Journal of Experimental Psychology: General
1999, Vol. 128, No. 3, 309–331

0096-3445/99/$3.00

Working Memory, Short-Term Memory, and General Fluid Intelligence: A Latent-Variable Approach

Randall W. Engle
Georgia Institute of Technology

Stephen W. Tuholski
Southern Illinois University at Edwardsville

James E. Laughlin
University of South Carolina

Andrew R. A. Conway
University of Illinois at Chicago

A study was conducted in which 133 participants performed 11 memory tasks (some thought to reflect working memory and some thought to reflect short-term memory), 2 tests of general fluid intelligence, and the Verbal and Quantitative Scholastic Aptitude Tests. Structural equation modeling suggested that short-term and working memories reflect separate but highly related constructs and that many of the tasks used in the literature as working memory tasks reflect a common construct. Working memory shows a strong connection to fluid intelligence, but short-term memory does not. A theory of working memory capacity and general fluid intelligence is proposed: The authors argue that working memory capacity and fluid intelligence reflect the ability to keep a representation active, particularly in the face of interference and distraction. The authors also discuss the relationship of this capability to controlled attention, and the functions of the prefrontal cortex.

> The term short-term memory is often used to refer to a concept quite similar to working memory. (Anderson, 1990, p. 150)
>
> Short-term memory is the type of memory we use when we wish to retain information for a short time to think about it. The short-term store has a working memory component, a sort of mental workspace or sketchpad in the mind, that is used to manipulate information in consciousness. (Seamon & Kenrick, 1994, p. 220)
>
> Working memory is a more complex construct than short-term memory, defined as the set of activated memory elements; there is no reason to doubt that working memory is based on that activated information along with central executive processes. (Cowan, 1995, p. 100)

What is one to conclude about the relationship between working memory (WM) and short-term memory (STM) based on the preceding quotes? The first depicts the two constructs as "similar," the second depicts WM as a subset of STM, and the third depicts just the opposite: that STM is a subset of WM. Ambiguity clearly exists in the field of cognitive psychology regarding the relationship between STM and WM. Given the centrality of these concepts to grand theories of cognition (e.g., Anderson, 1983; Cowan, 1995), remarkably little work has been done to resolve this confusion.

The question addressed here is "To what extent are the terms STM and WM different terms for the same construct and to what extent do they refer to different but more or less related constructs?" Although there has been considerable experimental work on each of the tasks that putatively reflect the two concepts, little work has been directed at the extent to which those tasks share conceptual and construct validity. We submit that these questions, although very important to mainstream cognitive psychology, cannot be answered by experimental studies alone, but require a combination of experimental and regression procedures. Even with tasks that are reported to be WM tasks (e.g., reading span, operation span, computation span, n back, etc.), there is little solid evidence that they, in fact, reflect a common construct. Even if they do reflect a single common construct, we do not know whether the tasks measure processes or structures that are different from those measured by tasks that putatively reflect STM (e.g., simple digit or word span). To answer the questions we set for ourselves here, we analyzed the unique and shared variance across tasks as well as the underlying factor structure of that variance and the extent to which theoretical conceptualizations of the underlying constructs are supported by structural models of the variance in those tasks.

We first discuss one view of the nature of STM and WM and justify why we think both constructs are necessary for a full understanding of cognitive phenomena. We then discuss measurement concerns about these two constructs and the

Randall W. Engle, School of Psychology, Georgia Institute of Technology; Stephen W. Tuholski, Department of Psychology, Southern Illinois University at Edwardsville; James E. Laughlin, Department of Psychology, University of South Carolina; Andrew R. A. Conway, Department of Psychology, University of Illinois at Chicago.

This work was supported by Air Force Office of Scientific Research Grants F49620-93-1-0336 and F49620-97-1-0041 and National Institute of Child Health and Human Development Grant RO1-HD27490-01A1. We express our gratitude to Heidi Dickinson, Mike Kane, Natalie Oransky, Virginia Rosen, and Rebecca Shisler for their help in collecting the data and to Timothy Salthouse and Christopher Hertzog for their helpful comments on the manuscript.

Correspondence concerning this article should be addressed to Randall W. Engle, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332-0170. Electronic mail may be sent to randall.engle@psych.gatech.edu.

tasks used to measure them. The study presented here attempts to differentiate STM and WM on an empirical level. It does so by measuring performance on tasks that, on the basis of research evidence and logic, can be argued to reflect the capacities of STM and WM. Thus, performance will be measured on tasks that tap STM capacity and tasks that tap WM capacity. We can then determine the extent to which those tasks give rise to unique and shared variance and the extent to which there is a need to posit two separate constructs identified as STM and WM. This is done through a series of analyses; the first is an exploratory factor analysis (EFA), followed by a series of confirmatory factor analyses (CFAs) and structural equation models (SEMs). We show that the STM construct and the WM construct predict higher level cognitive performance differentially. This is illustrated by examining the three-way relationship among STM capacity, WM capacity, and general fluid intelligence (*gF*) in one case and the relationship among STM, WM, and a large-scale test of verbal abilities (Verbal Scholastic Aptitude Test [VSAT]) in another case. There is growing evidence associating the construct of central executive or controlled processing capability with *gF* (cf. Duncan, Williams, Nimmo-Smith, & Brown, 1990). Thus, the connection between the constructs of *gF* and WM capacity is important, and specifying the nature of that connection is one of our goals.

Conceptual Difference Between STM and WM

Any treatment of the nature of STM and WM must start with Baddeley and Hitch's (1974) model because every other view discussed here is derived from it. They incorporated much of the earlier STM literature into their concept of slave systems: the articulatory or phonological loop and the visuospatial sketchpad. These terms have subsequently come to refer to a temporary store and a rehearsal mechanism for speech-based and visuospatial-based information, respectively (Baddeley, 1986). The central executive component of the Baddeley and Hitch model was loosely associated with controlled processing and attention but until recently (Baddeley, 1996) has received little empirical and theoretical notice.

Two very influential articles published in the 1980s touted the independence of STM and WM, particularly the assumption of a common resource for all tasks. Klapp, Marshburn, and Lester (1983) demonstrated a task, the missing digit task, that did not appear to rely on either the articulatory loop or central executive component. Brainerd and Kingma (1985) attacked this question from a developmental perspective. They demonstrated that children's performance on reasoning problems was statistically independent of their STM for critical facts on the problems, and that age-related changes in reasoning and STM in their study were independent.

Cowan (1988, 1995) made an important conceptual distinction between STM and WM. He argued that there is a single memory storage system that consists of elements at various levels of activation. The contents of the system can be thought of as long-term memory, with most of the elements being in a relatively inactive state. Some of the elements, however, may be in a higher state of activation (i.e., above some threshold of activation) but outside of the focus of attention. This information is also outside of conscious awareness but can, nevertheless, influence ongoing processing such as in subliminal perception or semantic priming. The elements that are active above resting baseline are considered to be in STM in the same sense that Hebb (1949) thought of activated units. Activation of the items decays rapidly, and a capacity limitation occurs because of the requirement to perform the processes that maintain the activation above threshold.

According to Cowan, another type of capacity limitation corresponds to what James (1890) referred to as primary memory and reflects the focus of attention. Maintaining units in this hyperactivated state requires controlled, limited-capacity attention, and the limitation reflects the capacity of what Baddeley and Hitch (1974) referred to as the central executive. Thus, STM is thought of as those items from long-term memory that are activated above some baseline. WM is thought of as the contents of STM plus the limited-capacity controlled-attention processes associated with the central executive that can be used to maintain some set of those STM units as the focus of attention.

STM is considered a subset of WM. STM is a simple storage component, whereas WM is that storage component as well as an attention component. This view is consistent with Baddeley and Hitch's (1974) original model, except the term STM is retained and viewed as consistent with the slave systems (the articulatory loop and the visuospatial sketchpad). Thus, Cowan's view is that the WM system consists of the contents of STM plus controlled attention. The attention component corresponds to the central executive in Baddeley and Hitch's (1974) model; to what Norman and Shallice (1986) referred to as the supervisory attentional system; and to what Posner and Snyder (1975) and Schneider and Shiffrin (1977) referred to as controlled attention. The central executive may also be related to the anterior attentional system proposed by Posner and Peterson (1990; see Wickelgren, 1997, for a review regarding the relationship between executive processes and the prefrontal cortex).

Measurement of STM and WM

Figure 1 shows a schematic of the underlying constructs relevant to our measurement model. The central executive component is the source of controlled attention that can be used to achieve activation of long-term traces through controlled retrieval, maintain activation through various means, or possibly even dampen activation through inhibition.[1] STM consists of those traces active above threshold with loss of activation as a result of decay or interference. Some very small number of those traces receive increased activation by becoming the focus of attention, a central executive function.

[1] The question of whether true cognitive inhibition occurs is controversial at this time. We have argued elsewhere (Engle, Conway, Tuholski, & Shisler, 1995) that if phenomena such as negative priming reflect true inhibition at the cognitive level, the resources of the central executive would be required.

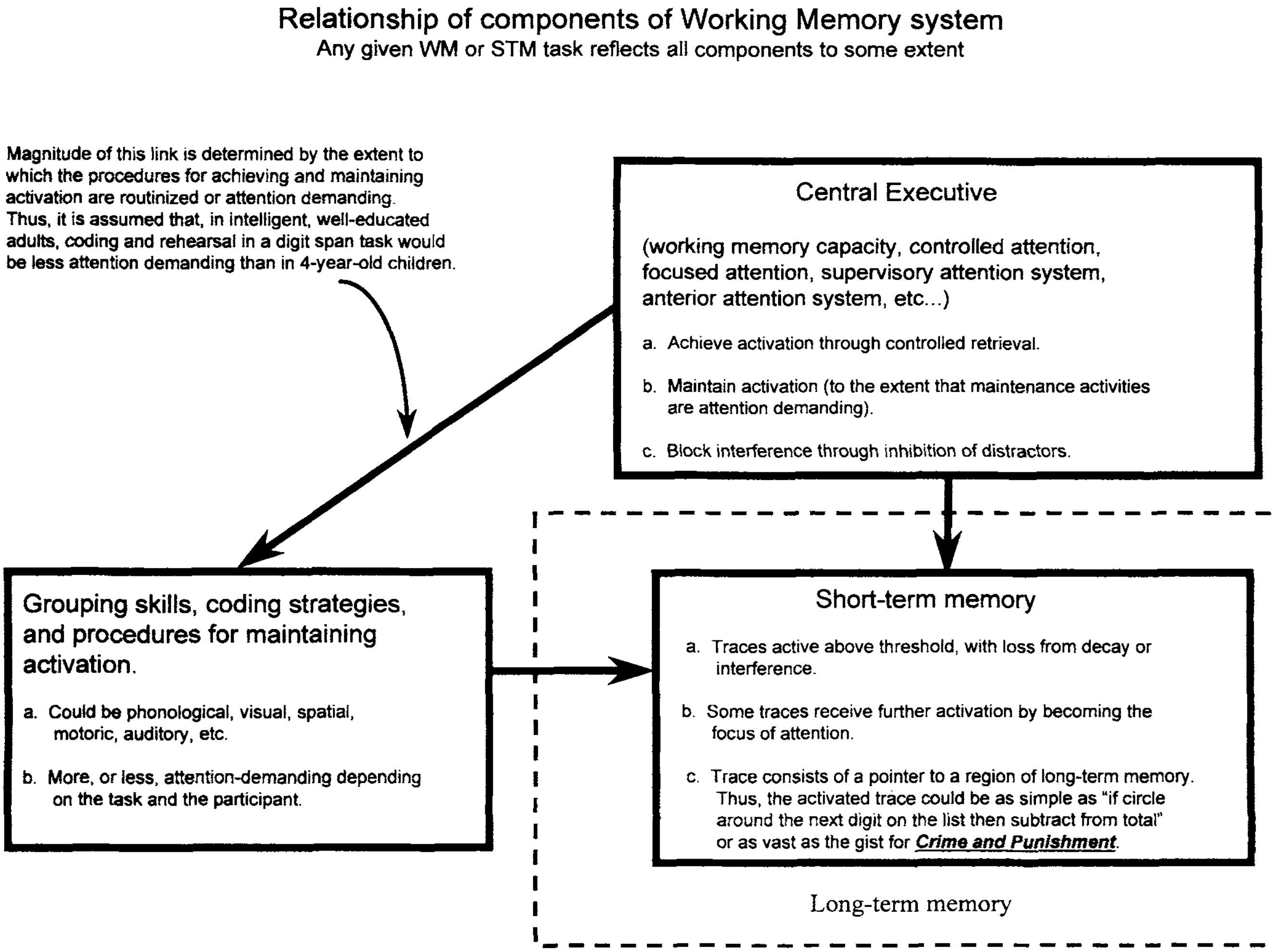

Figure 1. Relationship of the various components of working memory (WM) that are relevant to a measurement model (STM = short-term memory).

