Individual Differences in Working Memory Capacity and What They Tell Us About Controlled Attention, General Fluid Intelligence, and Functions of the Prefrontal Cortex

RANDALL W. ENGLE, MICHAEL J. KANE, AND STEPHEN W. TUHOLSKI


Working memory is a system consisting of those long-term memory traces active above threshold, the procedures and skills necessary to achieve and maintain that activation, and limited-capacity, controlled attention. The specific features of our model include:

1. Domain-free, limited-capacity controlled attention.

2. Domain-specific codes and maintenance (phonological loop and visuospatial sketchpad are two examples but the potential number of such codes is large).

3. Individual differences in both 1 and 2, but individual differences in capacity for controlled processing are general and possibly the mechanism for general fluid intelligence. Although people can, with practice and expertise, circumvent the abiding limitations of controlled attention in quite specific situations, the limitations reemerge in novel situations and even in the domain of expertise if the situation calls for controlled processing.

4. Limited-capacity, controlled processing is required for maintaining temporary goals in the face of distraction and interference and for blocking, gating, and/or suppressing distracting events.

5. The dorsolateral prefrontal cortex (PFC) and associated structures mediate the controlled processing functions of working memory. We also argue that individual differences in controlled processing represent differences in functioning of the PFC.

A number of intellectual influences have served to shape our thinking about working memory (WM) and its evolution as a construct separate from that of short-term memory (STM). In thinking about the nature of these constructs, both behavioral and biological, we need always be mindful of the tasks that we use to measure them. It is important to understand that tasks vary in validity as measures for their putative constructs. Further, there is no such thing as a pure measure of any construct, including short-term and working memory. This chapter will describe an attempt by our lab to determine construct validity for short-term and working memory measures.

Another influence is an approach to developmental psychology often called “neo-Piagetian,” although the ideas can easily be traced back to Baldwin (1894), one of Piaget’s early influences. Baldwin and others argued that memory span tasks reflect ability to maintain attention, a fundamental aspect of intellectual abilities, both for the developing human and across individuals at given stages of development (Case, 1985; Pascual-Leone & Bialer, 1994; Piaget, 1926). At least in studies with normal adults, however, simple digit and word span tasks do not consistently and reliably predict such mainstays of higher-level cognition as reading or listening comprehension (Perfetti & Lesgold, 1977).

Daneman and Carpenter (1980) developed a span measure called the reading span task that correlated reasonably well with language comprehension measures. Subjects read (or listened to) a list of 2 to 6 sentences. Afterward, the subject recalled the last word of each sentence. This critical measure, the number of sentence-final words recalled, is, on the face of it, very much like a simple word span task. Turner and Engle (1989) developed a similar task in which the subject solved a string of arithmetic operations and then read aloud a word that followed the string. After a series of such operation-word strings, the subject recalled the words.

Both the reading span task and the operation span task are really dual tasks that require the subject to do something (read a sentence or solve an operation string) and, separately and interleaved with this task, to keep track of an evolving and growing list of words. These span tasks, and others developed since, apparently reflect some ability that is fundamentally important to higher-level cognition because measures of working memory capacity reliably predict performance in a wide variety of real-world cognitive tasks. Significant relationships with measures of working memory capacity have been reported for reading comprehension (Daneman & Carpenter, 1980, 1983), language comprehension (King & Just, 1991; MacDonald, Just, & Carpenter, 1992), learning to spell (Ormrod & Cochran, 1988), following directions (Engle, Carullo, & Collins, 1991), vocabulary learning (Daneman & Green, 1986), notetaking (Kiewra & Benton, 1988), writing (Benton Kraft, Glover, & Plake, 1984), reasoning (Kyllonen & Christal, 1990), and complex learning (Kyllonen & Stephens, 1990; Shute, 1991).
Two questions have guided our work over the past 10 years. The first is “What accounts for individual differences on measures of working memory capacity?” More specifically, “What is measured by the complex tasks that is also important to higher-level cognitive tasks?” The second question is “What do results of studies on individual differences in WM capacity tell us about the nature of working memory in general?” Our attempts to answer these questions have used a combination of both regression studies and experimental or ANOVA-based studies using extreme groups. The extreme groups are individuals who score in the upper (high WM) and lower (low WM) quartile on a variety of working memory capacity tasks such as the reading span and operation span tasks.

We think of “working memory” as a system consisting of (a) a store in the form of long-term memory traces active above threshold, (b) processes for achieving and maintaining that activation, and (c) controlled attention. However, when we refer to “working memory capacity,” we mean the capacity of just one element of the system: controlled attention. We do not mean the entire working memory system, but rather the capabilities of the limited-capacity attention mechanism which Baddeley and Hitch (1974; see also Baddeley & Logie, Chapter 2, this volume) called the central executive. Thus, we assume that “working memory capacity” is not really about storage or memory per se, but about the capacity for controlled, sustained attention in the face of interference or distraction. This is conceptually similar to what Norman and Shallice (1986) called the Supervisory Attentional System and is related to what Posner and Snyder (1975) and Schneider and Shiffrin (1977) referred to as controlled attention. The central executive is also likely related to the anterior attention system proposed by Posner and Peterson (1990); see also Gevins, Smith, McEvoy, & Yu, (1997).

We have proposed that individual differences on measures of working memory capacity primarily reflect differences in capacity for controlled processing and, thus, will be reflected only in situations that either encourage or demand controlled attention (Conway & Engle, 1994; Engle, Conway, Tuholski, & Shisler, 1995; Rosen & Engle, 1997). Such situations include (c) when task goals may be lost unless they are actively maintained in working memory; (b) when actions competing for responding or response preparation must be scheduled; (c) when conflict among actions must be resolved to prevent error; (d) when there is value in maintaining some task information in the face of distraction and interference; (e) when there is value in suppressing or inhibiting information irrelevant to the task; (f) when error monitoring and correction are controlled and effortful; and (g) when controlled, planful search of memory is necessary or useful.

