Psychonomic Bulletin & Review
2005, 12 (5), 769-786

THEORETICAL AND REVIEW ARTICLES

Working memory span tasks: A methodological review and user's guide

ANDREW R. A. CONWAY
University of Illinois, Chicago, Illinois

MICHAEL J. KANE
University of North Carolina, Greensboro, North Carolina

MICHAEL F. BUNTING
University of Illinois, Chicago, Illinois

D. ZACH HAMBRICK
Michigan State University, East Lansing, Michigan

OLIVER WILHELM
Humboldt University, Berlin, Germany

and

RANDALL W. ENGLE
Georgia Institute of Technology, Atlanta, Georgia

Working memory (WM) span tasks—and in particular, counting span, operation span, and reading span tasks—are widely used measures of WM capacity. Despite their popularity, however, there has never been a comprehensive analysis of the merits of WM span tasks as measurement tools. Here, we review the genesis of these tasks and discuss how and why they came to be so influential. In so doing, we address the reliability and validity of the tasks, and we consider more technical aspects of the tasks, such as optimal administration and scoring procedures. Finally, we discuss statistical and methodological techniques that have commonly been used in conjunction with WM span tasks, such as latent variable analysis and extreme-groups designs.

Other than standardized instruments, such as intelligence test batteries, working memory (WM) span tasks, such as the counting span, operation span, and reading span tasks, are among the most widely used measurement tools in cognitive psychology. These tasks have come to prominence not only for their methodological merit, but also because theoretical advances in the study of human behavior since the cognitive revolution have placed WM as a central construct in psychology. Methodologically, WM span tasks have proven to be both reliable and valid measures of WM capacity (WMC), which we will document below. However, the larger factor in accounting for their increased use is simply that WM has become a widely useful, scientifically fruitful construct. It plays an important role in contemporary global models of cognition (e.g., J. R. Anderson & Lebiere, 1998; Cowan, 1995), and it is purportedly involved in a wide range of complex cognitive behaviors, such as comprehension, reasoning, and problem solving (Engle, 2002). Also, WMC is an important individual-differences variable and accounts for a significant portion of variance in general intellectual ability (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Conway, Kane, & Engle, 2003; Engle, Tuholski, Laughlin, & Conway, 1999; Kane et al., 2004; Kyllonen, 1996; Kyllonen & Christal, 1990; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). Furthermore, neuroimaging and neuropsychological studies have revealed that WM function is particularly dependent on cells in the prefrontal cortex of the brain, which has traditionally held a prominent status in the biological approach to studying complex goal-directed human behavior (Kane & Engle, 2002).

A diverse set of researchers is now using WM as a construct in research programs, as well as measures of WMC

We thank Chris Fraley for assistance in the effect size simulations and for helpful discussions about prior versions of the manuscript. A.R.A.C. is now at Princeton University. M.F.B. is now at the University of Missouri. Correspondence concerning this article should be addressed to A. R. A. Conway, Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544-1010 (e-mail: aconway@princeton.edu).

in the arsenal of research tools. Within psychology, discussions of WM are now common in almost all branches of the discipline, including cognitive, clinical, social, developmental, and educational psychology. For example, clinical research has demonstrated that WM is related to depression (Arnett et al., 1999) and to the ability to deal with life event stress (Klein & Boals, 2001) and is affected by alcohol consumption (Finn, 2002). Social psychologists have revealed that students under stereotype threat suffer reduced WMC and that WMC mediates the effect of stereotype threat on standardized tests (Schmader & Johns, 2003). Also, WMC is taxed and, subsequently, depleted as a result of interracial interaction for highly prejudiced individuals (Richeson et al., 2003; Richeson & Shelton, 2003). In neuropsychology, deficits in WMC may be a marker of early onset of Alzheimer's disease (Rosen, Bergeson, Putnam, Harwell, & Sunderland, 2002). Developmental research suggests that the development of WMC in children is central to the development of cognitive abilities in general (Munakata, Morton, & O'Reilly, in press) and that declines in WMC as a result of aging are central to general cognitive-aging effects (Hasher & Zacks, 1988). In short, recent research across the discipline implicates WM as a central psychological construct (for reviews, see Feldman-Barrett, Tugade, & Engle, 2004; Unsworth, Heitz, & Engle, 2005).

Although the WM *construct* has been successfully and appropriately exported from cognitive psychology to other disciplines, WM *tasks* have, in our opinion, suffered a bit in translation. Presumably because of their documented reliability and validity, *WM span tasks*, among all the available measures of WMC, have been embraced most strongly by researchers outside of cognitive psychology. Inevitably, as WM is exported to other scientific disciplines and as a more varied pool of investigators use WM span tasks, misconceptions and misuses are bound to increase. In particular, the literature presents inconsistent information regarding the reliability of WM span tasks, as well as inconsistent and, in our opinion, problematic procedures for administration and scoring.

We therefore believe that the time is right for a review of how and why WM span tasks came to dominate the measurement landscape of WM research and to provide guidelines for researchers who would like to use these tasks. That is, we perceive a need for a thorough review of all aspects of WM span tasks, from optimal administration and scoring procedures to assessment of reliability and validity. Another motivating factor is more self-serving. Each author of this article receives numerous inquiries about these tasks, including questions about administration, scoring, assessment of reliability, use of extreme-groups designs, and so forth, and so we devote considerable time to instructing other researchers about the proper use of these tasks. As such, an important purpose of the present article is to provide a "user's guide" for WM span tasks, such that any interested researcher can read this article, download the programs from our Web site (http://psychology.gatech.edu/renglelab/tasks.htm), and use them most appropriately.

We will begin with a historical overview of the development of WM span tasks, followed by guidelines for administration and scoring. The reliability and validity of the tasks will then be discussed. The tasks will then be contrasted with other empirical measures of WM function. Finally, we will discuss common research strategies that have been used in conjunction with these tasks, such as latent variable analysis and extreme-groups designs.

The review is limited to considering just three WM span tasks; counting span (Case, Kurland, & Goldberg, 1982), operation span (Turner & Engle, 1989), and reading span (Daneman & Carpenter, 1980). These three span tasks were chosen because there is much more data from these tasks than from any of the others (e.g., spatial WM span tasks; see Kane et al., 2004; Shah & Miyake, 1996), and the principles outlined here should generalize not only to other span tasks, but also to other measures of cognitive ability. Our hope is that this review will serve as an example of how to assess measurement instruments within any particular research domain. We also hope that it lays bare the importance of taking issues of measurement seriously in one's research program. We should also note at the outset that the particular tasks under review here have been used primarily to investigate individual differences in healthy young adults, and therefore, our recommendations apply best for similar purposes in similar populations. That said, we will indicate, where possible, how these tasks might be modified for other applications.

Finally, we note that the review is as theory neutral as possible with respect to a particular model of WM and/or the nature of individual differences in WMC, because our goal is to review the merits of WM span tasks as research tools for any researcher, with any theoretical stance. However, it is simply impossible to review the history of these tasks, discuss their validity, or even suggest how to score them without revealing some theoretical bias. Therefore, at the outset, we will briefly outline our theoretical approach to WM, for the simple purpose of warning the reader of any bias that may reveal itself later.

BRIEF THEORETICAL OVERVIEW OF WM AND WMC

We view WM as a multicomponent system responsible for active maintenance of information in the face of ongoing processing and/or distraction. Active maintenance of information is the result of converging processes—most notably, domain-specific storage and rehearsal processes and domain-general executive attention. Furthermore, the extent to which maintenance depends on domain-specific skills versus domain-general executive attention varies as a function of individual ability, task context, and ability × context interactions. For instance, a novice chess player will rely more on domain-general executive attention to maintain game information (e.g., recent moves or future positions) than on domain-specific skills (e.g., learned strategies and position patterns). In contrast, an expert chess player typically will rely more on domain-specific

processes and skills to maintain information. However, even the expert might need to call upon executive attention under some circumstances, such as playing the game in particularly demanding situations or under some sort of cognitive or emotional load.

Similarly, performance on WM span tasks depends on multiple factors, with domain-specific skills, such as chunking and rehearsal, facilitating storage and a domain-general capability allowing for cognitive control and executive attention. A critical aspect of our view, however, is that WM span tasks predict complex cognitive behavior across domains, such as reading comprehension, problem solving, and reasoning, primarily because of the general, executive attention demands of the tasks, rather than the domain-specific demands of the tasks (Conway & Engle, 1996; Conway et al., 2003; Engle, Cantor, & Carullo, 1992; Engle, Tuholski, et al., 1999; Turner & Engle, 1989). As such, we make a clear distinction between the traditional concept of short-term memory capacity (STMC) and WMC. STMC is thought to reflect primarily domain-specific storage, whereas WMC is thought to reflect primarily domain-general executive attention (Engle, Tuholski, et al., 1999). More specifically,

> By "executive attention" we mean an attention capability whereby memory representations are maintained in a highly active state in the face of interference, and these representations may reflect action plans, goal states, or task-relevant stimuli in the environment. Critical to our view is that, while the active maintenance of information can be useful in many situations, it is most necessary under conditions of interference. This is because in the absence of interference, task-relevant information or goals may be easily retrieved from long-term memory as needed. Under interference-rich conditions, however, incorrect information and response tendencies, are likely to be retrieved, and so such contexts set the occasion for the reliance on active maintenance of information. (Kane & Engle, 2002, p. 638)

Also, in claiming that executive attention is domain general, we make no distinction between verbal WMC and spatial WMC (or any other domain WMC for that matter). Kane et al. (2004) recently provided empirical support for this position, showing that verbal WM span tasks, such as counting, operation, and reading span, load on the same factor in a factor analysis as WM span tasks in which the tasks demand spatial processing and storage. More detailed theoretical reviews of our approach, as well as further empirical support for the claims presented here, can be found elsewhere (e.g., Conway et al., 2002; Conway & Kane, 2001; Conway et al., 2003; Engle, 2001, 2002; Engle, Kane, & Tuholski, 1999; Engle, Tuholski, et al., 1999; Kane, Bleckley, Conway, & Engle, 2001; Kane & Engle, 2002; Kane et al., 2004). Finally, because our view is concerned primarily with the domain-general attentional aspect of WM, our claims above are consistent with either structural (e.g., Baddeley, 1986) or functional (e.g., Nairne, 2002) models of storage.