STM

Cowan (1995) defined STM as completely and purely a subset of the WM system. Thus, at a conceptual level, shared variance in pure WM tasks and pure STM tasks should reflect the short-term component, with the residual variance in WM tasks reflecting the controlled attention or central executive component of WM. However, variance could be shared between STM and WM tasks at several other levels as well, depending on the specific mental procedures, skills, and strategies used to arrive at and maintain representations of materials used in the tasks. So, for example, if both WM and STM tasks used verbal materials in a serial recall task, then perceptual grouping or chunking skills, skill at phonological coding, and speed of rehearsal of phonological information would all contribute to shared variance between the WM and STM tasks. Likewise, skill at imaginal coding and speed of manipulating visual and spatial images for visual and spatial information would contribute to shared variance to the extent that both types of tasks made use of visuospatial materials and required similar mental procedures. However, to the extent that the two types of tasks required different procedures, there would be unique variance associated with individual differences in the domain-specific skills and abilities.

Another point at which variance would be shared between the two types of tasks is the extent to which both tasks rely on controlled attention. To the extent that the different mental procedures do rely on limited-capacity controlled attention, the variance would be unique to the specific domain. Even if the materials are from different domains and the procedures require different coding strategies, to the extent that the procedures are attention demanding, they would tap variance common to general WM capacity or central executive. For instance, we argue that tasks of STM capacity and tasks of WM capacity would reflect more common variance in children—unskilled at chunking and coding and with less routinized rehearsal—than in adults because these procedures are more attention demanding in children (Cowan, 1995, p. 98). Likewise, the nature of the participants tested in a given sample would determine the degree of covariation in tasks. If a very homogeneous sample of participants was tested with little variation in central executive capability, then such individual differences as did occur would likely reflect domain-specific skills and processes. If the sample includes a wider range of central executive capabilities, then that would likely be reflected in greater covariation in tasks reflecting the domain-free central executive and less covariation in domain-specific tasks.

It is trite but important to state that, to make inferences about concepts such as STM and WM, we are forced to use performance on tasks that are not pure reflections of those constructs. There are neither pure STM nor pure WM tasks. All such tasks rely on all three components of the model in Figure 1 to some extent. Individual differences in the performance of any task that does reflect the contents of STM as we have defined it here will reflect the true-score functioning of STM. However, it will also likely reflect individual differences in such factors as coding, grouping (particularly in the case of digits), manner and mode of rehearsal, and familiarity and knowledge of the items that are used in the memory task (particularly in the case of words). These factors are important in increasing or decreasing the amount recalled in an STM task and in the number of items that are active at any given time. However, to the extent that the measurement of STM capacity is influenced by these factors, the measurement would only generalize to other situations in which those factors also come into play. To the extent that STM tasks demand controlled attention, they will also reflect the central executive or WM capacity construct. Further, what is clearly an STM task for some participants might be primarily a WM task for others. This is likely true not just at different levels of development but also among individuals at a given stage of development depending on intellectual abilities and skill in the task.

WM

Measures of WM capacity presumably reflect the storage component and the central executive or attention component. One possibly confusing aspect of the measurement of WM capacity is that the component that distinguishes it from STM capacity is attention. Thus, WM capacity as distinct from STM capacity is about "memory" only indirectly. It reflects the ability to maintain the activation of knowledge units in the focus of attention. To quote Baddeley (1993), "the central-executive component of working memory does not itself involve storage, which produces the apparently paradoxical conclusion that not all working memory studies need involve memory" (p. 167).

Daneman and Carpenter (1980) developed the first demonstrably valid measure of WM capacity, where validity is judged by the correlation between the measure and higher level measures of cognition such as reading comprehension and reasoning. Their reading span task is actually a dual task that requires the participant to read or listen to a series of sentences and, separately, to keep track of the last word of each sentence so that the words can be recalled later. The span score is the maximum number of words that can be recalled perfectly. However, a wide variety of other complex measures have now been used to measure WM capacity (Kyllonen & Christal, 1990; Salthouse, Mitchell, Skovronek, & Babcock, 1989; Turner & Engle, 1989), and we have argued elsewhere (Cantor, Engle, & Hamilton, 1991) that these measures reflect a common mechanism.

Further, that mechanism is apparently of fundamental importance to higher level cognition because measures of WM capacity reliably predict performance in a wide variety of real-world cognitive tasks and ability measures. Significant relationships with measures of WM 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); note-taking (Kiewra & Benton, 1988); writing (Benton, Kraft, Glover, & Plake, 1984); reasoning (Kyllonen & Christal, 1990); and complex learning (Shute, 1991; Kyllonen & Stephens, 1990).

We have proposed that individual differences on measures of WM capacity primarily reflect differences in controlled-attention capability, particularly in situations involving interference or distraction. Thus, those differences will be reflected only in situations that either encourage or demand controlled attention (Conway & Engle, 1994; Engle et al., 1995; Rosen & Engle, 1997). We propose that controlled attention is necessary (a) when task goals may be lost unless they are actively maintained in WM; (b) where actions competing for responding or response preparation must be scheduled; (c) where conflict among actions must be resolved to prevent error; (d) where there is value in maintaining some task information in the face of distraction and interference; (e) where there is value in suppressing task-irrelevant information; (f) where error monitoring and correction are controlled and effortful; and (g) when controlled, planful search of memory is necessary or useful. We have argued elsewhere (Engle, Kane, & Tuholski, 1999) that the critical factor common to measures of WM capacity and higher level cognitive tasks is the ability to maintain a representation as active in the face of interference from automatically activated representations competing for selection for action and in the face of distractions that would otherwise draw attention away from the currently needed representation.

Our proposal, then, is that WM capacity reflects the amount of activation that can be applied to memory representations that are currently active to either bring them into focus or maintain them in focus or possibly, in the case of suppression, to dampen them from focus.[2] This is particularly important when retrieval of the needed information from a below-threshold state would be slow, difficult, or problematic because of interference. We view this attention

[2] "Focus" can be defined as exceeding a specific and high threshold of activation. Thus, any information active below that threshold would be defined as "outside" the focus of attention. Further, changing the threshold would simultaneously vary the number of units in focus and, inversely, the amount of activation available for each of those units. This idea is quite similar to the flashlight or zoom lens notion of attention proposed by Kahneman (1973).

capability as domain free; therefore, individual differences in this capability will reveal themselves in a wide variety of tasks.

It follows then that individual differences in WM capacity may influence performance on tasks designed to measure STM capacity to the extent that the STM tasks are also attention demanding. As discussed, it is not necessarily that individual differences in WM capacity account for all the individual differences in STM capacity. Other factors, such as rehearsal rate and coding strategies, may affect performance on STM tasks independently of the controlled attention component. Thus, individual differences on STM tasks, although strongly influenced by WM capacity, will reflect some variance independent of WM capacity. The argument is that, at a purely conceptual level,

> WM capacity = STM capacity + central executive or controlled attention + the error of measurement.

Statistically controlling for the variance shared between WM and STM tasks should leave a residual consisting of a component representing the central executive. We can question whether this latter component is responsible for the connection to *gF.*

Predictive Difference Between WM and STM Capacity

If STM and WM capacity actually reflect different cognitive limitations, then an important question is how those two limitations differentially impact performance on higher level cognitive tasks. Do they make independent contributions to performance, is one more important than the other, or does one mediate the contribution of the other? We address these questions as they pertain to the relationship between STM capacity, WM capacity, *gF,* and verbal ability as measured by the VSAT.

When we use the phrase "general intelligence" (*g*), we refer to a latent variable that results from the analysis of intercorrelations among multiple intelligence tests (i.e., Spearman, 1927). The variable *g* has been used to explain the ubiquitous finding that scores on a wide variety of cognitive tasks tend to correlate positively with each other, a phenomenon referred to as "positive manifold." A widely cited account of *g* is provided by Horn and Cattell's fluid-crystallized intelligence theory (Cattell, 1963, 1971; Horn, 1980; Horn & Cattell, 1967). *gF* refers to the ability to solve novel problems and adapt to new situations and is thought to be nonverbal and relatively culture free. Crystallized intelligence, *gC*, alternatively refers to acquired skills and knowledge and depends on educational and cultural background. Tests that measure *gF* include, but are not limited to, matrices and figural analyses, whereas tests that measure *gC* include vocabulary and general knowledge tests (Sattler, 1992). Tests such as the VSAT and Quantitative Scholastic Aptitude Test (QSAT) almost certainly reflect a combination of fluid and crystallized abilities.

Because *g* is a latent variable, great care must be taken to clarify what *gF* actually means. One approach is to conduct large-scale factor analytic studies, including multiple intelligence tests as well as tasks that tap some cognitive ability or mechanism that putatively contributes to *g*, and then examine the relationship between the latent variable that emerges from the intercorrelations among the intelligence measures (*g*) and the other tasks. Kyllonen and Christal (1990), for example, suggested that WM capacity might be the psychological mechanism responsible for *gF.* To test this possibility, they examined the relationship between WM capacity and reasoning ability because reasoning ability is considered a central aspect of *gF* (Carroll, 1989). Kyllonen and Christal (1990) found that correlations between the WM capacity factor and the reasoning factor ranged from .80 to .90. On the basis of this work, we predicted that WM capacity would contribute to a portion of the variance in *gF.*

Evidence suggests that STM capacity is also related to *g* (Bachelder & Denny, 1977a, 1977b; see Carroll, 1993, for a review). However, none of those studies examined the three-way relationship among STM capacity, *gF,* and WM capacity. We argued previously that, to the extent that the STM measures make use of controlled attention, individual differences in WM capacity would influence individual differences in STM. It is, therefore, possible that WM capacity will drive the relationship between STM capacity and *gF.* If that argument is true, then when individual differences in WM capacity are statistically controlled for, the relationship between STM capacity and *gF* should be diminished. Further, when the variance common to WM and STM is controlled for, there should remain a significant relationship between the residual of the WM component and *gF.* One interpretation of this result is that what is common to WM capacity and general fluid abilities is some aspect of controlled attention.

In contrast, if STM contributes to *gF* beyond the contribution by WM capacity, then statistically controlling for individual differences in WM capacity would affect the relationship between STM capacity and *gF.* There is also the possibility that WM and STM contribute differently to measures of *gF* and *gC.* For instance, some of the more common measures of STM are tasks that Baddeley and Hitch showed to benefit from use of speech-based coding or the articulatory loop. We argued previously that the VSAT reflects a combination of general fluid abilities and skills and abilities specific to the verbal domain. Thus, STM tasks may contribute unique variance to the VSAT above and beyond that contributed by measures of the central executive.

Some of these same questions were addressed in an earlier article from our lab. Cantor, Engle, and Hamilton (1991) tested 49 participants on simple digit and word span tasks, reading span tasks modified so that the to-be-recalled elements were digits or words, and probe recall tasks involving the cued recall of the first, second, or last three items in a list of nine digits or words. The experiment was not originally designed as a factor analysis study so the sample size was low. However, a factor analysis supplemental to the experiment suggested that the simple word and digit span tasks and the probe recall digit and word tasks loaded on one factor and the reading span-digit and reading span-word tasks loaded on another factor. We argued that these two factors represent the underlying constructs STM and WM, respectively. Stepwise regression showed that, although these two factors were highly related, the two factors each contributed unique variance to VSAT scores. The present study, using a sample size more appropriate for such analysis, extends the questions addressed by Cantor et al. to include the relationship between STM and WM and their relationship to *gF.*

Choice of Tasks

The choice of tasks was guided by both logic and previous empirical research. Tasks thought to be good STM tasks were (a) simple word span with dissimilar words, (b) simple word span with similar words, and (c) backward word span with dissimilar words. The tasks with similar and dissimilar words were used so that we could assess the role of Baddeley's (1986) phonological loop in the criterion measures, particularly the VSAT. As we stated earlier, all the memory tasks used here likely rely on all three components in Figure 1 to an extent. The backward span task is likely even more of a hybrid than the forward span tasks. Although Jensen and Figueroa (1975) categorized the backward span task as a test of higher level abilities, our research suggests it should fit with other tests of STM. For example, Rosen and Engle (1997) showed that the backward and forward word span tasks displayed similar effects of phonological similarity and similar patterns of correlation with the VSAT, suggesting the same emphasis on the articulatory loop for the two tasks. Likewise, the findings of Cantor et al. (1991) suggest that a simple transposition of order would be insufficient to move a task from the STM category to the WM category, at least in college student participants. Thus, the backward span task may make more demands on the controlled-attention component than the other two STM tasks, but the procedures for performing this task are likely to be proceduralized to an extent that would not make it a good test of controlled attention. More important, all three tasks, including backward span, can be performed with relatively little removal of attention from the representation of the list items. The WM tasks, in contrast, are characterized as dual tasks in that attention must be shifted back and forth between the representation of the list items and the so-called processing component of the task. As will be seen, inclusion of the backward span task as an STM task was supported by the exploratory factor analysis.

