Psychologica Belgica
2010, 50-3&4, 245-276.

DOES WORKING MEMORY TRAINING GENERALIZE?

Zach SHIPSTEAD, Thomas S. REDICK, & Randall W. ENGLE
Georgia Institute of Technology

Recently, attempts have been made to alter the capacity of working memory (WMC) through extensive practice on adaptive working memory tasks that adjust difficulty in response to user performance. We discuss the design criteria required to claim validity as well as generalizability and how recent studies do or do not satisfy those criteria. It is concluded that, as of yet, the results are inconsistent and this is likely driven by inadequate controls and ineffective measurement of the cognitive abilities of interest.

The capacity of working memory varies among individuals.
About 50% of the variance in general intelligence between individuals can be explained by differences in working memory capacity.
Kids with attention problems often have working memory deficits.
Working memory has been linked to academic success.
Stroke victims often suffer from impaired working memory.
Working memory is plastic. Like a muscle, it can be improved through exercise.

(www.cogmed.com; Archived at http://www.webcitation.org/5o3GZLNwQ)

While working memory capacity (WMC) is, in a literal sense, a measure of short-term information retention, its true characterization becomes apparent in relation to the phenomena with which it has been linked. Undoubtedly, WMC involves maintenance and retrieval, but, as is suggested by the above quote, it also reflects abilities beyond those that are easily mapped to standard notions of "memory". Such associations have made WMC the focus of a growing literature concerned with discovering whether constant, prolonged taxing of working memory (WM) will increase its capacity and strengthen its functions of maintenance and manipulation of information. Does such prolonged practice lead to improvements in performance of diverse measures of cognitive functions such as attentional focus and general reasoning abilities?

It should be no surprise that a concept which apparently reflects critical abilities such as attention (Engle, 2002) and reasoning (Engle et al., 1999) would serve as a keystone for real-world cognition. In recent years, researchers have found evidence that people who are high in WMC are less likely to

Randall W. Engle, School of Psychology, Georgia Institute of Technology, 654 Cherry Street NW, Atlanta, GA 30332, E-mail: randall.engle@gatech.edu. This work was supported by Office of Naval Research grant N00014-09-1-0129.

mind-wander when focus is needed (Kane et al., 2007a), are better multi-taskers (Buhner et al., 2006; Hambrick et al., 2010), are more skilled at engaging in effortful regulation of emotion (Schmeichel, Volokhov, & Demaree, 2008), and are more adept at mentally challenging endeavors such as acquiring rules of logic (Kyllonen, & Stephens, 1990) and learning a computer programming language (Shute, 1991). Indeed WMC is not a toy-construct of cognitive psychology that behaves well in the laboratory but has no obvious application to life in general. Rather, it is apparent that this measure predicts the ability to engage in appropriate behavior at the appropriate time in real-world tasks.

Fundamental to this relationship is the well-established association between WMC and controlled attention (Kane et al., 2007b). For instance, high WMC individuals are not only faster at performing the antisaccade task, which requires participants to suppress and redirect reflexive eye movements, but are less apt to make inappropriate eye movements in the first place (Kane et al., 2001; Unsworth, Schrock & Engle, 2004). The Flanker task (Eriksen & Eriksen, 1974), requires participants to report the middle portion of a display, in the face of conflict from the outer portions (e.g. →→←→→). High WMC participants are less likely to be affected by the outer items (Redick & Engle, 2006) but only after attention has been given ample time to engage (Heitz & Engle, 2007). Finally, the dichotic listening task (Cherry, 1953; Moray, 1959) requires participants to verbally shadow words that are spoken into one ear while ignoring what is said in the other. Here, high WMC individuals are less likely to be distracted or even notice when their name is unexpectedly spoken in the unattended ear (Conway, Cowan, & Bunting, 2001). However, when high WMC individuals are told to monitor the "unattended" ear for their name, they are actually more likely than low WMC individuals to hear it, without a loss of performance on the shadowing task (Coldflesh & Conway, 2007).

A more complex demonstration of the relationship between WMC and controlled attention is provided by Kane and Engle's (2003) study of WMC and performance in the Stroop task (MacLeod, 1991; Stroop, 1935), which requires participants to state the hue in which a word has been written. In this task, attentional demand is manipulated through the type of the words that are used. When the word is congruent to the hue (e.g. hue = blue, word = "BLUE"), participants can produce fast, accurate responses. However, when the word and hue are incongruent (e.g. hue = blue, word = "RED"), participants are relatively slow and less accurate in outputting their answers.

As might be predicted, given the links between WMC and attention, high WMC participants generally show less slowing in response to incongruent Stroop stimuli than do low WMC individuals (Kane & Engle, 2003). However, a second phenomenon becomes apparent in situations where the task is primarily composed of congruent stimuli. In most situations, the con-

flict that arises in response to incongruent Stroop trials reinforces the action which must be taken in order to respond accurately (e.g. ignore the word). Congruent stimuli, on the other hand, pose no challenge to conflict resolution, and may even allow for inadvertent responding based on word information (MacLeod & MacDonald, 2000). In situations where such trials are abundant, people who are low in WMC begin to make incorrect responses (e.g. "gree…r..blue") at a significantly higher rate than high WMC individuals. Kane and Engle (2003) therefore interpreted WMC-related differences in Stroop performance as reflecting both individual differences in the ability to resolve conflict between multiple streams of information and in the ability to maintain a connection to behavioral goals when external support is rare.

Recently, a line of research has evolved that questions whether WMC is fixed or whether it will increase in response to being constantly taxed by information. By the account of some researchers, WMC is not restricted to the relative differences people show in basic cognitive control. To these investigators, deficits in WMC are at the heart of several life-affecting phenomena, such as the lack of focus which accompanies attention deficit hyperactivity disorder (ADHD; Klingberg, Forsberg, & Westerberg, 2002), cognitive deficits following stroke (Westerberg et al., 2007) or general learning disability (Alloway & Alloway, 2009). In fact, some have gone so far as to argue that, for children with learning disabilities, WMC is the best predictor of future scholastic success (Alloway, 2009). For other researchers, the possibility that the executive functions of WMC might be improved as a result of training provides a rare opportunity to manipulate specific aspects of WM and observe the accompanying physiological and cognitive changes (Dahlin et al., 2008a; McNab et al., 2009).

Given that attempts to train cognitive abilities are neither new, nor without controversy (Barnett & Ceci, 2002; Carroll, 1993), claims that WMC or associated abilities have been successfully trained should not be taken lightly. A survey of the recent literature suggests reasons for optimism. However, there are also reasons to be cautious about the conclusions reached by those studies. Therefore, before reviewing the studies, we provide a degree of context. The sections that follow include a brief summary of our current opinions on WMC measurement and interpretation, a general synopsis of what a well-designed adaptive training study entails, as well as some of the basic assumptions behind these programs. This is followed by a discussion of the minimum requirements a researcher must meet in order to confidently state that a cognitive construct has been altered by training. Following the literature review, we raise questions as to whether it can be confidently stated that WMC has been successfully altered.

Working Memory Capacity and How Practice/Training Might Increase It

WMC measurement and mechanisms

Short-term memory (STM; e.g. memory storage) has long been measured by the simple-span task which requires participants to recall short lists of items (e.g. words, numbers, spatial locations; Engle & Oransky, 1999) in the same order as presented. A person's memory span is sometimes quantified as the longest list she can recall perfectly. Such simple span tasks, particularly scored this way, show an inconsistent pattern of predicting higher-order cognition. On the other hand, complex-span tasks (Daneman & Carpenter, 1980; Turner & Engle, 1989) show consistent and often sizable correlations with higher-order cognition (Unsworth & Engle, 2007a). Complex-span tasks add a degree of challenge by requiring participants to perform a processing task in between the presentation of each to-be-remembered item. For example, the operation span task (ospan; Turner & Engle, 1989) interleaves the presentation of each to-be-recalled item with a simple mathematical equation that must be solved. This secondary task causes to-be-remembered information to be removed from the focus of attention. Each time such an action occurs, a process of search-and-recovery is required to retrieve the needed information from inactive memory. It has been argued that the effectiveness of this process is what differentiates high and low WMC individuals on memory task performance (Unsworth & Engle, 2007a; 2007b). Unsworth & Engle (2007a; 2007b) have argued that simple span tasks at supra-span lengths that require search of inactive memory account for the same variance in higher order cognition as complex spans.