Our proposal, then, is that working memory capacity reflects the ability to apply activation to memory representations, to either bring them into focus or maintain them in focus, particularly in the face of interference or distraction. We have also argued recently that controlled processing capability is necessary in the case of suppression to either dampen activation of representations or otherwise remove them from focus (Engle et al., 1995). This attention capability is domain free and therefore individual differences in this capability will reveal themselves in a wide variety of tasks. We have also argued (Engle & Oransky, 1999; Kane & Engle, 1998) that differences in working memory capacity correspond to individual differences in the functioning of the prefrontal cortex.

But do the different tasks we refer to as measures of working memory capacity really reflect the same underlying construct? Further, do such tasks measure something different from traditional short-term memory tasks and, if so, what distinguishes them? Engle, Tuholski, Laughlin, and Conway (in press) addressed these questions in a recent study directed at Cowan’s (1995; see also Chapter 3, this volume) distinction between short-term memory and working memory and the relationship of these constructs to general fluid intelligence (gF). Cowan considers short-term memory a subset of working memory. Short-term memory is a simple storage component meaning those memory units active above some ambient baseline, whereas the term “working memory” refers to a system consisting of that storage component as well as an attention component. This view is consistent with Baddeley and Hitch’s (1974) original model, except the concept of short-term memory is retained and viewed as consistent with the slave systems (the phonological loop and the visuospatial sketchpad). Thus, Cowan’s view is that working memory consists of the contents of short-term memory plus controlled attention.

Engle et al. (in press) performed an analysis of the unique and shared variance in tasks thought to reflect STM and WM, the underlying factor structure of that variance, and the extent to which a theory of the two constructs is supported by structural models of the variance in the tasks. Figure 4.1 shows a schematic of a measurement model based on those constructs. Controlled attention can be used to achieve activation of long-term traces through controlled retrieval, to maintain activation through various means, or to dampen activation through inhibition. Short-term memory consists of those traces active above threshold, with loss of activation caused by decay and/or inhibition. The short-term traces could be primarily in the form of phonological features or visual features or, indeed, many other features; however, the traces obey the same principles of forgetting, interference, and so on, regardless of format. Some very small number of those traces receive increased activation through attention resulting from salience to the current task goal or from endogenous activation resulting from emotional salience.

Cowan defined short-term memory as a subset of working memory. Thus, at a conceptual level, variance shared between pure working memory tasks and pure short-term memory tasks should reflect the short-term component. The variance left over (or residual) in working memory tasks after removal of the variance shared by the two tasks should reflect the controlled attention or
Relationship of Components of Working Memory System

Any given WM or STM task reflects all components to some extent

Magnitude of this link is determined by the extent to which the procedures for achieving and maintaining activation are routinized or attention demanding. Thus, it is assumed that, in intelligent, well-educated adults, coding and rehearsal in a digit span task would be less attention demanding than in 4-year-old children.

Central Executive

- Working memory capacity, controlled attention, focused attention, supervisory attention system, anterior attention system, etc...

  a. Achieve activation through controlled retrieval.
  b. Maintain activation (to the extent that maintenance activities are attention demanding).
  c. Block interference through inhibition of distractors.

Short-term memory

- Consists of those traces that are active above threshold, with less due to decay or interference.

  a. Some traces involve further activation by recruiting the focus of attention.
  b. Trace consists of a pointer to a region of long-term memory. Thus, the activated trace could be as simple as “it circle around the next digit on the list” or as vast as the get for War and Peace.

Long-term memory

Grouping skills, coding strategies, and procedures for maintaining activation.

  a. Could be phonological, visual, spatial, motoric, auditory, etc.
  b. More, or less, attention demanding depending on the task and the subject.

Figure 4.1. A measurement model of the relationship among long-term memory, short-term memory, and the central executive.
words. We also tested all subjects on full versions of the Ravens and Cattell Culture Fair tests.1

One question was whether all the memory tasks reflected a single construct or whether a two-factor model (i.e., STM and WM) was necessary. A second question was whether, at the latent variable level, after variance common to WM and STM was removed, the WM residual variance (which should reflect controlled attention) would correlate with the residual for gF, and, whether the STM residual would also correlate with the gF. Of course, the STM residual should not correlate with the gF residual if it reflects only error as our logical analysis would suggest.

Confirmatory factor analysis/structural equation modeling was performed on the three tasks we thought would be good STM tasks and the three we thought would be good WM tasks. The goodness-of-fit results showed that the two-factor model was a significantly better fit than the single-factor model. Thus, the model required a separate construct for working memory capacity and for short-term memory even though the two constructs were strongly related. The next question was whether the two constructs showed any relationship to general fluid intelligence (gF) as indexed by the Ravens and Cattell Culture Fair tests. The analysis showed that the model required a strong connection between the working memory latent variable and general fluid intelligence but did not require a connection between short-term memory and gF.

If our arguments about the relationship among the constructs of working memory, capacity for controlled attention, and gF are correct, then we should be able to test whether controlled attention is related to general fluid intelligence. Starting with all the variance in complex working memory measures, we should be able to remove that variance owing to domain-specific materials and procedures. The residual variance in WM, after removal of the variance common to STM, should reflect the capacity for controlled processing. The WM residual should, therefore, strongly relate to the construct general fluid intelligence. Thus, the line of logic is as follows: (a) complex span tasks of working memory capacity reflect the construct short-term memory plus the construct controlled attention; (b) the construct controlled attention has a strong relationship with the construct general fluid intelligence; (c) there is little or no relationship between the construct short-term memory and the construct general fluid intelligence. Therefore, if, at the latent-variable level, we partial out the variance common to the constructs working memory and short-term memory, we would expect the residual variance in working memory capacity to reflect controlled attention. This method of removing com-

1 Several other tasks were also used, including the ARCD and Continuous Opposites tasks from the CAM battery (Ryllonen & Christal, 1990), keeping track task (Yntema, 1963), the primacy and recency portion of immediate free recall, and the random generation task. Scores were obtained on Verbal and Quantitative Scholastic Aptitude Test but those variables are not discussed here.

