Behavior Research Methods, Instruments, & Computers
2003, 35 (4), 550-564

The reliability and stability of verbal working memory measures

GLORIA S. WATERS
Boston University, Boston, Massachusetts

and

DAVID CAPLAN
Massachusetts General Hospital, Boston, Massachusetts

The psychometric properties of several commonly used verbal working memory measures were assessed. One hundred thirty-nine individuals in five age groups (18–30, 50–59, 60–69, 70–79, and 80+ years) were tested twice (Time I and Time II) on seven working memory span measures (alphabet span, backward digit span, missing digit span, subtract 2 span, running item span, and sentence span for syntactically simple and complex sentences), with an interval of approximately 6 weeks between testing. There were significant effects of age on all but two of the tasks. All the measures had adequate internal consistency. Correlations between performances at Time I and Time II were significant for all the tasks, other than the missing digit span task. The magnitude of the correlations was similar across the age groups and ranged from .52 to .81. Classification of subjects into discrete memory span groups on the basis of a single measure was highly inconsistent across testing sessions and tasks. Classification into upper and lower quartiles was more stable than using a cutoff score for group membership or than classification into high-, medium-, and low-span groups. Correlational analyses showed that there was a moderate relationship between performances on many of the span tasks. Confirmatory factor analysis suggested that six of the seven tasks reflected a common factor. Both test–retest reliability and stability of classification improved when a composite measure reflecting performance on several tasks was used.

Baddeley and Hitch developed the concept of working memory (WM) in a seminal paper in 1974. They argued that most cognitive tasks require the use of a WM system that not only stores small amounts of information for brief periods of time, as the older short-term memory (STM) system had been thought to do, but also simultaneously processes information. They carried out several experiments in which subjects were required to carry out a cognitive task, such as verbal reasoning, comprehension, or free recall, while simultaneously holding onto a memory load. Because the subjects were more affected by a concurrent memory load than by control tasks that simply required rehearsal, Baddeley and Hitch argued that reasoning and comprehension required WM, and not simply STM.

Because of the presumed role of WM in many cognitive tasks, there has been considerable interest in the extent to which individual differences in WM capacity may explain individual differences in other cognitive domains. Studies in which the role of WM in some aspect of cognitive performance has been evaluated have been carried out in a wide variety of populations, including children (e.g., Gathercole & Baddeley, 1990), college students (e.g., Just & Carpenter, 1992; Waters & Caplan, 1996), elderly individuals (e.g., Craik, Morris, & Gick, 1990; Park et al., 1996; Salthouse, 1990), and various brain-damaged populations (e.g., Caplan & Hildebrandt, 1988; Miyake, Carpenter, & Just, 1994; Tompkins, Bloise, Timko, & Baumgaertner, 1994; Waters & Caplan, 1997). However, the investigation of the role of WM in various cognitive tasks requires that WM be measured accurately. The focus of this article is on the measurement of WM capacity in both young and elderly individuals. Below, we will review the literature on various means that have been used to measure WM capacity and on the psychometric properties of these measures.

Measurement of WM Capacity

Many tasks have been developed that attempt to measure WM capacity. The common element in these tasks is that, in contrast to simple span tasks in which the material is usually repeated verbatim after some delay, these tasks involve both a processing and a storage component. Numerous WM span tasks have been developed that require the subject to perform an operation on each item or on the list as a whole and then to repeat the list or a particular item back. Examples of these tasks include tasks

This work was supported by Grant AG00966 from the National Institute on Aging. Correspondence concerning this article should be addressed to G. S. Waters, Department of Health Sciences, Boston University, 635 Commonwealth Ave., Boston, MA 02116 (e-mail: gwaters@bu.edu).

that require subjects to repeat back a series of words after arranging them in alphabetical order (alphabet span; Craik, 1986) or to repeat a series of digits in reverse order (backward digit span task; Botwinick and Storandt, 1974) or after subtracting 2 from each (subtract 2 span task; Salthouse, 1988b). One of the most widely used of such tasks is Daneman and Carpenter's (1980) reading or listening span task, in which subjects read aloud or listen to increasingly longer sequences of sentence and then recall the final words of all of the sentences in each sequence.

Other WM tasks have been developed in which subjects are required to process and store a list of items and, at some point in the presentation of the list, to retrieve a previously presented item that has some relationship to the item currently being presented. The *n*-back task (Welford, 1958) and the running item task (Talland, 1968) are examples of such tasks. Other tasks require the subject to report which item is missing upon hearing the list for a second time (e.g., missing digit; Talland, 1965). Finally, in yet other tasks, subjects are required to perform an operation while simultaneously storing items that are unrelated to the operation (e.g., operation word span task; Turner & Engle, 1989).