HISTORY OF WM SPAN TASKS AND OVERVIEW OF ADMINISTRATION PROCEDURES

WM span tasks, such as counting span, operation span, and reading span, were designed from the perspective of Baddeley and Hitch's (1974) theory of WM, which stressed the *functional importance* of an immediate-memory system that could briefly store a limited amount of information in the service of ongoing mental activity. That is, a WM system would be unlikely to evolve for the sole purpose of allowing an organism to store or rehearse information (such as a phone number) while it was doing nothing else. A more adaptive system would allow the organism to keep task-relevant information active and accessible in memory during the execution of complex cognitive tasks. WMC measures were, therefore, created to require not only information storage and rehearsal (as do "simple" measures of STMC, such as digit span or word span), but also the simultaneous processing of additional information (Case et al., 1982; Daneman & Carpenter, 1980; Turner & Engle, 1989). Such WM span tasks interleave the presentation of to-be-remembered target stimuli, such as digits or words, with the presentation of a demanding, secondary processing task, such as comprehending sentences, verifying equations, or enumerating an array of shapes.

The reading span task was the first task developed with the purpose of jointly tapping the storage and processing functions of WM (Daneman & Carpenter, 1980). The task is essentially a simple word span task, with the added component of the comprehending of sentences. Subjects read sentences and, in some cases, verify the logical accuracy of the sentences, while trying to remember words, one for each sentence presented. The sentences are presented in groups that typically range in size from two to six (we refer to a group of sentences as one *item*).[1] Word recall is prompted at the completion of an item.

In the original version of reading span (Daneman & Carpenter, 1980, Experiment 1), subjects were required to read aloud, at their own pace, sentences presented on index cards, while remembering the last word of each sentence for later recall. After a series of sentences, the subject recalled the to-be-remembered words in the order in which they had been presented. There were 15 items, 3 each consisting of two, three, four, five, and six sentences that were 13–16 words in length, and they were presented in ascending order (i.e., from smallest to largest). Increasingly larger items were presented until the subject failed to recall all 3 items of a given size. At this point, the experiment was terminated. A subject's reading span was the level at which he or she could correctly recall 2 of the 3 items. For example, if a subject were to successfully recall at least 2 out of 3 two-word items, the experiment would continue for the subject to attempt 3-word items. If the subject were then to successfully recall only 1 out of 3 of the 3-word items, the experiment would terminate, and the subject's reading span would be 2.

Daneman and Carpenter (1980) added a true–false component to the task in their Experiment 2. The subjects indicated the veracity of each sentence they read by responding *true* or *false* within 1.5 sec of each sentence's presentation. Here, the sentences were drawn from general knowledge quiz books and covered multiple domains, including the biological and physical sciences, literature, geography, history, and current affairs, and were selected to be of moderate difficulty (e.g., "You can trace the languages English and German back to the same roots"). Although Daneman and Carpenter (1980) did not monitor the subjects' accuracy on the true–false component, the subjects believed it was an important part of the task. This prevented the subjects from adopting a strategy of focusing on the final words without devoting much attention to reading the sentences.

The reading span task used by Engle and colleagues is somewhat different from Daneman and Carpenter's (1980) original version, but the fundamental premise of the task is unchanged. For instance, Turner and Engle (1989) created a version of reading span that consisted of fewer overall items (12 items, 3 each consisting of two, three, four, and five sentences). Also, whereas Daneman and Carpenter (1980) tested subjects on the veracity of the sentences, Turner and Engle tested subjects on whether the sentences were semantically and syntactically correct (e.g., "The grades for our finals will classroom the outside posted be door"). Turner and Engle administered their reading span task to small groups of subjects, rather than individually. The sentences were projected on a screen via an overhead transparency, and the experimenter used a sheet of paper to keep all but the current sentence hidden. The stimuli were simultaneously presented auditorially by means of a prerecorded cassette, which served to pace the subjects through the task. Thus, the subjects heard the sentences while they read them aloud. This procedure varied considerably from the self-paced reading in Daneman and Carpenter's (1980) version of the task and afforded some opportunity for error. Fast readers, especially, might devise strategies to devote more time to memorizing (i.e., coding and rehearsing) the to-be-remembered stimuli. Recognizing this potential problem, Turner and Engle tested subjects individually in their Experiment 2. They also monitored accuracy on the sentence verification component of the task, and subjects who scored below 80% were excluded from all analyses. This criterion further helped ensure that attention was paid to the processing component of the task.

The version of reading span that we have most commonly used is one in which the to-be-remembered word is different from the last word, or any word, in the sentences (see, e.g., Engle, Tuholski, et al., 1999), and so each sentence is followed by an unrelated word. Subjects still read the sentences aloud and verify whether the sentence is semantically or syntactically correct. Here, however, they are charged with remembering the unrelated word and not the last word of the sentence. We made this change because individual differences in reading ability could lead to differences in the ability to generate the words at test on the basis of the gist of the sentence (rather than on the basis of episodic recall). In our latest version of the task, subjects no longer remember words; instead, they remember isolated letters that follow each sentence (Kane et al., 2004).

Does the Secondary Task Need to Involve Reading?

Daneman and Carpenter (1980) argued that reading is an integral component of their span task, and in order to predict reading ability, a WM span task must make use of reading strategies. Daneman and Carpenter (1980) took the position that WMC is strategy specific, subsequently known as the *task-specific view* (cf. Engle et al., 1992). On this view, subjects who have developed effective strategies for the processing component of the task will have greater capacity to devote to storage. Turner and Engle (1989), however, hypothesized that WMC is independent of the specific nature of the processing component of the span task. A highly demanding processing component is necessary to engage the processing functions of WM and draw out individual differences in task performance. Turner and Engle showed that they could predict reading ability with a WM span task that does not involve the reading of sentences. Their task, the operation span task, required that subjects solve mathematical operations while trying to remember words.

There were 84 mathematical operation strings in Turner and Engle's (1989) first operation span task. Each string consisted of a mathematical equation with two arithmetic operations on one side of the equation and a stated solution on the other side of the equation. The first operation was a simple multiplication or division problem and was followed by a simple addition or subtraction operation. The stated solution was correct on half of the trials. The following are examples of a correct and incorrect equation, respectively: $(9/3) - 2 = 1$ and $(9/3) - 2 = 6$.

Turner and Engle's (1989) operations replaced the sentences in Daneman and Carpenter's (1980) task, but otherwise the task demands were largely unchanged. In the first published version of the task, 12 items were presented, with 3 items each consisting of two, three, four, and five operation word pairs, presented in ascending order. The stimuli consisted of mathematical operations followed by a to-be-remembered word, drawn from the same normed set of common four- to six-letter words as that for Turner and Engle's version of the reading span task. The task was administered to small groups of subjects (Experiment 1) and individually (Experiment 2) by the same means as the reading span task. When the operation word string was presented, the subjects read the operation aloud and verified whether the stated solution was correct or incorrect. They then read aloud and remembered the word for later recall. All intermediate calculations were done silently and without the aid of pencil and paper. Engle et al. (1992) developed the version of the operation span task currently used in our laboratories. The primary difference

from earlier versions is the manipulation of presentation order. Rather than presenting reading span and operation span items in ascending order (items with fewer elements first), which permitted the subjects to anticipate the number of words that they would be asked to remember on any given trial, Engle et al. (1992) randomized the presentation order, effectively eliminating reliance on any strategies that come from knowing the size of the memory set.[2] This modification has the added benefit of deconfounding item size and buildup of proactive interference, since recent studies have shown that proactive interference builds from trial to trial in WM span tasks (Bunting, in press; Lustig, May, & Hasher, 2001; May, Hasher, & Kane, 1999). It also results in a wider range of scores than does the traditional ascending approach. A potential risk of this approach, however, is that the early presence of difficult items may discourage some subjects, particularly those who are less able, such as children, the elderly, or patients. (Our advice to researchers working with such populations is to stress to the subject that perfect recall is not expected in these tasks.)

Case et al.'s (1982) counting span task has also frequently been used to measure WMC, particularly in school-aged children (the simplicity of the processing component—i.e., counting—makes this task ideal for a variety of populations, including patients, the elderly, and nonnative English speakers). On the surface, the commonalities between the counting span task and the reading span task or the operation span task are not easily apparent, but the underlying structure is the same for all three tasks. Whereas most versions of reading span and operation span tasks require subjects to remember words, the counting span task involves counting shapes and remembering the count totals for later recall. In Case et al.'s version of the counting span task, subjects orally counted (and pointed their finger at) the green dots presented against a white background. Yellow dots, interleaved with the green dots, disrupted the visual patterns of the green dots (Case et al. did not report a range of values for the number of green or yellow dots presented). The task presented three items of each size from one to five, in ascending order.

In a version of the task designed for adults by Engle, Tuholski, et al. (1999), the stimuli consisted of three items of each size from one to eight. The visual displays were made more complex by placing the target shapes among a field of distractors that shared either the same shape or the same color (and so counting required conjunctive search, à la Treisman & Gelade, 1980). Each display consisted of a random arrangement of three to nine dark blue circles, one to nine dark blue squares, and one to five light blue circles. The subject was to count all of the dark blue circles without pointing. When the final shape was counted, the subject repeated the total, to signal being finished. The experimenter immediately presented the next display, and the subject commenced counting immediately. The subject recalled the total number of dark blue circles from each display in serial order.