Figure 4.2. Path model with the variance common to STM and WM removed as common and the relationships among the various constructs. The curved lines represent correlations between the construct gF and the residual for STM and WM.
The Engle et al. (in press) latent-variable study yields two important conclusions: (a) working memory and short-term memory are highly related but separable constructs, and (b) when we partial out the variance common to STM and WM, the link between the residual of WM and gF, which should theoretically be controlled attention, is high and highly significant. This lends support to the idea that the component of the working memory tasks that is important to higher-order functioning is controlled attention.

As we said earlier, the reading span and operation span are really dual tasks and so should tap the subject's capability to sustain, divide, or switch attention between these task components. One component can be thought of as a processing task, reading the sentence, or solving the arithmetic string, and the other can be thought of as a storage task, recalling the gradually lengthening list of words or digits. One possible alternative reason that these tasks correlate with higher-order measures of cognition is that individuals differ in skill on the processing component as a result of experience; this processing efficiency frees up resources to be used to rehearse the items in the storage task.\(^2\) Engle, Cantor, and Carullo (1992) measured the time it took subjects to perform the processing component of the reading span and operation span tasks. They then used these times as measures of processing efficiency and/or the extent to which subjects traded off time on the processing component for time on the storage component. The processing-skill explanation of the relationship between WM measures and measures of higher-order cognition suggests that when the measure of processing time is partialled out of the relationship between complex span and reading ability, the correlation should significantly decrease if not disappear. However, Engle et al. showed that partialling out the processing-time measure led to no decrease in the correlation between span and the Verbal Scholastic Aptitude Test.

A study by Conway and Engle (1996) approached this same issue differently. Instead of statistically controlling for processing skill, they attempted to equate, across subjects, the processing demands of the span task. The logic was quite simple. If the relationship between the WM measure and reading comprehension is a result of trade-off between the processing and storage components, then when the difficulty of the processing component is equated, the correlation between the span measure and reading comprehension should disappear. On the other hand, if the correlation between the WM measure and comprehension is a result of a controlled attention capability above and beyond the trade-off, then equating subjects on the difficulty of the processing task should have no effect on the correlation. Conway and Engle pretested subjects on operations of the type used in the operation span task but differing greatly along a dimension of difficulty. The pretest determined the point on the difficulty dimension at which each subject solved operations accurately 75%, 85%, and 95% of the time. Each subject then received an operation-word span task in which the sets of operations were created specifically for that subject. One set was created separately at each of the 75%, 85%, and 95% levels, specifically for each subject. One question was whether the correlation between the operation span score and reading comprehension would disappear since the procedure equated the processing skill component. The second was whether the correlations between the resulting span scores and reading comprehension would differ as the processing component of the span test varied in difficulty from 75% to 95% accuracy. The answer to both questions was a resounding no. The correlation between the span score and Verbal Scholastic Aptitude test for the easy 95% condition was .62 and for the difficult 75% condition the correlation was .54. This compares to a correlation of .59 for the same sample of subjects on an operation span test in which all subjects receive the same set of operations. Conway and Engle also measured the time each subject took to perform the operation-processing component in the various tasks and statistically removed this solution time from the aforementioned correlations. The effects were virtually identical, with a partial correlation of .60 for the easy condition and .52 for the difficult condition. Clearly, skill on the processing component had no bearing on the significant and high correlation between this measure of working memory capacity and a measure of reading ability.

These findings show, quite convincingly, that the correlation between measures of working memory capacity and higher-order cognitive tasks is not a result of skill in the specific processing components of the WM tasks. These findings are hard to reconcile with the views of Ericsson and Delaney (Chapter 8, this volume) and O'Reilly, Braver, and Cohen (Chapter 11, this volume) that individual differences in WM capacity represent differences in skill resulting from experience. The findings do support our view that the critical feature of the tasks accounting for the correlations with higher-order cognition is some aspect of controlled attention, and that it represents a rather abiding characteristic of the individual.

We have recently completed two longitudinal studies with young children that further support this conclusion. In one study, children were given a series of WM tasks and an age-appropriate listening-comprehension test at ages 3.5, 4.5, and 5.5. The memory measures for each year predicted the subsequent year's comprehension scores; however, the comprehension scores for each year did not predict the subsequent year's memory measures. Moreover, the WM scores at age 3.5 accounted for 27.5% of the variance in comprehension at 5.5. Interestingly, the WM scores were better predictors of comprehension at age 5.5 than at 3.5 or 4.5, suggesting that controlled attention is not as important for comprehension at younger ages but becomes more important by age 5.5.

A second longitudinal study investigated the relationship between working memory capacity and inhibition in children. Adele Diamond and her

\(^2\) It should be noted that this makes the storage component even more like the short-term memory latent variable in the Engle, Tuholski, Laughlin, and Conway analysis.
colleagues (Gerstadt, Hong, & Diamond, 1994) have argued that development of the ability to inhibit or suppress prepotent or predisposed behaviors corresponds with the development of the frontal lobes. They performed a study in which children were shown pictures of the moon and stars and were required to say “day” and a picture of a sun and were required to say “night.” Though the children could perform the task initially, performance deteriorated over the 16 trials. Performance improved over the ages at which the frontal lobes are thought to be completing their development. The conclusion was that maturation of the frontal lobes was necessary to inhibit the prepotent response and allow performance of the lower-strength response instead. In our second longitudinal study, children were tested on measures of WM capacity and a “Stroop-like” test in which they were shown line drawings of a traditional mother (long skirt, hair in bun, etc.) and asked to call it “dad” and a drawing of an adult male (pants, facial hair, etc.) and asked to call it “mom.” Scores on the WM tasks measured at age 3.5 predicted Stroop accuracy 2 years later (with correlations ranging from .34 to .40; span scores given at the same age as the Stroop task correlated with Stroop from .33 to .38). These studies suggest that relatively simple measures, which we argue reflect capacity for controlled attention, are valid measures for very young children and that they reflect an abiding ability important to higher-level cognition.