Performance on WM tasks is often expressed as a continuous measure. Many researchers have reported subjects' WM spans, where WM span may be defined as the longest sequence of items for which recall of all or the majority of items was correct (e.g., Daneman & Carpenter, 1980). Other researchers test all subjects to a common span size and calculate the total number of items correctly recalled on the task (e.g., Tompkins et al., 1994; Waters & Caplan, 1996). However, in many studies, WM is treated as a categorical measure. In these studies, subjects are typically divided into high-, medium-, and low-span groups, using an absolute cutoff score (e.g., MacDonald, Just, & Carpenter, 1992), or are divided into high- and low-span groups on the basis of the upper and lower quartiles of performance.

Psychometric Properties of WM Tasks

Despite their widespread use, several basic psychometric properties of WM tasks remain incompletely characterized. These properties include the internal consistency of many tasks, their test–retest reliability, and the stability of categorization of subjects into WM groups over short time periods. Internal consistency refers to the homogeneity of a measure and is frequently assessed by determining the extent to which items or subtests correlate with the total score. Test–retest reliability refers to the extent to which scores obtained in one testing session correlate with those obtained in another. Stability of subject classification refers to the issue of whether classification of subjects into discrete WM groups (e.g., high, medium, or low) is stable over time and/or different WM tasks.

Below, we review the results of studies that have explored the psychometric properties of WM tasks. In general, estimates of internal consistency tend to be higher than estimates of test–retest reliability. Stability of subject classification into WM groups seems to have been poor in many studies.

Internal consistency. Table 1 shows some representative studies that have examined internal consistency as measured by split-half reliability. The tasks used in these studies include Daneman and Carpenter's (1980) reading

Table 1
Internal Consistency as Measured by Split-Half Reliability for Several Working Memory Tasks

Study	*N*	Task	Estimate
Salthouse & Babcock (1991)		computation span	
Study 1	227		.90
Study 2	233		.84
Salthouse & Babcock (1991)		listening span	
Study 1	227		.86
Study 2	233		.86
Park et al. (2002)	345	computation span	.91
Park et al. (2002)	345	reading span	.88
Waters & Caplan (1996)		Daneman & Carpenter reading span	
Time 1	96		.82
Time 2	44		.78
Waters & Caplan (1996)		Waters & Caplan reading span	
Time 1	96		.95
Time 2	44		.92
Tirre & Peña (1992)	283	reading span	
Word recall			.95
Sentence verification			.67
Klein & Fiss (1999)		Turner & Engle operation span	
Time 1	33		.78
Time 2	33		.81
Time 3	33		.83

span task (described above), a variant of this task in which subjects listened to sentences (Salthouse & Babcock, 1991) or read them silently (Waters & Caplan, 1996) and made judgments about them, the computation span task, and the operation span task. In the computation span task, subjects saw an equation on the screen, selected the correct answer from three alternatives, and then stored in memory the last digit from the equation. The subjects were required to recall the final digits after completing a designated number of problems, with the number of problems increasing across trials. In the operation span task, subjects were presented with a series of simple arithmetic operations with an answer followed by a one-syllable word (e.g., [9 × 1)−9 = 1 back]. They were required to respond verbally whether the answer following the equal sign was true or false and then to say the word that followed the operation. Waters and Caplan (1996) and Klein and Fiss (1999) tested college students, Tirre and Peña (1992) tested U.S. Air Force personnel, Salthouse and Babcock tested elderly subjects, and Park et al. (2002) tested subjects across the life span. As can be seen in Table 1, internal consistency, as estimated by split-half reliability, ranges from .67 to .95 for these tasks.

Test–retest reliability. Fewer studies have examined test–retest reliability of WM measures. Backward digit span is measured on the Wechsler Adult Intelligence Scale, and test–retest reliability is reported as part of the standardization for this test. Waters and Caplan (1996) examined test–retest reliability after an interval of approximately 3 months for 44 college students who were tested on Daneman and Carpenter's (1980) reading span test and on the version of the task outlined above, in which the subjects made judgments about the acceptability of the sentences. MacDonald, Almor, Henderson, Kempler, and Andersen (2001) examined test–retest reliability for a version of Daneman and Carpenter's task in which the stimuli were modified to allow for testing on two separate occasions separated by a week. Klein and Fiss (1999) tested college students across three administrations of the operation span test. As can be seen in Table 2, test–retest reliability ranges from .41 to .83 for these tasks.

Stability of subject classification. Waters and Caplan (1996) also investigated the stability of subject classification (the extent to which individuals who are categorized at a particular level retain their status across time) for Daneman and Carpenter's (1980) task and for their variant of the sentence span task. The subjects were divided into high, medium, and low WM span groups on the basis of their scores on Daneman and Carpenter's task at Time 1 and Time 2. Of the 44 subjects who participated in the follow-up study, 41% changed in terms of their classification at Time 2, with equal numbers of subjects improving and declining. Klein and Fiss (1999) also examined whether individuals classified as high or low span would maintain this categorization across the testing intervals, using the operation span test. They found a classification error rate of 10% for Time 1–Time 2 and 5% for Time 2–Time 3. Stability of subjects' classification was thus much higher than that reported by Waters and Caplan (1996).