Summary of Critical Task Components

The reading span, operation span, and counting span tasks share an underlying structure and are implemented in much the same way. The tasks are designed to force WM storage in the face of processing (or distraction), in order to engage executive attention processes. As such, the following procedural recommendations should be kept in mind when these tasks are administered (again, note that these recommendations are most appropriate for studies involving healthy young adults).

Immediate and vigilant stimulus presentation. A critical feature of the processing component of WM span tasks is that it interferes with rehearsal. Substantial delays between stimulus presentations may, therefore, permit rehearsal of the to-be-remembered stimuli, thereby making the task more a measure of STM storage than of WM/executive functioning. Indeed, Friedman and Miyake (2004) found that an experimenter-paced version of reading span correlated more strongly with verbal SAT ($r = .49$) and reading comprehension ($r = .55$) than did a subject-paced version of the task (rs = .18 and .28, respectively). Moreover, several studies have documented, via partial correlations, that the time subjects spend on the processing or storage components of self-paced span tasks (or very generously paced tasks) can suppress the correlation between span and ability (Engle et al., 1992; Friedman & Miyake, 2004; Turley-Ames & Whitfield, 2003). When WM span tasks are administered, then, each stimulus subsequent to the first stimulus in an item should be presented immediately upon completion of the preceding stimulus, and subjects should be instructed to begin acting upon stimuli immediately.

Individual administration. Versions of the reading span task and the counting span task have been designed for either single-subject or group sessions. Administering the task to more than 1 subject at a time, however, introduces greater potential for error. Subjects must adequately attend to the processing component of the task in order for the processing component to disrupt rehearsal. When a WM span task is administered to groups, it is more difficult for the experimenter to observe whether the subject is attending to the processing task. Moreover, subjects who are more skilled in the processing task may complete it more quickly than do others in the same session, leaving them more time to rehearse the target stimuli.

Sufficient item size. We have described multiple versions of the reading span task, for which the range of item sizes varied from two to five or six. Other published versions have used an even smaller range (from two to four; e.g., May et al., 1999). Whereas larger item sizes considerably increase the running time of the experiment, insufficient item sizes create the potential for ceiling effects among those subjects in the upper end of the performance distribution. We consider the range from two to five elements per item to be adequate for most college student populations (on the basis of distributions from Conway et al., 2002, Engle, Tuholski, et al., 1999, and Kane et al., 2004).

SCORING

The scoring of WM span tasks is a neglected topic in the research literature, and this is unfortunate because different scoring procedures not only may affect the rank order of subjects, but also may have implications for data analyses. Generally, scoring measures of cognitive behavior are considered to be straightforward and simple processes. However, in dual-task situations, such as WM span tasks, there are two sources of data: one from the processing component of the task and one from the storage component. In operation span, for example, multiple data points might be collected, such as accuracy on the math problems, time spent processing the math problems, and recall of the words.[3]

Correlational evidence from studies on adults supports the common procedure of not considering processing performance in the WM span score. First, processing accuracy is typically close to ceiling, because task instructions emphasize processing-task accuracy to ensure that subjects are attending to the secondary task. Second, despite this near-ceiling accuracy, performance on the processing component usually correlates *positively* with performance on the storage component: Subjects who recall the most target items also perform most accurately on the processing task (Kane et al., 2004; Waters & Caplan, 1996). Thus, there is typically no evidence for processing/storage trade-offs.

In the traditional scoring of WM span tasks, the subject is assigned a quasi-absolute span score (e.g., Daneman & Carpenter, 1980; Waters & Caplan, 1996). The task begins with an item consisting of two elements and continues until the subject's accuracy falls below a threshold. Once this threshold is reached, testing is discontinued, and the last item size recalled with a specific probability (say, four out of five items) is the span score. The underlying assumption here is that items with a given load or demand "meet" a person with a given ability. The person is either able to solve an item or not, and so item difficulty and person ability are on the same scale. If item difficulty exceeds the subject's ability, the probability of correct response is low, and if the subject's ability exceeds item difficulty, the probability of correct response is high.

A problem with these absolute scoring methods is that the difficulty of a span item may vary on many dimensions, thus threatening span reliability across different tasks (or different versions of the same task). For example, other things being equal, longer sentences in a reading span task should decrease the quantity of recalled words (see Towse & Hitch, 1995; Towse, Hitch, & Hutton, 1998, 2002). Similarly, the display duration for individual sentences, or the semantic similarity of the stimuli, could have an influence on recall performance (see Copeland & Radvansky, 2001). Thus, there are several ways in which various instantiations of the same WM span measure might yield different "span" values for the same person. A second problem in studies in which absolute scores are used is that, by simply estimating the item size at which a subject falls below a given threshold (and then ending the task), information on all other trials is discarded. Here, the scores can take only one of very few values, usually somewhere between 2 and 6, greatly limiting the sensitivity of the measure (see Oberauer & Süß, 2000). We therefore suggest that absolute span scores are inappropriate for individual-differences research.

Consider, instead, a much simpler scheme. Correct responses to individual elements within an item are assigned one number, and all other responses are assigned a different number (e.g., correct = 1 and incorrect = 0), with no distinction among different types of errors. That is, errors are not classified as omissions or commissions, as more or less erroneous, or as indicating any particular cognitive process. For items of various sizes, there are varying numbers of observations. For an item with six elements, there are three times as many responses as for an item with two elements. Other things being equal, then, we may assume that items with more elements are more reliable indicators than are those with fewer elements, because longer items rely on more instances of the behavior of interest.

In a next step, the data are aggregated, and here there are several possible procedures from which to choose. To illustrate, consider the performance of the fictional but realistic subject depicted in Table 1 (correct serial recall of individual elements is presented for various items across various tasks). This person is performing rather well, but even on items with only two elements, not all the elements are correctly recalled in serial order. Moreover, on items imposing a higher memory load, the person sometimes recalls fewer elements than on items with a smaller load. Also note that the person is performing somewhat differently on the different tasks.

In order to assign a score to this person, some decisions need to be made. Should credit be given if the recall of elements was correct but there were errors on the associated processing component of the task? Should full or partial credit be given if some, but not all, of the elements were recalled in the correct serial position?

Table 1
Results From Three Working Memory Span Tasks for Person A

No. of Elements	Item No.	Counting Span	Operation Span	Reading Span
2	1	2	2	2
	2	2	1	2
	3	2	2	2
3	1	3	3	3
	2	3	3	3
	3	3	3	2
4	1	4	3	2
	2	4	2	3
	3	2	3	4
5	1	5	4	4
	2	5	4	2
	3	3	4	3

Note—Each cell represents the number of elements recalled correctly for that item.

Should a higher weight be assigned to items with a higher memory load (i.e., more elements)? In the WM literature, these questions generally have been raised implicitly, if at all, but from a psychometric perspective it is crucial that these scoring decisions are theoretically and empirically informed.

Our own answer to the first question is straightforward. We assign credit to elements recalled despite errors made on the processing component of the task (for the reasons we discussed previously). That said, we do strive to ensure that accuracy on the processing component of the task is near perfect. If accuracy on the processing component of the task falls below a certain level (typically, 85%), the entire data set for that subject is discarded.

The latter decisions require more discussion. The first question is whether partial-credit scoring, in which credit is given to partly correct items, is superior to all-or-nothing scoring, in which credit is given only to completely correct items (i.e., where all elements are recalled in the correct serial position). Considering our fictional subject in Table 1, a partial-credit procedure would give some points for items with a memory load of five, but the all-or-nothing scoring procedure would not. The second question is whether all items should count the same or whether those with a higher memory load should contribute more to the overall score. Counting all items equally is done by scoring each item as a *proportion* of correctly recalled elements per item, regardless of item size (e.g., recalling one element from a two-element item would count as much as recalling two elements from a four-element item—i.e., .50). These proportions are then averaged. In contrast, giving a higher weight to items with a higher load is done by computing the mean of all correctly recalled *elements* (irrespective of item size). Note that these two decisions, one regarding partial versus all-or-nothing credit, and one regarding equal weighting (or *unit weighting*) versus load weighting, are orthogonal to each other. Consequently, we will consider four scoring procedures that cross these solutions: partial-credit unit scoring (PCU), all-or-nothing unit scoring (ANU), partial-credit load scoring (PCL), and all-or-nothing load scoring (ANL).

The results for our fictional subject, for the four scoring procedures, are summarized in Table 2. For unit scoring, PCU expresses the mean proportion of elements within an item that were recalled correctly, and ANU expresses the proportion of items for which all the elements were recalled correctly. For load-weighted scoring procedures, PCL represents the sum of correctly recalled elements from all items, regardless of whether the items are perfectly recalled or not (also without respect to serial order within items), and ANL represents the scoring method we have most often used, reflecting the sum of the correctly recalled elements from only the items in which all the elements are recalled in correct serial order.

Although load scoring is rather uncommon in psychometrics, it is perhaps the most frequently applied method for span measures. This tradition can be traced as far back as Ebbinghaus (1897), who used a digit span task as an individual-differences measure and applied a partial-credit load-weighted scoring procedure. However, load-weighted scoring is rarely used in psychometrics, because there simply is no good reason to assign a greater weight to harder items. That is, all items within a task, such as WM span, are supposed to measure the same underlying ability, such as storage in the face of concurrent processing; they just discriminate at different points along the ability distribution. Moreover, a typical consequence of load-weighted scoring is positive skew: Individual differences in the upper half of the ability distribution are inflated, relative to the lower half of the ability distribution. Obviously, normal distributions are to be preferred in correlational studies.