Working Memory, General Fluid Intelligence, and Controlled Attention

The behavioral evidence discussed thus far certainly suggests that individual differences in working memory and in general fluid intelligence are significantly related to one another and that they probably do not depend on differences in general knowledge or specific procedural skills. Our discussion of latent-variable results and individual-differences work on WM tasks themselves tentatively suggests that controlled attention abilities lie at the heart of individual differences in working memory capacity. We briefly review some further evidence below. We will then focus our discussion on a recent set of studies from our lab that explicitly ties together WM capacity, general fluid intelligence, and controlled-attention constructs.

The evidence that people who differ on indices of working memory show systematic differences in other controlled attention measures may be found from studies of both perceptual and memory interference. In perceptual interference tasks, subjects typically must identify visual target stimuli that appear simultaneously amidst salient visual distractors. In memory interference tasks, subjects are required to recall some subset of previously presented information while disregarding similar prior information. In both classes of tasks, attention must be selectively directed toward some information sources (external or internal) and away from others. In both classes of tasks, high-span subjects outperform low-span subjects. For example, in a perceptual interference task (negative priming), high-span subjects appear to inhibit visual distractors during selection whereas low spans do not (Conway, Tuoholski, Shisler, & Engle, 1998).

Furthermore, on selective-attention tasks involving memory distraction, low spans experience more proactive and retroactive interference than do high spans (Conway & Engle, 1994; Kane & Engle, 1997; Rosen & Engle, 1998). In verbal fluency tasks, where subjects generate exemplars from a category such as “animals,” low spans recall few exemplars compared to high spans. This is particularly true across long retrieval periods where competition (distraction) from already-recalled, high-probability exemplars should be high (Rosen & Engle, 1997). Finally, performance on interference and attention-switching tasks is severely disrupted by the imposition of a simultaneous working memory load—that is, by a divided attention requirement (Engle et al., 1995; Kane & Engle, 1997; Law, Morrin, & Pellegrino, 1995; Roberts, Hiart, & Lenon, 1994).

Furthermore, people who differ in psychometric intelligence also show differences in controlled attention capabilities that are similar to those seen across subjects who differ in working memory. For example, correlations between gF and controlled-visual-search tasks are high under variably mapped conditions, where specific stimuli may appear as targets or distractors across trials and so automaticity is unlikely to develop. Correlations between gF and visual search are low, however, under consistently mapped conditions, where specific stimuli appear only as targets or as distractors, and so automaticity is likely to develop (e.g., Ackerman, 1988). Thus, performance under controlled processing, as opposed to automatic processing, correlates with psychometric gF. Furthermore, individuals with low scores on psychometric gF tests are more disrupted by perceptual distractors than are those who score high on gF tests (e.g., Blanco & Alvarez, 1994; Witkin & Goodenough, 1981). Low-gF individuals have more difficulty initiating shifts of attention from one stimulus location or modality to another when the shifting rules must be maintained in working memory (e.g., Duncan, 1995; Duncan, Emslie, Williams, Johnson, & Freer, 1996). Finally, scores on tests of psychometric gF predict the ability to divide attention among simultaneous tasks or among competing stimuli, with high-gF individuals being the more effective “timesharing” (e.g., Hunt, 1980; Morrin, Law, & Pellegrino, 1994; Pellegrino, Hunt, & Yee, 1989; Stankov, 1983).

Thus, individuals with poor working memory capabilities, and those of lower psychometric intelligence, do indeed show similar patterns of performance to each other. When tasks demand that subjects selectively focus attention amidst external or internal sources of distraction, or that subjects shift attention according to memorized rules, or that subjects divide their attention between different stimuli or tasks, working memory capacity and psychometric gF scores are good predictors of performance.
But do we have any direct support for this idea that capability for controlled processing is the critical element common to the working memory tasks, tests of general fluid intelligence, and tests of higher-order functions such as reading comprehension? A recent set of studies performed by Tuholski (1997) might provide that support. He used a speeded-counting task that, though quite ancient (Jevons, 1871), was recently repopularized by Trick and Pylyshyn (1993). The task requires the subject to simply count from 1 to \( n \) presented objects as quickly as possible. The time to count objects is a nonlinear function of the number of objects with the slope of the function nearly flat over the first 3 to 4 objects but rising steeply from 4 to \( n \). The argument is that subjects retrieve the number of objects from 1 to 4 automatically (i.e., they subitize) and as a single pattern. However, counting the number of objects of number 4 and higher requires controlled processing, hence the steeper slope of the counting time function (Trick & Pylyshyn, 1993). Tuholski reasoned that, if high and low working memory capacity subjects differ in their capacity for controlled attention, that difference should be reflected in the time to count from 4 to \( n \) objects but not in the time to subitize from 1 to 4 objects. Further, if controlled attention is related to general fluid intelligence, the slope of the counting function for more than 4 objects should correlate with a measure of \( gF \). An independent prediction is that high and low WM subjects might differ in the size of the pattern they could subitize, with high spans possibly able to pick up a larger number of objects in the initial encoding than are low WM subjects. If that were true, then we should see the flat subitizing function extending over more targets for high WM subjects than for low WM subjects.

Tuholski (1997) tested 60 subjects who had been previously tested on the operation span task and the Cattell Culture Fair test. For the counting task, 1 to 12 vertical yellow bars were arranged randomly on a computer screen and the subject counted them silently, as quickly as possible without sacrificing accuracy, and then gave an oral response of the total, which activated a speech trigger. The 15 subjects who scored in the upper quartile on the operation span task were selected as high WM subjects and the 15 scoring in the bottom quartile were selected as the low WM subjects. The size of the subitizing range was determined by doing a series of incremental analyses that tested linear and quadratic trends as the number of targets increased from 1 to \( n \). The subitizing span was indicated when the quadratic trend became significant. Thus, if the quadratic trend was not significant over the range 1 to 3 but was significant over the range 1 to 4, this indicated the subject had subitized 3 objects but not 4. This analysis showed that high WM subjects had a subitizing span of 3.35 and low WM subjects had a subitizing span of 3.25, quite remarkably similar and refuting the possibility that high and low WM individuals differed in the size of the initial pattern they were able to perceive. As shown in Figure 4.3, the slope over the subitizing range, presumed to reflect automatic processing, was flat and not different for the two groups. However, high and low WM subjects did differ dramatically over the counting range presumed to require controlled attention.