To be clear, many different tasks can be used to measure WMC, the critical component is that the task challenges the limits of immediate awareness. It is at this boundary that accurate recall requires controlled, effortful cognition. For instance, Unsworth & Engle (2007a) recently demonstrated that both simple- and complex-span tasks reflect the mechanisms of WM that are critical to higher-order cognition (e.g. memory updating, maintenance, and controlled processing). However, the association becomes apparent in different ways. Simple span tasks were shown to be maximally predictive of general fluid intelligence (Gf; i.e. novel reasoning ability) when the lists had a supra-span length of at least 5-6 items, thus exceeding the 4±1 items (Cowan, 2001) that can be held in the focus of attention. People attempting to recall supra-span lists will lose access to some items requiring a search of recently-active-but-now-inactive memory. It is in these cases that accurate recall of items in their proper order begins to reflect aspects of WM that are linked with controlled higher-order cognition. By such logic, complex-span tasks should reach their predictive potential with shorter lists, as the interleaved

processing component of these procedures instigates this process from the start. This is indeed the case. Unsworth and Engle (2007b) found complex span tasks to be maximally predictive of Gf with lists of only two items.

Adaptively training WMC

From the perspective of the above theory, it is easy to understand how "adaptive" training programs might facilitate changes to the functionality of WM. As with simple-span tasks, the typical adaptive-span task requires individuals to remember lists of items such as letters/numbers (verbal WM), or spatial sequences/movements (visuo-spatial WM) and repeat the items in a pre-specified order (generally forward, although sometimes backward). The difference is that adaptive-span tasks actively adjust list length to find the point at which an individual experiences difficulty in recalling items. If the user becomes more accurate at recalling lists of this length, the sequence is extended. This action provides a constant stress on the boundaries of WMC. If, as the quote that began this article asserts, WM is like a muscle, it is under these circumstances that one would expect strengthening.

Many of the studies reported here utilize commercially available versions of adaptive-span tasks such as Cogmed Working Memory Training program (Cogmed, 2006) or JungleMemory (Alloway & Alloway, 2008), while others have developed software in-house (see Tables 1 & 2, p. 276, for a summary). These in-house programs tend to be varied in purpose and function and thus will be discussed in turn. Most adaptive-span-researchers assume that adaptive-span training affects relatively general mechanisms of cognitive control such as the ability to maintain and manipulate information over short periods (e.g. Kingberg, et al., 2002; Holmes et al., 2009), the processing capacity of domain-free attention (e.g. Chein, & Morrison, in press; Jaeggi et al., 2008), or acquisition-and-retrieval of new information (Alloway & Alloway, 2009). As a product of this sustained demand, it is assumed that a multitude of tasks which tap these broad abilities will also benefit. A minority of studies (Dahlin et al, 2008a, 2008b; Persson & Reuter-Lorenz, 2008) begin with the premise that the effects of training will be rather specific and the benefit of stressing WM will be limited to tasks which share certain critical features (e.g. both tasks require memory to be updated regularly).

Minimum Criteria for Claiming Cognitive Abilities Have Been Altered

Barnett and Ceci (2002) reviewed the literature on cognitive training and admonished the field that training programs must attempt to clarify the meaning of "transfer of training" to other tasks or situations. For the current

discussion we use "near-transfer" to refer to increased performance on tasks that are highly similar to the tasks on which training occurred. Higher scores on simple-span tasks following training on an adaptive n-back task would be an example of near-transfer. This type of transfer is quite common and will only be discussed as necessary. More relevant to the notion that WMC-training will benefit cognitive function is the concept of "far-transfer." By far-transfer, we refer to post-training performance improvements on tasks which are not of the same nature or appearance as the tasks on which participants have been trained. Improvement on measures of fluid abilities such as the Raven or measures of attention control such as the Stroop task following adaptive-span training would be an example of far-transfer. The point that is critical to far-transfer is the assumption that two different tasks share an underlying processing component (e.g. both tasks require memory updating, conflict resolution, or attention control) that is tapped regardless of performance circumstances or visual context.

It should be noted that the primary focus of the current article is the claim that training WMC will have benefits to aspects of people's lives, beyond making them expert performers of a given task (Chase & Ericsson, 1982). For ease of exposition, most of our discussion revolves around far-transfer effects. The reader should not assume this to be a tacit acceptance that performance on near-transfer tasks necessarily signals a change in the general capacity of working memory. This point will be expanded in the General Discussion.

Validity of a training experiment

The studies reported here are viewed in relation to what Campbell and Stanley (1963) refer to as the "pretest-posttest control group design." This type of experiment involves (1) random assignment to training and control groups, (2) an initial measurement of the abilities of interest (pretest), (3) some form of intervention that differs between groups and (4) a final measurement of the abilities of interest (posttest). The primary reason for using such a design is that it protects against threats to internal validity: Researchers who properly utilize this design can safely assume that the changes in performance occurring between the pretest and the post test were caused by the experimental manipulation. Specifically, *randomly* assigning participants to an *experimental* and *control* group eliminates the internal confounds of history, maturation, testing, instrumentation, regression to the mean and the interaction of participant selection with any of these variables.

- *History* refers to specific events that occur between the pre- and post-test, which might otherwise explain a change in scores. For instance, the

act of going to school might have an effect on measures of intelligence. Having a randomly-assigned control group protects against this possibility, particularly when participants are a mix of community members and students. Campbell and Stanley (1963) note that the simple act of having a control group does not control for history, as experimental sessions have events within them. When experiments are conducted in group settings, these authors advise the sessions contain a mixture of control and experimental participants to cancel out the effects of random events.

- *Maturation* refers to changes in participants that occur simply with the passage of time. A critic of a training study might challenge that, rather than attributing improved test scores to training, aging (in the case of children) or time since disturbance (e.g. stroke victims) may serve as reasonable explanations. However, the control group provides protection against these hypothetical events by serving as a baseline for changes that occur, absent of training.
- *Testing* refers to the potential for pretesting to lead to inflated posttest scores, as participants have been sensitized to the nature of test materials. Inclusion of a control group allows the experimenter to examine performance changes above and beyond this effect.
- *Instrumentation* refers to changes that may occur in a given instrument's calibration. Here we specifically refer to the possibility that the particular cognitive construct that a transfer-task initially measures (e.g. attention, intelligence, focus) might change with repeated administrations. The pretest-posttest design prevents a change-in-instrumentation from serving as an explanation for an increase in test scores. In other words, inclusion of a control group ensures that, regardless of what ability a posttest is measuring, between-group difference in test score increases can be attributed to the training program.
- *Regression to the mean* is the tendency for participants who scored extremely low at pretest to score somewhat better at posttest while participants who scored extremely high at pretest tend to score somewhat lower at posttest. Regression to the mean is most easily understood as a statistical artifact that arises from pretest and posttest scores sharing a less than perfect correlation (Cohen, Cohen, West, & Aiken, 2003). Assume a score can range from -1 to +1 and that the correlation of the pre- and posttest is .75. It can be easily demonstrated that the extreme scores will be predicted to show the largest pretest-posttest movement toward the mean score (e.g. posttest score prediction = pretest score × .75). As this change is independent of the experimental manipulation, randomization should ensure it will be present in both the training and control group.
- *Interaction of selection* with any other experimental variable. An experimental manipulation may have different effects on children with ADHD

than it does on young college students. Assuming this potential interaction is not of interest to the study, it may be cancelled out through random assignment of participants to conditions.

As has already been mentioned, the simple inclusion of a control group is not what protects the internal validity of an experiment. This design requires the control group to be treated in as similar a manner to the training-group as is possible. A common violation of this necessity occurs within the adaptive-training literature when "no-contact" control groups are used. Because no-contact control groups only interact with the experimenters at pretest and posttest, use of these groups does not constitute a valid pretest-posttest experiment, as history has not been controlled. That is, the training-group's contact with the experimenters may serve as a viable explanation for their change in scores. To illustrate this point, consider a recent report by McCarney et al. (2007) examining the tendency for people's task performance to change in response, not only to the knowledge that they are being observed, but to the amount of attention they are given. This is the classic ***Hawthorne effect***. In the McCarney study, two groups of dementia-diagnosed adults were given placebos but were led to believe the drugs were Ginko biloba supplements that would serve as a memory aid. In addition, the participants were tested on assessments of cognitive functioning and quality of life every two months for six months. The critical difference was that one group received rather comprehensive assessments while the second group was only fully examined at six months. At the time of posttest, the comprehensive-assessment group tested as being higher in cognitive function and reported a higher quality of life. Neither group knew they were taking placebos and the experimenters were blind to this as well. In effect, the only aspect of the experiment that can account for the between-groups difference is the extra attention one group received.