Table 2
Results From Four Scoring Procedures for Three Working Memory Span Tasks of One Fictitious Subject

Scoring Procedure	Counting Span	Operation Span	Reading Span
Partial-credit unit scoring	(1 + 1 + 1 + 1 + 1 + 1 + .5 + 1 + 1 + .6) = 11.1/12 = **.93**	(1 + .5 + 1 + 1 + 1 + .75 + .5 + .75 + .8 + .8 + .8) = 9.9/12 = **.83**	(1 + 1 + 1 + 1 + 1 + .67 + .5 + .75 + 1 + .8 + .4 + .6) = 9.72/12 = **.81**
All-or-nothing unit scoring	(1 + 1 + 1 + 1 + 1 + 1 + 0 + 1 + 1 + 0) = 10/12 = **.83**	(1 + 0 + 1 + 1 + 1 + 0 + 0 + 0 + 0 + 0) = 5/12 = **.42**	(1 + 1 + 1 + 1 + 1 + 0 + 0 + 0 + 1 + 0 + 0 + 0) = 6/12 = **.50**
Partial-credit load scoring	(2 + 2 + 2 + 3 + 3 + 3 + 4 + 4 + 2 + 5 + 5 + 3) = 38/42 = **.90**	(2 + 1 + 2 + 3 + 3 + 3 + 3 + 2 + 3 + 4 + 4 + 4) = 34/42 = **.81**	(2 + 2 + 2 + 3 + 3 + 2 + 2 + 3 + 4 + 4 + 2 + 3) = 32/42 = **.76**
All-or-nothing load scoring	(2 + 2 + 2 + 3 + 3 + 3 + 4 + 4 + 0 + 5 + 5 + 0) = 33/42 = **.79**	(2 + 0 + 2 + 3 + 3 + 3 + 0 + 0 + 0 + 0 + 0 + 0) = 13/42 = **.31**	(2 + 2 + 2 + 3 + 3 + 0 + 0 + 0 + 4 + 0 + 0 + 0) = 16/42 = **.38**

To empirically compare these various scoring procedures in action, we reanalyzed the data from Kane et al. (2004). In a study with 236 subjects from both university and community samples, Kane et al. (2004) administrated three verbal WM tasks: operation span (12 items of two to five elements), counting span (15 items of two to six elements), and reading span (12 items of two to five elements). Table 3 presents internal consistencies, as indicators of task reliability, for the four scoring methods in these three tasks. Partial-credit scoring procedures show a clear advantage, and within these, unit-weighted scoring has a slight advantage over load-weighted scoring.

The correlations among the scores for all four scoring procedures within a task rely on the same initial information (whether or not an individual element was recalled correctly). Consequently, the correlations among scoring procedures necessarily are very high. Within the all-or-nothing and partial-credit scoring procedures, all correlations are .98 or higher for the three tasks. However, correlations between all-or-nothing scores and partial-credit scores within a task are substantially lower (although still high, ranging from .87 to .93). On the basis of this information there is no relevant difference between load weighting and unit weighting, once one commits to partial-credit versus all-or-nothing scoring. However, correlating partial-credit scores with all-or-nothing scores shows substantial deviations from perfect correlations within all of the tasks, and so researchers' decisions regarding these options should be considered carefully and justified theoretically.

In summary, established procedures of assigning absolute spans have various shortcomings, and so scoring procedures that exhaust the information collected with a task should be used instead. Because empirical results favor partial-credit scoring, we prefer it over all-or-nothing scoring. Our preference between unit-weighted and load-weighted procedures is less strong. The empirical results—including approximation of normal distributions—do not strongly favor one of these procedures over the other. However, one might favor unit-weighted scoring because it follows established and sound procedures from psychometrics.

RELIABILITY

Reading span, operation span, and counting span have been administered to literally thousands of subjects in over a hundred independent studies. One conclusion that can be drawn from this body of research is that measures obtained from these tasks (span scores) have adequate reliability. That is, irrespective of what WM span tasks are *supposed* to measure, evidence suggests that they measure, with reasonable accuracy, whatever it is that they *actually* measure. For example, estimates of reliability based on internal consistency, such as coefficient alphas and split-half correlations, which reflect the consistency of participants' responses across a test's items at one point in time, are typically in the range of .70–.90 for span scores, where values can range from 0 (*no reliability*) to 1 (*perfect reliability*). As a specific example, with a sample size of 236, Kane et al. (2004) observed coefficient alphas of .78 for reading span, .80 for operation span, and .77 for counting span. This indicates that subjects who responded with the correct answer for one set of span stimuli in these tasks (e.g., equation word pairs in operation span) tended to respond with the correct answer on the others (and vice versa). Therefore, span scores were reliable in the sense that there was consistency in responding across items within the task at one point in time. Internal consistency estimates of similar magnitudes have been reported in a number of other large-scale studies, including Conway et al. (2002), Engle, Tuholski, et al. (1999), Hambrick and Engle (2002), Miyake, Friedman, Rettinger, Shah, and Hegarty (2001), and Oberauer, Süß, Schulze, Wilhelm, and Wittmann (2000).

Evidence also suggests that WM span tasks are reliable in the sense that the rank order of span scores are stable *across* time. In adults, test–retest correlations of approximately .70–.80 have been observed for operation span and reading span, over minutes (Turley-Ames & Whitfield, 2003), over weeks (Friedman & Miyake, 2004; Klein & Fiss, 1999), and even over 3 months (Klein & Fiss, 1999).[4] In children, Hitch, Towse, and Hutton (2001) found operation span to be slightly less reliable over a year (.56), but reading span was more acceptable (.71). Importantly, Hitch et al. also found that the original administrations of operation and reading span predicted number skills and verbal skills measured 1 year later and that the second administrations of the span tasks accounted for very little additional variance beyond the first.

Although two studies have shown less adequate test–retest reliability for the reading span task, ranging from .50 over weeks to .40–.65 over months (MacDonald, Almor, Henderson, Kempler, & Andersen, 2001; Waters & Caplan, 1996), the observation that span scores correlate strongly with various other measures provides additional evidence for their reliability. This is because the correlation between two measures is limited by reliability. More specifically, the correlation between any two measures (x and y) cannot exceed the square root of the product of their reliabilities. That is,

$$r_{xy} \leq \text{sqrt}\left(r_{xx} \cdot r_{yy}\right).$$

Therefore, when span scores from one WM span task correlate with span scores from another span task, the implication is that the span scores *must* have some degree of reliability—given that if a measure has zero reliability,

Table 3
Internal Consistency for Three Working Memory Span Tasks With Four Scoring Procedures

Task	PCU	ANU	PCL	ANL
Counting span	.768	.668	.763	.673
Operation span	.814	.698	.808	.701
Reading span	.788	.697	.776	.699

Note—PCU, partial-credit unit scoring; ANU, all-or-nothing unit scoring; PCL, partial-credit load scoring; ANL, all-or-nothing load scoring.

correlation with all other measures will *necessarily* be zero. To illustrate, in Kane et al.'s (2004) study, the correlation between operation span and reading span was .69. This indicated that for either task, reliability must have been at least .69. As other examples, Conway and Engle (1996) and Lehto (1996) found correlations among operation span tasks that varied in processing difficulty in a range from .70 to .80, suggesting, again, that the reliability of operation span is at least .70.

One way that this evidence for reliability can be understood is in terms of *classical test theory*, first proposed by Spearman (1904). Briefly, the basic assumption of classical test theory is that a single score on a test—an *observed score* (x)—consists of two components. The *true score* (t) is assumed to reflect stable aspects of the trait (or traits) that the test measures, whereas *error* (e) is conceptualized as a random fluctuation in scores. That is,

$$x = t + e.$$

Correspondingly, the total variance of scores on a test (σ_x^2) is decomposed into *true-score variance* (σ_t^2) and *error variance* (σ_e^2). That is,

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2.$$

Finally, reliability (r_{xx}) is interpreted as the proportion of the total variance that is attributable to true-score variance:

$$r_{xx} = 1 - \left(\sigma_e^2 / \sigma_x^2\right).$$

Or conversely, the proportion of the total variance that is attributable to error variance is equal to one minus the reliability:

$$\sigma_e^2 / \sigma_x^2 = 1 - r_{xx}.$$

Within this framework, it can be demonstrated that span scores are influenced more by stable true scores than by error. Consider again the coefficient alphas observed by Kane et al. (2004) for reading span, operation span, and counting span: .78, .80, and .77, respectively. A coefficient alpha—the average reliability resulting from all possible split-half correlations for a test—indexes error due to factors operating at a given point in time, including momentary fluctuations of attention or mood, fatigue, and so forth. Thus, it follows that the proportion of the total variance in scores due to such random factors was 22% for reading span (1 − .78), 20% for operation span (1 − .80), and 23% for counting span (1 − .77).

In sum, as with any psychological instrument, no WM span task is free of measurement error. In other words, no WM span task is perfectly reliable. However, it is clear that WM span tasks do a reasonable job of measuring accurately whatever it is that they measure. As evidenced by acceptable reliability estimates, as well as by moderate to strong correlations with other measures, it is evident that span scores are influenced by something stable, with a minor contribution of error due to random fluctuations in scores. Next, we will consider the question of what this stable something is.

VALIDITY

Much of cognition is ballistic, in that one thought leads to the next through automatic activation. Attention is often captured by events in the environment and by thoughts that intrude into consciousness. Those perceptions and thoughts, in turn, lead inexorably to other thoughts. However, the solution to life's problems often requires that such automatically elicited thoughts, associations, and captured attention be resisted and thought be directed or controlled. We have argued that this ability to control attention and thought represents the common construct measured by tests of WMC. The evidence is quite clear that there are abiding individual differences in the ability to control attention and thought and that those differences are reflected by WM span tasks. It is also becoming clear that, in addition to the abiding individual differences in ability to control cognition, a host of other variables, from drunkenness to depression, also influence this ability, and those variables also manifest their effect on WM span tasks.

