Hussey et al., 2017; Loosli et al., 2016; Oelhafen et al., 2013). That is, although these studies sought to manipulate PI, ultimately subjects that repeatedly practiced the high- and low-PI versions of the tasks did not show a different amount of improvement across training. Unfortunately, the study (Bomyea et al., 2015) most comparable to ours did not present training data analyses. Note that we do not have a clear method for checking our PI manipulation based on the training data. Even if we restrict our analysis to session one performance, because our task was adaptive and list length in the operation-mix varied randomly with the to-be-remembered stimuli, it is difficult to draw a comparison between our PI conditions as was done in previous single-session research (e.g., Blalock & McCabe, 2011; Bunting, 2006). These findings, in conjunction with the previously described WM and PI research that motivated our study, suggest that manipulations that affect performance in specific ways within a one-off assessment of WM do not necessarily have the same effect when the WM test is practiced over multiple sessions. While the reasons for this difference are not known, understanding this discrepancy is important if researchers hope to know what changes with repeated practice in WM training applications.

Our transfer results are extremely narrow, similar to our previous research using complex span tasks as WM training (Foster et al., 2017; Harrison et al., 2013). Transfer from one type of working memory training (e.g., complex span) to other types of working memory tasks (e.g., *n*-back, change detection) is not consistently observed (Redick & Lindsey, 2013; Soveri et al., 2017). Even the transfer to running letter span was marginal, which was somewhat surprising given that Harrison et al. (2013) observed clear transfer to running letter span. The main difference is that the visual-search control group in our study significantly improved from pretest to posttest, which did not occur in Harrison et al. (2013). In fact, when comparing the operation-letters and operation-mix groups against each other, the group × session interaction was significant, $F(1,55) = 5.240$, $p = 0.026$, $n_{\mathrm{p}}^2 = 0.087$. This analysis also facilitates the comparison to Bomyea et al. (2015), who found that their high-PI reading span training group showed significantly greater transfer to operation span, compared to their low-PI reading span training group that alternated memoranda across trials. With our design, we obtained near-transfer results consistent with Bomyea et al. (2015), but by presenting the training session data and examining cognitive transfer with multiple measures, we were able to conclude that improvements in the ability to deal with PI are not the likely explanation for the pattern of training and transfer data.

Although the focus of many WM training studies has been whether or not it works to increase fluid intelligence (Au et al., 2015), reading comprehension is an ecologically and academically relevant outcome that has a strong relationship with individual differences in WM (Daneman & Merikle, 1996). Given the interesting results of Chein and Morrison (2010), who reported significant transfer from complex span training to the Nelson-Denny reading comprehension test in young adults, our lack of far transfer here could be viewed as a failure to replicate. However, other WM training studies have also failed to produce transfer to reading comprehension as assessed by the Nelson-Denny test (Carretti, Borella, Zavagnin, & de Beni, 2013; Gropper, Gotlieb, Kronitz, & Tannock, 2014; Payne & Stine-Morrow, 2017; Sprenger et al., 2013, Experiment 2; Thompson et al., 2013). Notably, many of the studies not observing transfer to the Nelson-Denny test used active-control groups, whereas Chein and Morrison (2010) used a passive-control group. In addition, the significant reading comprehension transfer reported in Chein and Morrison (2010) was based on using a one-tailed *t* test, and as reported in Sprenger et al. (2013), a Bayesian re-analysis of the results indicates "the data are equally supportive of both the alternative and null hypotheses" (p. 659). Thus, despite the robust relationship between verbal complex span tasks and reading comprehension (meta-analytic zero-order correlations with global and specific language comprehension measures ranged from $r = 0.30$ to 0.52; Daneman & Merikle, 1996), improvements with training on operation span did not lead to improvements on reading comprehension.

Limitations and future directions

We note some potential criticisms with the current study. First, 10 WM training sessions were administered, in contrast to, for example, Chein and Morrison (2010) in which subjects completed 20 training sessions. Thus, one possibility is that the current subjects did not receive enough training to produce far transfer to reading comprehension. However, the literature indicates no clear pattern for a relationship between amount of training and transfer to reading comprehension, nor to far transfer more generally. For example, in a recent meta-analysis, Melby-Lervåg et al. (2016) statistically compared greater versus fewer training sessions as a moderator variable, and observed no evidence that studies with more training sessions produced greater near or far transfer.