In terms of training studies that utilize no-contact control groups, such a phenomenon may manifest itself at posttest by better performance by the training-group participants, who in some cases have already been measured 20-30 times by the experimenters. In addition, there is the concern about demand characteristics (Orne, 1962) which refer to the tendency for participants to behave in accordance with the perceived expectations of the experimenter. In this case, members of a no-contact group can readily recognize that they are not receiving treatment and are thus not expected to show improvements. Again, effort at posttest provides a plausible alternate explanation to training effects.

Although a researcher may provide reasonable justification for not believing that Hawthorne effects or demand characteristics are present, we again state that history is not controlled. As such, critics have a built-in reason for

doubting experimental outcomes. In short, use of a no-contact control group guarantees that a study must be replicated with appropriate controls before it can be said to have internal validity (see Tables 1 & 2 for a summary of control groups used in the reviewed experiments).

Generalizing the results of a valid experiment to cognitive abilities

Although it is important to be able to state that a training intervention caused an increase in test scores, the true goal of training is not to change test scores. The goal of training is to change the generalizable cognitive abilities of the individual. Therefore, one must be able to confidently state that a change in scores reflects a change in the intended faculty. Within the adaptive-training literature, the most prominent threat to this type of generalizability is an assumption that the results of single tasks (e.g. Raven's Progressive Matrices; RPM, [Raven, 1995b]; Stroop task) can be interpreted as pure measures of abstract hypothetical constructs (e.g. Gf; attention).

Kim and Mueller (1978) point out that a test score can be decomposed into two sources of variance. First is the variance that this test shares with other purported tests of the same cognitive ability (e.g. other tests of mathematical skill, or other tests of reasoning). This is the information that is utilized when performing a factor analysis. It is of primary interest to the researcher. The second source is unique or error variance. This is variance that is not shared among measurements of a given ability and is removed by factor analysis.

Error variance may be further dividend into two components (Kim & Mueller, 1978). First is random error, which is due to random measurement issues. When the above guidelines for obtaining internal validity are followed, random error cancels out with large samples. The second component is unique or systematic error. To say this is error is to say it reflects aspects of a task that are not related to the ability of interest. For instance, the Stroop task is intended to measure attention. However, performance on the Stroop should also be affected by the acuity of one's color perception as well as one's knowledge of a given language. A colorblind participant will experience difficulty in distinguishing between red and green hues. The word-hue interference can only arise in people who are literate in a given language. In neither case do these performance-affecting phenomena reflect attention. Therefore, "attention" is not performance on the Stroop task. Attention is performance that is common to the Stroop task *and* the antisaccade task (Hallett, 1978), *and* flanker tasks (Eriksen & Eriksen, 1974) *and* the dichotic listening task (Cherry, 1953; Moray, 1959).

As a particularly relevant example, consider matrix tasks such as Raven's Progressive Matrices (RPM; Raven, 1995b) and BOMAT (Hossiep et al., 1999). These tasks present a series of abstract figures and require partici-

pants to select from among several options the one that logically completes the sequence. The puzzles are arranged such that they increase in difficulty with each answer given. The novel reasoning required by these types of tasks makes them valid predictors of Gf. Jensen (1998) estimates RPM to have a correlation of about .80 with general intelligence. If that estimate is accurate, roughly 64% of the variability in people's performance of RPM can be attributed to general intelligence. While this is a strong relationship, it illuminates the reality that performance on this task is multiply determined. If RPM is the only intelligence measure used in an experiment, 36% of the variance in what is referred to as "Gf" will be accounted for by other factors (e.g. response speed or spatial memory). Thus, even when training can be said to have caused an increase in RPM scores, the degree to which this effect reflects a change in Gf or a change in some other variable remains unknown.

Moody (2009) has noted that these non-Gf factors likely make their largest contribution to performance on test items that do not present a particular challenge to reasoning, often early items. Of course, test-retest effects would extend this possible confound to other items as well. Moody elaborates that if training has equipped one group of participants with the ability to better keep track of the matrix items in memory (as opposed to scanning the display repeatedly), they might advance through these low-Gf items more rapidly than the control group. When time constraints are present, the trained group may show relatively higher Gf scores, simply because early responding provided them with a longer period to solve the difficult items. Thus, although their reasoning abilities have not necessarily changed, the trained participants show relative improvement on the test.

Due to such complications, generalizability of results requires experimenters to use multiple tests for each ability of interest. Although it might not always be possible to extract factor scores (thus doing away with all non-construct error), combining the results of multiple measures into a single composite score will help cancel out unwanted task-specific changes, thus leading to more robust measures. Unfortunately, within the training literature, this practice is quite rare.

A Review of Recent WMC-Training Studies

Now that a degree of context has been established, we review the adaptive WMC-training literature. It was previously stated that this is a rather diverse literature. As such, the review has been divided into therapeutic and non-therapeutic research (see Tables 1 & 2 for further subdivisions).

WMC training as a therapeutic device

Children with ADHD and/or low WM

Adaptive-span research showed early promise when Klingberg et al. (2002) demonstrated its possible therapeutic value in treatment of ADHD. In this study 14 children who had previously been diagnosed with ADHD were divided into two groups. The adaptive-task group trained for 5-6 weeks on adaptive visuo-spatial, backward-digit and letter-span tasks as well as a reaction time task. The second group spent the same period of time performing non-adaptive visuo-spatial, digit- and letter-span tasks.

Relative to the control group, the adaptive-task-trained children showed significant gains in performance on pre-tested measures of reasoning (Raven's Coloured Progressive Matrices; RCPM; Raven, 1995a) and attention (Stroop accuracy). Furthermore, children trained on the adaptive task showed a sharp drop in the number of head movements recorded during the performance of a separate task. Thus objective evidence was provided that the therapy was alleviating ADHD-specific symptoms.

Klingberg et al. (2005) attempted to extend these findings with a larger sample (53 ADHD diagnosed children) and a second post-training examination which occurred 3 months after completion of training. Although the posttest that immediately followed training provided a replication of the previous study with the training-group showing relative gains on RCPM and Stroop accuracy, these effects were nonexistent at the 3-month follow-up. Klingberg et al. note the lack of group differences at the second post-training examination was not due to scores regressing. Rather the training-group had little room for additional improvement on RCPM or Stroop at the initial posttest. By the 3-month exam, their scores remained steady while the control group showed improved scores. Perhaps more meaningful to the therapeutic perspective, at neither post-training examination did the adaptive-training and control groups show any reduction or differences in the number of head movements made. Klingberg et al. did report subjective measures of ADHD symptoms which were provided by the children's parents and teachers. These scores were split, with parents of training-group children reporting a significant drop in symptoms (relative the reports of control group parents), but the teachers reporting no change. Thus it is difficult to argue that this study provided conclusive evidence that adaptive-span training reduced ADHD-specific behaviors.

Recently Holmes et al. (2009) examined the conjunctive effects of adaptive-span training and "stimulant medication" on the WMC of children who had a previous diagnosis of ADHD. Changes in working memory were assessed via the Automated Working Memory Assessment (AWMA; Alloway,

2007) which provides several measures of verbal STM, visuo-spatial STM, verbal WM and visuo-spatial WM. For present purposes, analysis of changes in these scores is complicated due to the researchers' inconsistent administration of these measures at each of four testing sessions. For the sake of simplicity, we reference only the subscales that were used in each session (one of each type). Far-transfer effects on IQ/reasoning were examined via Wechsler Abbreviated Scales of Intelligence (WASI, Wechsler, 1999).

As a baseline for the effect of medication on WMC, children in this study were pre-tested twice: Once after being off medication for at least 24 hours and again after returning to it. Only visuo-spatial WM showed improvement in response to a return to medication. After completion of 20-25 sessions of adaptive training, a third examination occurred, at which all four of the STM/WM tests showed significant increases relative to the most recent assessment. Finally, a 6-month follow-up revealed these changes to be fairly stable as only the verbal STM measure showed signs of regression.

From a far-transfer standpoint, these performance changes were not accompanied by any signs of change in IQ. These scores remained stable through the first three examinations and the WASI was not administered at the 6-month follow-up. Holmes et al. (2009) interpret these results as revealing an effect of adaptive-span training on the WMC of ADHD children above-and-beyond that of stimulants. However, this interpretation is limited by lack of a control group to account for test-retest effects in AWMA scores as well as the lack of significant change in performance of a far-transfer task (which would have validated the meaning of increased AWMA scores).