Tuholski then looked at correlations among the slope of counting time over the range 4 to 12, the slope of the subitizing function, operation span, and the Cattell Culture Fair test for all 60 subjects. First, the correlation between operation span and Cattell was .52, a strong correlation between a measure of working memory capacity and a nonverbal test of spatial reasoning and general fluid intelligence. The slope of the counting function correlated with both operation span (−.32) and with the Cattell measure of \( gF \) (−.26) but the subitizing slope correlated with neither.

Why do high and low WM subjects differ in counting time from 4 to 12? What is it that requires controlled attention? We think that the difficult part in counting beyond the subitizing range is keeping track of what you have already counted and that high WM subjects are better able to keep active the tags that indicate an object has been counted. In other words, this is another example of a task that puts a premium on maintaining a representation in an active and easily accessible state and individuals with greater capability for controlled processing are better able to do that.

In a second study, Tuholski did a manipulation that should lead to controlled counting even over the subitizing range. He included distractors that either shared physical features (orientation or color) with the targets, the con-
and controlled-attention capabilities. The research has used methods such as surgery and/or single-cell recording with nonhuman primates, cognitive testing of human patients with brain damage, and cognitive testing of healthy humans under various brain-imaging techniques. Each one of these methods has inherent limitations, and the PFC is clearly a heterogeneous region both structurally and functionally. Nevertheless, we believe that the confluence from the various methods provides an emerging story of working memory and PFC functioning (for recent reviews see Duncan, 1995; Goldman-Rakic, 1987; Pennington, 1994; see also O'Reilly et al., Chapter 11, this volume).

We first discuss experiments that demonstrate the role of the PFC in working memory functions associated with tasks that require simultaneous (or alternating) storage and processing of information. Next, we review findings that suggest PFC involvement in more “classic” controlled-attention tasks, such as those requiring the maintenance, focusing, or shifting of attention. Finally, we consider evidence that the PFC is critical to tasks that most effectively tap general fluid intelligence or gF.

**Working Memory and the PFC**

With macaques, working memory research has focused on so-called delay tasks (such as delayed response, delayed alternation, and delayed matching-to-sample). On each trial, a stimulus (an object or stimulus location) is briefly presented and is then removed from view for some delay duration. Following the delay, the monkey must recognize the stimulus from among distractors. The stimuli are typically chosen from a small pool and tested repeatedly across trials, which would create a high potential for proactive interference. Monkeys with surgical lesions to dorsolateral PFC, or with temporary “lesions” from cortical cooling or electrical stimulation, show chance levels of recall on these delay tasks. Deterioration in performance is found even with delays of only a few seconds (e.g., Fuster & Bauer, 1974; Jacobsen, 1936; Mishkin & Pribram, 1955, 1956).

Several brain areas that are anatomically connected with the dorsolateral PFC also appear to be important to delay-task performance. For example, lesions to certain parietal, hippocampal, and subcortical regions produce deficits that mirror those produced by PFC lesions (e.g., Koch & Fuster, 1989; Koijima & Goldman-Rakic, 1982, 1984; Watanabe & Niki, 1985). Such findings make sense, because these structures are significantly interconnected with the PFC. Further, as we argued above, there are no such tasks as “pure” WM and STM tasks and, thus, no complex cognitive task is likely to rely on a single brain region in isolation. Delay-task performance thus appears to involve integrated neural networks that operate together to retain information across temporal delays.

One might be wary of our reference to these delay tasks as working memory tasks. After all, on the surface they appear to have more in common with human short-term memory tasks, for they seem to require only information
storage, and not much simultaneous processing. However, these tasks typically demand that the monkey’s attention be drawn away from the to-be-recalled information during the delay either by placing a physical barrier between the subject and the stimuli, by requiring subjects’ eyes to remain fixed at a neutral location, or by requiring subjects to extend their hand toward a neutral stimulus. Thus, just as in span tasks inspired by the Baddeley and Hitch (1974) model, these delay tasks require subjects to remember stimuli to which they are only intermittently permitted to attend.

This point is made all the more clear by an experiment by Malmo (1942) that demonstrated that macaques with large PFC lesions showed delay-task deficits only when the testing chamber remained brightly lit during the delay. When the chamber was darkened during the delay, thus reducing visible distractions, no deficits were evident. It appears, then, that the PFC is critical to maintaining the memorial activation of information across shifts in attention.

The delays used in human research are typically longer than those used with nonhuman primates (i.e., 10 to 20 s). Nevertheless, humans with PFC damage show more forgetting in delay tasks than do healthy subjects or control patients with comparable damage to posterior brain areas (Chorover & Cole, 1966; Verin et al., 1993).

Imaging studies with healthy human subjects indicate that working memory tasks elicit significant increases in dorsolateral PFC activity compared to non-working memory tasks. For example, in PET studies using short-term memory tasks that merely require subjects to retain information across short delays, there is limited evidence of frontal lobe activation. Where activation has been observed, it was not centered in the dorsolateral PFC, but in more ventral areas such as Broca’s area and/or in premotor cortex and supplementary motor area (e.g., Dupont et al., 1993; Smith, Jonides, & Koepp, 1996).

In contrast, other imaging studies that required subjects to retain information across longer delays, or to shift attention between storage and processing functions, such as in the “N-back” task, did show PFC activation. In the N-back task subjects must monitor a long sequence of visual or auditory stimuli and continuously report the stimulus that occurred n (1 to 3) stimuli ago in the sequence. Such tasks clearly involve more information storage and processing than do short-delay tasks, and so it is no surprise that significant dorsolateral PFC activation is seen during the performance of these tasks (e.g., Cohen et al., 1994).