Other tasks were chosen as prototypical WM tasks: (a) operation span with words (Turner & Engle, 1989); (b) a modified version of the reading span task (Daneman & Carpenter, 1980); and (c) the counting span, based on a modified version of a task first used by Case, Kurland, and Goldberg (1982). All of these tasks are dual tasks requiring processing and storage (i.e., the calculation of arithmetic strings, reading of a sentence or controlled counting of a quantity of objects, and the retention of words or digits to recall). Our logic was that the tasks required the shifting of attention, alternately, between the representation of the list of items to be recalled and the processing component of the task. A variety of other tasks were used because either logic or research suggested they might reflect WM capacity or central executive functioning: (a) keeping-track task (Yntema, 1963; Zacks, 1982); (b) the ABCD task from the CAM4 battery developed by Kyllonen and his colleagues (e.g., Kyllonen & Christal, 1990); (c) the continuous opposites task from the CAM4 battery; and (d) the random generation task developed by Baddeley (1996). Participants also performed an immediate free recall of 12-word lists. Two scores were generated from the free-recall data, one for primary memory and another for secondary memory (cf. Tulving & Colotla, 1970). Cohen and Sandberg (1977) argued that rehearsal of the middle items in a list is suppressed by the processing of the items at the end of the list. In some sense then, retention of the middle items in a free-recall list reflects the ability to keep those items active and accessible in the face of inattention and distraction from the recency items. This quality could make the secondary memory component a reasonable WM task while the primary memory component could share variance with the STM tasks.

In addition to the memory tasks, participants performed the Raven's Standard Progressive Matrices (Raven, Court, & Raven, 1977) and Cattell Culture Fair Test (Cattell, 1973), both of which are argued to be good measures of *gF* (Carroll, 1993; Snow, Kyllonen, & Marshalek, 1984). Scores for the VSAT and QSAT were obtained from university records. We assumed that the Scholastic Aptitude Test (SAT) measures reflected both general fluid ability and crystallized verbal and quantitative abilities. Thus, the SAT scores represented our best estimate of *gC.*

We were interested in a number of issues, and thus we will use a number of statistical techniques to address them. We first performed an EFA on the memory tasks. This analysis is crucial because it allows us to identify tasks that may not belong in our later analyses as we had supposed a priori. For the later analyses, scores on all the memory tasks and the two intelligence tasks were entered into CFAs and SEMs. One issue concerns the pattern of correlations among the memory measures. If STM capacity and WM capacity are indeed different constructs and our tasks are successful at tapping STM capacity and WM capacity, then a two-factor CFA solution should fit the data better than a one-factor solution. Tasks that putatively tap STM capacity (e.g., simple word span) should load on the STM factor, and tasks that putatively tap WM capacity (e.g., reading span) should load on the WM factor.

We then used SEMs to examine the three-way relationship among the three latent variables: STM, WM, and *gF.* We predicted that the latent constructs for both STM and WM would be related to *gF.* However, on the basis of previous work suggesting the predictive power of WM capacity, we predicted that WM capacity would drive the relationship between STM capacity and *gF.* Therefore, if we statistically controlled for individual differences in WM, the relationship between STM and *gF* would be diminished. In contrast, if we controlled for individual differences in STM, the relationship between WM and *gF* would remain unchanged. In an SEM, this would be manifest as a significant path between WM and *gF*, but a nonsignificant path between STM and *gF*. We also asked whether, if all variance common to STM and WM is removed, the residual of the variance in the WM component, which, theoretically, is controlled attention plus measurement error, is correlated with *gF* above and beyond the relationship between the common factor and *gF*. In other words, if the effects of the STM latent variable are removed from the relationship between the WM latent variable and the *gF* latent variable, is the latter relationship still significant?

We next examined the relationship among the STM and WM factors and VSAT and QSAT. Here the STM factor was predicted, based on the findings of Cantor et al. (1991), to contribute variance to VSAT above and beyond that shared with the WM factor. A further analysis examined the contribution of the phonological loop (as indexed by the difference between the scores for forward span-similar and forward span-dissimilar) to verbal abilities as indexed by VSAT.

Method

Participants

One hundred thirty-three participants completed all tasks in the study. All participants were undergraduates at the University of South Carolina and received course credit or $5 per hour for their participation in the study. The University of South Carolina (Columbia) is a comprehensive state university and enrolls students over a wide range of ability levels. For example, the VSAT scores in the present sample of 133 students ranged from 300 to 800 of a possible 200 to 800.

Tasks

Tasks were administered over the course of 3 days. The order in which tasks were run on each day was counterbalanced so that each task was run in each position (first, second, third, etc.) equally often. Table 1 outlines the different task orders.

In retrospect, it was probably unwise to counterbalance order of task, because it creates the potential for Participant *X* order effects, which would increase error and reduce the power. However, as is obvious later, the data still allowed us to make rather strong conclusions.

On average, approximately 1 week passed between the administration of Day 1 and Day 2 tasks and as much as a month or more between the administration of Day 2 and Day 3 tasks. Day 1 and Day 2 tasks were administered to 1 participant at a time and took 1 and 1.5 hr to complete, respectively. Day 3 tasks were administered in groups of 1 to 5 participants and took approximately 1 hr to complete. All of the WM and STM tasks were conducted on a computer with color monitor and were programmed in MEL. Each of these tasks was performed with the participant tested alone in a room except for the experimenter, who monitored performance on each task.

Recalling words was crucial in many of our tasks, which are described next. For each task that required recall of words, a separate pool of one- or two-syllable high-frequency words was constructed. No words were repeated within or between tasks for a given participant.

Table 1
Order in Which Tasks Were Performed

Day & task	Orders					
Day 1	a	b	c	d	e	
OSPAN	1	5	3	2	4	
RSPAN	2	3	1	4	5	
FSPAND	3	2	4	5	1	
FSPANS	4	1	5	3	2	
BSPAN	5	4	2	1	3	
Day 2	a	b	c	d	e	f
KTRACK	1	5	3	2	6	4
IFR	2	1	4	3	5	6
ABCD	3	4	5	6	1	2
CONTOP	4	3	6	5	2	1
RAND	5	6	2	1	4	3
CSPAN	6	2	1	4	3	5
Day 3	a					
CATTELL	1					
RAVENS	2					

Note. The numbers in each column indicate the order of tasks for that particular day. For example, if a participant was run in Day 1 (a), he or she would first perform OSPAN, then RSPAN, FSPAND, FSPANS, and finally BSPAN. OSPAN = operation span; RSPAN = reading span; FSPAND = forward span, dissimilar; FSPANS = forward span, similar; BSPAN = backward span; KTRACK = keeping track; IFR = immediate free recall; CONTOP = continuous opposites; RAND = random generation; CSPAN = counting span.

WM Tasks

Operation span (OSPAN). Participants saw individual operation-word strings like those that follow centered on the monitor of the computer. They read aloud and solved the math problems, each of which was followed by a lowercase word, and, after a set of these operation-word strings, they recalled the words. For example, in the following set size of three operation-word strings, the participant would read aloud "IS (8/4) − 1 = 1? The participant would answer "yes" if the equation was correct or "no" if the equation was incorrect and then would read aloud the word "bear." On hearing the word "bear," the experimenter would press a key that caused presentation of the next string. This procedure allows adequate time for each individual to process the operation and word but serves to reduce the time for rehearsal. After the last operation-word string in the set, in this case the third string, the participant would see a set of question marks centered on the screen. The question marks cued participants to write down the words that followed the operation strings in the correct serial order.

IS (8/4) − 1 = 1? bear
IS (6 × 2) − 2 = 10? beans
IS (10 × 2) − 6 = 12? dad
???

The number of operation-word strings (set size) presented before the recall cue varied from two to six, with three trials of each set size. Set size varied in the same randomly chosen order for each participant. Thus, the participant could not know the number of words to be recalled until the question marks appeared. Participants who were less than 85% accurate on the arithmetic portion of the task were dropped from the study. The few participants who did not achieve 85% accuracy were removed before testing in Days 2 and 3. The OSPAN score was the cumulative number of words recalled from perfectly recalled trials. That is, the score on this and other tasks discussed later consisted of an accumulation of the number of items from those trials on which the participant recalled all the items in the correct order.

Reading span (RSPAN). Participants read aloud sentences that were shown centered on the monitor while trying to remember unrelated words printed at the end of the sentence. For example, in the following set of three sentences, participants were shown one sentence at a time on the monitor. They read the sentence aloud and then read aloud the word in capital letters. At that point, the experimenter pressed a key that caused the next sentence to be immediately presented. After the last sentence in each set, the participant saw the question marks, which served as a cue that the participant should write down the capitalized words in the correct serial order.

For many years, my family and friends have been working on the farm. SPOT
Because the room was stuffy, Bob went outside for some fresh air. TRAIL
We were fifty miles out at sea before we lost sight of the land. BAND
???

The number of sentence-word combinations (set size) presented before the recall cue varied from two to six, with three trials of each set size. Set size varied in the same randomly chosen order for each participant. After recalling the words, participants were asked a comprehension question about one of the sentences, chosen at random, such as "Did Bob go outside?" The comprehension questions were used to ensure that participants attended to the sentences. Participants with comprehension performance below 85% were dropped before Day 2 testing. The RSPAN score was the cumulative number of words recalled from perfectly recalled trials.

Counting span (CSPAN). The experimenter pressed a key that caused presentation of the initial display. Each display consisted of randomly arranged dark blue circles, dark blue squares, and light blue circles on the monitor. Participants counted the number of dark blue circles aloud and repeated the digit corresponding to the final tally. For example, if the display contained three dark blue circles, the participant would say aloud "one-two-three-three." When the "three" was repeated, the experimenter pressed a key that caused immediate presentation of the next display, and counting was to begin immediately. The number of targets per display varied from three to nine, with three trials of each. The number of color distractors (light blue circles) varied from one to five, and the number of shape distractors (dark blue squares) was one, three, five, seven, or nine. After two to eight displays, a recall cue (RECALL) was presented, at which point participants wrote down the number of targets in each of the previous displays, in the serial order in which they occurred. It should be noted that the memory component of this task is essentially a digit span task with counting of objects interwoven with "presentation" of the digits to be recalled. The experimenter monitored counting performance, and all participants with more than 15% errors were to be eliminated, but few errors in counting occurred and no participants were dropped. The CSPAN score was the cumulative number of digits recalled from perfectly recalled trials.

STM Tasks

Forward span-dissimilar (FSPAND). Participants were presented visually with nonrhyming words at the rate of one word per second. Each word was shown in lowercase letters centered on the screen. The number of words per trial incremented from two to seven, with three trials of each. The participants read aloud each word as it appeared on the monitor. After the final word, a recall cue (???) was presented that prompted participants to write down the words in correct serial order. The words were one- and two-syllable high-frequency words, and no words were repeated within the task or used in any other task. The FSPAND score was the cumulative number of words recalled from perfectly recalled trials.

Forward span-similar (FSPANS). This task was identical to FSPAND with one exception. As in FSPAND, the number of words per trial incremented from two to seven, with three trials of each. In this task, however, all the words on each trial rhymed with each other (e.g., dog, hog, bog, log). However, the words did not rhyme or repeat across trials. The words were one- and two-syllable high-frequency words, and no words were repeated within the task or used in any other task.

Backward Span (BSPAN). This task was identical to FSPAND with one exception. In this task, participants were asked to recall the nonrhyming words in the reverse order in which they were presented. Thus, if a participant was presented with "den, pole, car," the correct recall sequence was "car, pole, den."

gF Tasks

Cattell's Culture Fair Test (CATTELL). The CATTELL Test (Cattell, 1973) is composed of four separate and timed paper-and-pencil subtests. Per standard instructions, participants were allowed 2.5 to 4 min to complete each subtest. When time expired for a subtest, participants were instructed to stop working on that subtest and begin the next. At no point were participants allowed to go back to work on previous subtests.

In the first subtest, *Series,* participants saw 13 incomplete, progressive series of abstract shapes and figures, along with 6 alternatives for each, and selected the alternative that best completed the series. In the second subtest, *Classifications,* participants saw 14 problems composed of abstract shapes and figures, and selected the two out of the five that differed from the other three. Figures and shapes differed in size, orientation, or content. The third subtest was *Matrices,* which is similar to the Raven's test described later. In the matrices task, participants were presented with 13 incomplete matrices containing four to nine boxes that had abstract figures and shapes as well as an empty box and six choices. Participants had to infer the relationships among the items in the matrix and choose an answer that correctly completed each matrix. In the final subtest, *Conditions,* participants saw 10 sets of abstract figures consisting of lines and a single dot along with five alternatives. The participants had to assess the relationship among the dot, figures, and lines, and choose the alternative in which a dot could be placed according to the same relationship. For example, if the dot had been placed inside a circle but outside a square, the participant needed to choose the alternative in which a single dot could be placed inside a circle but outside of a square. The CATTELL score was the sum of all correct answers across all four subtests.

Raven's Progressive Matrices (RAVENS). In this standardized test (Raven, Court, & Raven, 1977), participants were presented with 60 patterns-matrices composed of abstract shapes, lines, and nonverbal figures, each of which was missing a piece. For each pattern, six choices (pieces) were presented. The participants had to choose the piece that fit best in the empty space. Participants had as much time as they needed to complete this task. The total number of correct choices was used as the RAVENS score.

Additional Tasks

Keeping track (KTRACK). This task is quite old (Yntema, 1963) and has been used to study situations in which the participant keeps track of the most recent level of several changing variables (Zacks, 1982). In the present version, participants saw words that were exemplars from one to six different categories (metals, animals, colors, distances, countries, relatives). Category exemplars were selected from the Battig and Montague (1969) norms. Each word appeared at the center of the monitor screen in lowercase letters at the rate of one word every 2 s. Before presentation of the words, one to six category names were presented on the center of the screen until the participant cleared them. Participants performed 3 trials of each size, for a total of 18 trials. When they were ready, participants signaled the experimenter to begin the trial.