Scoring method. A factor that may affect the reliability and stability of WM measures is the method used to score the task. Miyake, Emerson, and Friedman (1999) have claimed that highly discrete measures of individual differences, such as the traditional reading span scores, are nonoptimal and result in low statistical power, since they reduce variance by not capturing subtle differences that may exist among individuals. In addition, they pointed out that the division of subjects into WM span groups raises problems, because this method treats all the members of a group as identical and so reduces power. They claimed that more continuous ways of scoring WM span tasks, such as the total number of words recalled, are better, since these methods increase power. However, several studies have suggested that these measures are highly correlated. Turner and Engle (1989) calculated two different scores: the traditional measure of memory span (the maximum size of the set for which the subject recalled the words correctly on the majority of trials) and the total memory span (the sum of the number of correctly recalled words across the entire task). They did not report the results for each scoring method but reported that the two different types of scores led to the same conclusions, in a study in which the question of

Table 2
Test–Retest Reliability for Several Working Memory Tasks

Study	*N*	Task	Estimate
Wechsler (1981)	*	backward digit span	.83
Waters & Caplan (1996)	44	Daneman & Carpenter reading span	.41
Waters & Caplan (1996)	44	Waters & Caplan reading span	.65
MacDonald et al. (2001)	38	Daneman & Carpenter reading span	.52
Klein & Fiss (1999)	33	Turner & Engle operation span	
Time 1, Time 2			.67
Time 2, Time 3			.73
Time 1, Time 3			.81

*According to the manual, the reliability coefficient is based on double-testing studies of samples ranging in size from 48 to 80 individuals at four age groups.

whether the measurement of WM capacity is task dependent was investigated. Waters and Caplan (1996) found a correlation of .91 for these two types of measures for Daneman and Carpenter's (1980) reading span task and correlations ranging between .93 and .95 for these two types of scores for their variant of the reading span task, in which subjects made acceptability judgments. Klein and Fiss (1999) found that the relationship between the two methods used to score the operation word span task were .89, .92, and .91 at the three testing times.

Overall, these studies demonstrate variable test–retest reliability for several commonly used measures of WM. Whether test–retest reliability improves when subjects' item or span scores, rather than their classification, are used is not clear. None of the studies that have provided data on this issue has studied older subjects.

Relationship Among WM Tasks

A related set of questions that arises about WM tasks is how performance compares across different tasks. Given the number of WM tasks that have been developed, this is important operationally if researchers are to use a generalizable measure of WM. Although it is often assumed that performance on all WM tasks reflects a common mechanism, it is possible that different tests measure at least partially separate cognitive capacities (Salthouse, 1990) and that performance dissociates over different tasks.

There are also few studies of this subject. In their original study, Daneman and Carpenter (1980) demonstrated a high correlation between their original task, in which the processing operation consisted of reading sentences aloud, and reading and listening span tasks, in which the processing operation consisted of making judgments about the acceptability of written or orally presented sentences. In other studies, a wider range of tasks have been explored, using both correlational and factor analytic approaches.

The results of several studies of college students seem to provide evidence for the distinction between STM and WM tasks. Cantor, Engle, and Hamilton (1991) tested 49 undergraduates on two STM span tasks (digit and word span), two WM span tasks (two operation span tasks; Turner & Engle, 1989), and two probe recall tasks. Factor analysis identified two separate factors that the authors associated with STM, measured by the span and probe recall tasks, and with WM, measured by the operation span tasks. Engle, Tuholski, Laughlin, and Conway (1999) tested 133 college students on 11 tasks, some of which were thought to reflect STM and some WM. Factor analysis indicated that an operation span task, a reading span task, a counting span task, a keeping track task, a reasoning task, and the secondary memory component of a free recall task loaded on a single factor, which the authors labeled "working memory," whereas a continuous opposites task, forward span tasks, and a backward span task loaded on a second, "short-term memory" factor. Waters and Caplan (1996) performed factor analysis on the results of a study in which 99 college students were tested on a variety of tasks thought to measure verbal and spatial STM and WM. The analysis suggested groupings of tests into factors that corresponded to digit-related tasks (digit span, self-ordered number generation, and externally ordered number generation), spatial tasks (self-ordered design and externally ordered design), sentence processing in span tasks, and recall in sentence span tasks. These results could be interpreted as being consistent with Engle et al.'s description of the factors found in their study, in that the digit tasks may reflect an STM factor, the recall in sentence span tasks a WM factor, the spatial tasks a visual-spatial STM, and the sentence processing component of the sentence span tasks a separate language factor. However, there are empirical differences between the loadings of tests on factors in the two studies. For example, the random number generation task loaded on the same factor as the digit span task in Waters and Caplan's (1996) study but did not load on any factor in Engle et al.'s study.