Second, the training materials in both training conditions used only alphanumeric materials, in contrast to visuo-spatial memory content, or a combination of alphanumeric and visuo-spatial stimuli. Thus, we may have limited the opportunity for 'core', domain-general WM processes to be trained and, thus, minimized the chance that far transfer could be observed. We find this critique lacking on at least two accounts. First, recent meta-analyses have examined training content as a possible moderator of transfer, with no evidence that using verbal, visuo-spatial, or a combination of the two types of training affects the degree of transfer (Au et al., 2015; Schwaighofer et al., 2015). Second, our previous study (Harrison et al., 2013) administered both operation and symmetry span training tasks, in what could be considered as quite similar to the high-PI condition in the current research. Despite the use of both verbal and visuo-spatial training stimuli, Harrison et al. observed only narrow transfer to other memory span tasks. Finally, our far transfer tasks were verbal-based (fluency, reading comprehension), so the use of alphanumeric training stimuli seemed most appropriate.

Third, our verbal working memory transfer tasks only included letters as stimuli, whereas including additional verbal working memory transfer tasks with words or digits may have been helpful to demonstrate the specificity of any strategy gains by the operation-letters training group. This would be consistent with the approach advocated by Gathercole, Dunning, Holmes, and Norris (2019) in their recent process-specific framework.

Fourth, one might argue that if the low-PI condition produced transfer, it would be difficult to determine if the causal mechanism is the reduction in PI versus the increase in the variability of training. We strongly disagree, because the PI-as-cause account would not predict far transfer for the low-PI condition relative to the high-PI condition, whereas the variability-as-cause account would predict far transfer for the low-PI condition relative to the high-PI condition.

Finally, although the sample sizes used in our training and control groups were larger than those typically used in the literature (Bogg & Lasecki, 2015; Melby-Lervåg et al., 2016), we did not conduct a formal power analysis before the study. In their review of existing meta-analyses, Soveri et al. (2017, cf. Table 1) reported that near transfer to WM exhibited effect sizes between Cohen's $d = 0.12$–0.79, depending on the exact inclusion criteria (e.g., age, type of training, type of control group, category of task outcomes). The meta-analysis by Melby-Lervåg et al. (2016) reported a Hedge's $g = 0.15$ for reading comprehension as the outcome when WM training was compared against an active-control group. However, to the degree that meta-analyses over-represent published studies (which are more likely than unpublished studies to contain significant results) and/or include small-sample studies, these meta-analytic effect size estimates may be inaccurate for determining sample sizes.

Despite these issues with determining the relevant effect size to detect, we used G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) to conduct the same power analysis three times, using a range of f values to establish necessary sample sizes. Using labels from Cohen (1988, pp. 285–286), we used $f = 0.25$ (medium effect), $f = 0.10$ (small effect), and $f = 0.175$ (halfway between a medium and small effect). Given our view of the training literature, perhaps $f = 0.175$ represents the effect one might expect in terms of near transfer, when including an active-control group. Other than setting desired power at 0.80, we used default G*Power 3 settings. The power analysis results indicated that our final sample size of $N = 86$ does achieve at least 0.80 power to detect transfer effects of $f = 0.25$ and $f = 0.175$, but not to detect transfer effects of $f = 0.10$.

We agree that, ceteris paribus, larger sample sizes for training studies are desirable. Pragmatically, however, there is a real cost (time and money) to conducting such studies; in the current study, subject compensation alone was $14,089 USD. We anticipate that future studies will continue to focus on large-sample studies to improve the precision of the effect size estimates obtained as the debate about WM training efficacy continues.

In addition, recent studies have examined the role of individual differences in predicting the amount of training and transfer (e.g., Foster et al., 2017). In fact, our research (Gunn, Gerst, Wiemers, Redick, & Finn, 2018; Wiemers, Redick, & Morrison, 2018) has shown that individuals higher in cognitive ability at pretest show larger training gains on complex span training tasks like those used in the current study. Foster et al. (2017) demonstrated that despite the larger training gain for high-WM subjects, there was no difference in the amount of transfer when compared to low-WM subjects. Although the current study was not designed to examine individual differences in relation to training and transfer, future studies should address this aspect of WM training.