A separate study by Holmes, Gathercole, & Dunning (2009) involving children who had pretested as low in WMC, avoided one of the above complications by including a control group who performed a non-adaptive version of Cogmed software. Following a minimum of 20 sessions, the training-group did in fact show larger relative gains on AWMA scores, as well as near-transfer to a WM task that involved memory for instructions. However, this was not accompanied by far-transfer effects to WASI scores, nor to measures of word reading (Wechsler Objective Reading Dimensions reading subtest; WORD) or mathematical reasoning (Wechsler Objective Number Dimensions subtest; WOND). A 6-month follow-up revealed that the training-group's AWMA scores were mostly intact, along with some signs of increased IQ. Unfortunately, these researchers (Holmes, Gathercole, & Dunning, 2009) reintroduced the confounds of the previous study (Holmes et al., 2009) by excluding the control group from the follow-up. Therefore the limited IQ increases might be attributed to confounds such as test-retest effects or general childhood maturation.

Alloway & Alloway (2009) take a somewhat different philosophical approach in assuming that the far-transfer benefits of adaptive-span training

will be revealed though improved storage-and-retrieval of new information over the long term. Fifteen children who had been identified as having "learning disabilities" were given pre- and post-training tests pertaining to learning and academic attainment (vocabulary and numerical operations subtests of the WASI, respectively) as well as one test of verbal WM. Adaptive training with 8 children involved components of the commercially available JungleMemory Training Program (Alloway & Alloway, 2008). These tasks involved (1) memory for and later use of word endings, (2) mental rotation of letters and (3) sequential memory of mathematical problem solutions. The control group (7 children) received what was described as "targeted learning support."

Relative to the control group, the adaptive-training-group did have statistically higher scores on the three transfer tasks following 8 weeks of training. However, given the small sample size, confidence in this difference would be greatly aided by demonstration that the trained group showed significant within-group increases in scores from pretest to posttest. Unfortunately, Alloway & Alloway (2009) do not provide an analysis of either group's pre- vs. posttest scores, nor do they provide raw data.

Discussion of children with ADHD and/or low WM

It is worth noting that the results of Klingberg et al. (2005) provide an example of the potential complications associated with the use of single tasks to measure entire mental constructs. These researchers reported a disappearance of far-transfer effects at a three month follow-up, due to increased control group scores. While this does not bring into question the role of the intervention on the training-group's initial increase in scores (i.e. the experiment was internally valid), the fact that the control group eventually closed the gap does raise questions about whether the far-transfer tasks remained appropriately difficult across repeated administrations. It has already been mentioned that when test items do not present a challenge to the mental ability of interest, unintended factors (e.g. faster responding based on strategic use of spatial memory) can begin to drive performance. Certainly, this was the case by the third assessment (unless the cognitive abilities of the control group spontaneously improved). As such Klingberg et al. (2005), have little evidence on which to base the conclusion that the training-group's early improvements in test scores represent long-term changes to cognitive abilities. This concern may have been mitigated through more extensive testing of Gf using multiple measures.

A more serious challenge to any claim that adaptive-span training affects reasoning or learning ability relates to the lack of far-transfer in the studies of Holmes et al. (2009) and Holmes, Gathercole, & Dunning (2009). In Holmes

et al., IQ was measured via WASI which includes multiple measures of non-verbal reasoning ability and learned verbal ability, while Holmes, Gathercole, & Dunning (2009) additionally included a reading subscale of the WORD and a mathematical subscale of the WOND. Thus, the measures used in these studies provide a much more robust picture of changes to reasoning and/or learning abilities that can be expected to follow adaptive-span training. To recapitulate the findings of these two studies, the only signs of far-transfer occurred at the 6-month follow-up of Holmes, Gathercole, & Dunning. These increases neither directly followed training, nor were compared to a control group, thus can be attributed to virtually all of Campbell and Stanley's (1963) threats to internal validity. This is a stark contrast to the results of the previously mentioned studies that reported far-transfer to Gf (Klingberg et al., 2002; Klingberg et al., 2005) or to learning ability (Alloway & Alloway, 2009) using single measures of the constructs.

As for alleviation of ADHD-specific symptoms, little objectively collected data has been reported. Of the three studies involving ADHD, only two included measures of the ADHD symptoms that might have been alleviated by training (e.g. Klingberg et al, 2002; Klingberg et al., 2005). Of those two, only Klingberg et al. (2002) demonstrated a quantifiable decrease in an ADHD-related behavior. Holmes et al. (2009), on the other hand, chose not to include a specific measure of ADHD-related behavior. In this case, one might argue that WM dysfunction is contributing to the children's cognitive difficulties and improved AWMA scores reflect the first step to correcting the issue. It is therefore worth noting that low-WMC was not among the criteria for inclusion in this study.

Adaptive-span training and recovery from stroke

Recently, Klingberg and associates have expanded their research interests to include Cogmed as a tool for stroke recovery (Westerberg et al., 2007). In a pilot study, 9 patients performed 5 weeks of at-home adaptive-span training along with weekly feedback from a certified psychologist via telephone. The no-contact control group (9 patients), on the other hand, had no training or feedback for the 5 week period.

Far-transfer results in this study are mixed. The adaptive-span group did show performance increases that were significantly larger than those of control patients on PASAT (rapid summation of numbers; Gronwall, 1977) and Ruff 2&7 (searching for the numbers 2 and 7 among various distractors; Ruf et al., 1992). Westerberg et al. (2007) interpreted this as evidence of improved attention. However, this interpretation is contradicted by a lack of training effect for the Stroop task, which served as a third measure of attention. RPM was the only measure of transfer to reasoning abilities. As with the

Stroop task, no change in performance was found.

Additionally, far-transfer was absent for two memory tasks. The first task repeatedly presented participants with a fixed sequence of 10 words until they performed one correct recall (maximum of 10 presentations). The second test, which required participants to attempt to free-recall as many of these 10 words as possible, occurred 30 minutes later. The treatment and control groups did not show differential improvement for either task. This is a curious finding to which we will return in the General Discussion.

Discussion of adaptive-span training and recovery from stroke

Focusing on stroke-related symptoms, the intended effect of adaptive-span training is unclear. Researchers did include a questionnaire on which participants rated their daily cognitive functioning. However, the design of this experiment makes this measure a particular candidate for a Hawthorne-type effect (e.g. McCarney et al., 2007). The training-group performed daily training sessions, and were given weekly progress reports by a psychologist. The no-contact control group, on the other hand, did not perform a daily task and was contacted by the researchers for pre- and posttests only. For subjective measures (such as self-report of cognitive function) to be valid or meaningful, it would be preferable that members of the control group not be treated differently.

Beyond that measure, objective stroke-related symptoms are not directly addressed. Perhaps most critical to a study (Westerberg et al., 2007) which states that "stroke-induced deficits in WM and attention are often severe and result in impairments to vocational performance and social functioning (p. 21)", low WMC was not listed among the inclusion/exclusion criteria. Additionally, direct comparisons of this variable were not drawn between the included participants and a healthy population. In light of this, it is worth noting that Westerberg et al. (2007) point out that lack of far-transfer on Raven's performance was likely due to a ceiling effect. Participants averaged 15.5 out of 18 correct answers at pretest, indicating there was little room for improvement (similar to Klingberg et al., 2005). Coupled with reported mean pre-test IQ score of 102, these results indicate that, at least from the perspective of reasoning abilities, these patients were not impaired. As with Holmes et al. (2009), the line between the cognitive abilities that are being trained and objectively measured symptoms of the disorder is not apparent.

WMC training and healthy populations

WMC training in healthy populations (see Table 2) involves adaptive-span training that is focused on both on behavioral research and the physiological changes that accompany training. An additional line of research can be categorized as attempting to train the capacity of WM not through adaptive-span training but through training based on adaptive-memory-updating. This research involves many differences in technique and philosophy and thus will be considered separately.

Adaptive-span training and healthy populations: Behavioral research

To our knowledge the only published attempt to train healthy participants strictly using adaptive complex-span tasks is that of Chein and Morrison (in press). Two tasks were used in the training procedure. A verbal WM task required participants to remember a series of letters. In between the presentation of each letter, participants were required to make word/non-word judgments on strings of letters (e.g. "brick" = word; "blick" = non-word). A spatial WM task required memory for several positions on a 4 × 4 grid. After each position was revealed, participants were required to judge the symmetry of a picture presented within an 8 × 8 grid. Sequence lengths were automatically adjusted based on accuracy. Training occurred 5 days a week for four weeks.

Relative to a no-contact control group, adaptive-training participants did show significant increases in "temporary-memory" scores (a composite of the two trained complex-span tasks along with simple-span versions of each). However, neither group showed increased performance in Gf (Raven's Advanced Progressive Matricies; RAPM) and did not have different increases in spatial-reasoning (ETS Surface Development and Paper Folding tests; Ekstrom et al., 1976) abilities. One-tailed t-tests revealed that the training-group did show a larger decrease in Stroop interference, relative to the control group, as well as a relatively larger increase in reading comprehension scores. Chein and Morrison (in press) conclude that these last two findings are consistent with the view that their training program had an effect on domain-general attention, however this interpretation is complicated by the results of the between-groups, repeated-measures ANOVAs the researchers performed on their data. These analyses did not show significant interactions between group (training vs. control) and assessment time (pretest vs. posttest) for either Stroop ($p = .13$) or reading comprehension ($p = .08$). Such results constrain confidence in the meaningfulness of the planned comparisons.