Participants kept track of (i.e., remembered) the last exemplar of each category that was presented. Thus, if a participant was presented with "distances" and "countries," the only two words he or she needed to remember were the last exemplar of a distance and a country. To illustrate, if the list of words presented on this hypothetical trial were "iron," "dog," "gold," "mile," "United States," "meter," "silver," "cat," "uncle," "France," and "brother," the participant would recall "meter" and "France." For each trial, there were one to four exemplars from each to-be-remembered category as well as words from other categories, totaling 15 words per trial. We varied the number of words presented from the to-be-remembered category so that participants would be unable to predict which word was the last exemplar. After the last word, a ??? recall cue was presented, which signaled the participant to write down the last exemplar of each of the target categories shown at the beginning of the trial. A trial was scored as correct if all correct exemplars were recalled. The KTRACK score was the cumulative number of words recalled from correctly recalled trials.

Immediate free recall. In this task, 12 lowercase words appeared at the center of a computer screen at the rate of one word per second. Participants read each word aloud as it appeared on the screen. After the 12th word, a ??? recall cue was presented. At recall, participants wrote the words down in any order they wished, but they were encouraged to write the last few words from the list first. Each participant performed 10 trials. Two scores were taken from these data, following a procedure described by Tulving and Colotla (1970). For each word recalled, the number of words between its recall and its presentation was counted. If there were seven or fewer words between the recall and presentation of a word, that word was said to be recalled from primary memory. If there were eight or more words between recall and presentation of a word, that word was said to be recalled from secondary memory. The Immediate Free Recall Primary Memory (IFRPM) score was the average number of words recalled by each participant from primary memory. The Immediate Free Recall Secondary Memory score was the average number of words recalled by each participant from secondary memory.

ABCD task. ABCD is a verbal reasoning task and a subtest of the CAM4 battery, which Kyllonen and Christal (1990) developed. Participants were presented with three different premises that described a *spatial* situation (i.e., The DOG is before the CAT). After the third premise, different situations were presented for the participant to choose from. The eight different situations depicted the two categories (in the present case, animals and furniture) and two exemplars of each (dog-cat and lamp-couch) in a variety of spatial juxtapositions. The participant's task was to select the situation that the three premises described. This task would appear to rely on the verbatim speech-based representation of the words and the ability to manipulate each word spatially to form a representation that could be used for the recognition test. Each premise was presented for 5 s, at which point it was replaced by the next premise. After the offset of the final premise, participants had to choose one of eight spatial situations that was consistent with all three premises. Participants had 15 s to make a response. If they did not respond in 15 s, that trial was scored as incorrect. To illustrate the task further, an example trial is presented next.

The participant could be presented with the following premises:

The ANIMALS are before the FURNITURE.
The DOG is before the CAT.
The LAMP is before the COUCH.

After the last premise, eight choices were given to the participant such as

1. dog cat lamp couch
2. cat dog lamp couch
3. dog cat couch lamp
4. cat dog couch lamp
5. lamp couch dog cat
6. couch lamp dog cat
7. lamp couch cat dog
8. couch lamp cat dog

In this example, the correct answer is 1. Participants received accuracy feedback after each trial and summary feedback after each of three blocks of eight trials. The ABCD score was the percentage of correct responses.

Continuous opposites (CONTOP). In this task, also a subtest of the CAM4 battery, participants saw three to eight words in capital letters at the rate of one word every 2 s. The participants' task was then to recognize the last three words or their opposites, depending on the color of the words. When a word was presented in white, participants were told to remember that word. When a word was presented in red, they were told to remember the opposite of that word (i.e., the word BAD presented in red should be remembered as GOOD). The words used in this task were "BAD," "GOOD," "DRY," "WET," "FAST," "SLOW," "NEW," and "OLD." Right–wrong feedback was provided after each trial, along with summary feedback after each block of eight trials. The CONTOP score was the percentage of correct responses.

Random generation (RAND). In this task, participants randomly generated numbers (1–9) in synchrony with a recorded tone, which occurred every 1.5 s. Each time the tone sounded, participants said a number, at which point the experimenter keyed in the response to a computer for later scoring for a total of 122 responses. Before beginning, participants were instructed on what "random" meant. We told them to imagine they were placing their hand into a bag that had nine balls, each with a number on it, and that each time a tone sounded, they should imagine removing a ball, reading its contents, and placing it back in the bag. To determine the level of randomness in each participants' output, we implemented an algorithm similar to one used by Baddeley (1996). The score was the difference between the observed and expected probability that any given digit would follow any other digit by chance summed over digits and over the testing period. For each participant, a 9×9 matrix was constructed to derive the conditional probability that any given digit would follow any other digit. The value of each cell was the squared residual between an expected probability and the observed proportion. The score was derived in two different ways. One was to sum across each cell of the matrix, which allowed scores greater than 1.0, and the other was to average across all cells of the matrix, which had a range of 0 to 1. The two methods correlated .81, and the former is presented here. The methods do not correlate perfectly because when calculating the expected probabilities, the former method took into account the frequency with which each key was pressed, whereas the latter method assumed each key was pressed equally often. A score of 0 represented perfect randomness, and higher scores reflected less randomness.

Results

Preliminary Analyses

Correlations among all variables are presented in Table 2. Several of the memory tasks consisted of three different presentations at each set size or list length (e.g., from two to seven items for recall). By combining the first presentation of all the sets of different lengths into a single score, the second presentation into a single score, and the third presentation into a single score, we obtained three subscores for each task, and these were used to compute Cronbach's α as a measure of reliability. When available, these are shown at the bottom of Table 2. Another, if less desirable, estimate of reliability is the multiple R^2 for each task or test, which is also shown at the bottom of Table 2. This measure reflects the degree of relationship between a variable and all other variables in the study. The presumption is that if a task is reliable and also reflects some covariation with the other tasks, the multiple R^2 would be high. As expected, there is a general pattern of positive intercorrelation among measures. Two exceptions to this trend are the RAND and IFRPM measures, both of which correlate less than .2 with the other measures. Before testing our hypotheses with CFA and SEM, we performed a series of EFAs to assess how well RAND and IFRPM associated with our other memory measures.

Each EFA was estimated with the maximum-likelihood procedure and rotated to final solution with a promax rotation, which is an oblique-correlated solution. Because of its obvious lack of association with the other measures, RAND was dropped from all further analyses. The IFRPM measure was retained for one CFA, described next, although that analysis also demonstrated that it did not share a significant amount of variance with any of our measures. The EFA showed two factors with eigenvalues greater than 1 (4.45 and 1.09), and a scree test also supported a two-factor solution. The two unrotated factors accounted for 40% and 10% of the variance, respectively. The two rotated factors accounted for 30% (10% unique) and 32% (12% unique), reflecting the high correlation of the two factors. The two-factor maximum-likelihood EFA factor solution, which included all memory tasks except RAND and IFRPM, is presented in Table 3. Note that the EFA provides preliminary evidence for the notion that two factors are needed to describe our memory data. The only tasks with loadings above .5 were the three we had chosen to reflect the WM construct and the three we had chosen to reflect the STM construct. If we consider those loadings above .3, then all tasks loaded as predicted except CONTOP, which loaded on the STM factor. This hypothesis is tested directly with CFA next.

CFAs and SEMs

CFA was implemented to address our measurement question, "Are WM and STM distinguishable, or are they the same construct?" We chose CFA to address this question because it allows us to test explicitly which of these "models" is most consistent with our data. Support for a particular model in CFA is based on the pattern of correlations obtained among observed variables. Briefly, observed measures hypothesized to tap a particular factor or latent variable should correlate at least moderately among themselves (a reflection of convergent validity) and less so with measures hypothesized to tap a different factor or latent variable (a reflection of discriminant validity). Thus, for example, if OSPAN, RSPAN, and CSPAN primarily reflect WM as hypothesized, then they should be more highly correlated among themselves than they are with FSPAND, FSPANS, and BSPAN, which are hypothesized to reflect primarily STM. If a particular model is found to fit data well under CFA, then it means that the constructs hypothesized by the model reflect an adequate degree of convergent and divergent validity.

SEM was used to address our question of whether WM and STM differentially relate to higher order cognitive functioning such as general *gF* and verbal comprehension. The SEM analysis takes CFA one step further by allowing hypotheses to be considered about how underlying factors specifically influence each other. That is, SEM allowed us to specify and test a specific pattern of relationships among latent variables. In our case, it allowed us to evaluate the relative contributions of STM and WM factors in explaining variation in *gF* and SAT scores.

For both types of modeling, we fit covariance matrices using the maximum-likelihood procedure in CALIS (SAS 6.11: SAS Institute, 1990). The parameter estimates that are provided in CFA and SEM models are similar to regression weights and, as such, can be reported in either unstandardized or standardized form. As in regression analyses, standardized weights are estimated by transforming all measures to the same scale, whereas unstandardized weights result from potentially different scales for all variables. Because the magnitude of unstandardized coefficients are hard to interpret (as they are in general regression procedures), only standardized estimates are reported. Following the recommendations of Hoyle and Panter (1995), we evaluated model fit by examining a combination of absolute and incremental fit statistics. The absolute fit statistics included the traditional chi-square test of "exact" model fit, the chi-square test of "close" model fit (Browne & Cudeck, 1993), the goodness-of-fit index (GFI) and adjusted GFI (AGFI: Jöreskog & Sörbom, 1981), and the root-mean-square error of approximation (RMSEA: Steiger & Lind, 1980). The incremental fit statistics included the type 2 Tucker and Lewis index (TLI: Tucker & Lewis, 1973), and the type 3 comparative fit index (CFI: Bentler, 1989). It is important to note that, for the chi-square tests, the hypothesis being tested assumes either an exact model fit or an acceptably close model fit. Thus, a good-fitting model is indicated by nonsignificant results from these tests. Because experimental psychologists are generally trained with classical hypothesis testing in which rejecting the null hypothesis is desirable, we will repeat when necessary that nonsignificant chi-square indicates a well-fit model. For the GFI, AGFI,

Table 2
Correlation Matrix for All Measures

Target measures	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
WM																
1. OSPAN	—															
2. RSPAN	.51	—														
3. CSPAN	.47	.32	—													
STM																
4. BSPAN	.35	.35	.43	—												
5. FSPAND	.37	.35	.31	.54	—											
6. FSPANS	.38	.31	.31	.44	.59	—										
Other measures																
7. KTRACK	.36	.22	.34	.30	.32	.27	—									
8. IFRSM	.46	.36	.37	.29	.43	.39	.34	—								
9. IFRPM	.29	.21	.26	.21	.29	.25	.21	−.02	—							
10. ABCD	.34	.28	.45	.36	.36	.34	.23	.28	.25	—						
11. CONTOP	.36	.29	.31	.37	.38	.45	.33	.28	.33	.53	—					
12. RAND	−.18	−.16	−.14	−.10	−.17	−.06	−.15	−.05	−.26	−.01	−.10	—				
Fluid intelligence																
13. CATTELL	.28	.24	.29	.19	.05	.20	.31	.21	.06	.44	.30	.06	—			
14. RAVENS	.34	.28	.32	.27	.19	.21	.37	.24	.18	.43	.38	−.05	.68	—		
SAT																
15. VSAT	.49	.35	.36	.31	.47	.49	.38	.36	.22	.39	.36	−.12	.45	.47	—	
16. QSAT	.46	.25	.42	.34	.36	.44	.40	.37	.29	.41	.46	−.23	.59	.61	.74	—
Cronbach's alpha	.69	.53	.75	.69	.74	.61	.50									
Multiple R^2	.61	.40	.36	.45	.64	.51	.22	.31	.13	.30	.34	.04	.56	.83	.48	.60

Note. OSPAN = operation span; RSPAN = reading span; CSPAN = counting span; BSPAN = backward span; FSPAND = forward span, dissimilar; FSPANS = forward span, similar; KTRACK = keeping track; IFRSM = Immediate Free Recall Secondary Memory; IFRPM = Immediate Free Recall Primary Memory; CONTOP = continuous opposites; RAND = random generation; SAT = Scholastic Aptitude Test; VSAT = Verbal Scholastic Aptitude Test; QSAT = Quantitative Scholastic Aptitude Test; WM = working memory; STM = short-term memory. The mean and standard deviation for each measure with all other measures, except IFRPM and RAND, are shown at the bottom of the table.

Table 3
Factor Loadings for the Maximum-Likelihood EFA Factor Solution

Measure	Factor 1	Factor 2
OSPAN	**0.7693**	*0.0074*
CSPAN	**0.6094**	*0.0543*
RSPAN	**0.5378**	*0.0842*
IFRSM	0.4257	*0.2303*
KTRACK	0.3721	*0.1651*
ABCD	0.3691	*0.2559*
FSPAND	*0.0213*	**0.8262**
FSPANS	*0.0673*	**0.6782**
BSPAN	*0.1844*	**0.5244**
CONTOP	*0.2927*	0.3573

Note. OSPAN = operation span; RSPAN = reading span; CSPAN = counting span; BSPAN = backward span; FSPAND = forward span, dissimilar; FSPANS = forward span, similar; KTRACK = keeping track; IFRSM = Immediate Free Recall Secondary Memory; CONTOP = continuous opposites. All loadings below .3 are set in italics, those above .3 are set in regular type, and those above .5 are set in boldface font.