Performance on WM span tasks correlates with a wide range of higher order cognitive tasks, such as reading and listening comprehension (Daneman & Carpenter, 1983; Daneman & Merikle, 1996), language comprehension (King & Just, 1991), following oral and spatial directions (Engle, Carullo, & Collins, 1991), vocabulary learning from context (Daneman & Green, 1986), note taking in class (Kiewra & Benton, 1988), writing (Benton, Kraft, Glover, & Plake, 1984), reasoning (Barrouillet, 1996; Kyllonen & Christal, 1990), hypothesis generation (Dougherty & Hunter, 2003), bridge playing (Clarkson-Smith & Hartley, 1990), and complex-task learning (Kyllonen & Stephens, 1990).

WM span measures predict performance on lower level attention and perception tasks as well. For instance, in comparing individuals who score in the upper and lower quartiles on such tasks, lower quartile individuals (1) have difficulty resisting the attention capture of an exogenous cue in the antisaccade task (Kane et al., 2001; Unsworth, Schrock, & Engle, 2004), (2) have difficulty constraining their attention to discontiguous regions of space (Bleckley, Durso, Crutchfield, Engle, & Khanna, 2004), (3) are slower to constrain their focus of attention in a flanker task with incompatible distractors (Heitz & Engle, 2004), (4) make many more errors in a Stroop task (Kane & Engle, 2003), and (5) are more vulnerable to proactive interference (Kane & Engle, 2000). All of these findings point to the idea that the central construct measured by WM span tasks is the ability to control attention and thought.

Whereas the literature described above demonstrates the importance of the construct measured by WM span tasks for what might be thought of as *cold* cognition, more recent studies point to the importance of this construct for *hot* cognition as well. Barrett et al. (2004) reviewed a large literature connecting individual differences on WM span measures with dual-process theories of mind and

the ways in which those differences should be manifest in studies of emotional control and social cognition. Brewin and Beaton (2002), thinking of individual differences in WMC as a causal factor, showed that high WM span individuals were better at suppressing intrusive thoughts than were low WM span individuals. However, reductions in WMC can also be thought of as a result of events that control attention and thought. For instance, Klein and Boals (2001) found that individuals who reported more life event stress scored lower on operation span than did low life event stress individuals. The interpretation was that stressful events captured attentional resources, which reduced ability to perform the WM span task.

Perhaps the most elaborate example in which a span task has been used to measure the effect of a hot cognition variable on WM resources is an elegant article by Schmader and Johns (2003) on stereotype threat. Such a threat occurs when a relevant stereotype or social stigma is primed in the context of a performance situation and leads to a reduction in performance. Schmader and Johns reasoned that stereotype threat might have its effect on subsequent performance through reduction in available WMC. They had women complete both a WM span task and a standardized math test under stereotype threat or nonthreat conditions. Women in the stereotype threat condition did worse on the WM span task and on the standardized math test. More important, a mediation analysis supported the contention that reduction in WMC was responsible for the decreased performance on the math test.

As this mediation analysis suggests, WMC is sometimes viewed as a *cause* and sometimes as an *effect*. Early investigations of immediate memory performance almost always treated memory capacity as an outcome variable—that is, the dependent measure in an experimental design. In contrast, more recent work, particularly individual-differences studies, has tended to treat WMC as a more stable trait and to use scores on WM span tasks as predictors of some other outcome measure (e.g., using WMC to predict intelligence). Until the cognitive and biological mechanisms underlying performance of WM tasks are better understood, we argue that treating WMC as either a cause or an effect is warranted as long as one recognizes the limitations of any one investigation (e.g., recognizing the inherent limitations of correlational data). That said, we do view WM span scores as reflecting both stable interindividual variation and more state-dependent intra-individual variation. Although some researchers might take issue with the former argument about stability, there is direct evidence of stability from the strong test–retest reliability data discussed earlier, as well as indirect evidence of stability, which comes from the fact that WM span scores are strongly related to general fluid intelligence, which itself is relatively stable across the lifespan (Conway et al., 2003).

In summary, measures of WMC, such as counting, operation, and reading span, show considerable construct validity insofar as they predict performance on a wide array of tasks for which control of attention and thought are important. Importantly, construct validity implies not only *convergent*, but also *discriminant* validity. We have documented above the convergent validity of WM span tasks, such that they correlate extremely well with each other and correlate well with performance on tests of more complex cognition that purportedly depend upon WM. In addition, WM span tasks reveal *discriminant* validity in that they *do not* predict performance on tasks that appear to reflect relatively automatic processing, such as the prosaccade condition in the antisaccade task (Kane et al., 2001) or recall/recognition in the absence of interference (Cantor & Engle, 1993; Conway & Engle, 1994; Kane & Engle, 2000; Rosen & Engle, 1997, 1998). Finally, WM span tasks also diverge from more traditional "simple" span tasks in their *predictive* validity—that is, in their ability to successfully predict complex cognition. It is to this distinction between WM span tasks and other tests of immediate memory that we will now turn.

WM SPAN TASKS VERSUS OTHER WM TASKS

Ample empirical work has demonstrated the importance of the processing demand of WM span tasks by contrasting their predictive utility with STM span tasks, which present only to-be-recalled items and no additional processing task. In short, this research shows that WM span tasks tend to be stronger predictors of general intellectual ability than are STM span tasks, and STM span tasks account for no unique variance in general cognitive ability after variance related to WMC is accounted for (Conway et al., 2002; Engle, Tuholski, et al., 1999). Subjects in these studies completed multiple tests of verbal STM, all using word stimuli, as well as reading span, operation span, and counting span tasks. Confirmatory factor analysis and structural equation modeling techniques demonstrated that the variance shared among the STM tasks was closely related to the variance shared among the WM tasks; that is, the constructs of STM and WMC were correlated. However, the correlations were not strong enough to suggest that STM and WM tasks measure the same construct, and moreover, only the WMC construct shared unique variance with standardized tests of Gf. These studies will be discussed in more detail below (see the Latent Variable Analysis section), but for now they indicate that in order to most effectively measure WMC, a task must include a demanding secondary task to compete with information storage.

Questions remain open, however, regarding the required structure of a WMC task. Must the processing task present stimuli additional to those in the storage task, or can additional processing be required on the target items themselves? Must the presentation of target memory items alternate regularly with the presentation of the secondary processing task, or can the entire set of memory items be presented together, prior to a processing task? As well, the specific research goals of the investigator might influence the choice of WMC task. For example, to what extent

is it important to the question at hand to tap primarily domain-specific storage processes versus domain-general executive attention?

There is mixed evidence regarding the first question, of whether the secondary task must present actual stimuli to be processed, rather than simply requiring some mental transformation of the target memory items. Engle, Tuholski, et al. (1999) tested subjects on the backward word span task, in which target words were recalled in the reverse order from that in which they had been presented, in addition to traditional WM and STM span tasks. Note that the backward span's "processing" requirement was only a mental transformation, and not consideration of new stimuli, to interfere with the primary storage task. Factor analyses showed that backward word span grouped itself with the STM tasks, rather than with the WM span tasks, indicating that a mental transformation alone is not enough to turn an immediate-memory task into a WMC task (see also Hutton & Towse, 2001). In contrast, Oberauer et al. (2000) found that simple *transformation span* tasks seemed to measure the same construct as did WM span. They tested subjects in a backward digit span task and an *alpha span* task (Craik, 1986), in addition to reading span and counting span tasks. Alpha span required recall of target words in alphabetical order, rather than in their presentation order, and so, like backward span, presented a secondary processing task without secondary stimuli. Here, the correlation between reading and counting span ($r = .66$), was only slightly stronger than their correlations with the transformation tasks (mean $r = .60$), suggesting that all of these span tasks reflect a single construct. The source of the discrepancy between the Engle and the Oberauer findings is not obvious, so further research must determine the importance of interfering *stimuli*, in addition to interfering processing, to the measurement of WMC.

With respect to the second open question about the structure of WM tasks, the limited research regarding the regular interleaving of memory and processing stimuli in WMC tasks does not suggest it to be a critical variable. For example, Kane and Engle (2000) found that subjects identified as having high or low WMC (via a quartile split on the operation span task) performed quite differently on a Brown–Peterson-like task as proactive interference built throughout the task. The subjects recalled three consecutive lists of 10 words each, with all words drawn from the same category (e.g., *animals*) and with recall of each list preceded by a demanding processing task. Here, then, the secondary task followed, rather than alternated with, the target memory items. High- and low-span subjects recalled equivalent numbers of items on the first list, but low-span subjects recalled fewer items than did high-span subjects on subsequent lists. Thus, Brown–Peterson tasks may tap the WMC construct after several stimulus lists are presented and recalled, allowing proactive interference to challenge retrieval. Indeed, Oberauer et al. (2000) and Oberauer, Süß, Wilhelm, and Wittmann (2003) tested subjects in Brown–Peterson-like tasks, using 15–30 lists of target digits or words, and found that they correlated with WM span measures with mean *r*s of .59. In fact, their WM span, transformation span, and Brown–Peterson tasks (backward digit and alpha span) all loaded onto a single factor: a unitary WMC construct.

Although WM span, transformation span, and Brown–Peterson tasks are structurally heterogeneous in subtle ways, they all present subjects with lists of two to seven target items (often in a predictable sequence of list lengths) and require recall of each list, in turn, following some interfering task.[5] It may not be surprising, then, that they appear to measure the same underlying construct. However, other candidate WMC tasks in the literature, such as running span, keeping-track, and *n*-back tasks, present quite different cognitive demands. Specifically, these more dynamic tasks of immediate memory require subjects to monitor a continuous stream of stimuli, often of uncertain length, and to respond according to only a subset of the stimuli presented. The subjects in these tasks must, therefore, continuously update their mental representation of the target items while also dropping now-irrelevant items from consideration. So, like WM span tasks, some demanding processes are required in addition to storage.