Another somewhat unique study was that of Shavelson et al. (2008) which is one of the few reported here to feature untrained complex-span tasks in

measurement of near-transfer (ospan; reading span; Daneman & Carpenter, 1980). These researchers randomly assigned 37 middle school children (mean age 13.5) to perform 25 sessions of adaptive Cogmed training or a non-adaptive control version along with computerized science lessons. Other near-transfer tasks involved a digit-based simple span task along with Cogmed's span-board task, which is described in the supplemental materials of McNab et al. (2009) as requiring participants to replicate a sequence of cube illuminations, presented via computer monitor. RPM served as the only measure of far-transfer (e.g. Gf).

The results of this study are straightforward. The training-group showed significant improvement in the simple-span tasks (digit-span and span-board) above and beyond that of the control group, but these benefits did not extend to the complex-span tasks or RPM. In this case, adaptive training led to no discernable signs of transfer, beyond tasks which were highly similar to those on which participants were trained.

Children with learning disabilities tend to be the focus of adaptive-span training in younger populations. However, the study by Thorell et al. (2009) does provide data on healthy preschool children. Participants from four schools spent five weeks either performing adaptive-span training, adaptive-response-inhibition training, playing commercially available video games or serving as passive controls. The inclusion of an inhibition-training-group allowed researchers to explore the possible pliability of cognitive mechanisms beyond WMC. This group performed three specially designed training tasks involving go/no-go (e.g. response only made to specific stimuli, otherwise response withheld), stop-signal (e.g. inhibit an ongoing response; Logan & Cowan, 1984), and flanker tasks (e.g. respond to the middle of 5 stimuli, ignore the others; Eriksen & Eriksen, 1974). Difficulty was adjusted through time allowed to respond.

In terms of inhibition training, the results of this study were quite clear. Relative to the control groups, the adaptive-inhibition group showed no signs of transfer to performance of other tasks. The adaptive-span group, on the other hand did show far-transfer in the form of fewer omitted responses during the performance of two attention tasks which required monitoring for specific stimuli (continuous performance task and go/no-go). Neither group showed improvement in a Stroop-type task, nor in a task which required withholding inappropriate responses (go/no-go), in a problem solving task (block design subtest of WPPSI-R; Wechsler, 1995) or in overall reaction time when performing the go/no-go task.

A final study reporting healthy-population data was reported in Experiment 2 of Klingberg et al.'s (2002) ADHD article (Experiment 1 was described above). After an average of 26 days of adaptive-span training, 4 healthy college students did show test-retest improvements in both reasoning

and attention as measured by the test battery (RAPM and Stroop). However, the generalizability of these effects is called into question by the decision of Klingberg et al. to compare the results relative to the placebo-trained ADHD-diagnosed children from Experiment 1. The researchers argue that, as change scores (rather than raw scores) were examined, pre-test differences are inconsequential. Regardless, beyond the small sample-size and non-random assignment, the incompatible nature of the control group (both in age and, ostensibly, cognitive functioning) complicates any attempt to generalize these data to the population at large.

Discussion of behavioral research

Of the studies reviewed, far-transfer in non-preschool populations is restricted to the experiment of Chein and Morrison (in press) who found some signs that adaptive-span training benefits to attention and reading comprehension. However, as the results of the main analyses did not reach significance, these findings should be followed up in future research. Beyond this study, Shavelson et al. (2008) reported no transfer to performance of complex-span tasks or RPM, while Klingberg et al. (2002; Experiment 2) is complicated by the use of inappropriate controls.

Thorell et al. (2009) did find a potential benefit of adaptive-span training as the trained preschool children showed signs of improved attentional focus in the form of fewer omitted responses to the appearance of specific stimuli. It is worth noting that in this study, participants were not randomly assigned to specific training programs, rather four different schools each received one of the interventions. The children at the school that received adaptive-span training also committed more pretest-omissions on these tasks than the children at any of the other schools. At posttest, their performance had essentially pulled even with the other schools. As such it is difficult to tell whether their improvements represent true training-based improvement or a regression-to-the-mean effect. A replication with true randomization would eliminate this confound.

Adaptive-span training and healthy populations: Physiological research

Changes in WM-related brain activity following adaptive-span training were examined in a 2004 fMRI study, by Olsen, Westerberg, & Klingberg. Two studies involved training participants (n = 3 and 7, respectively) on adaptive Cogmed tasks for several weeks in between scans. Across sessions, increased activation when performing a WM task (relative to a control task) was reported in the frontal and parietal cortices. The increased parietal activity replicated across studies, however, right-frontal activation increases were

reported in Experiment 1 and increased left-frontal activation was seen in Experiment 2. Among several possible explanations for this change, Olsen et al. offer that it may be due to participants being trained on WM tasks which contained verbal components in Experiment 1 (adaptive letter and digit span-tasks), while Experiment 2 training involved only visuo-spatial tasks.

Behavioral data, relative to an 11 person no-contact control group, were also reported. For Experiment 1, Olsen et al. (2004) report that the trained group showed significant improvement on the span-board, Stroop task and RAPM, but do not inform the reader whether this was relative to the control group or a test-retest effect (a second analysis of this data [Westerberg & Klingberg, 2007] was similarly vague). In the second experiment, the trained group showed test-rest improvements on all transfer tasks (near-transfer: span-board, digit-span; far-transfer: Stroop task), but only showed gains above the non-trained control group on the Stroop task (i.e. far-transfer only). Performance on the in-scanner WM task (on which the physiological changes were based) only showed improvement in Experiment 2.

In a more recent study McNab et al. (2009) utilized fMRI and positron emission tomography (PET) to identify five cortical regions of interest and examine changes to the density of dopamine D1 and D2 receptors, following five weeks of adaptive-span training. As with Olsen et al. (2004), changes were relative to early scans, rather than relative to a control group.

Results of McNab et al. (2009) indicate an effect of training on D1 receptors in select regions of right ventrolateral prefrontal, right dorsolateral prefrontal and left and right posterior cortices. Data were interpreted in terms of a negative, linear correlation such that decrease in D1 binding potential was associated with increased scores on a composite of five near-transfer memory tasks. However, the exact nature of this association remains somewhat vague, as the statistical model that was actually fitted to the cortical areas of interest (a more complex quadratic relationship) was not given thorough discussion. Of relevance to the current discussion, far-transfer tasks were administered (RPM and an attention task), but the results were not reported.

Discussion of Physiological research

From a physiological perspective, the results of Olsen et al. (2004) and McNab et al. (2009) may prove to be valuable models of brain-related changes following extended performance of a given activity. However, the significance of these findings to behavioral changes following WMC-training is unclear. Foremost, neither study included a control group in their physiological assessment, meaning that cortical changes that are the results of (1) changes to WM in general, (2) training-specific learning, and (3) test-retest effects cannot be dismissed.

Test-retest effects are particularly problematic for Experiment 2 of Olsen et al. (2004). In this study participants did show improved performance on the WM task on which the physiological data was based. However, this performance was not relative to a control group, therefore, any claims of training-related changes to general WM-functionality requires additionally accounting for the lack of near-transfer to non-scanner WMC tasks (which were analyzed relative to a no-contact control group). In short, attributing the physiological data to general changes in WM is hindered by equivocal behavioral results.

McNab et al. (2009) did show a statistical relationship between physiological changes and behavioral performance. However, a closer look at their data reveals limitations in the statements that can be made about the relationship of this finding to general cognitive function. Although the researchers report a non-significant correlation between D1 binding potential and performance on WMC tasks before training ($p = .08$ for a quadratic regression model), it is not reported whether training strengthened or attenuated this relationship. Rather, the regression model they use reports only how *change* in D1 binding potential following training predicts a *change* in WMC scores. As such, the data could represent a change in general cognitive functioning or it could represent changes that accompany extensive experience with a specific WMC task. This concern may have been assuaged had the results of the far-transfer tasks been reported. Unfortunately this information was not made available in the main report or the supplemental materials.

Adaptive-training of WM updating

Another developing line of research in WMC training can be categorized as "memory updating". Some researchers have begun their exploration under the assumption that constant updating of information is heavily taxing on limited-capacity attention. As described by Jaeggi et al. (2008), tasks which rely on attention (such as those which measure WMC and Gf) should benefit from strengthening the system. Other researchers (Dahlin et al., 2008a; Dahlin et al., 2008b; Persson & Reuter-Lorenz, 2008) take a more conservative approach. Rather than looking for general-resource-based transfer, these investigators have focused on mechanisms which are specific to the performance of memory updating tasks.