TLI, and CFI indices, we follow the general guideline that "good"-fitting models are indicated by a value of .90 or more. Finally, for interpreting the RMSEA statistic, we follow the recommendation that values of .05 or less indicate a "good"-fitting model (Browne & Mels, 1992; Steiger, 1989).

Descriptive statistics for all variables are presented in Table 4. Note that in some cases a moderate level of skewness or kurtosis was reflected in the univariate distributions. Because the maximum-likelihood estimation procedure and chi-square tests assume multivariate normality, we took several steps to evaluate whether our results might be sensitive to violation of this assumption. First, we examined Mardia's (1970) normalized test for multivariate kurtosis. As noted by Hoyle and Panter (1995), excessive multivariate kurtosis can be a potentially serious problem for normal theory estimators such as maximum likelihood. The Mardia statistic, which follows an approximate standard normal distribution, turned out to be .88 for our data, which indicated that there was not a significant degree of multivariate kurtosis present in the data. Second, we ran the CFA and SEM analyses described later twice. The first time we used the original raw scores on our measures, and the second time we used transformed scores introduced to reduce the level of skewness and kurtosis in the univariate distributions. When we compared the results of the two sets of analyses, we found no instance in which a chi-square test or a fit statistic differed in interpretation. Further, there was very little difference in the magnitude of standardized parameter estimates. Thus, we concluded that our results were not likely sensitive to any departure from the multivariate normal assumption. To facilitate interpretation, we report only the analyses on raw scores.

Table 4
Descriptive Statistics for All Measures

Measure	*M*	*SD*	Skew	Kurtosis
OSPAN	14.6	7.5	0.86	0.94
RSPAN	10.6	5.9	0.69	1.38
CSPAN	32.7	15.9	0.80	0.16
BSPAN	28.2	9.2	0.63	0.51
FSPANS	28.1	7.9	0.72	0.40
FSPAND	36.0	9.7	0.57	0.52
KTRACK	15.3	5.9	0.87	1.16
IFRSM	1.9	0.9	0.97	1.49
ABCD	81.4	16.6	−1.37	1.81
CONTOP	78.7	18.7	−0.99	0.88
RAND	41.8	12.8	1.44	2.53
RAVENS	50.7	5.9	−1.29	2.15
CATTELL	27.4	4.3	−0.63	0.60
VSAT	552.2	90.4	0.04	0.09
QSAT	555.9	83.5	0.25	−0.09

Note. OSPAN = operation span; RSPAN = reading span; CSPAN = counting span; BSPAN = backward span; FSPAND = forward span, dissimilar; FSPANS = forward span, similar; KTRACK = keeping track; IFRSM = Immediate Free Recall Secondary Memory; CONTOP = continuous opposites; RAND = random generation; VSAT = Verbal Scholastic Aptitude Test; QSAT = Quantitative Scholastic Aptitude Test.

Can We Distinguish Two Memory Systems?

One of our interests was to evaluate whether there was evidence for separate STM and WM constructs. To provide a clear test of the two-construct hypothesis, we fit separate one-factor and two-factor CFA models to the six target memory tasks we discussed previously (WM: OSPAN, RSPAN, and CSPAN; STM: FSPAND, FSPANS, and BSPAN). In the two-factor model, referred to as model A_2, OSPAN, RSPAN, and CSPAN were linked to one latent variable, whereas FSPAND, FSPANS, and BSPAN were linked to a second latent variable. The two latent variables were allowed to correlate freely with each other. The one-factor model, A_1, was the equivalent of setting the relationship between the STM and WM factors to a perfect correlation of 1.0. Thus, the one-factor model is consistent with a single general memory system, whereas the two-factor model is consistent with the idea that the two constructs are distinguishable.

Model fit statistics for the two analyses are shown in Table 5. Parameter estimates for the two-factor model are displayed in Figure 2a. In all figures, circles represent latent variables, and squares represent observed or manifest variables. Numbers on paths (path coefficients) leading from latent variables to observed variables indicate the degree to which an observed variable is influenced by a particular latent variable. Paths leading to observed variables that are not attached to latent variables reflect residual or the degree to which the observed variable is influenced by unique factors.

To reiterate, a well-fitting model is one in which (a) chi-square tests are nonsignificant, (b) the RMSEA estimate is below .05, and (c) the GFI, AGFI, TLI, and CFI estimates are above .90. Clearly from Table 5, the one-factor model (A_1) did not provide an adequate fit to the data. Six of the seven fit statistics did not meet the established criteria for a good-fitting model (only the GFI index met criterion). In

Table 5
Fit Statistics for CFA and SEM

Model	*df*	χ^2 (exact)	*p*(exact)	*p*(close)	GFI	AGFI	TLI	CFI	RMSEA
A_1	9	32.59	**0.00**	**0.00**	0.92	**0.81**	**0.82**	**0.89**	**0.14**
A_2	8	10.96	0.20	0.41	0.97	0.93	0.97	0.99	0.05
B_1	45	69.51	0.01	**0.21**	0.92	**0.88**	0.93	0.94	**0.06**
B_2	37	48.59	0.10	0.49	0.94	0.90	0.96	0.97	0.05
C_1	25	24.05	0.52	0.85	0.96	0.94	0.99	0.99	0.01
C_2	24	23.51	0.49	0.83	0.96	0.94	0.99	0.99	0.01
D	15	19.60	0.19	0.47	0.97	0.92	0.97	0.99	0.05
E	23	31.39	0.11	0.43	0.95	0.91	0.97	0.98	0.05

Note. *p*(exact) and *p*(close) reflect the significance levels for the χ^2(exact) and χ^2(close) tests. CFA = confirmatory factor analysis; SEM = structural equation model; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; TLI = Tucker and Lewis Index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation. Values in boldface indicate those that fail to meet criteria necessary to support the model for that row. Values in regular print meet the criteria for support of the model.

sharp contrast, all seven fit statistics indicated the two-factor model fit the data well. Further, because these two models were hierarchical, we were able to test whether the two-factor model fit the data significantly better than the one-factor model. This test was highly significant, $\chi^2(1) = 22.64$, $p < .01$, supporting the conclusion that the two-factor model fits better than the one-factor model. Thus, there appears to be strong support for the hypothesis that the three target WM tasks and three target STM tasks reflect different latent variables. Although the evidence points to two distinguishable latent variables, it should be noted that the estimated correlation between these two factors is quite high (.68). We address this strong relationship in the General Discussion.

How Well Do the Rest of the Tasks Measure STM and WM?

We next explored how the remaining tasks measured the STM and WM factors identified previously. To accomplish this, we fit a two-factor CFA model to the six target variables and five additional measures (ABCD, CONTOP, IFRSM, KTRACK, and IFRPM). We maintained the integrity of the two factors identified in our first analysis by setting the target measure–latent variable paths to the estimated values obtained in that analysis. The remaining five tasks were allowed to vary freely on either factor. Fitting the model in this way allowed us to estimate more cleanly how each additional task measured STM and WM without distorting the character of these latent variables. As discussed previously, our hypothesis was that the ABCD, CONTOP, IFRSM, and KTRACK tasks would load more highly on WM than STM, and that the IFRPM task would show the converse pattern.

Model fit statistics for this CFA (the Model B_1 results) are shown in Table 5. Although four of the seven fit statistics indicated a good-fitting model, three just missed the criterion of good fit (chi-square exact test, AGFI, and RMSEA). On examination of parameter estimates and correlations among the observed measures, it became clear that the IFRPM task was not a good indicator of either STM or WM. It did not significantly load on either latent variable, and its pattern of correlations with the other 10 tasks was quite low. We thus revised the original analysis by dropping the IFRPM task from the model.

Fit statistics for the revised model, the Model B_2 results, are also presented in Table 5. Parameter estimates for this model are displayed in Figure 2b. All seven fit statistics indicate the revised model fits the data very well. In line with our hypotheses ABCD, IFRSM, and KTRACK did significantly load on the WM factor but not on the STM factor. Contrary to our hypothesis, however, CONTOP significantly loaded on the STM factor and not the WM factor. Thus, among our original 12 tasks, we have identified (a) three target measures (OSPAN, RSPAN, and CSPAN) and three additional tasks (ABCD, IFRSM, and KTRACK) that primarily tap WM or central executive capacity, (b) three target measures (FSPAND, FSPANS, and BSPAN) and one additional task (CONTOP) that primarily tap STM, and (c) two tasks (IFRPM and RAND) that do not clearly associate with either latent variable.

How Do WM and STM Relate to gF?

To evaluate our hypothesis on the relationship between the two memory constructs and *gF*, we tested two alternative SEMs based on the six target memory measures and our two measures of *gF* (the RAVENS and CATTELL tests). Again, to maintain the integrity of the two original memory factors, we fixed the target measure–latent variable paths to the estimated values obtained for the original two-factor model. Further, for both SEM models, we allowed RAVENS and CATTELL to load freely on a third factor (the *gF* factor). The difference between the two models centered on the paths between the memory and *gF* latent variables. In the first model, we incorporated our hypothesis that WM would influence *gF* but that STM would not. This was accomplished by specifying a path from WM to *gF* but not between STM and *gF* (to be referred to as model C_1). In the second

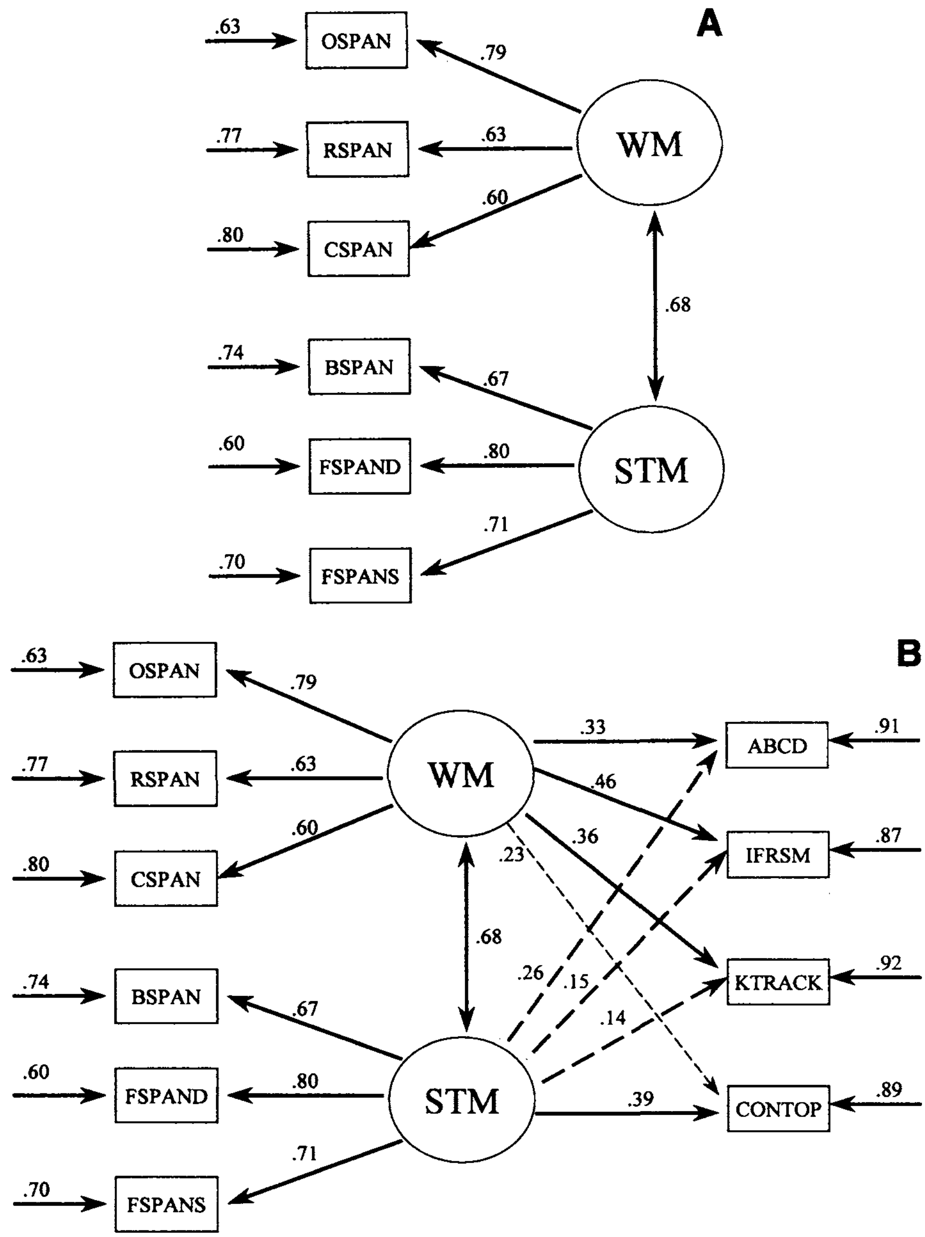

Figure 2. (a) Path model for two-factor model (A_1). All paths are significant at the .05 level. (b) Path model for two-factor model with additional tasks (B_2). Paths significant at the .05 level are indicated by solid lines. OSPAN = operation span; RSPAN = reading span; CSPAN = counting span; BSPAN = backward span; FSPAND = forward span, dissimilar; FSPANS = forward span, similar; KTRACK = keeping track; IFRSM = Immediate Free Recall Secondary Memory; CONTOP = continuous opposites; WM = working memory; STM = short-term memory.