Effects of Age on the Relationship Between WM Tasks

The data reviewed above on the relationship between different WM measures have been obtained in young subjects. It is generally assumed that there are at least moderate declines in WM capacity with age. Consistent with this view, several studies have shown differences between younger and older subjects on measures of WM capacity (e.g., Stine & Wingfield, 1987). However, others have failed to show such a difference (e.g., Hartley, 1986). One possible reason for the discrepant results concerns differences in the reliability and nature of the WM tasks used. However, there has been little research on the psychometric properties of, and on the relationship between, different WM measures in the elderly. This information is particularly important to have in older subjects, given the widespread use of the concept of WM in the cognitive aging literature and given that variability in performance tends to be greater in older than in younger subjects.

Light and Anderson (1985) found correlations of .27 and .33 in two studies in which the relationship between backward digit span and reading span in elderly subjects ranging in age from 56 to 80 years was investigated. Dobbs and Rule (1989) found a median correlation of .14 between several different WM measures in a group of subjects ranging in age from 30 to 99 years. Salthouse (1988b) found average correlations among backward digit span, missing item span, subtract 2 span, and computation span of about .40 in a group of elderly subjects. However, very few subjects were tested in these studies.

Park and her colleagues (Park et al., 2002; Park et al., 1996) and Hultsch and his colleagues (Hultsch, Hertzog, & Dixon, 1990; Hultsch, Hertzog, Small, & Dixon, 1999; Hultsch, Hertzog, Small, McDonald-Miszczak, & Dixon, 1992) have provided information about the relationship between WM measures in much larger samples of elderly individuals as a part of their longitudinal studies of memory and aging. Park and colleagues measured

WM with three tasks: backward digit span task, a reading span task, and a computation span task. They found that correlations between the tasks ranged from .42 to .63 in one study (Park et al., 1996) and from .46 to .62 in another (Park et al., 2002). WM was initially measured in Hultsch's longitudinal study (Hultsch et al., 1992) by two tasks: a nonverbal number tracking task, in which subjects saw a series of digits and had to hold in memory the last, second to last, or third to last digit, and a sentence construction task, in which subjects read a series of sentences, each of which had one underlined word, and at the end of a set of sentences were asked to recall the sentence that was formed by the underlined words. Confirmatory factor analyses suggested that the two measures form a WM factor, although the loading for the nonverbal measure was low. In subsequent years of the study, he added Salthouse's reading span and computation span measures (Salthouse & Babcock, 1991) to the battery, and confirmatory factor analyses once again showed that the four measures formed a WM factor.

There has also been very little research in which performance has been compared across various WM tasks as a function of age. An interesting question is whether the magnitude of the age difference is task dependent. One might expect there to be across-task differences in the magnitude of the age effect, since different WM tasks appear to differ in the type and amount of processing required while information is being stored (Salthouse, 1990). However, in two studies of the operation span task, Salthouse and Babcock (1991) did not find a difference in the magnitude of the age difference when subjects simply had to remember target digits, as compared with when they had to solve arithmetic problems while remembering the digits. Light and Anderson (1985) and Gick, Craik, and Morris (1988) found larger age differences in a word span task in which the words were presented in isolation than in a task in which the words to be recalled were presented in the context of a sentence span task. On the other hand, Wingfield, Stine, Lahar, and Aberdeen (1988) found larger age differences in the recall of words presented in the context of a sentence span task than of words presented in isolation. Together, these data do not provide strong support for the idea that age-related differences in WM are dependent on the amount or type of processing involved in the processing task. Rather, they seem to argue against the notion that differences in processing efficiency can account for age differences in WM capacity. However, very few tasks have been studied.

In summary, in several studies, the issue of whether the many tasks used in the literature to measure WM correlate with one another has been investigated. The results of several large studies suggest that they do not correlate very well and that more than one factor contributes to performance on these tasks. However, there are discrepancies across studies and unresolved questions about the nature of the factors that have been identified in some studies. In very few studies have researchers looked at how WM tasks correlate in the aging population. This issue is important from a practical point of view and also bears on the question of whether WM tests reflect a single construct.

The Present Study

The goal of the present study was to investigate several frequently used measures of WM capacity in the elderly population. Seven WM tasks were chosen for study. These tasks were backward digit span (Botwinick & Storandt, 1974; Hayslip & Kennelly, 1982; Hooper, Hooper, & Colbert, 1984), running item span (Parkinson, 1980; Talland, 1968), missing item span (Fozard, Nuttall, & Waugh, 1972; Salthouse, 1988b; Talland, 1965), subtract 2 span (Salthouse, 1988b), alphabet span (Craik, 1986), and sentence span for two different sets of materials that differed in terms of complexity (Gick et al., 1988; Hartley, 1986; Light & Anderson, 1985). These tasks were chosen for several reasons. All have been shown to be sensitive to the effects of aging on performance. Some tasks were chosen for their widespread use in the cognitive aging literature as a measure of WM capacity—for example, backward digit span and sentence span. The tasks were also chosen to represent a range of types of verbal stimulus materials (i.e., numbers, words, or sentences), a range of different processing demands in the processing component of the task (e.g., arrange items in a new order, make a judgment about the acceptability of a sentence, or perform an operation on each item and then recall the items), and a range of difficulty in terms of the processing operation required (e.g., acceptability of syntactically simple vs. complex sentences).