The training-task used by Jaeggi et al. (2008) is based on the n-back (Kirchner, 1958) which requires participants to attend to a constantly changing stream of information (e.g. letters, spatial locations, etc.) and make a specific response each time the currently presented item was presented n-items-ago. In the task of Jaeggi et al., *n* is adjusted adaptively in response to performance. An additional wrinkle required participants to divide their

attention between two streams of information: the location of squares on a computer monitor and letters that were presented in an auditory manner.

According to the predictions of Jaeggi et al. (2008), this constant taxing of attention should lead to pretest-posttest improvements in both WMC and Gf, relative to an untrained control group. This hypothesis was only half confirmed, as the trained group did not show a differential increase in performance of a complex-span task (reading span; Daneman & Carpenter, 1980) which had been included as criterion measure. Gf (as measured by the BOMAT), on the other hand, showed an apparent dose-dependent effect. A group that was trained for 12 sessions showed marginal differences in Gf scores relative to the untrained control group. However, groups which had been trained for 17 or 19 days showed significant differences, relative to the no-contact control group.

While this increase in BOMAT scores cannot be attributed to an increase in the span of WM, a recent report by this group (Studer et al., 2009) also eliminates the need for a dual-task component. Again utilizing a no-contact control group, Studer et al. included two experimental conditions. One group was trained on the dual n-back of Jaeggi et al. (2008) while the other group performed a single-task n-back. Following 20 training sessions (dose effect was not examined) both groups showed significant transfer effects on both RAPM and BOMAT.

Dahlin et al. (2008a) took a less general approach to training, making the assumption that training on an updating task should only transfer to other tasks which require an act of updating. Furthermore, they reasoned that the behavioral overlap should show common brain activation that is not shared with non-updating tasks. Training involved several variations of the running-span task (Pollack, Johnson, & Knaff, 1959) in which participants needed to remember the most recent four items from lists of varying lengths. The adaptive feature of these tasks was the length of the list. At the easiest level, list length varied between 4-7 items, at the hardest level it ranged between 5-15. Participants also trained on the keep-track task (Yntema, 1963) which requires memory for the most recently presented instances of pre-specified categories within an ever changing list of exemplars. Brain activation changes were judged relative to the control group; however as with Jaeggi et al. (2008), the control group was no-contact.

Results of the first experiment were in line with the predictions. The trained group showed significantly greater relative increases on a 3-back version of the n-back task, but not on the Stroop. fMRI scanning revealed activation in the left striatum that was common to both the trained task and the 3-back, but absent during performance of the Stroop. Moreover, this common activation for updating tasks was absent in a second experiment involving older adults, as was far-transfer to 3-back performance. The authors (Dahlin et

al., 2008a) interpret these results as indicative of the striatum's role in updating the contents of WM (Awh & Vogel, 2008; McNab & Klingberg, 2008; Monchi, Ko, & Strafella, 2006), a process which is ostensibly less relevant in Stroop performance.

These limited-transfer results were replicated in a separate study (Dahlin et al., 2008b) using the same training program. Far-transfer was seen in the n-back as well in a memory task which required three trials of the free recall of 18 words. Only words that were not recalled on one trial were re-presented on the next (Buschke, 1973). Importantly, transfer was not seen for several memory-span and reasoning tasks. As with Dahlin et al. (2008a), transfer effects were not seen for older adults.

Discussion of adaptive-training of WM updating

Taken in isolation, the results of Jaeggi et al. (2008) and Studer et al. (2009) indicate that repetitive performance of the n-back leads to increased scores on measures which reflect, among other things, general reasoning ability. Taken together critical shortcomings emerge. First, as WMC scores did not change in the study of Jaeggi et al. (2008), and were not measured in the Studer et al. (2009), the mechanism of change that is brought about by n-back training remains unclear. Second, these researchers have yet to show that the Gf-transfer effect exists above and beyond that which might be expected from a control group that performs a non-adaptive n-back task and thus has the same amount of contact with the researchers as does the training-group. Finally, Gf is measured through performance on single tasks, rather than composite scores which are less vulnerable to task-specific error. The previously discussed criticism that experience with adaptive-WM tasks may train non-Gf aspects of matrix tasks was initially formulated by Moody (2009) in response to Jaeggi et al. (2008) and particularly references a decision by the authors to only allow 10 minutes for performance of the BOMAT, rather than the typical 45 minutes. Moody notes that the BOMAT requires keeping track of a sequence of 15 items, and thus seems particularly suited to show an effect of improved memory or improved memory strategies when performed under strict time constraints.

Dahlin et al. (2008a, 2008b) seem to have had a degree of success in using a more conservative approach to training. However, their only replicated transfer effect involved the n-back, which, like the running span, involves keeping track of recently presented items. Future studies should attempt to show that a broader range of tasks which involve memory updating are similarly affected.

Non-adaptive-training of WM updating

Although somewhat outside of the realm of studies under current discussion, one of the more interesting examples of far-transfer was provided by Persson and Reuter-Lorenz (2008). Rather than attempting to constantly strain the boundaries of immediate awareness, these researchers have focused on manipulating the degree to which participants have to deal with competing memory representations within WM. Similar to Dahlin et al. (2008a, 2008b), WMC would not be viewed as a matter of how much information a person can maintain at any given time, but instead a matter of how well relevant information is retained and irrelevant information is discarded. Although the study (Persson & Reuter-Lorenz, 2008) involved two control groups, for the sake of simplicity we focus on the one which was most similar to the training-group. Both training and control groups performed the same three tasks for 10 separate sessions. Two tasks involved recognition of faces or letters (respectively), and the third task was a 3-back version of the n-back.

On the two recognition memory tasks, participants saw a central fixation cross which was surrounded by four to-be-remembered items. After a 3 second delay in which the items were not visible, one probe-item was shown. Participants simply indicated whether or not it was a part of the most recent set. The critical difference between the two groups occurred on trials in which the probe did not belong. For the interference-training-group, two thirds of these trials featured probes which were drawn from recent trials (e.g. Nelson et al., 2003). Responding "no" involved differentiating between memory for the most recent items and memory of other recent trials. That is, they could not simply rely on familiarity when making their judgment. The control group, on the other hand, rarely saw non-matching probes that were drawn from recent trials. For this group, "no" responses rarely involved conflicting memory representations. Thus, judgments could generally be based on familiarity alone.

The n-back task was conceptually similar. Both groups simply responded as to whether the currently presented item had been presented 3 items ago. The critical difference was that the training-group, three quarters of all "no" trials featured a letter that had been presented two, four or five trials ago (e.g. Gray, Chabris, & Braver, 2003). The control group never encountered such lures.

Three tasks were selected for demonstration of transfer effects: paired-associates, item-recognition and verb-generation. Paired-associates featured lists of 8 cue words paired with one highly associated word each. After studying the list, participants saw each cue word repeated in random order. Time to recall the associated word was the dependent variable. Non-interference trials featured cue words that always had the same associate, while interfer-

ence trials featured cue words which had previously been associated with another word (e.g. current trial: queen-king, previous trial: queen-crown). Item-recognition was similar to the recognition memory task that had been used during training, except that this time words were used (rather than letters or faces). In verb-generation, participants were shown a noun and required to press a key when they had generated an associate. High-interference conditions featured nouns that possessed several associates, low-interference conditions featured nouns that possessed one clear associate. Response time differences between the conditions were compared pretest vs. posttest.

The results were consistent for all three tasks. Despite the fact that both groups performed essentially the same tasks during training, only the group that was consistently subjected to proactive interference showed significant pretest vs. posttest improvement on the transfer tasks. We note that Unsworth and Engle (2007b) have recently argued that resolving memory interference may be a critical difference between low-and high-WMC individual's ability to locate items in memory. As such, this study could provide a much more concise model for future attempts to effect change in WMC.

General Discussion of Adaptive WMC Training

There is little doubt that adaptive-span training consistently improves performance on tasks which measure simple retention of short lists. This finding is so common it was rarely discussed in the above review. However, as the quote that began this article implies, the goal of working memory training is not to increase retention of short lists per se. The goal is to alter cognitive function, particularly Gf and attention. Evidence of such changes has been mixed.