model, referred to as model C_2, we allowed a direct path between STM and *gF.*

Fit statistics for these two SEMs are shown in Table 5. Parameter estimates for the two models are displayed in Figure 3a and b. On the basis of the fit statistics, both models provide an excellent fit. In both models, the estimated path between WM and *gF* is statistically significant and quite high. There are two ways to evaluate the hypothesis that no direct path need be specified between STM and *gF.* First, we can test whether the second model (which incorporates the path) provides a significantly better fit than the first model (which does not incorporate the path). This test is clearly not significant, $\chi^2(1) = 0.54$. Additionally, we tested whether the link between STM and *gF* in the second model was significant with a *t* test. It was not ($t = -0.72$), confirming that the link did not need to be there. Thus, model C_1 is better

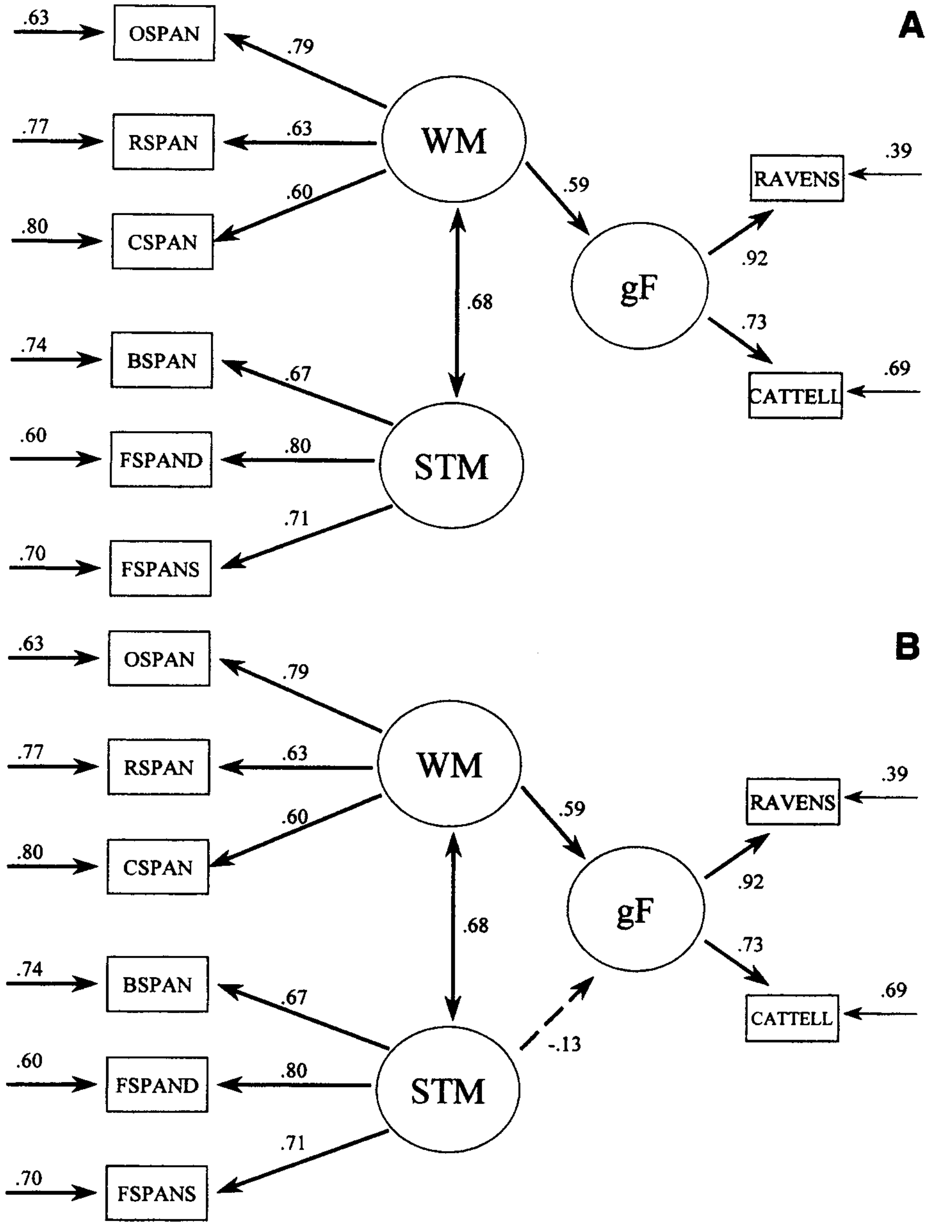

Figure 3. (a) Path model for Model C_1. All paths are significant at the .05 level. (b) Path model for Model C_2. Paths significant at the .05 level are indicated by solid lines. OSPAN = operation span; RSPAN = reading span; CSPAN = counting span; BSPAN = backward span; FSPAND = forward span, dissimilar; FSPANS = forward span, similar; WM = working memory; STM = short-term memory; *gF* = fluid intelligence.

on the basis of parsimony, and it supports our hypothesis that WM, but not STM, is directly linked to *gF*.

As an alternative way to express these relationships between WM, STM, and *gF*, recall our previous conceptualization of WM as

$$\text{WM} = \text{STM} + \text{central executive}.$$

If that is true, then factoring out the variance common to the WM and STM latent variables should leave a residual that represents the central executive, presumed to reflect primarily controlled attention. Further, because we have argued that it is actually the central executive component that accounts for the relationship between WM and *gF*, that residual should show a significant correlation with the *gF* latent variable. Restating this argument in modeling terms, if a common latent variable is removed from WM and STM and is then linked to *gF*, it should reflect the degree of STM variance in all three constructs. The variance in WM that is

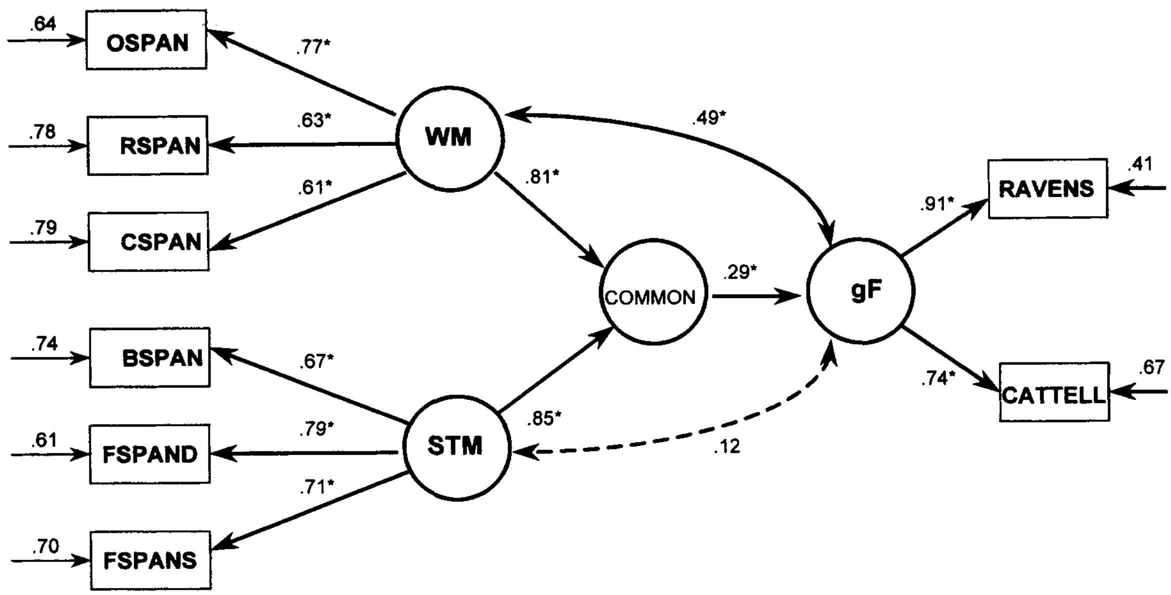

Figure 4. Path model for Model D. Significant paths are indicated by an asterisk. OSPAN = operation span; RSPAN = reading span; CSPAN = counting span; BSPAN = backward span; FSPAND = forward span, dissimilar; FSPANS = forward span, similar; WM = working memory; STM = short-term memory; *gF* = fluid intelligence.

explained by this common memory factor (i.e., the WM residual) should reflect both central executive and error, whereas the STM residual should reflect just error.[3] Thus, if we fit such a common memory factor model to our data, we would expect to find a significant residual correlation between WM and *gF* but not between STM and *gF*. Fit statistics for this model are presented in Table 5, Model D, and parameter estimates are shown in Figure 4. The model was a good fit on all indicators and shows a significant correlation of .49 between the residual of WM (after the STM component has been removed) and *gF*. The correlation between the residual of the STM component and *gF* was not significant, however.[4]

The correlation of .49 between the residual of WM (after controlling for STM) and *gF* probably underestimates the relationship between the derived central executive component and *gF*. Following our earlier argument, WM should influence STM to the extent that STM tasks are attention demanding and unproceduralized. Thus, the common factor in Figure 4, which correlated with *gF* at .29, also presumably includes some influences of controlled attention removed from the residual of WM. This analysis supports the idea that the central executive component of WM is responsible for the relationship among WM tasks such as OSPAN, RSPAN, and CSPAN, and measures of general *gF* above and beyond variance common to STM.

How Do WM and STM Relate to SAT Performance?

Our final analysis explored the relationship between the two memory factors and performance on the VSAT and QSAT. Our thinking was that both standardized tests reflect a combination of general fluid and general crystallized abilities. To the extent that STM reflects skills specific to the domain and materials in common with one of the tests, STM should account for unique variance in the test independently of WM, but WM should account for the *gF* component above and beyond STM. As in the preceding analysis, we used SEM to evaluate the direct paths among STM, WM, VSAT, and QSAT (referred to as Model E in Table 5). For the memory measures, we again set target measure–latent variable paths to the estimated values obtained in our original analysis. The SAT measures were included as manifest variables, and direct paths were estimated between each of the memory factors and each of the SAT scales.

The fit statistics for this SEM are shown in Table 5, and parameter estimates are displayed in Figure 5. All seven fit

[3] See Kliegl and Mayr (1992) and Salthouse (1996) for treatments on the use of the analytical method used here in which specific links are tested after effects from a common factor are controlled.

[4] At the risk of being further redundant, we also looked at the zero-order correlations and semipartial correlations among the factor scores. The zero-order correlations were as follows: *gF*-STM = .24, *gF*-WM = .40, and STM-WM = .54. When semipartials were calculated by alternately removing variance caused by STM from the WM-*gF* relationship, the correlation was a significant .32. When the variance resulting from WM was semipartialed from the relationship between STM and *gF*, the correlation was .03. Again, the conclusion was that the relationship between WM and *gF* is real and not mediated by STM, and that any relationship between the tasks contributing to STM and *gF* is mediated by WM.

statistics indicate the model was an excellent fit of the observed covariances. For the VSAT, both memory factors showed a significant path; the WM path was slightly higher. For the QSAT, only the WM factor showed a significant path. This finding is consistent with earlier work (Cantor et al., 1991) that demonstrated that an STM component, derived from primarily verbal tasks, contributes variance above and beyond that of WM for measures of verbal abilities.

As a rather minimal assessment of the role of the phonological loop in our measures of *gF* and *gC*, we examined the correlation of VSAT, QSAT, CATTELL, and RAVENS with the size of the phonological similarity effect, which was computed with the following formula (Logie, Della-Salla, Laiacona, & Chalmers, 1996):

$$((\text{FSPAND} - \text{FSPANS})/\text{FSPAND}) \times 100.$$

None of the measures correlated significantly with the phonological similarity effect (all $rs < .12$). Logie et al. (1996) reported that, at the level of participant, the phonological similarity effect is not highly reliable. Similar to Logie et al.'s data, we found that 20 of 133 participants did not show the effect (i.e., they scored equal to or higher on FSPANS than on FSPAND). We do not have estimates of reliability for this measure, but can assume that unreliability of the effect most likely contributes to its inability to predict performance on our higher order measures.

General Discussion

This study addressed two questions: Are STM and WM identifiably distinct constructs? If so, do they differentially relate to other higher order constructs such as *gF* and verbal abilities? The study used factor analysis and latent-variable modeling and provided rather clear answers to both questions.

What Is the Relationship Between STM and WM?