More specifically, the running-memory span task (N. S. Anderson, 1960; Pollack, Johnson, & Knaff, 1959; Waugh, 1960) presents stimuli in lists of unknown length, and subjects must recall only the last *n* items (the prespecified, variable memory load). Thus, the subjects retain only the most recent *n* items that are presented and continuously drop items from the maintenance/rehearsal set once the list length exceeds *n*. Similarly, the keeping-track task (Yntema & Mueser, 1960, 1962) presents a list of items, of unknown length and from *n* categories (the memory load), and subjects retain only the most recent exemplar of each category. Finally, the *n*-back task (Kirchner, 1958; Mackworth, 1959; Moore & Ross, 1963) presents a list of items in which the subject must continuously report whether each item matches the one that had appeared *n* items ago in the stream (*n* typically ranges from 1 to 4). In a two-back task, for example, subjects must continuously maintain the last 2 items in the list, updating this memory set with each new item and dropping out the least recent one.

Unfortunately, very little research has contrasted these dynamic WM tasks with other WMC or STMC tasks. We know of no relevant studies on running span, but the very similar keeping-track task does appear to be a valid index of WMC. Engle, Tuholski, et al. (1999) used exploratory and confirmatory factor analyses to test whether a number of different immediate-memory tasks tapped either the STMC or the WMC. The keeping-track task had reasonably high loadings on the WMC factor (consisting of WM span tasks), low loadings on the STMC factor, and correlations with fluid intelligence scores of similar magnitude to those in the WM span task. Similarly, Oberauer and colleagues (Oberauer et al., 2000; Oberauer et al., 2003) found strong correlations among WM span tasks,

Brown–Peterson tasks, and versions of the keeping-track task developed by Salthouse (Salthouse, 1995; Salthouse, Babcock, & Shaw, 1991), in which different numbers of screen locations, rather than taxonomic categories, are monitored for the most recent items presented there. Thus, the keeping-track task and, perhaps by analogy, the running-memory span task currently appear to be valid measures of WMC, along with WM span and Brown–Peterson tasks.[6] Although the *n*-back task is arguably the current gold standard measure of WMC in the cognitive neuroscience literature (for a review, see Kane & Engle, 2002), almost no behavioral research has been conducted to validate it. The only study that has compared *n*-back with other immediate-memory tasks (Dobbs & Rule, 1989) found the two-back task to correlate more strongly with simple digit span than with a Brown–Peterson task (*r*s = .27 and .14, respectively). Given that the two-back task correlated only modestly with a one-back task (r = .38), its correlation with digit span may be considered relatively high. Thus, the *n*-back task may be a more appropriate indicator of the construct measured by STMC, rather than by WMC tasks, but more research is obviously needed.

Finally, we should note that the particular research goals of the investigator typically influence which tasks are used to measure WMC. For instance, some researchers are more interested in testing hypotheses about the mechanisms underlying storage and, as a result, may pay little attention to the attentional aspects of WM tasks, whereas others are more interested in testing hypotheses about the mechanisms underlying executive attention and cognitive control and, as a result, may pay little attention to the storage demands of the task. Although this state of affairs is understandable, it is also quite problematic, because tasks as diverse as operation span and *n*-back are referred to as *WM tasks* in the literature, as if they come from the same class, yet very little data exists to assess the extent to which they tap similar constructs.

LATENT VARIABLE ANALYSIS

All of the above-mentioned measures of WMC, including operation span, reading span, and counting span, suffer from the fact that no single task is a perfect measure of the construct it ostensibly represents. For example, the operation span task measures WMC but, most likely, also taps mathematical ability, motivation, and word knowledge, among other factors. Similarly, the reading span task measures WMC but, certainly, also verbal ability. Also, as was noted in the section on reliability above, despite the strong reliability of WM span tasks, they are not *perfectly* reliable. Thus, despite being valid and strongly reliable measures of WMC, WM span tasks are not perfect or process pure. Given such imperfection, an optimal research strategy is to administer multiple WM span tasks and then use the average (or weighted average) of scores on all the tasks as the measure of WMC (note that this is the same logic frequently applied by experimental psychologists; rather than measuring performance with one trial, they assess performance with multiple trials, and then a measure of central tendency is taken).

Latent variable analysis is a statistical approach in which multiple measures of a construct are administered and then a latent variable is derived from the common variance among those measures. For example, in our own work, we have derived a latent variable, which we label *working memory capacity*, from the common variance among counting span, operation span, and reading span (Conway et al., 2002; Engle, Tuholski, et al., 1999; Kane et al., 2004). Conceptually, the latent variable represents only the variance that is common among the three tasks and removes task-specific factors. Statistically, the latent variable is a predictor of the manifest (task) variables, and the strength of each predictor is a function of the correlations among the manifest tasks and the individual reliabilities of the tasks. Ideally, a set of manifest variables predicted by the same latent variable will have good reliability and be moderately to strongly correlated.

Counting span, operation span, and reading span, as a group, are particularly suited for latent variable analysis, because they are all moderately correlated with one another, suggesting that they are indeed tapping a similar construct, yet are not mere replicas of one another. From a measurement standpoint, this is an ideal situation; when a construct is measured with imperfect tools, it is best to use multiple, reliable measures that do not replicate one another. Correlations among the three span tasks typically range from .40 to .60, suggesting that they are indeed tapping some common process or ability but also suggesting that they are not identical (it is also worth noting here that the correlations are not diminished when spatial WM span tasks are considered; see Kane et al., 2004).

The main benefit of latent variable analysis is that a more "pure" measure of WMC can be derived from three span tasks than from one. As such, the predictive power of latent variables is better than that of individual manifest variables. To demonstrate this point, we analyzed the correlations between individual span tasks and Raven's advanced progressive matrices (a prototypical measure of general fluid intelligence) and contrasted these correlations with those between latent variables derived from the span tasks and Raven's matrices. The results of this analysis are reported in Table 4. The correlations between the latent variables and Raven's matrices are considerably higher than the correlations between the manifest variables and Raven's matrices. Also note that the correlations from the latent variable analyses are much more stable across studies than are the correlations from the manifest variables.

Another benefit of having multiple measures per construct is that multivariate outliers can be detected and the impact of such outliers can be controlled. For example, suppose that a subject is particularly anxious about math and, therefore, performs very poorly on the operation span task, despite the fact that he or she may "truly" have an above-average WMC. In this scenario, the subject might score well above average on counting span and reading span but well below average on operation span. Given that

Table 4
Correlations Between Multiple Measures of Working Memory Capacity and Raven's Advanced Progressive Matrices

Task	Engle, Tuholski, et al. (1999) (N = 133)	Conway et al. (2002) (N = 120)	Kane et al. (2004) (N = 236)
Counting span	.32	.38	.25
Operation span	.34	.20	.32
Reading span	.28	.15	.30
Latent variable	.44	.40	.37

Note—Kane et al. (2004) used only the odd-numbered items from the Raven's and implemented a strict time limit, which is not customary. These modifications to Raven's might account for the slightly lower correlations in that study.

all three tasks tend to be correlated, this subject's multivariate profile would be flagged as an outlier. Multivariate outliers can be detected by calculating the Mahanalobis distance (or d^2) for each case in a data set. The d^2 represents the distance a case is from the centroid of a multivariate distribution. For example, Figure 1 represents the three-way relationship between counting span, operation span, and reading span from Kane et al. (2004). Each line in the figure represents the Mahanalobis distance for one case. Furthermore, assuming multivariate normality, one can calculate the probability with which a case with such a distance would be observed. Outliers can be detected and deleted on the basis of this information.

A final benefit of latent variable analysis (and factor analysis, for that matter) is that latent variable scores (i.e., factor scores) can be calculated for each subject and then used as a dependent measure in experimental designs. For example, in Kane et al.'s (2004) study, some subjects were college students, whereas others were community dwellers. If one were interested in group comparisons, factor scores could be created for each individual in the data set, and these scores could be entered into an analysis as the dependent measure.

In sum, latent variable analysis is a powerful research strategy that has proven particularly beneficial in examining individual differences in WMC. Of course, these methods are expensive and time consuming. Multiple measures of each construct are needed, and large samples are required. When tasks are chosen for a latent variable analysis, it is important to select tasks that are reliable and also reveal moderate to strong correlations with one another. The overall correlation matrix for the study also should indicate both convergent and discriminant validity, meaning that measures that purportedly measure the same construct should converge on one another and diverge from the rest of the pack. In determining sample size, there are many factors to consider, such as the number of tasks used, the quality of each univariate distribution, and the number of parameters being estimated in the latent variable model. Although there are no hard and fast rules for determining sample size for a latent variable design or for structural equation modeling, it has been suggested

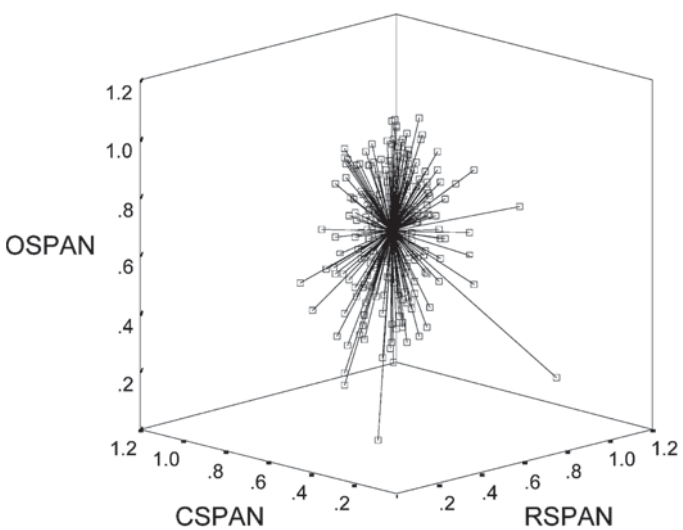

Figure 1. The three-way relationship between counting span (CSPAN), operation span (OSPAN), and reading span (RSPAN) from Kane et al. (2004), N = 236. Each line represents Mahanalobis distance for each case in the data set. The *x*-, *y*-, and *z*-axes represent scores on OSPAN, RSPAN, and CSPAN, respectively.

that in any situation more than 100 subjects are necessary and 10 times the number of manifest variables is also desired (Kline, 1998).