In terms of intellectual abilities, six studies report improvements in reasoning (Jaeggi et al., 2008; Klingberg et al., 2002; Klingberg et al., 2005; Olsen, et al., 2004; Studer et al., 2009) or learning (Alloway & Alloway; 2009) as measured by individual tasks. On the other hand, five studies reported no increase in reasoning abilities (Chein & Morrison, in press; Dahlin et al., 2008b; Shavelson, 2008; Thorell et al., 2009; Westerber et al., 2007) using similar methods. A recurring theme in this article has been the reality that no tasks are process-pure and as such, higher scores on a test may reflect a change to the mental construct of interest, or they may reflect task specific learning. The best way to avoid this concern is to include a battery of tests that converge on a common measurement goal. Thus, the most telling results are those of Holmes et al. (2009) and Holmes, Gathercole, and Dunning (2009). These studies involved the most comprehensive measures of reasoning and learning of all studies reported. In both cases, little to no

far-transfer effect was found.

As for improvements of attention following adaptive-WM training, four studies reported that participants' performance in the Stroop task increased above and beyond that of the control group (Klingberg et al., 2002; Klingberg et al., 2005, Chein & Morrison, in press, Olsen et al., 2004), two reported improvement in non-Stroop attention tasks, coupled with no improvement in the Stroop (Thorell et al., 2009; Westerberg et al., 2007) while one study reports no improvement on the Stroop and no other attention tasks (Dahlin et al., 2008a). None of the studies reported an attempt to comprehensively measure far-transfer to attention via composites of several independent measures. WMC and attention are certainly highly related concepts (Kane et al., 2007b), and despite the mixed results, it does seem plausible that adaptive-span training may prove to be effective as a method of attention training. However, it is worth noting that this may be a somewhat indirect effort as some researchers have demonstrated that attention can be directly trained using more traditional attention measures such as flanker tasks (Rueda et al., 2005; Tang & Posner, 2009).

Beyond the problem of noisy results across studies, a more basic concern exists. Across all reported studies involving adaptive-span training, not a single attempt was made to demonstrate that improved performance on training and near-transfer tasks represented actual changes to WMC in general. Rather, increased span scores are assumed to represent increased WMC. As such, one major question has been avoided: Are span scores valid measures of WMC following extensive practice on span tasks?

Chase and Ericsson (1982) report an experimental participant who began a training routine with a span score of 7 digits, however, over the course of 264 sessions, was able to accurately recall spans of up to 82. Upon questioning the participant, the researchers discovered that practice had not given him an excessively large memory capacity, rather he had learned to compress sequences of digits by mapping the numbers to pre-existing knowledge (e.g. athletic records, years, ages, etc.). When Chase and Ericsson tripled the rate of digit presentation, thus circumventing the strategy, the participant's span dropped to 8-9 digits. Moreover recent research indicates that rather than enhancing the relationship between WMC and higher order cognition, strategies tend to obscure it (Turley-Ames & Whitfield, 2003) and are prone to be ineffective when unusual stimuli are encountered (e.g. snowflakes rather than numbers; Maguire et al., 2003).

Strategy learning aside, several already-discussed findings of the above studies are worth further consideration. First, following training, the stroke patients of Westerberg et al. (2007) did show statistically higher simple-span scores (+1 to 2 items) relative to control participants. This was not associated with increased performance in a task which involved immediate serial mem-

ory for 10 items, nor was it associated with later recall of these same items. Westerberg et al. interpreted these findings as suggesting that the Cogmed intervention specifically targets WM, and not memory in general. However, given previous arguments that WM is most important when searching for memory representations which reside outside of one's immediate awareness (Unsworth & Engle, 2007a, 2007b), and given assertions of others that WMC predicts storage and later retrieval on information (Alloway & Alloway, 2009), Westerberg et al. seem to have a strikingly conservative view of WM which is limited to performance on short lists over unbroken periods of time.

Second, it would seem that if adaptive-WM training was in fact increasing WMC, transfer to complex-span tasks would be a common finding. However, beyond the AMWA scores reported in the study of Holmes, Gathercole and Dunning (2009), this type of transfer is rare. Chein and Morrison (in press) do report that complex-span task performance improved following training, however the researchers tested WM using the same tasks on which participants were trained. Shavelson et al. (2008), on the other hand trained participants on Cogmed software and only administered their complex-span tasks (ospan; running span) at pre- and posttest. Although the training participants did show numerical increases in these tasks, it was not significantly different from the performance of control subjects. Similarly, among the tasks that did *not* show transfer effects following adaptive-running-span training of Dalin et al. (2008b) was the computation-span task (Salthouse & Babcock, 1991) which requires participants to solve arithmetic problems while keeping the final digit of each problem in memory for later recall. Finally, Alloway & Alloway (2009) reported that children who were trained using JungleMemory software showed larger increases in scores on a complex-span task relative to a control group. However, these researchers also report that the two groups had marginally different scores at the beginning of training ($p = .10$). This leaves open the possibility that pre-existing group differences might have interacted with training (e.g. Campbell & Stanley, 1963). That is, higher and lower WMC individuals may have reacted differently to adaptive span training. Unfortunately, as raw scores were not reported, this point remains speculative.

Conclusions

Although many studies have been conducted with the intent of training WMC, it seems that there is still a lot to be learned about the behavioral ramifications of such interventions. Tables 1 and 2 indicate that several studies have inadequately controlled for threats to internal validity. Likewise, it is rare to see researchers attempt to produce stable measures of changes to

the cognitive abilities that are of interest to the studies. Most important, it seems that basic questions regarding whether changes on span scores following training represent actual changes to WMC, or whether they represent task-specific learning, have yet to be addressed. Future work using appropriate experimental design (training and active-control groups) and measurement (multiple indicators of constructs) should help answer these questions.

References

Alloway, T. P. (2007). *Automated working memory assessment*. Oxford: Harcourt.

Alloway, T. P. (2009). Working memory, but not IQ, predicts subsequent learning in children with learning difficulties. *European Journal of Psychological Assessment, 25*, 92-98.

Alloway, T. P., & Alloway, R. G. (2008). *JungleMemory Training Program*. Memosyne Ltd, UK.

Alloway, T. P., & Alloway, R. G. (2009). The efficacy of working memory training in improving crystallized intelligence. *Nature Precedings,* <http://hdl.handle.net/10101/npre.2009.3697.1>

Awh, E., & Vogel, E. K. (2008). The bouncer in the brain. *Nature Neuroscience, 11*, 5-6.

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.

Buhner, M., König, C. J., Prick, M., & Krumm, S. (2006). Working memory dimensions as differential predictors of the speed and error aspect of multitasking performance. *Human Performance, 19*, 253-275.

Buschke, H. (1973). Selective reminding for analysis of memory and learning. *Journal of Verbal Learning and Verbal Behavior, 12*, 543-550.

Campbell, D. & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand-McNally.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytical studies*. New York: Cambridge University Press.

Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation,* Vol. 16 (pp. 1-58). New York: Academic Press.

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

Cherry, F. (1953). Some experiments on the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America, 25*, 975-979.

Cogmed (2006). *Cogmed Working Memory Training*. Cogmed America Inc.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiples regression/correlation analysis for the behavioral sciences (3rd ed). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Colflesh, G. J. H., & Conway, A. R. A. (2007). Individual differences in working memory capacity and divided attention in dichotic listening. *Psychonomic*

Bulletin & Review, *14*, 699-703.

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

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

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

Dahlin, E., Stigsdotter Neely, A., Larsson, A., Bäckman, L. & Nyberg, L. (2008a). Transfer of Learning After Updating Training mediated by the Striatum. *Science, 320*, 1510-1512.

Dahlin, E., Nyberg, L., Bäckman, L. & Stigsdotter Neely, A. (2008b). Plasticity of Executive Functioning in Young and Old Adults: Immediate Training Gains, Transfer, and Long-Term Maintenance. *Psychology & Aging, 23*, 720-730.

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

Engle, R. W., & Oransky, N. (1999). The evolution from short-term to working memory: Multi-store to dynamic models of temporary storage. In R. Sternberg (Ed.), *The Nature of Cognition* (pp. 514-555). Cambridge, MA: MIT Press.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General, 128*, 309-331.

Ekstrom, R. B., French, J. W., Harman, M. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*, 143-149.

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

Gronwall, D. M. A. (1977). Paced Auditory Serial-Addition Task: A measure of recovery from concussion. *Perceptual and Motor Skills, 44*, 367-373.

Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research, 18*, 1279-1296.

Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., Randy Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology*. (in press.)

Heitz, R. P. & Engle, R. W. (2007). Focusing the spotlight: Individual differences in visual attention control. *Journal of Experimental Psychology: General, 136*, 217-240.

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental Science, 12*, F9-F15.