In answer to the first question, the results demonstrated that STM and WM are two distinguishable albeit highly related constructs. EFA resulted in a two-factor solution, with tasks thought to be WM tasks generally loading on one factor and tasks thought to be STM tasks loading on the other. SEMs using the three target WM tasks and three target STM tasks demonstrated that (a) a model with WM and STM as separate but highly correlated latent variables met accepted criteria for good fit and (b) a model that represented WM and STM as a single latent variable did not provide an acceptable fit. Three of four additional tasks thought to be WM tasks did, in fact, show significant paths to the WM latent variable, whereas one showed a significant path to the STM latent variable. Thus, there was ample evidence to support the conclusion that WM and STM reflect distinct, if highly correlated, constructs.

The second question posed by the study was whether the two constructs relate differentially to a higher order construct such as *gF*. A latent variable for *gF* was based on

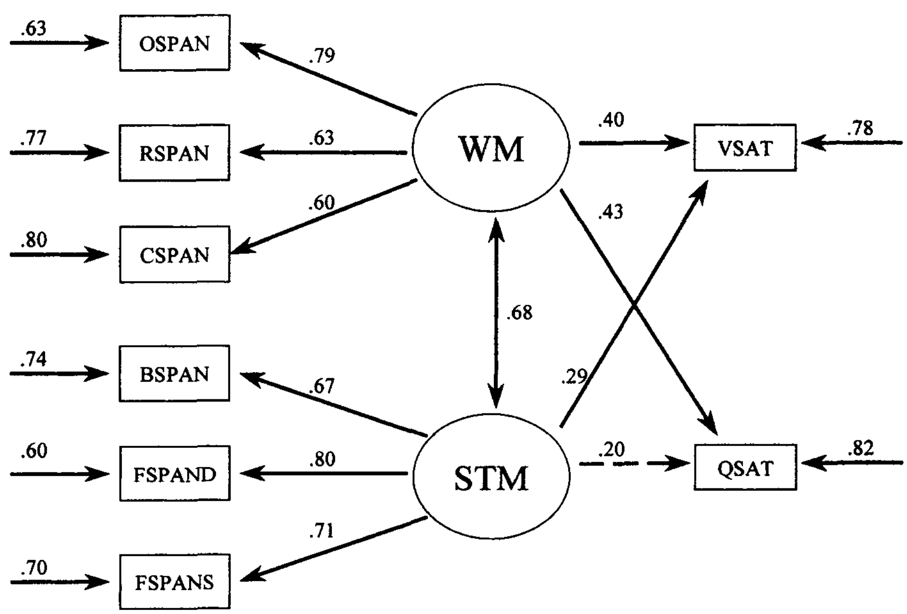

Figure 5. Path model for Model E. Paths significant at the .05 level are indicated by solid lines. OSPAN = operation span; RSPAN = reading span; CSPAN = counting span; BSPAN = backward span; FSPAND = forward span, dissimilar; FSPANS = forward span, similar; VSAT = Verbal Scholastic Aptitude Test; QSAT = Quantitative Scholastic Aptitude Test; WM = working memory; STM = short-term memory.

scores from Raven's matrices and the Cattell Culture Fair Test. The best fitting model required a connection between WM and *gF*, but no such link was necessary between STM and *gF*. If the WM latent variable is responsible for any connection between STM and *gF*, then when the variance common to the STM and WM latent variables is removed there should be no connection between STM and *gF*, but the connection between WM and *gF* should remain. After the variance common to WM and STM latent variables was accounted for, there was still a strong correlation between the residual from the WM latent variable and *gF* but no such link between STM and *gF*. Finally, a significantly fitting model demonstrated that the STM latent variable shared a relationship with VSAT above and beyond that attributable to the WM latent variable.

Our results support the conclusion that WM and STM should be thought of as distinct but highly related. This strong degree of overlap is consistent with Baddeley and Hitch's (1974) general model of WM and with Cowan's (1988, 1995) theory of STM and WM and of controlled attention. To restate, Cowan (1988, 1995) argued for a system limited in capacity in two ways. One, with Hebb's (1949) ideas as its foundation, is based on the decay rate and rehearsal rate. This effectively limits the number of traces active above threshold at a given time. The second limitation, based on James's (1890) ideas, is in capability for controlled and focused attention. There is a rather severe limit in the number of memory elements that can be attended to at one time. We argued above that WM tasks reflect both limitations, but that tasks typically thought of as reflecting STM primarily reflect the former. If performance on STM and WM tasks relies on the same memory system, then performance on the two types of tasks should be highly correlated, particularly to the extent that (a) STM tasks rely on central executive-based controlled processing and (b) WM tasks rely on the same materials and coding-rehearsal formats used in the STM tasks.

What is important is the notion that, although both STM and WM tasks rely on the same memory system, WM tasks primarily engage the central executive to maintain the activation of information relevant to the current task, particularly when the participant is confronted with distraction from internal and external events and when there is interference from other competing information. It is this differential reliance on controlled attention that makes the two constructs different empirically and theoretically. Our study is consistent with the position that STM is a subset of a general WM system that includes an attention-based central executive (Baddeley, 1996; Cowan, 1995). Presumably, WM tasks make more demands on the central executive or controlled-attention component than do the STM tasks. This is particularly salient when the dual-task nature of WM capacity tasks induces greater distraction of attention from the representations in subtasks needed for ongoing processing and when the task induces a level of interference or response competition. By our logic, removing the variance common to STM and WM tasks left a residual consisting primarily of the central executive or controlled-attention component, and, as the analysis of the model in Figure 4 showed, this component correlated with *gF* (.49). The conclusion, then, is that the primary factor contributing to the relationship between measures of WM and *gF* is controlled attention.[5] This nicely supports the ideas proposed by Duncan, Emslie, Williams, Johnson, and Freer (1996) that *gF* and the functions of the prefrontal cortex reflect controlled processing capability, and supports our own suggestion that the dimension we have referred to as high and low WM capacity is a reflection of the functioning of the normal prefrontal cortex (Engle, Kane, & Tuholski, 1999; Engle & Oransky, 1999).

Thus, we are left with a position consistent with that taken by Cowan (1988, 1995). STM and WM are strongly related, although (a) they are represented by separate factors and (b) they are differentially related to higher order cognitive abilities. We argue that these findings are the result of a system that includes memory elements and a central executive. Memory elements will vary in level of activation, and the number of such elements that can be the focus of attention, which derives from the central executive, is quite small. Further, distraction of attention away from elements in the focus will allow their activation to drop below the level that permits rapid retrieval, which is characteristic of the active state. If there is relatively little interference from competing information, this does not present much hindrance to ongoing processing. Retrieval based on a trace in the inactive state is slower than that based on an active trace but could still be successful. The less interference from competing traces, the greater is the chance of successful and speedy retrieval. When I park my car in a shopping mall lot that I am visiting for the very first time, even when I lose rapid access to the trace representing where I parked, I can likely retrieve it based on an inactive long-term trace. The retrieval is slower than when based on an active trace, but in most real-world tasks that is a small cost. Under conditions of interference, however, retrieval based on a trace that has been allowed to slip into the inactive state is more problematic. When I park in the lot of a mall I visit and park in regularly, if I lose rapid access to the trace representing my current parking location, retrieval is made difficult because of competition from the traces representing all the previous times I parked in this lot. Thus, under conditions of high interference, it is particularly important that the goals and details of the current task be maintained in the active state. "Good" WM tasks bring all these conditions together. They make sustained attention to representations of information, which are necessary to perform the current task, very important. This is particularly true under conditions that would allow attention to be captured by distracting events (either internally or externally generated) under conditions of interference.

[5] This analytical position would have been strengthened considerably if we had included tasks that are more direct and putative reflections of the construct "controlled attention." However, that was not done and should be an important part of a future study. Thus, our conclusion that "controlled attention" is the critical common factor associating WM capacity and *gF* is the result of a logical analysis and is, at best, an educated conjecture.

STM refers to the activated elements in this memory system, whereas WM is the greater system composed of the activated elements as well as the controlled attention functions of the central executive. The controlled attention functions of the central executive are necessary for those processes required to maintain the activation of memory units and to focus, divide and switch attention as well as those processes to block inappropriate actions and to dampen activation through inhibition.

A dichotomous view of STM and WM tasks is much too simplistic. Although WM and STM are distinguishable constructs, fitting with the arguments made by Klapp et al. (1983) and by Brainerd and Kingma (1985), tasks that putatively measure one construct likely also reflect the other to some degree. For example, as we noted early in this article, even WM tasks make use of phonological or visuospatial coding, STM tasks may call on the resources of the central executive, and tasks that are WM tasks for some people (e.g., children) may be primarily STM tasks for others (e.g., adults) because of differential reliance on the central executive. It is also likely that individuals will vary in the weight that a given task will have on the STM and WM constructs depending on their differential vulnerability to interference (cf. Hasher & Zacks, 1988). A related point is that it is also probable that what is primarily an STM task for some individuals could be a WM task for others, even of the same age, which would increase the error variance in a study such as the one reported here. Given this possibility, the cleanliness of the results and strength of conclusions permitted in the present study are rather remarkable.

One alternative to the view we have proposed here is to interpret individual differences in WM capacity and the STM-WM distinction in terms of storage and processing (as opposed to storage and controlled attention). Although this seems plausible on the surface, we have demonstrated rather serious problems with this explanation of the relationship between measures of WM capacity and higher order cognition. Engle, Cantor, and Carullo (1992) used a moving window presentation of both the operation span and reading span, and used the time to present the elements of the tasks as a reflection of processing. Partialing the time to process the elements of the span tasks from the relationship between the span score and reading comprehension did not diminish the correlation at all. So controlling for processing did not change the relationship between the complex span scores and higher level cognition. Likewise, manipulating the difficulty of the processing component does not affect that relationship. Conway and Engle (1996) pretested individuals on their arithmetic skills before having them perform the operation span task. Each individual received a span task that adjusted the difficulty of the operation component so that it would provide a performance level of 75%, 85%, or 95% correct. A strong view of the idea that WM capacity is a function of storage plus processing suggests that the correlation between the recall of the items at the end of each operation and reading comprehension should drop to 0. In fact, the correlation was unchanged from an unadjusted version of the operation span. Controlling for processing does not make the correlation go away, and manipulating difficulty of processing and controlling it do not make the correlation go away. For the storage plus processing view to explain the results presented here, the processing involved in solving arithmetic operations, reading and understanding sentences, and performing controlled counting of objects must be similar and must be similar to the processes necessary to perform the spatial reasoning tasks in RAVENS and CATTELL. Such a view of "processing" strikes us as so general and vacuous as to be useless.

Comments on the Present Methodology

Our methodology allowed us to assess not only whether STM and WM are different but also which tasks are good measures of each construct. Early in this article, we identified three target STM tasks (FSPAND, FSPANS, BSPAN) and WM tasks (OSPAN, RSPAN, CSPAN) for our initial investigation regarding the one-construct or two-constructs question. From there, we added tasks to our analysis to determine which construct better identified them. By doing so, we discovered that KTRACK, ABCD, and IFRSM were adequate measures of WM. Interestingly, we found no evidence whatsoever that the IFRPM and RAND tasks related to STM or WM. This latter finding was especially surprising given Baddeley's (1996) work involving the random generation task, although differences in methodology may account for the discrepancy between our results and his. In our study, participants were asked to generate a random number every 1.5 s, and we found that the "randomness" of output did not relate to our other memory, *gF* or SAT measures. However, Baddeley (1996) demonstrated that decreasing the amount of time allowed between responses also decreases the amount of randomness in output. It may be that our participants had more time than necessary to generate random sequences, and as such, the RAND task was not a good measure of central executive functioning. An interesting question for the future is whether or not the RAND task would fit with the WM tasks if participants were given less time to generate responses. It should be noted that much of the work supporting random generation as a central executive task is based on using the task in conjunction with a second task, which is generally the primary task. However, it is likely that many different tasks that individually are not good central executive tasks would create stress on the central executive when combined with another task that could not be compiled with the first. D'Esposito et al. (1995) demonstrated that when two different tasks, neither of which caused activation of the prefrontal cortex, were combined into a dual-task requirement, the prefrontal cortex was activated. Thus, random generation may be one of myriad tasks that, when combined with another task, create stress on the central executive.

Before discussing the relationships between STM and WM with *gF* and verbal abilities, we would like to comment on the methodology used in this study and how it could be useful to other psychologists. We submit that the methodology used here provides a powerful way to understand the hypothetical constructs that are responsible for performance on a wide variety of tasks. For example, if a researcher wanted to know whether performance on a cognitive task was better explained by WM or processing speed, they could set up a study in which participants were tested on tasks that were good measures of both constructs and observe on which factor the task of interest falls. Salthouse (1991) provided a good example of this approach in answering questions about the effects of aging on cognition. The act of setting up this type of research makes one think very carefully about what a WM or processing speed task may be. In doing so, it provides a fresh perspective on the construct as well as interesting empirical findings that theories of cognition must be able to explain. Thus, if a cognitive neuroscientist were interested in exploring the brain structures associated with central executive functioning, we recommend determining whether the task chosen to reflect the central executive loads substantially with other tasks we refer to here as WM tasks.

Central Executive Account of Fluid Intelligence

Our study provided very strong evidence that WM capacity is related to *gF*, supporting the conclusions of Kyllonen and Christal (1990). By adding STM tasks into the analysis, however, we are able to theorize about the specific aspect of WM that drives the relationship between WM and *gF*. That is, we are able to determine that the central executive or controlled-attention component of WM is responsible for the strong relationship.