Conclusion

The current study examined the hypothesis that an increased resistance to PI underlies WM training and transfer improvements. However, the training and transfer results were not consistent with a PI account. Transfer was limited to tasks that shared very similar stimuli and task features—serial-order memory for letters. Our results cast further doubt on the utility of WM training as a tool to produce far transfer to correlated yet structurally dissimilar outcomes.

Acknowledgements The research reported here was funded by the Office of Naval Research (Award # N00014-12-1-1011) to RWE. While working on this manuscript, TSR was supported by the National Institutes of Health (Award # 2R01AA013650-11A1). The research described here was presented at the 2015 annual meetings of the Association for Psychological Science and the Midwestern Psychological Association. We thank Devlin Bertha, Chandani Bhatt, Haley Brower, Caleb Carriere, Taylor Daniel, Kent Etherton, Andrea Grovak, Anoop Javalagi, Yun Qi Lim, Sarika Srivastava, and Michael White for assistance with data collection and scoring. We thank Nash Unsworth and Matt Robison for helpful comments on an earlier draft of the manuscript.

Compliance with ethical standards

Conflict of interest All authors declare they have no conflict of interest.

Ethical standards The study described here was approved by the local university Institutional Review Board.

Informed consent Informed consent was obtained from all individual subjects included in the study.

References

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review, 22,* 366–377. https://doi.org/10.3758/s13423-014-0699-x.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128,* 612–637. https://doi.org/10.1037/0033-2909.128.4.612.

Blalock, L. D., & McCabe, D. P. (2011). Proactive interference and practice effects in visuospatial working memory span task performance. *Memory, 19,* 83–91. https://doi.org/10.1080/09658211.2010.537035.

Bogg, T., & Lasecki, L. (2015). Reliable gains? Evidence for substantially underpowered designs in studies of working memory training transfer to fluid intelligence. *Frontiers of Psychology., 5,* 1589. https://doi.org/10.3389/fpsyg.2014.01589.

Bomyea, J., Stein, M. B., & Lang, A. J. (2015). Interference control training for PTSD: A randomized controlled trial of a novel computer-based intervention. *Journal of Anxiety Disorders, 34,* 33–42. https://doi.org/10.1016/j.janxdis.2015.05.010.

Borella, E., Carretti, B., & Pelegrina, S. (2010). The specific role of inhibition in reading comprehension in good and poor comprehenders. *Journal of Learning Disabilities, 43*(6), 541–552. https://doi.org/10.1177/0022219410371676.

Bunting, M. (2006). Proactive interference and item similarity in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(2), 183–196. https://doi.org/10.1037/0278-7393.32.2.183.

Burgess, G. C., Gray, J. R., Conway, A. R. A., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General, 140,* 674–692. https://doi.org/10.1037/a0024695.

Carretti, B., Borella, E., Zavagnin, M., & Beni, R. (2013). Gains in language comprehension relating to working memory training in healthy older adults. *International Journal of Geriatric Psychiatry, 28*(5), 539–546. https://doi.org/10.1002/gps.3859.

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. https://doi.org/10.3758/PBR17.2.193.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Cowan, N. (2016). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24,* 1158–1170. https://doi.org/10.3758/s13423-016-1191-6.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51,* 42–100. https://doi.org/10.1016/j.cogpsych.2004.12.001.

Crannell, C. W., & Parrish, J. M. (1957). A comparison of immediate memory span for digits, letters, and words. *The Journal of Psychology, 44,* 319–327. https://doi.org/10.1080/00223980.1957.9713089.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3,* 422–433. https://doi.org/10.3758/BF03214546.

Dunning, D. L., & Holmes, J. (2014). Does working memory training promote the use of strategies on untrained working memory tasks? *Memory & Cognition, 42,* 854–862. https://doi.org/10.3758/s13421-014-0410-5.