In sum, evidence from a variety of tasks and subject populations clearly shows that PFC lesions cause considerable difficulties on working memory tests, and that PFC structures increase their activity while the subject is performing working memory tests but not during short-term memory tests. The existing data suggest that the dorsolateral PFC, and regions to which it is networked, are critical to working memory functions, that is in behavioral tasks that require retention of information across shifts of attention. Therefore, we think it is reasonable to contemplate the possibility that individual differences in working memory capacity among normal individuals are mediated through individual differences in PFC functioning.

**Controlled Attention and the PFC**

Attention is complex and multidimensional, and so it may be naive to presume that all controlled processing is accomplished by a single brain region. Indeed, the neuroscience of attention has suggested broad networks of cortical and subcortical regions that participate in various attentional functions (e.g., Posner & Peterson, 1990). But, though current theories may not all agree on the precise function of the PFC, most theories would suggest it has an integral role in controlled attention. Below we briefly review the empirical evidence for PFC involvement in maintenance, focusing, and shifting attention, most of which includes human studies using neuropsychological testing or imaging.

Regarding attention maintenance, experiments with brain-damaged humans have shown that patients with PFC lesions show marked difficulty with vigilance tasks. On those tasks that require subjects to maintain attention to long series of stimuli while in search of rare targets, patients with PFC damage miss more targets and commit more false alarms than do control patients (e.g., Salinas & Denes, 1982; Critch, Shallice, & McCarthy, 1987). In visual and auditory detection tasks with healthy humans, PFC areas (along with some parietal areas) show increased activation compared to baseline conditions (Cohen et al., 1988; Pardo, Fox, & Raichle, 1991). Furthermore, PFC activation changes are also seen during habituation. Here, one might expect that if the PFC is critical to attention maintenance, then its role (and hence its activation) should decrease as a habituated context demands less and less attention. Indeed, this appears to be the case. Warach et al. (1992) found decreased PFC (and inferior parietal) blood flow across three 14-min rest periods in which the testing context became less novel. No changes in blood flow were evident in temporal, occipital, or superior parietal areas across these rest periods.

As in the working memory and intelligence literatures, the PFC’s selective attention functions may be measured in perceptual and memory interference tasks. In perceptual interference tasks, nonhuman primates with PFC lesions clearly demonstrate increased perceptual distractibility. They are often hyperactive in brightly lit, but not in dimly lit, environments (e.g., Isaac & DeVito, 1958; Pettigrew, Zable, & Harlow, 1948), and they have difficulty performing previously learned visual discrimination tasks when visual or auditory distractors are presented simultaneously with the target stimuli (Grueniger & Pribram, 1969).

Humans with PFC damage also show a heightened susceptibility to perceptual distractors, but it may only be evident when a simultaneous memory load is also imposed. For example, PFC damage reliably impairs performance in the classic Stroop (1935) color-word interference task only in discrete-trial
procedures with a random sequence of trials on which some color-word combinations match and some mismatch (e.g., Vendrell et al., 1995). It is also under these same Stroop conditions that PFC areas show evidence of activation in imaging studies with healthy subjects (e.g., Bench et al., 1993). In contrast, if patients perform the Stroop task in which all color-word combinations either match or mismatch within a block, PFC lesions sometimes do (Perret, 1974; Richer et al., 1993), but often do not, lead to significant deficits (e.g., Ahola, Vikki, & Servo, 1991; Butters, Kaszniaik, Gillsky, Eslinger, & Schacter, 1994; Shallice & Burgess, 1991). It thus appears that PFC involvement in Stroop performance will be found consistently only when the task demands must be constantly maintained in working memory (see Richer et al., 1993).

Memory interference paradigms also suggest that the PFC is important to the control of selective attention. In proactive interference tasks, in which subjects must recall only the most recent of a series of memory lists, patients with PFC damage have difficulty in limiting recall to the target list. Frontal-lesion patients recall fewer target words as the successive lists proceed, and they make more intrusion errors from prior lists than do controls (e.g., Freedman & Cermak, 1986; Parkin, Leng, & Stanhope, 1988; Van der Linden, Bruyer, Roland, & Schils, 1993). Imaging data also point to the PFC as crucial to interference resistance, because dorsolateral and anterior PFC areas show increased activity in healthy subjects when they attempt to recall words under high-interference conditions but not under low-interference conditions (Uhl et al., 1990; Uhl, Podreka, & Deecke, 1994).

Finally, increased interference susceptibility may be inferred from the performance of PFC-lesioned patients in verbal fluency tasks. When subjects are asked to recall category exemplars for periods of longer than 5 min, patients with PFC lesions recall fewer instances than do patients with nonfrontal lesions (e.g., Milner, 1964; Pendleton, Heaton, Lehman, & Hulihan, 1982; Perret, 1974). The longer recall period likely places a premium on controlled, effortful search and increases the interference from previously recalled items (Rosen & Engle, 1997). Additionally, imaging studies show significant left-PFC activation in healthy adults during long-duration fluency tests (e.g., Cuenod et al., 1995; Frith, Friston, Liddle, & Frackowiak, 1991). However, when recall periods are 2 min or less, recall can likely occur more on the basis of automatic spreading activation and PFC lesions appear to have little effect on recall (e.g., Joanette & Goulet, 1986; Newcombe, 1969; Shallice & Burgess, 1991).

Research on attention shifting has also implicated the PFC in attentional control. Some of the "evidence," however, is based more in clinical folklore than in fact. For example, the Wisconsin Card Sorting Test (Berg, 1948; Grant & Berg, 1948) and the Trail Making Tests (Armitage, 1946) are two neuropsychological tests that are widely believed to tap attentional-shifting capabilities and to be sensitive to frontal lobe damage (for reviews see Reitan & Wolfson, 1994, 1995). In fact, whereas some studies show Wisconsin Card Sorting Test deficits in patients with PFC injury (e.g., Drew, 1974; Milner, 1963; Nelson, 1976), many others do not (e.g., Anderson, Damasio, Jones, & Tanel, 1991; Grafman, Jones, & Salazar, 1990; Shallice & Burgess, 1991). Furthermore, no studies to date indicate substantial Trail Making Test deficits for frontal-lesion patients compared to nonfrontal-lesion patients (e.g., Reitan & Wolfson, 1995; Shallice & Burgess, 1991). Clearly, any conclusions about the role of the PFC in attention shifting should not be based on findings from these tests.