As we noted early in this article, other studies have demonstrated a relationship between STM and intelligence (i.e., Bachelder & Denny, 1977a, 1977b), but our data suggest that this relationship is driven by the central executive component. In our study, we found a nonsignificant path from STM to *gF*, which demonstrated that STM did not explain any unique variance in *gF* above and beyond that explained by the WM latent variable. Further, when the variance common to STM and WM is treated as a separate latent variable, the correlation between the residual variance in the WM latent variable and *gF* is highly significant (.49).

Thus, the present study leaves us with an interesting account of *gF*. Consistent with the ideas presented by Kyllonen and Christal (1990), Stankov (1983; Crawford & Stankov, 1983), and Duncan (1993; Duncan et al., 1996; and Duncan, Williams, Nimmo-Smith, & Brown, 1990) we conclude that the central executive and *gF* are intimately related. The relationship between controlled attention and *gF* was demonstrated by Duncan et al. (1990) using a real-world task in which participants were measured on components of automobile driving as well as the Cattell Culture Fair Test. They found that those components that were more susceptible to interference, and hence more likely to be performed under controlled attention, were strongly related to CATTELL scores, whereas components that were not susceptible to interference did not relate to CATTELL performance.

Further evidence for the relationship between controlled attention and intelligence comes from Ackerman (1988). He demonstrated that performance on novel tasks correlated strongly with measures of general abilities (which included measures of reasoning = *gF*) to the degree that performance required controlled attention. Specifically, Ackerman's (1988) work demonstrated a strong relationship between task performance and general abilities when the participants were not practiced at the task. As the amount of practice on a task increased (and presumably the need for controlled processing decreased), the correlation between task performance and measures of general abilities decreased.

The present study does not address brain structures in any specific way, but it does permit some speculation based on existing literature. If controlled attention and *gF* are as intimately related as the literature and our data suggest, we should be able to find evidence for overlap in regions of the brain thought to be important for the two functions. In fact, cognitive neuroscience data do suggest that the two constructs are strongly related. Duncan, Burgess, and Emslie (1995) demonstrated that the frontal lobes were critically involved in *gF* but not *gC*. They tested frontal patients who scored high on a standard measure of intelligence (Wechsler Adult Intelligence Scale—Revised [WAIS-R]) as well as normal controls who were matched on WAIS-R scores and posterior damaged controls (not matched on WAIS-R). Of interest to the present study was performance on the Cattell Culture Fair Test. Duncan et al. found that normal and brain-damaged controls scored significantly higher on the Cattell Culture Fair Test than frontal patients, despite the fact that the three groups were matched on the WAIS-R. The conclusions from this study were straightforward; the frontal lobes, which for years have been thought to be related to executive functioning and controlled attention (Luria, 1966; Norman & Shallice, 1980), are also implicated in *gF*.

An emerging literature using brain-imaging converges on the notion that *gF* and controlled attention are reflected in similar brain areas. For example, Prabhakaran, Smith, Desmond, Glover, and Gabrieli (1997, cited in Wickelgren, 1997) used functional magnetic resonance imaging to demonstrate that performing the Raven's Progressive Matrices Test activates areas of prefrontal and parietal cortex, which are also activated in WM tasks. Prabhakaran et al. concluded that "strong links between WM and fluid reasoning occur because the tasks measuring those processes are, in fact, measuring common neural systems" (p. 60).

Another finding that addresses the relationship between STM and WM was the demonstration by D'Esposito et al. (1995) that two tasks, neither of which caused activation in the dorsolateral prefrontal cortex when performed alone, did lead to prefrontal activation when performed together. This lends support to the idea that an important factor in what makes a valid WM capacity or central executive task is scheduling task components that compete for stages of processing or attention switching between the components, particularly where one of those components consists of the storage of one or more items (Baddeley, 1993).

Role of Short-Term Memory

If WM is as intimately related to *g* as we have argued, is there a place for STM with regard to higher level cognitive functioning? We think so. Our data suggest that STM does contribute unique variance to verbal abilities, as measured by the VSAT. Importantly, STM did not contribute unique variance to quantitative abilities, as measured by the QSAT (see Figure 4). We think that this finding is very informative with regard to the nature and utility of STM.

The tasks that we used as STM tasks were strongly verbally oriented. That is, we asked participants to remember lists of words and recall them in order of presentation or in the reverse order. Performance on these tasks did relate to VSAT performance above and beyond that explained by WM capacity. This finding suggests that STM, as we measured it, reflects something important about verbal abilities. We argue that WM or controlled attention is not modality specific, whereas STM is. WM tasks should predict performance on a wide variety of cognitive tasks. However, the ability of STM tasks to do so will depend on the materials used in the task and the nature of the coding and rehearsal encouraged by the task. That is, we should be able to devise STM tests that are not verbally based and that would predict QSAT above and beyond the ability of WM tasks to do so. Further, these STM tasks would not contribute to VSAT.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 3,* 288–318.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1990). *Cognitive psychology and its implications.* New York: Freeman.

Bachelder, B. L., & Denny, M. R. (1977a). A theory of intelligence: I. Span and the complexity of stimulus control. *Intelligence, 1,* 127–150.

Bachelder, B. L., & Denny, M. R. (1977b). A theory of intelligence: II. The role of span in a variety of intellectual tasks. *Intelligence, 1,* 237–256.

Baddeley, A. D. (1986). *Working memory.* Oxford, England: Clarendon Press.

Baddeley, A. D. (1993). Working memory or working attention? In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control: A tribute to Donald Broadbent* (pp. 152–170). Oxford, England: Oxford University Press.

Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology, 49,* 5–28.

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

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology, 3,* 1–29.

Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles: BMDP Statistical Software.

Benton, S. L., Kraft, R. G., Glover, J. A., & Plake, B. S. (1984). Cognitive capacity differences among writers. *Journal of Educational Psychology, 76,* 820–834.

Brainerd, C. J., & Kingma, J. (1985). On the independence of short-term memory and working memory in cognitive development. *Cognitive Psychology, 17,* 210–247.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–142). Newbury Park, CA: Sage.

Browne, M. W., & Mels, G. (1992). *RAMONA PC user's guide.* Columbus: Ohio State University, Department of Psychology.

Cantor, J., Engle, R. W., & Hamilton, G. (1991). Short-term memory, working memory, and verbal abilities: How do they relate? *Intelligence, 15,* 229–246.

Carroll, J. B. (1989). Factor analysis since Spearman. Where do we stand? What do we know? In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Learning and individual differences: Abilities, motivation and methodology* (pp. 43–67). Hillsdale, NJ: Erlbaum.

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

Case, R., Kurland, M. D., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33,* 386–404.

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

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

Cattell, R. B. (1973). *Measuring intelligence with the Culture Fair tests.* Champaign, IL: Institute for Personality and Ability Testing.

Cohen, R. L., & Sandberg, T. (1977). Relation between intelligence and short-term memory. *Cognitive Psychology, 9,* 534–554.

Conway, A. R. A., & Engle, R. W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General, 123,* 354–373.

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory, 4,* 577–590.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin, 104,* 163–191.

Cowan, N. (1995). *Attention and memory: An integrated framework.* Oxford, England: Oxford University Press.

Crawford, J., & Stankov, L. (1983). Fluid and crystallized intelligence and primacy/recency components of short-term memory. *Intelligence, 7,* 227–252.

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

Daneman, M., & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 9,* 561–584.

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language, 25,* 1–18.

D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature, 378,* 279–281.

Duncan, J. (1993). Selection of input and goal in the control of behaviour. In A. Baddeley & L. Weiskrantz (Eds.), *Attention:*

Selection, awareness and control (pp. 53–71). Oxford, England: Clarendon Press.

Duncan, J., Burgess, P., & Emslie, E. (1995). Fluid intelligence after frontal lobe lesions. *Neuropsychologica, 33,* 261–268.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology, 30,* 257–303.

Duncan, J., Williams, P., Nimmo-Smith, I., & Brown, I. (1990). The control of skilled behavior: Learning, intelligence and distraction. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance 14* (pp. 323–341). Cambridge, MA: MIT Press.

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

Engle, R. W., Carullo, J. J., & Collins, K. W. (1991). Individual differences in working memory for comprehension and following directions. *Journal of Educational Research, 84,* 253–262.

Engle, R. W., Conway, A. R. A., Tuholski, S. W., & Shisler, R. J. (1995). A resource account of inhibition. *Psychological Science, 6,* 122–125.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 102–134). Cambridge, England: Cambridge University Press.

Engle, R. W., & Oransky, N. (1999). Multi-store versus dynamic models of temporary storage in memory. In R. Sternberg (Ed.), *The nature of cognition* (pp. 514–555). Cambridge, MA: MIT Press.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193–225). New York: Academic Press.

Hebb, D. O. (1949). *The organization of behaviour.* New York: Wiley-Interscience.

Horn, J. L. (1980). Concepts of intellect in relation to learning and adult development. *Intelligence, 4,* 285–317.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26,* 107–129.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Panter (Ed.), *Structural equation modeling: Concepts, issues, and applications.* Thousand Oaks, CA: Sage.

James, W. (1890). *The principles of psychology.* New York: Henry Holt.

Jensen, A. R., & Figueroa, R. A. (1975). Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. *Journal of Educational Psychology, 67,* 882–893.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: National Education Resources.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Kiewra, K. A., & Benton, S. L. (1988). The relationship between information processing ability and notetaking. *Contemporary Educational Psychology, 13,* 33–44.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30,* 580–602.

Klapp, S. T., Marshburn, E. A., & Lester, P. T. (1983). Short-term memory does not involve the "working memory" of information processing: The demise of a common assumption. *Journal of Experimental Psychology: General, 112,* 240–264.

Kliegl, R., & Mayr, U. (1992). Commentary on Salthouse (1992). "Shifting levels of analysis in the investigation of cognitive aging." *Human Development, 35,* 343–349.

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

Kyllonen, P. C., & Stephens, D. L. (1990). Cognitive abilities as determinants of success in acquiring logic skill. In *Special Issue: Frontiers of learning and individual differences: Paradigms in transition. MOE-Project Learning Abilities Measurement Program.* Brooks Air Force Base, TX: US Air Force Human Resources Lab.

Logie, R. H., Della-Salla, S., Laiacona, M., & Chalmers, P. (1996). Group aggregates and individual reliability: The case of verbal short-term memory. *Memory and Cognition, 24,* 305–321.

Luria, A. R. (1966). *Higher cortical functions in man.* London: Lavistock.

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology, 24,* 56–98.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

Norman, D. A., & Shallice, T. (1980). *Attention to action: Willed and automatic control of behavior* (report no. 8006). San Diego: University of California Center for Human Information Processing.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation* (Vol. 4, pp. 1–18). New York: Plenum Press.

Ormrod, J. E., & Cochran, K. F. (1988). Relationship of verbal ability and working memory to spelling achievement and learning to spell. *Reading Research and Instruction, 28,* 33–43.

Posner, M. I., & Peterson, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13,* 25–42.

Posner, M. I., & Snyder, C. R. (1975). Attention and cognitive control. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Prabhakaran, V., Smith, J. A. L., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1997). Neural substrates of fluid reasoning: An fMRI study of neocortical activation during performance of the Raven's Progressive Matrices Test. *Cognitive Psychology, 33,* 43–63.

Raven, J. C., Court, J. H., & Raven, J. (1977). *Standard progressive matrices.* London: H. K. Lewis.

Rosen, V. M., & Engle, R. W. (1997). Forward and backward serial recall. *Intelligence, 25,* 37–47.

Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging.* Hillsdale, NJ: Erlbaum.

Salthouse, T. A. (1996). Constraints on theories of cognitive aging. *Psychonomic Bulletin and Review, 3,* 287–299.

Salthouse, T. A., Mitchell, D. R., Skovronek, E., & Babcock, R. L. (1989). Effects of adult age and working memory on reasoning and spatial abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 507–516.

SAS Institute. (1990). *SAS 6.11* [Computer software]. Cary, NC: Author.

Sattler, J. M. (1992). *Assessment of children* (3rd ed.). San Diego: Sattler.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing I: Detection, search & attention. *Psychological Review, 84,* 1–66.

Seamon, J. G., & Kenrick, D. T. (1994). *Psychology.* Englewood Cliffs, NJ: Prentice Hall.

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

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 47–103). Hillsdale, NJ: Erlbaum.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.

Stankov, L. (1983). Attention and intelligence. *Journal of Educational Psychology, 75,* 471–490.

Steiger, J. H. (1989). *EZPATH. A supplementary module for SYSTAT and SYGRAPH.* Evanstan, IL: SYSTAT.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Tulving, E., & Colotla, V. A. (1970). Free recall of trilingual lists. *Cognitive Psychology, 1,* 86–98.

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

Wickelgren, I. (1997). Getting a grasp on working memory. *Science, 275,* 1580–1582.

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

Zacks, R. T. (1982). Encoding strategies used by young and elderly adults in a keeping track task. *Journal of Gerontology, 37,* 203–211.

Received July 25, 1997
Revision received April 28, 1998
Accepted August 5, 1998 ■