Emery, L., Hale, S., & Myerson, J. (2008). Age differences in proactive interference, working memory, and abstract reasoning. *Psychology and Aging, 23*(3), 634–645. https://doi.org/10.1037/a0012577.

Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 18,* 972–992. https://doi.org/10.1037/0278-7393.18.5.972.

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science, 208*(4448), 1181–1182. https://doi.org/10.1126/science.7375930.

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,* 175–191. https://doi.org/10.3758/BF03193146.

Foroughi, C. K., Monfort, S. S., Paczynski, M., McKnight, P. E., & Greenwood, P. M. (2016). Placebo effects in cognitive training. *Proceedings of the National academy of Sciences of the United States of America, 113*(27), 7470–7474. https://doi.org/10.1073/pnas.1601243113.

Foster, J. L., Harrison, T. L., Hicks, K. L., Draheim, C., Redick, T. S., & Engle, R. W. (2017). Do the effects of working memory training depend on baseline ability level? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 1677–1689.

Gathercole, S. E., Dunning, D. L., Holmes, J., & Norris, D. (2019). Working memory training involves learning new skills. *Journal of Memory and Language, 105,* 19–42. https://doi.org/10.1016/j.jml.2018.10.003.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6*(3), 316–322. https://doi.org/10.1038/nn1014.

Gropper, R. J., Gotlieb, H., Kronitz, R., & Tannock, R. (2014). Working memory training in college students with ADHD or LD. *Journal of Attention Disorders, 18*(4), 331–345. https://doi.org/10.1177/1087054713516490.

Gunn, R. L., Gerst, K. R., Wiemers, E. A., Redick, T. S., & Finn, P. R. (2018). Predictors of effective working memory training in those with alcohol use disorders. *Alcoholism: Clinical & Experimental Research, 42,* 2432–2441. https://doi.org/10.1111/acer.13892.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science, 24,* 2409–2419.

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82,* 511–518. https://doi.org/10.1037/h0076767.

Hussey, E. K., Harbison, J. I., Teubner-Rhodes, S. E., Mishler, A., Velnoskey, K., & Novick, J. M. (2017). Memory and language improvements following cognitive control training. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(1), 23–58. https://doi.org/10.1037/xlm0000283.

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 of the United States of America, 105,* 6829–6833. https://doi.org/10.1073/pnas.0801268105.

Jolles, D. D., & Crone, E. A. (2012). Training the developing brain: A neurocognitive perspective. *Frontiers in Human Neuroscience, 6,* 76. https://doi.org/10.3389/fnhum.2012.00076.

Jonides, J., & Nee, D. E. (2006). Brain mechanisms of proactive interference in working memory. *Neuroscience, 139*(1), 181–193. https://doi.org/10.1016/j.neuroscience.2005.06.042.

Kane, M. J., Bleckley, M. K., Conway, A. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130,* 169–183. https://doi.org/10.1037/0096-3445.130.2.169.

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), 749–777. https://doi.org/10.1037/0278-7393.32.4.749.

Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology, 24,* 781–791. https://doi.org/10.1076/jcen.24.6.781.8395.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4,* 863. https://doi.org/10.3389/fpsyg.2013.00863.

Loosli, S. V., Falquez, R., Unterrainer, J. M., Weiller, C., Rahm, B., & Kaller, C. P. (2016). Training of resistance to proactive interference and working memory in older adults: A randomized double-blind study. *International Psychogeriatrics, 28*(3), 453–467. https://doi.org/10.1017/S1041610215001519.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305. https://doi.org/10.1037/h0025105.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General, 130,* 199–207. https://doi.org/10.1037/0096-3445.130.2.199.

Lustig, C., Shah, P., Seidler, R., & Reuter-Lorenz, P. A. (2009). Aging, training, and the brain: A review and future directions. *Neuropsychology Review, 19,* 504–522. https://doi.org/10.1007/s11065-009-9119-9.

May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory & Cognition, 27,* 759–767. https://doi.org/10.3758/BF03198529.

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": Evidence from a meta-analytic review. *Perspectives on Psychological Science, 11,* 512–534. https://doi.org/10.1177/1745691616635612.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40–48. https://doi.org/10.1037/0021-843X.110.1.40.