However, there is limited evidence from other tests that are conceptually related to the Wisconsin Card Sorting Test that PFC damage impacts performance when attention must be shifted away from a previously learned discrimination in favor of a new discrimination. These findings suggest that simpler measures of attention shifting may be selectively sensitive to PFC lesions (e.g., Delis, Squire, Bihlre, & Massman, 1992; Harlow & Dagnon, 1943; Owen, Roberts, Polkey, Sahakian, & Robbins, 1991). In addition, an oral version of the Trail Making Test, in which the subject must count aloud an alternating progression of letters and numbers, and so must also maintain the last reported letters and numbers in working memory, appears to discriminate PFC-lesioned subjects from nonfrontal-lesioned subjects (Ricker, Axelrod, & Houtler, 1996). Finally, tasks in which subjects must shift their visual attention from one spatial location to another on demand (and in opposition to a salient cue) indicate that patients with PFC damage perform more poorly than do patients with nonfrontal damage (e.g., Duncan et al., 1996; Pierrot-Deseilligny, Rivaud, Gaymard, & Agid, 1991).

A growing body of work using varied methodologies and subject populations thus supports the notion that the PFC is an important structure to the functioning of attention. As inferred from patients with PFC damage, and from imaging studies with healthy subjects, the PFC is an active participant in attention maintenance, selection, and shifting. Again, it seems plausible that individual differences in these attention functions among normal, healthy individuals are mediated by individual differences in PFC functioning.

**General Intelligence and the PFC**

If working memory and controlled attention rely on intact PFC functioning, then the behavioral research we described earlier surely suggests that general intelligence should also depend heavily on PFC structures. Indeed, patients with PFC damage show problems with complex, everyday cognitive activities (e.g., Lezak, 1983; Lucia, 1966; Shallice & Burgess, 1991). It may be surprising, then, that clinical and experimental reports indicate that PFC injury has little effect on intelligence as defined by conventional psychometric IQ tests (e.g., Ackery, 1937; Hebb, 1945; Hebb & Penfield, 1940). Duncan (1995) attempted to resolve this paradox by arguing that broad IQ test batteries assess crystallized knowledge as well as general fluid intelligence whereas novel reasoning tests such as the Raven's Progressive Matrices or the Cattell
Culture Fair have very high \( g_f \) loadings (e.g., Carroll, 1993; Snow, Kyllonen, & Marshalek, 1984). WAIS-R subtests such as Vocabulary and Information, for example, more accurately reflect \( g_f \) at the time of learning, and not necessarily at the time of testing. Averaging across high-\( g_f \) and low-\( g_f \) subtests in conventional IQ tests may thus dilute any real effect that PFC lesions have on \( g_f \).

To test this idea, Duncan, Burgess, and Emslie (1995) matched three patients with frontal lobe lesions (of mixed etiologies) to healthy controls on age and on their overall WAIS or WAIS-R scores, with these scores ranging from 126 to 130. These subjects were then administered the Cattell Culture Fair test, as a measure of \( g_f \). All of the frontal-lesion patients showed a significant drop in "intelligence" (22–38 points) from their WAIS to their Cattell scores. In contrast, their matched controls showed equivalent or higher Cattell scores than WAIS scores. Moreover, the Culture Fair IQs of the frontal-lesion patients were significantly lower by 23 to 60 points, or 3 standard deviations) than those of their controls. Finally, a group of 5 patients with posterior cortex damage, with significantly lower WAIS or WAIS-R scores than the other subject groups, showed IQ patterns similar to those of the control subjects.

Although these findings involved few subjects and therefore demand replication, they do provide rather striking preliminary evidence that the PFC is significantly involved in the performance of tasks that load highly onto psychometric \( g_f \).

These clinical data are consistent with a recent imaging study with healthy adults (Prabhakaran, Smith, Desmond, Glover, & Gabrieli, 1997; see also Risberg, Maximilian, & Prohovnik, 1977). Seven subjects solved Raven's Progressive Matrices problems while undergoing fMRI scans. Activation patterns were compared across different subtypes of Raven's problems: those that required only figural reasoning and those that additionally required more abstract analytical reasoning. Figural reasoning activated the right dorsolateral PFC (and some networked posterior areas). However, analytical reasoning activated dorsolateral PFC bilaterally (as well as other PFC areas and posterior areas bilaterally). Thus, when the task required more general, less-specialized reasoning, greater bilateral PFC activation was observed.

The intelligence findings reviewed earlier, in concert with those concerning working memory/attention, then, converge to tell a consistent story. //The PFC may be the critical brain structure mediating functions of, and individual differences in, working memory, controlled attention, and general intelligence in normal, healthy individuals.//

Speculations, Suggestions, and Conclusions

Do individual differences in working memory, attention, and general intelligence arise from individual differences in PFC functioning? In this chapter we first discussed correlational and experimental studies indicating that working memory and general intelligence constructs are intimately linked - if not iso-

morphic. We then reviewed neuropsychological evidence suggesting that working memory, attention, and \( g_f \) are all subserved by the PFC and the various posterior regions to which it is networked.