We sought to characterize several basic psychometric properties (internal consistency, test–retest reliability, and classificatory stability) of these tests, their correlation across tests, and their factor structure in this population. In addition to the practical significance of this work for the area of cognitive aging, these studies provide information about whether different WM tests measure a single cognitive function and allow us to begin to characterize the function(s) they measure.

METHOD

Subjects

A total of 139 individuals, 64 males and 75 females, were recruited through advertisements posted in the university, churches, and synagogues and newsletters to seniors and were paid for their participation. They were divided into five age groups: 18–30 years (*n* = 27), 50–59 years (*n* = 29), 60–69 years (*n* = 28), 70–79 years (*n* = 27), and 80+ years (*n* = 28). All the participants were required to have English as their mother tongue and at least a high school education. Elderly subjects were required to report that they were aging normally and living independently.

Procedure and Materials

All the subjects were pretested on a battery of neuropsychological tests, to rule out any evidence of cognitive decline or dementia. These background measures included the Mini-Mental State Exam (MMSE; for subjects 50 years and older; Folstein, Folstein, & McHugh, 1975), the Logical Memory I and II subtests of the Wechsler Memory Scale–Revised (WMS–R; Wechsler, 1987), the vocabulary subtest of the Wechsler Adult Intelligence Scale–Revised (WAIS–R; Wechsler, 1981), the reading vocabulary subtest of the Nelson–Denny Reading Test Form A (Nelson & Denny, 1960), and the Boston Naming Test (BNT; Goodglass & Kaplan, 1972). Owing to time limitations, one 50-year-old subject was not tested on the BNT or on the Nelson–Denny vocabulary test. In addition, two 60-year-old and one 80-year-old subjects were not tested on the Nelson–Denny vocabulary test.

The subjects were tested individually in four sessions of approximately 60–90 min each in Phase I of the experiment and then were retested approximately 1.5 months later in two sessions in Phase II of the experiment. In Phase I, the subjects were tested on the WM measures outlined below, as well as on measures of language processing efficiency as a part of another study. In Phase II, the subjects were retested on only the WM measures.

Five subjects were unavailable for testing in Phase II of the experiment. Of these, 1 was in the 18- to 30-year-old group, 2 were in the 50- to 59-year-old group, and 2 were in the 80+ group. In addition, in Phase II, one 50- to 60-year-old subject was not tested on running item span, one 70- to 80-year-old subject was not tested on WM span for simple sentences, and two 70- to 80-year-old and two 80+ year old subjects were not tested on the alphabet span task, owing to time constraints. Finally, one subject who was 80+ years old was tested only to span in Phase II on the backward digit span and subtract 2 span tasks and so did not contribute to the analysis of the data by items.

WM Measures

WM capacity was tested using the seven WM span tasks described below. The order of presentation of the tasks was randomized across subjects. For all but the sentence span tasks, items were presented at the rate of one per second. Testing began at span size 2 and continued through span size 8 for all but the running item and sentence span tasks. Owing to time limitations, for the sentence span tasks, testing began with span size 2 and was discontinued at the span size at which the subject could no longer recall the sentence-final words in the correct serial order on two out of five trials. As will be outlined below, list lengths 2 to 8 were randomized for the running item task.

For all the tasks, there were five trials at each span size. The subjects were required to repeat all of the items in a trial in the correct serial order to obtain credit for the trial. They were instructed to indicate, by saying "blank," if they knew that an item had been presented in a particular serial position but could not remember what the item was. Span was defined as the longest list length for which the subjects correctly recalled all of the items in the correct serial order on three out of five trials. An additional 0.5 was added if the subjects were correct on two out of five trials at the next span size. The total percentage of items recalled correctly across the seven span sizes was also calculated for all but the sentence span and the running item tasks. The data from the sentence span tasks could not be analyzed in this way, since not all the subjects were tested at all the span sizes. Previous work by Waters and Caplan (1996) with college students, using exactly the same materials as those used here, had shown that there was a very high correlation between span and item scoring for this task ($r = .93$). The data from the missing item span task were not amenable to a separate analysis at the level of percentage of items.

Alphabet span. This task required that the subjects repeat a series of words after rearranging them in alphabetical order. The stimuli consisted of monosyllabic words of moderate frequency. The words presented on each trial were semantically and phonologically dissimilar, and no words were repeated within a trial.

Backward digit span. In this task, on each trial, the subjects were required to repeat a series of digits in reverse order of presentation. The stimuli for this task, as well as for the missing digit and subtract 2 tasks outlined below, were digits drawn from the digits 1 to 9 and presented randomly.