EXTREME-GROUPS DESIGNS

Extreme-groups designs refer to situations in which a continuous variable is categorized and only categories representing the upper and lower ends of the distribution of the continuous variable are represented. In the case of WM span tasks, the most common extreme-groups design is one in which the upper and lower quartiles of a distribution of WM span scores are categorized as high and low span, respectively. The process of categorizing a continuous variable is considered problematic among most statisticians for obvious reasons. First, information and power are lost, because less variability is captured by categories than by a continuum. Second, subjects who are not equal on some ability or trait are treated as if they are equal. Third, subjects can easily be misclassified, due to measurement error. We will not review these issues here. A more in-depth discussion of the problems associated with categorization can be found in Cohen (1983).

Despite these problems, extreme-groups designs are common in the WM literature (e.g., the present authors have collectively published over a dozen experiments with extreme-groups designs). In this section, we will justify the use of extreme-groups designs but will caution researchers to use them only in certain situations and to be aware of the interpretive problems that they can create.

First, it is necessary to explicitly state that, according to all theoretical accounts of WM that we are familiar with, WMC is assumed to be normally distributed in a population of healthy subjects and, therefore, should be measured with an instrument that can produce a continuous normal distribution. In terms of creating a normal distribution, WM span tasks are successful, particularly if they are scored with the partial-credit unit-weighting method, as was discussed above. For example, using this scoring procedure, Kane et al. (2004) observed normal distributions (as indicated by standard deviation, skew, and kurtosis) for each counting span, operation span, and reading span. The Kane et al. (2004) distributions will serve as a good reference point for other researchers who would like to use these tasks, because they were created using our most recent versions of the span tasks (which are available on our Web site) and because the sample was quite large ($N = 234$) and represents both college student and nonstudent populations. Thus, if a researcher is concerned that he or she is suffering from a restriction of range in WMC (e.g., if the research is being conducted with college students from an elite institution), he or she could check the distribution against the ones observed in Kane et al. (2004).

As was mentioned above, the ideal research approach is to sample the entire range of WMC. This is especially true if the goal of the research is to estimate, in the population, the magnitude of the relationship between WMC and performance of some other task (or in some experimental context). However, if the goal is simply to test whether a relationship exists or does not exist, a more efficient approach is to compare extreme groups. As Underwood (1975) has suggested, most nomothetic psychological theories make at least tacit predictions about individual differences; thus, confirmed predictions regarding individual differences may give a theory a "go-ahead signal," whereas failed predictions may refute the theory, in which case "there is no alternative but to drop the line of theoretical thinking" (p. 130). Hence, Underwood famously concluded that "individual differences may indeed be used as a crucible in nomothetic theory construction." Critical to our argument is that it is often the mere presence or absence of a relationship that is in question, rather than the magnitude of the relationship.

Of course, one hazard of extreme-groups designs is that the observed span × treatment interaction might be a Type 1 error due to the sampling of extreme groups or might be an overestimation of the true relationship between WM span and the treatment. In order to illustrate this point, we simulated experiments in which the entire distribution of span was used and calculated the effect size (R^2) for a simulated span effect. We then simulated the same effect, using an extreme-groups design, in which the upper and lower quartiles were used as groups and an ANOVA was conducted. Figure 2 demonstrates the extent to which the extreme-groups design overestimates the "true" effect size (i.e., the effect size that would have been observed had the experiment been run on the continuum). As the figure illustrates, extreme-groups designs tend to modestly overestimate effect size, particularly for moderate effects.

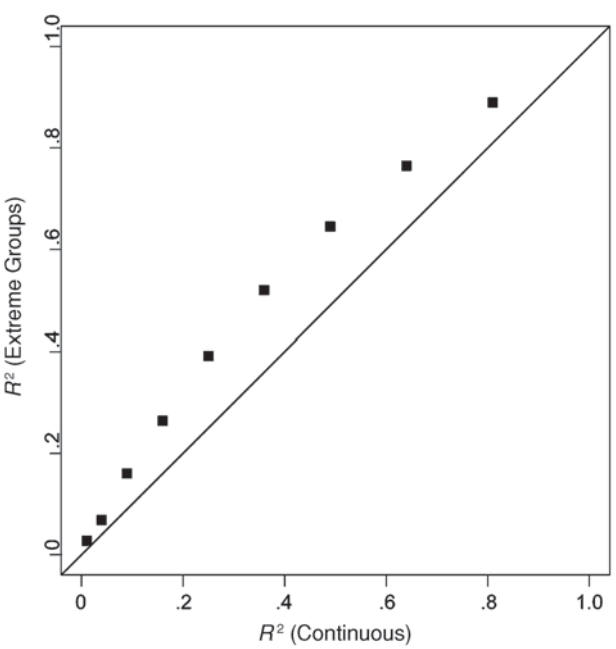

Figure 2. Simulated effect sizes (R^2) for situations in which the entire distribution of span is used (continuous; *x*-axis), as compared with situations in which extreme groups are used (*y*-axis).

Table 5
Number of Subjects Classified as High Span on One Span Task and as Low Span on Another Span Task

	High Span on:		
Low Span on:	Counting Span	Operation Span	Reading Span
Counting span	0	4	3
Operation span	2	0	2
Reading span	1	3	0

Note—Data are from Conway, Cowan, Bunting, Therriault, and Minkoff (2002). $N = 120$ (therefore, the maximum number of misclassifications per cell is 30).

Thus, extreme-groups designs can be cost efficient, but they can pose interpretive problems. We therefore caution researchers to use extreme-groups designs sparingly and to be aware of the problems documented here. We also advocate replicating effects that are observed in extreme-groups designs, as well as the use of converging methods to provide support for the theoretical conclusions derived from extreme-groups design experiments.

An additional positive aspect of extreme-groups designs is worth mentioning here. When the upper and lower quartiles of a distribution of WM span scores are classified as high and low span, respectively, subjects are rarely misclassified as high when they should have been classified as low, and vice versa. To demonstrate this point, we reanalyzed the data from Conway et al. (2002). Table 5 illustrates how many subjects would have been classified as high span on one span measure and as low span on another. As the table illustrates, a very small percentage of subjects (approximately 8%) would have been misclassified had a quartile design been used with these distributions. To further illustrate this point, we examined the consistency of quartiles across the three span tasks in Conway et al. (2002). In this analysis, we first created four quartiles on the basis of a *z* score composite representing the average of counting, operation, and reading span. We then examined how many subjects were classified in the correct quartile (i.e., the same quartile they were classified by with the composite score) when only one span task was considered and then when two span tasks were considered. The results of this analysis are presented in Table 6. This analysis suggests that quartile efficiency (i.e., the extent to which a subject is classified in the correct quartile) is significantly better when two span tasks are considered than when just one span task is considered. On the basis of this analysis, we recommend using at least two span tasks to assess WMC, whenever possible.

Finally, we would like to note here that median split designs are not acceptable, for two simple reasons: (1) There is no reason to categorize subjects when the entire continuum has been sampled, and (2) misclassification of subjects is more likely in a median split design than in an extreme-groups design. Recall that in the Conway et al. (2002) data, as illustrated in Table 5, only 15 out of a possible 180 classifications were mismatches (8% of the cases) when classifications were based on quartile splits. A parallel analysis examining the consistency of median split classifications with the same data reveals that 25% of the cases were mismatches.

CONCLUSION

WM is a central construct in cognitive psychology. Furthermore, WMC is an important individual-differences variable in differential approaches to understanding human behavior. WM span tasks, such as counting span, operation span, and reading span, are reliable and valid measures of WMC. They have proven to be extremely useful research tools in cognitive psychology and, more recently, in other branches of psychology. In an attempt to maximize the future utility of these tasks, we have documented here all the relevant information we have gathered in our 15 years experience with the tasks.

Our hope is that the review above has demonstrated the reliability and validity of WM span tasks. As well,

Table 6
Consistency of Quartiles When One or Two Working Memory (WM) Span Tasks Are Used to Assess WM Capacity, Relative to a Standard in Which Three WM Span Tasks Are Used

Quartile	*z* Score Composite	CSPAN	OSPAN	RSPAN	CSPAN & OSPAN	CSPAN & RSPAN	OSPAN & RSPAN
1	30	19	22	21	24	24	27
2	30	7	16	13	20	20	20
3	30	15	18	11	21	23	18
4	30	21	19	21	26	27	25
Efficiency for all quartiles		.52	.63	.55	.76	.78	.75
Efficiency for upper and lower quartiles		.67	.68	.70	.83	.85	.87

Note—Each cell represents the number of subjects classified in the correct quartile (i.e., the same quartile in which they were classified by the *z* score composite). *Efficiency* refers to the proportion of subjects classified correctly. CSPAN, counting span; OSPAN, operation span; RSPAN, reading span.

we hope that other researchers will now be able to use these tasks to their full potential, by adopting the optimal administration and scoring procedures. We also encourage other researchers to conduct latent variable analyses whenever possible. We caution researchers to use extreme-groups designs sparingly and with great care. Finally, we do not suggest that these WM span tasks are, or should be, the *gold standard* measures of WMC. Instead, we hope that our explicit documentation of WM span tasks, their genesis, and their development will inspire new task development and analysis, which hopefully will result in even better measurement of WMC and cognitive abilities in the future.

REFERENCES

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Anderson, N. S. (1960). Poststimulus cuing in immediate memory. *Journal of Experimental Psychology*, **60**, 216-221.

Arnett, P. A., Higginson, C. H., Voss, W. D., Bender, W. I., Wurst, J. M., & Tippin, J. M. (1999). Depression in multiple sclerosis: Relationship to working memory capacity. *Neuropsychology*, **13**, 546-556.

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

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.

Barrett, L. F., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin*, **130**, 553-573.