It is therefore very tempting to conclude that individual differences in working memory/controlled attention and general intelligence are mediated strictly by individual differences in PFC functioning or efficiency. We must be careful, however, because in the absence of other supporting evidence, the fact that PFC damage causes working memory/attention dysfunction does not necessarily dictate that working memory/attention dysfunction reflects PFC damage (or differences). This result could instead reflect the fact that the PFC is heavily interconnected to other important processing regions. For example, research with macaques shows that specific lesions to posterior or subcortical brain areas to which the PFC is networked produce PFC-lesion-like syndromes in delay task performance. Likewise, human neuropsychological research indicates that damage to some posterior brain areas (or widely diffuse brain damage) elicits PFC-lesion-like impairments on some working memory and attention tasks. Finally, we know from human brain-imaging studies that some posterior brain areas anatomically connected to the PFC are highly activated during the performance of working memory, attention, and psychometric \( g_f \) tasks.

Such findings do not rule out the importance of the PFC to working memory/attention and \( g_f \), or to individual and group differences therein. Rather, the PFC is not uniquely important to these, or probably to any other functions. The PFC appears to be a necessary brain structure in the emergence of working memory/attention behaviors, but is not the only important structure for these functions (see also O'Reilly et al., Chapter 11, this volume). Thus, individual or group differences that are observed in working memory/attention capabilities may reflect specific differences in PFC functioning. But we need to be cautious and point out the possibility that they reflect discrete differences in posterior cortical areas or, indeed, diffuse variations across many brain regions. An interesting question for further research, then, is whether imaging techniques suggest individual differences in brain structure or volume to be more pronounced in anterior than in posterior cortex. If there is more variation across frontal areas than across parietal or temporal areas, this would tentatively support the notion that individual differences in working memory/attention and \( g_f \) arise from individual variations in frontal lobe functioning.

We also suggest that further behavioral work be aimed at the generality of working memory and controlled attention. We find the evidence for a unitary working memory/attention system to be compelling. We have reviewed findings in which widely varying working memory and \( g_f \) tests predict a broad range of attentional capabilities and in which PFC lesions impair performance on a vast array of working memory and controlled-attention tests. We also offer the following findings as further support for a unitary working mem-
Working Memory and Controlled Attention

There is clearly empirical support for a unitary working memory. In truth, however, the working memory/attention system is probably neither entirely unitary nor entirely separable into domain-specific systems. Instead, we suggest that working memory/attention may be organized similarly to intelligence (e.g., Carroll, 1993; Kyllonen, 1996; Snow et al., 1984), that is, as a hierarchical structure with a general domain-free factor overarching several subordinate domain-specific factors. Just as in intelligence research, general working memory factors appear to account for too much variance to be ignored. However, in some studies (e.g., Shah & Miyake, 1996), significant variance is left to be explained beyond that accounted for by a general factor. We suggest that the behavioral, neuropsychological, and neuroanatomical evidence supports such a hierarchical view of working memory/attention. The specific factors correspond primarily to the domain of to-be-stored information, but the general factor transcends the domain of processing. Our answers to the eight questions posed by the editors of this volume reflect that view (see Table 4.1).

Table 4.1. The Eight Great Questions

(1) Basic Mechanisms and Representations in Working Memory
Encodings and representations are as varied as the formats for perception, emotion, and thought. Maintenance is through exogenous activation from focus of attention, of which rehearsal is one form. However, some knowledge units achieve activation through endogenous emotional salience or as a result of a goal that has reinforcement or emotional salience. Retrieval results from automatic spreading activation but can be supported and guided by planned, effortful search – which is controlled.

(2) The Control and Regulation of Working Memory
The particular units that are active above threshold (i.e., STM) are a result of a variety of factors including the recency with which they were activated, importance of the unit to the task goal, etc. Some units may have a salience tag, which leads to some thoughts coming to consciousness regardless of the immediate task goal. These units typically have a strong emotional color to them. Except for such endogenous activation, maintaining the contents of working memory, particularly in the face of automatically elicited thoughts such as those just mentioned, requires controlled attention.

(3) The Unitary versus Non-Unitary Nature of Working Memory
The myriad representational formats, controlled attention, and the procedures and skills for maintaining activation constitute a system. What we have called working memory capacity or capacity for sustaining attention is not just unitary but is domain free.

continued
Table 4.1, continued

(4) The Nature of Working Memory Limitations
Much, if not all, retrieval is automatic in nature, but controlled processing is necessary to deal with the results of that retrieval. Procedural skills for manipulating language and spatial/visual representations certainly differ across individuals as does episodic and semantic knowledge. Thus, people will differ in knowledge and the skills for manipulating that knowledge. However, people also differ in their capacity for sustaining, maintaining, and shifting attention. This leads to differences in the ability to maintain and to inhibit or suppress activation. The difference in working memory capacity or controlled attention is isomorphic with general fluid intelligence.

(5) The Role of Working Memory in Complex Cognitive Activities
Individual-differences studies have implicated working memory or controlled-attention capacity in nearly every activity on this list. However, individuals can differ in these tasks because of differences in PFC-mediated controlled processing and/or differences in the procedural skill and knowledge necessary to perform the task.

(6) The Relationship of Working Memory to Long-Term Memory and Knowledge
Working memory is a system consisting of LTM units activated above threshold plus controlled attention. The limitations on controlled processing capability can, at least to some extent, be circumvented by expert knowledge and practice on the task. However, even slight changes in the task demands can lead to reemergence of the limitations.

(7) The Relationship of Working Memory to Attention and Consciousness
WM = STM (activated portion of LTM) + controlled attention.

(8) The Biological Implementation of Working Memory
There is considerable evidence supporting the role of the dorsolateral prefrontal cortex of the frontal lobes in the functions of controlled attention, primarily maintenance and suppression. It is also likely the case that the anterior attention system proposed by Posner and Petersen (1990) is involved in these functions. Further, we believe that the patterns of performance we observe between high and low working memory capacity subjects is strikingly similar to differences in psychometric general fluid intelligence and to the differences between intact normals and patients with prefrontal damage. The so-called slave systems would be mediated by the structures appropriate to the domain. Thus, the speech-based coding would be mediated speech centers in the brain.

REFERENCES


---

**Working Memory and Controlled Attention**


Tuholski, S. W. (1997). Examining the relationships among attentional resources, working memory, and fluid intelligence. Unpublished doctoral dissertation, University of South Carolina, Columbus, SC.