Missing digit span. The subjects were read a string of digits. The experimenter then reread the string in a different random order, with one item omitted. The subject was required to report the missing item.

Subtract 2 span. The subjects were required to repeat a random sequence of digits after subtracting 2 from each.

Running item span. In this task, the subjects were required to recall the final items in a list of an unknown length. On each trial, the subjects were presented with a list that was 9, 11, 14, 15, or 17 digits long. They were asked to recall the last 2, 3, 4, 5, 6, 7, or 8 digits presented on each of five trials. The list lengths presented and the number of digits the subjects were asked to recall were randomized across trials.

Sentence span. Sentence span was measured using a modified version of Daneman and Carpenter's (1980) task. The methods and the materials were taken from Waters, Caplan, and Hildebrandt (1987, Experiment 2A). The subjects were presented with a series of sentences on the video screen of a computer. They were required to read each sentence silently and to make a judgment about its acceptability by pushing the right response key if the sentence was acceptable and the left if it was unacceptable. They were instructed to perform the sentence task very accurately and then to perform as well as they could on the recall task. The subjects were tested on this task with two different sets of stimulus materials that varied in syntactic complexity. The syntactically simple sentences consisted of sentences of the cleft subject form (e.g., *It was the gangsters that broke into the warehouse*) and the complex sentences consisted of subject–object relatives (e.g., *The meat that the butcher cut delighted the customer*). Half of the sentences of each type were acceptable, and half were unacceptable. The stimuli were constructed so that recognition of the acceptability of a sentence required a syntactic analysis. The stimuli were blocked by sentence type and were divided into five sets of sentences at each of the span sizes two, three, four, five, and six.

RESULTS

Background Characteristics

Table 3 shows the characteristics of the subjects in the five age groups. One-way analyses of variance (ANOVAs), followed by Tukey post hoc tests, were carried out for all of the background measures other than the WMS–R to determine whether there were any significant differences across the age groups. Since the 18- to 30-year-old subjects were not tested on the MMSE, the analysis for this task compared performance across the remaining four age groups. The data from the Logical Memory I (immediate) and II (delayed) subtests of the WMS–R were analyzed in a 5 (age group) × 2 (immediate vs. delayed) ANOVA.

The groups did not differ statistically in terms of number of years of education [$F(4,134) = 2.4$, $MS_e = 7.7$, n.s.]. On the WMS–R, there was no effect of age [$F(4,134) = 2.2$, $MS_e = 1,176.9$, n.s.] or delay [$F(1,134) = 1.2$, $MS_e = 122.9$, n.s.] and no interaction between these factors [$F(4,134) = 2.0$, $MS_e = 122.9$, n.s.]. In addition, there were no significant differences across the age groups on the Nelson–Denny vocabulary subtest [$F(4,130) = 1.6$, $MS_e = 452.5$, n.s.].

There were significant effects of age on the MMSE [$F(3,108) = 7.79$, $MS_e = 2.03$, $p < .001$], the vocabulary subtest of the WAIS–R [$F(4,134) = 3.08$, $MS_e = 8.3$, $p < .05$], and the BNT [$F(4,133) = 5.19$, $MS_e = 15.3$, $p < .001$]. Post hoc tests showed that on the MMSE, the scores of

Table 3
Subject Characteristics

	Age Group (Years)									
	18–30 (*n* = 27)		50–59 (*n* = 29)		60–69 (*n* = 28)		70–79 (*n* = 27)		80+ (*n* = 28)	
Characteristic	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Age*	20.9	2.4	53.9	2.6	64.7	2.9	74.5	2.8	83.4	3.7
Education*	14.1	1.4	15.0	2.7	14.8	2.8	14.2	2.9	14.2	3.4
Mini-Mental State Exam*	–		28.8	0.9	29.1	1.2	28.1	1.7	27.4	1.8
Wechsler Memory Scale–R										
Logical Memory I*	29.7	5.1	25.1	5.9	27.4	5.5	28.2	5.9	23.4	5.6
Logical Memory II*	25.9	6.5	20.0	6.6	24.2	6.3	22.3	6.1	19.0	5.9
WAIS Vocabulary†	13.6	2.2	13.9	2.1	13.0	3.9	15.1	2.9	15.2	2.9
Nelson–Denny Vocabulary‡	73.9	22.4	84.0	21.4	85.0	21.4	87.5	16.7	85.3	19.5
Boston Naming Test* (/60)	54.2	3.5	56.4	2.9	54.4	3.9	55.3	3.6	51.9	5.3

*Mean scores. †Mean standard scores. ‡Mean percentiles.

the two younger age groups (50–59 and 60–69 years) were significantly higher than those of the oldest age group (80+ years). On the vocabulary subtest of the WAIS–R, the scores of the oldest group (80+ years) were significantly *higher* than those of the 60- to 69-year-old subjects. In contrast, on the BNT, the scores of the oldest group (80+ years) were significantly *lower* than those of the 50- to 59- and the 70- to 79-year-old subjects. Thus, there were a few minor differences across the age groups, some of which favored younger subjects (MMSE and BNT) and some of which favored older subjects (WAIS–R vocabulary subtest). It was not the case that, overall, the older subjects were more cognitively impaired than the younger subjects, and none of the older subjects met the criteria for dementia on the basis of their background test scores.