Barrouillet, P. (1996). Transitive inferences from set-inclusion relations and working memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 1408-1422.

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.

Bleckley, M. K., Durso, F. T., Crutchfield, J. M., Engle, R. W., & Khanna, M. M. (2004). Individual differences in working memory capacity predict visual attention allocation. *Psychonomic Bulletin & Review*, **10**, 884-889.

Brewin, C. R., & Beaton, A. (2002). Thought suppression, intelligence, and working memory capacity. *Behaviour Research & Therapy*, **40**, 923-930.

Bunting, M. F. (in press). Proactive interference and item similarity in working memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

Cantor, J., & Engle, R. W. (1993). Working-memory capacity as long-term memory activation: An individual-differences approach. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 1101-1114.

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.

Clarkson-Smith, L., & Hartley, A. A. (1990). The game of bridge as an exercise in working memory and reasoning. *Journal of Gerontology*, **45**, P233-P238.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, **3**, 249-253.

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

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.

Conway, A. R. A., & Kane, M. J. (2001). Capacity, control and conflict: An individual differences perspective on attentional capture. In C. Folk & B. Gibson (Eds.), *Attraction, distraction and action: Multiple perspectives on attention capture* (pp. 349-372). Amsterdam: Elsevier.

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

Copeland, D. E., & Radvansky, G. A. (2001). Phonological similarity in working memory. *Memory & Cognition*, **29**, 774-776.

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

Cowan, N., Towse, J. N., Hamilton, Z., Saults, J. S., Elliott, E. M., Lacey, J. F., et al. (2003). Children's working-memory processes: A response-timing analysis. *Journal of Experimental Psychology: General*, **132**, 113-132.

Craik, F. I. M. (1986). A functional account of age differences in memory. In F. Klix & H. Hagendorf (Eds.), *Human memory and cognitive capabilities* (pp. 409-422). Amsterdam: North-Holland.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & 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, & Cognition*, **9**, 561-584.

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

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

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology & Aging*, **4**, 500-503.

Dougherty, M. R. P., & Hunter, J. (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, **31**, 968-982.

Ebbinghaus, H. (1897). Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern [On a new method of testing mental abilities and its application with school children]. *Zeitschrift für die Psychologie*, **13**, 401-459.

Engle, R. W. (2001). What is working memory capacity? In H. L. Roediger, III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297-314). Washington, DC: American Psychological Association Press.

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

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., 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: Mechanisms of active maintenance and executive control* (pp. 102-134). New York: Cambridge University Press.

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

Finn, P. R. (2002). Motivation, working memory, and decision making: A cognitive–motivational theory of personality vulnerability to alcoholism. *Behavioral & Cognitive Neuroscience Reviews*, **1**, 183-205.

Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory & Language*, **51**, 136-158.

Hambrick, D. Z., & Engle, R. W. (2002). Effects of domain knowledge, working memory capacity, and age on cognitive performance: An investigation of the knowledge-is-power hypothesis. *Cognitive Psychology*, **44**, 339-384.

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). San Diego: Academic Press.

Heitz, R. P., & Engle, R. W. (2004). *Focusing the spotlight: Individual differences in visual attention control.* Manuscript submitted for publication.

Hitch, G. J., Towse, J. N., & Hutton, U. (2001). What limits children's working memory span? Theoretical accounts and applications for scholastic development. *Journal of Experimental Psychology: General*, **130**, 184-198.

Hutton, U. M. Z., & Towse, J. N. (2001). Short-term memory and working memory as indices of children's cognitive skills. *Memory*, **9**, 383-394.

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

Kane, M. J., & Engle, R. W. (2000). Working memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 333-358.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, **9**, 637-671.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, **132**, 47-70.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General*, **133**, 189-217.

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 & Language*, **30**, 580-602.

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

Klein, K., & Boals, A. (2001). The relationship of life event stress and working memory capacity. *Applied Cognitive Psychology*, **15**, 565-579.

Klein, K., & Fiss, W. H. (1999). The reliability and stability of the Turner and Engle working memory task. *Behavior Research Methods, Instruments, & Computers*, **31**, 429-432.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.

Kyllonen, P. C. (1996). Is working memory capacity Spearman's *g*? In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 49-75). Mahwah, NJ: Erlbaum.

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. *Learning & Individual Differences*, **2**, 129-160.

Lehto, J. (1996). Are executive function tests dependent on working memory capacity? *Quarterly Journal of Experimental Psychology*, **49A**, 29-50.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General*, **130**, 199-207.

MacDonald, M. C., Almor, A., Henderson, V. W., Kempler, D., & Andersen, E. S. (2001). Assessing working memory and language comprehension in Alzheimer's disease. *Brain & Language*, **78**, 17-42.

Mackworth, J. F. (1959). Paced memorizing in a continuous task. *Journal of Experimental Psychology*, **58**, 206-211.

May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory & Cognition*, **27**, 759-767.

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

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

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British Journal of Psychology*, **81**, 111-121.

Munakata, Y., Morton, B., & O'Reilly, R. C. O. (in press). Developmental and computational approaches to variation in working memory. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory*. Oxford: Oxford University Press.

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology*, **53**, 53-81.

Oberauer, K., & Süß, H.-M. (2000).Working memory and interference: A comment on Jenkins, Myerson, Hale, and Fry (1999). *Psychonomic Bulletin & Review*, **7**, 727-733.

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity: Facets of a cognitive ability construct. *Personality & Individual Differences*, **29**, 1017-1045.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, **31**, 167-193.

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

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., et al. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience*, **6**, 1323-1328.

Richeson, J. A., & Shelton, J. N. (2003). When prejudice does not pay: Effects of interracial contact on executive function. *Psychological Science*, **14**, 287-291.

Rosen, V. M., Bergeson, J. L., Putnam, K., Harwell, A., & Sunderland, T. (2002). Working memory and apolipoprotein E: What's the connection? *Neuropsychologia*, **40**, 2226-2233.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, **126**, 211-227.

Rosen, V. M., & Engle, R. W. (1998). Working memory capacity and suppression. *Journal of Memory & Language*, **39**, 418-436.

Salthouse, T. A. (1995). Differential age-related influences on memory for verbal-symbolic and visual-spatial information? *Journal of Gerontology,* **50B**, P193-P201.

Salthouse, T. A., Babcock, R. L., & Shaw, R. J. (1991). Effects of adult age on structural and operational capacities in working memory. *Psychology & Aging*, **6**, 118-127.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality & Social Psychology*, **85**, 440-452.

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

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, **15**, 201-293.

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, **30**, 261-288.

Towse, J. N., & Hitch, G. J. (1995). Is there a relationship between task demand and storage space in tests of working memory capacity? *Quarterly Journal of Experimental Psychology*, **48A**, 108-124.

Towse, J. N., Hitch, G. J., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory & Language*, **39**, 195-217.

Towse, J. N., Hitch, G. J., & Hutton, U. (2002). On the nature of the relationship between processing activity and item retention in children. *Journal of Experimental Child Psychology*, **82**, 156-184.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, **12**, 97-136.

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

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

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist*, **30**, 128-134.

Unsworth, N., Heitz, R. P., & Engle, R. W. (2005). Working memory capacity in hot and cold cognition. In R. W. Engle, G. Sedek, U. Hecker, & D. N. McIntosh (Eds.), *Cognitive limitations in aging and psychopathology* (pp. 19-43). New York: Oxford University Press.

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

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *Quarterly Journal of Experimental Psychology*, **49A**, 51-74.

Waugh, N. (1960). Serial position and the memory-span. *American Journal of Psychology*, **73**, 68-79.

Yntema, D. B., & Mueser, G. E. (1960). Remembering the present states of a number of variables. *Journal of Experimental Psychology*, **60**, 18-22.

Yntema, D. B., & Mueser, G. E. (1962). Keeping track of variables that have few or many states. *Journal of Experimental Psychology*, **63**, 391-395.

NOTES

1. For our discussion of the administration and scoring of WM span tasks, it is important to distinguish between elements, items, and tasks. *Elements* are the individual stimuli that have to be recalled. This is the lowest level of observation. The recall of an element is either correct or incorrect. *Items* include various numbers of elements. Usually, there is an experimental manipulation of how many elements an item includes. The number of elements in an item has often been labeled *set size*, and typically varies between two and seven. A number of homogeneous items make up a *task*. In psychometrics, a task is usually called a *test*. We prefer the term *task*, in order to maintain continuity with the experimental, rather than the psychometric, origins of this research tradition.

2. Engle, Tuholski, et al. (1999) also changed the stimuli by adding the word *is* before the operation word string. Stating the operation in the form of a question emphasized that the subjects were to indicate aloud whether or not the stated solution was correct.

3. Additional data can also be obtained from the recall portion of WM span tasks. Although it is not a very common procedure, evidence suggests that overall recall durations, as well as the pause time between recall of words, can add independent variance to that offered by span scores in predicting ability (Cowan et al., 2003).

4. Klein and Fiss (1999) found that scores on the operation span task markedly increased from the first to the second administration of the test, indicating a practice effect. This result is important for researchers interested in training or intervention-type manipulations that require a pretest/posttest design with a control group.

5. Most WM tasks require serial recall. However, the original reading span task was essentially a constrained free recall task, in that subjects could recall the sentence-final words in any order, as long as they did not recall the last word first. More recently, Friedman and Miyake (2004) instructed subjects to recall in order, but if they could not, then to just avoid recalling the last word first. Thus, strict serial recall is not always required.

6. There is some evidence that running-memory span may have more in common with STM than with WMC: It is highly vulnerable to the effects of articulatory suppression and background speech (Morris & Jones, 1990), which are hallmarks of the operation of the phonological loop, or verbal STM system (see Baddeley, 1986).

(Manuscript received June 18, 2004;
revision accepted for publication February 24, 2005.)