Holmes, J., Gathercole, S. E., Place, M., Dunning, D. L., Hilton, K. A., & Elliott, J. G. (2009). Working memory deficits can be overcome: Impacts of training and medication on working memory in children with ADHD. *Applied Cognitive Psychology*. DOI: 10.1002/acp.1589.

Hossiep, R., Turck, D. & Hasella, M. (1999). *Bochumer Matrizentest: BOMAT-Advanced-Short Version*. Götingen: Hogrefe.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CN: Praeger.

Jaeggi, S. M., Buschkuehl, M., Jonidas, 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.

Kane, M. J., Bleckley, K. M., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. Journal of Experimental Psychology: General, 130, 169-183.

Kane, M. J., Brown, L. E., Little, J. C., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007a). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science, 18*, 614-621.

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007b). Variation in working memory capacity as variation in executive attention and control. In A.R.A. Conway, C. Jarrold, M. J. Kane, A. Miyake, and J. N. Towse (Eds.), *Variation in Working Memory* (pp. 21-48). NY: Oxford University Press.

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*, 47-70.

Kim, J., Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it*. Beverly Hills, CA: SAGE Publications, Inc.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology, 55*, 352-358.

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

Klingberg, T., Forssberg, H. & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology, 24*, 781-791.

Kyllonen, P. C., & Stephens, D. L. (1990). Cognitive abilites as determinants of success in acquiring logic skill. *Learning and Individual Differences, 2*, 129-160.

Logan, G. D. & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review, 91*, 295-327.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*, 163-203.

MacLeod, C. M., & MacDonald, P. A. (2000). Inter-dimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences, 4*, 383-391.

Maguire, E. A., Valentine, E. R., Wilding, J. M., & Kapur, N. (2003). Routes to remembering: The brains behind superior memory. *Nature Neuroscience, 6*, 90-95.

McCarney, R., Warner, J., Illife, S., van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne Effect: A randomized, controlled trial. BMC Medical Research Methodology, 7(30).

McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience, 11*, 103-107.

McNab, F., Varrone, A., Farde, L., Jucaite, A., Bystritsky P., Forssberg, H., Klingberg, T. (2009). Changes in cortical dopamine D1 receptor binding associated with cognitive training. *Science, 323*, 800-802.

Monchi, O., Ko, J. H., Strafella, A. P. (2006). Striatal dopamine release during performance of executive functions: A [^{11}C] raclopride PET study. *NeuroImage, 33*, 907-912.

Moody, D. E. (2009). Can intelligence be increased by training on a task of working memory? *Intelligence, 37*, 327-328.

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

Nelson, J.K., Reuter-Lorenz, P.A., Sylvester, C.-Y.C., Jonides, J., & Smith, A.D. (2003). Dissociable neural mechanisms underlying response-based and familiarity-based conflict in working memory. *Proceedings of the National Academy of Sciences, USA, 100*, 11171-11175.

Olesen, P., Westerberg, H., Klingberg, T. (2004). Increased prefrontal and parietal brain activity after training of working memory. *Nature Neuroscience, 7*, 75-79.

Orne, M. T. (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. American Psychologist, 17, 776-783.

Persson, J., & Reuter-Lorenz, PA (2008). Gaining control: Training of executive function and far transfer of the ability to resolve interference. *Psychological Science*, *19*, 881-889.

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

Raven, J. C. (1990). *Advanced progressive matrices*. Oxford, UK: Oxford Psychological Press.

Raven, J. C. (1995a). *Coloured progressive matrices*. Oxford, UK: Oxford Psychologists Press.

Raven, J. C. (1995b). *Standard progressive matrices*. Oxford, UK: Oxford Psychologists Press Ltd.

Redick, T.S., and Engle, R.W. (2006). Working memory capacity and Attention Network Test performance. *Applied Cognitive Psychology, 20*, 713-721.

Rueda, M. R., Rothbard, M. K., McCandliss, B. D., Saccomanno, L., & Posner, M. I. (2005). Training, maturation, and genetic influences on the development of executive attention. *Proceedings of the National Academy of Sciences, 102*, 14931-14936.

Ruff, R. M., Neimann, H., Allen, C. C., Farrow, C. E., Wylie, T. (1992). The Ruff 2 and 7 selective attention test: A neuropsychological application. *Perceptual and Motor Skills, 75*, 1311-1319.

Salthouse, T. A. & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psycology, 27*, 763-776.

Shavelson, R. J., Yuan, K., Alonzo, A. (2008). On the impact of computer training on working memory and fluid intelligence. In D. C. Berliner & H. Kuermintz (Eds). *Fostering change in institutions, environments, and people*: *A festschrift in honor of Gavriel Salomon* (pp. 35-48). Routeledge, New York.

Schmeichel, B. J., Volokhov, R., & Demaree, H. A. (2008). Working memory capacity and the self-regulation of emotional expression and experience. *Journal of*

*Personality and Social Psychology, 95,*1526-1540.

Shute, V. (1991). Who is likely to acquire programming skills? *Journal of Educational Computing Research, 7*, 1-24.

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

Studer, B. E., Jaeggi, S. M., Buschkuehl, M., Su, Y.-F., Jonides, J., Perrig, W. J. (2009). Improving fluid intelligence – Single n-back is as effective as dual n-back. Poster session presented at the 50th annual meeting of *The Psychonomic Society*, Boston, MA.

Tang, Y., & Posner, M. I. (2009). Attention training and attention state training. *Trends in Cognitive Sciences, 13*, 222-227.

Thorell, L. B., Lindqvist, S., Bergman, S., Bohlin, G., Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science, 11*, 969-976.

Turley-Ames, K.J., & Whitfield, M.M. (2003). Strategy training and working memory task performance. *Journal of Memory and Language, 49,* 446-468.

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

Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 1302-1321.

Unsworth, N., & Engle, R.W. (2007a). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order ability. *Psychological Bulletin, 133*, 1038-1066.

Unsworth, N. & Engle, R.W. (2007b). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*, 104-132.

Wechsler, D. (1993). *Wechsler Objective Reading Dimensions (WORD).* New York: Psychological Corporation.

Wechsler, D. (1996). *Wechsler Objective Number Dimensions (WOND).* New York: Psychological Corporation.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence (WASI).* London: Harcourt Assessment.

Wechsler, S. (1995). *WPPSI-R. Wechsler Preschool and Primary Scale of Intelligence – Revised (Psykologiförlaget AB, Stockholm, trans).* New York: Psychological Corportation.

Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Östensson, M.-L., Bartfai, A., & Klingberg, T. (2007). Computerized working memory training after stroke – A pilot study. *Brain Injury, 21*, 21-29.

Westerberg, H., Klingberg, T. (2007). Changes in cortical activity after training of working memory – a single-subject analysis. *Physiology & Behavior*, doi:10.1016/j.physbeh.2007.05.041.

Yntema, D. B. (1963). Keeping track of several tings at once. *Human Factors, 5*, 7-17.

Table 1
WMC training as a therapeutic device

Authors	Training Task	Control Group
Children with ADHD and/or low WM		
Klingberg et al. (2002)	Cogmed	Non-Adaptive Task
Klingberg et al. (2005)	Cogmed	Non-Adaptive Task
Holmes et al. (2009)	Cogmed	None
Holmes, Gathercole, & Dunning (2009)	Cogmed	Non-Adaptive Task
Alloway & Alloway (2009)	JungleMemory	Targeted Instruction
Adaptive-span training and recovery from stroke		
Westerberg et al., 2007	Cogmed	No-Contact Control

Note. When not specified by the methods section of a given article, use of Cogmed software was verified against information available at www.cogmed.com

Table 2
WMC training and healthy populations

Authors	Training Task	Control Group
Adaptive-span training and healthy populations: Behavioral research		
Chein and Morrison (in press).	Complex-span	No-Contact Control
Shavelson et al. (2008)	Cogmed	Non-Adaptive Task
Thorell et al. (2009)	Cogmed/Inhibition	Multiple
Klingberg et al.'s (2002)	Cogmed	Children w/ADHD
Adaptive-span training and healthy populations: Physiological research		
Olsen, Westerberg, & Klingberg (2004)	Cogmed	Mixed
McNab et al. (2009)	Cogmed	None
Adaptive-training of WM updating		
Jaeggi et al. (2008)	Dual n-back	No-Contact Control
Studer et al. (2009)	Dual/single n-back	No-Contact Control
Dahlin et al. (2008a)	Running Span	No-Contact Control
Dahlin et al. (2008b)	Running Span	No-Contact Control
Non-adaptive-training of WM updating		
Persson and Reuter-Lorenz (2008)	Memory Interference	Non-interference

Note. When not specified by the methods section of a given article, use of Cogmed software was verified against information available at www.cogmed.com