Effects of Age and Phase on WM Scores

Span Measures

Table 4 shows the mean span scores for the five age groups on the seven WM span tasks when tested in Phase I and Phase II. The data for each task were analyzed in a repeated measures ANOVA with independent age groups and phase as the repeated measures. In addition, we performed a linear contrast on the age group factor in order to assess effects associated with the ordered nature of its categories.

There were significant effects of age for five of the seven tasks: alphabet span [$F(4,125) = 10.6$, $MS_e = 1.1$, $p < .001$], backward digit span [$F(4,129) = 4.3$, $MS_e = 2.7$, $p < .01$], subtract 2 span [$F(4,129) = 3.2$, $MS_e = 2.1$, $p < .01$], simple sentence span [$F(4,128) = 11.3$, $MS_e = 2.2$, $p < .001$], and complex sentence span [$F(4,129) = 3.5$, $MS_e = 2.9$, $p < .01$]. The effect of age was nonsignificant for the missing digit span task [$F(4,129) = 2.3$, $MS_e = 0.50$, n.s.] and for the running item span task [$F(4,128) = 2.0$, $MS_e = 2.1$, n.s.]. There was a significant effect of phase in the analysis of the subtract 2 [$F(1,129) = 4.5$, $MS_e = 0.45$, $p < .05$], running item [$F(1,128) = 5.4$, $MS_e = 0.53$, $p < .05$], and simple sentence span [$F(1,128) = 4.5$, $MS_e = 0.47$, $p < .05$] tasks. In all cases, span scores were higher in Phase II than in Phase I. In all of the

Table 4
Working Memory Span Scores in Phase I and Phase II

	Age Group (Years)									
	18–30		50–59		60–69		70–79		80+	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Alphabet I	4.7	0.76	4.3	0.87	4.1	0.79	3.8	0.68	3.6	1.00
Alphabet II	4.9	0.80	4.3	0.75	4.4	0.82	3.9	0.75	3.5	0.97
Backward digit I	6.2	1.20	6.2	1.50	6.0	1.40	5.3	1.30	5.6	1.40
Backward digit II	6.7	1.10	6.4	1.20	5.8	1.40	5.5	1.20	5.5	1.30
Missing digit I	6.6	0.53	6.6	0.56	6.8	0.44	6.6	0.55	6.3	0.72
Missing digit II	6.7	0.64	6.5	0.49	6.6	0.80	6.5	0.81	6.3	0.80
Subtract 2 I	5.7	1.10	5.5	1.10	5.2	1.10	5.0	1.20	5.0	0.96
Subtract 2 II	6.2	0.95	5.5	1.40	5.2	1.20	5.2	1.20	5.2	1.10
Running item I	4.4	1.10	4.1	1.10	3.8	1.10	3.9	1.00	3.5	1.40
Running item II	4.4	0.97	4.4	1.40	4.0	1.10	3.9	1.10	3.8	1.10
Sentence (simple) I	4.5	0.86	3.4	1.30	3.0	1.20	2.9	1.30	2.6	1.30
Sentence (simple) II	4.5	0.88	3.7	1.20	3.3	0.98	2.8	1.10	2.9	1.40
Sentence (complex) I	2.9	1.30	2.8	1.60	2.3	1.20	2.2	1.30	2.2	1.00
Sentence (complex) II	3.4	1.50	2.9	1.40	2.4	1.20	2.3	1.20	2.2	1.20

analyses, the age × phase interaction was not significant. Linear contrasts on the effect of age were significant for all seven of the tasks [alphabet span, $F(1,125) = 40.9$, $p < .001$; backward digit span, $F(1,129) = 15.0$, $p < .001$; missing digit span, $F(1,129) = 5.1$, $p < .05$; subtract 2 span, $F(1,129) = 11.1$, $p < .01$; running item span, $F(1,128) = 7.7$, $p < .01$; simple sentence span, $F(1,128) = 39.6$, $p < .001$; complex sentence span, $F(1,129) = 12.0$, $p < .001$]. These results suggest that there are effects of age on the WM span tasks, with span scores decreasing across the five age groups tested. Consistent with this, Pearson product–moment correlation coefficients showed that age was significantly negatively correlated with all of the span measures other than the missing digit task (alphabet = −.47, backward digit span = −.25, missing digit span = −.11, subtract 2 span = −.28, running item span = −.21, simple sentence span = −.47, and complex sentence span = −.25).