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review, 18,* 46–60. https://doi.org/10.3758/s13423-010-0034-0.

Nelson-Denny. (1993). *Reading Comprehension test*. Rolling Meadows: The Riverside Publishing Company.

Oelhafen, S., Nikolaidis, A., Padovani, T., Blaser, D., Koenig, T., & Perrig, W. J. (2013). Increased parietal activity after training of interference control. *Neuropsychologia, 51*(13), 2781–2790. https://doi.org/10.1016/j.neuropsychologia.2013.08.012.

Payne, B. R., & Stine-Morrow, E. A. L. (2017). The effects of home-based cognitive training on verbal working memory and language comprehension in older adulthood. *Frontiers in Aging Neuroscience, 9,* 256. https://doi.org/10.3389/fnagi.2017.00256.

Persson, J., & Reuter-Lorenz, P. A. (2008). Gaining control: Training executive function and far transfer of the ability to resolve interference. *Psychological Science, 19*(9), 881–888. https://doi.org/10.1111/j.1467-9280.2008.02172.x.

Persson, J., & Reuter-Lorenz, P. A. (2011). Retraction of "Gaining control: Training executive function and far transfer of the ability to resolve interference". *Psychological Science, 22*(4), 562. https://doi.org/10.1177/0956797611404902.

Redick, T. S. (2015). Working memory training and interpreting interactions in intelligence interventions. *Intelligence, 50,* 14–20. https://doi.org/10.1016/j.intell.2015.01.014.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28,* 164–171. https://doi.org/10.1027/1015-5759/a000123.

Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review, 20,* 1102–1113. https://doi.org/10.3758/s13423-013-0453-9.

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,* 359–379. https://doi.org/10.1037/a0029082.

Redick, T. S., Shipstead, Z., Wiemers, E. A., Melby-Lervåg, M., & Hulme, C. (2015). What's working in working memory training? An educational perspective. *Educational Psychology Review, 27,* 617–633. https://doi.org/10.1007/s10648-015-9314-6.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General, 126*(3), 211–227. https://doi.org/10.1037/0096-3445.126.3.211.

Schwaighofer, M., Fischer, F., & Buhner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist, 50*(2), 138–166. https://doi.org/10.1080/00461520.2015.1036274.

Shipstead, Z., & Engle, R. W. (2013). Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 277–289. https://doi.org/10.1037/a0028467.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. https://doi.org/10.1177/0956797611417632.

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin & Review, 24,* 1077–1096. https://doi.org/10.3758/s13423-016-1217-0.

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Chrabaszcz, J. S., … et al. (2013). Training working memory: Limits of transfer. *Intelligence, 41,* 638–663. https://doi.org/10.1016/j.intell.2013.07.013.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137–149. https://doi.org/10.3758/BF03207704.

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 137–151. https://doi.org/10.1037/a0020365.

Thompson, T. W., Waskom, M. L., Garel, K.-L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., … et al. (2013). Failure of working memory training to enhance cognition or intelligence. *PLoS One, 8*(5), e63614. https://doi.org/10.1371/journal.pone.0063614.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5.

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2011). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *The Quarterly Journal of Experimental Psychology, 64*(3), 447–466. https://doi.org/10.1080/17470218.2010.505292.

Wiemers, E. A., Redick, T. S., & Morrison, A. B. (2018). The influence of individual differences in cognitive ability on working memory training gains. *Journal of Cognitive Enhancement*. https://doi.org/10.1007/s41465-018-0111-2.

Wright, D. B. (2006). Comparing groups in a before-after design: When *t* test and ANCOVA produce different results. *British Journal of Educational Psychology, 76,* 663–675. https://doi.org/10.1348/000709905X52210.

Yoon, J.-S., Ericsson, K. A., & Donatelli, D. (2018). Effects of 30 years of disuse on exceptional memory performance. *Cognitive Science, 42,* 884–903. https://doi.org/10.1111/cogs.12562.

Young, C. W., & Supa, M. (1941). Mnemic inhibition as a factor in the limitation of the memory span. *The American Journal of Psychology, 54,* 546–552. https://doi.org/10.2307/1417204.