In order to facilitate comparison across tasks, the scores of the older adults for each measure were converted into units of young standard deviations by subtracting the mean performance of the young group from the mean of the old group and then dividing the difference by the standard deviation of the young scores, as suggested by Salthouse (1990). The results for Phase I and Phase II are shown in Table 5. These results suggest that the age difference was much larger for some tasks than for others. In particular, in Phase I, the age difference was largest on the alphabet span task and on the sentence span task for syntactically simple sentences. In Phase II, large differences were seen for these tasks, as well as for the backward digit span task, the subtract 2 span task, and the sentence span task for syntactically complex sentences. These scores tended to be much larger for the 70- and 80-year-old subjects than for the 50- and 60-year-old subjects, showing that, as compared with the young subjects, the decrement in performance was much larger for these groups.

Table 5
Units of Young Standard Deviations

		Age Group (Years)			
		50–59	60–69	70–79	80+
Measure	Phase	(n = 29)	(n = 28)	(n = 27)	(n = 28)
Alphabet	I	−0.576	−0.878	−1.344	−1.583
	II	−0.687	−0.612	−1.225	−1.697
Backward digit	I	−0.086	−0.300	−0.817	−0.563
	II	−0.282	−0.792	−1.125	−1.121
Missing digit	I	−0.083	0.226	−0.106	−0.583
	II	−0.209	−0.129	−0.181	−0.593
Subtract 2	I	−0.141	−0.477	−0.654	−0.655
	II	−0.613	−0.975	−1.024	−0.976
Running item	I	−0.380	−0.499	−0.418	−0.717
	II	−0.020	−0.401	−0.496	−0.597
Sentence (simple)	I	−1.033	−1.612	−1.699	−1.931
	II	−0.967	−1.359	−1.891	−1.810
Sentence (complex)	I	−0.130	−0.563	−0.629	−0.603
	II	−0.367	−0.707	−0.811	−0.844

Item Measures

Table 6 shows the data from the four tasks that were scored for percentage of items correctly recalled. The data for each task were analyzed in a repeated measures ANOVA with independent age groups and phase as the repeated measures. In addition, we performed a linear contrast on the age group factor in order to assess effects associated with the ordered nature of its categories.

Once again, there was a significant effect of age group on all four tasks: alphabet span [$F(4,125) = 8.8$, $MS_e = 0.020$, $p < .001$], backward digit span [$F(4,128) = 4.2$, $MS_e = 0.029$, $p < .001$], subtract 2 span [$F(4,128) = 3.5$, $MS_e = 0.024$, $p < .001$], and running item span [$F(4,128) = 3.1$, $MS_e = 0.024$, $p < .05$]. There was an effect of phase on the subtract 2 span task [$F(1,128) = 6.7$, $MS_e = 0.002$, $p < .01$] and on the running item span task [$F(1,128) = 11.2$, $MS_e = 0.002$, $p < .01$]. For both tasks, scores were higher in Phase II than in Phase I. The interaction between phase and age was not significant on any of the four tasks.

Linear contrasts on the effect of age were significant for all four tasks (alphabet span, $F = 29.6$, $p < .001$; backward digit span, $F = 15.2$, $p < .001$; subtract 2 span, $F = 12.3$, $p < .001$; running item span, $F = 12.0$, $p < .001$). Once again, these results suggest that there are effects of age on the WM span tasks, with item scores decreasing across the five age groups tested.

Relationship Between Span and Item Measures

The correlation between the span and the item measures for each of the four tasks for which item scores were collected were calculated. The correlations for the group as a whole at Phase I and Phase II, respectively, were .72 and .70 for alphabet span, .84 and .81 for backward digit span, .80 and .87 for subtract 2 span, and .73 and .76 for running item span. There were no systematic differences across the age groups in the magnitude of these correlations.

Relationship Among the WM Tasks

Given that the pattern of performance was virtually identical across the two phases of the study, the data from Phase I and Phase II were averaged for each subject for each task in order to investigate the relationship among the WM tasks. The values below the diagonal in Table 7 show the correlations among the seven span measures. As can be seen in the table, there were moderate significant correlations between all of the measures, other than the missing digit measure. Correlations between the missing digit and other measures were much smaller and, in some cases, nonsignificant.

These correlations could have been due to the fact that the span measures were all correlated with age, since as was noted above, there were significant correlations between age and all of the span scores other than missing digit. We, therefore, also calculated correlations among the tasks with the effect of age partialled out. Values above

Table 6
Working Memory Scores in Phase I and Phase II (% Items)

	Age Group (Years)									
	18–30		50–59		60–69		70–79		80+	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Alphabet I	60.5	9.8	53.9	11.7	51.5	10.2	46.8	8.9	48.5	15.0
Alphabet II	64.4	8.7	53.7	11.6	54.3	10.5	48.4	10.6	47.6	13.1
Backward I	80.2	12.5	78.9	13.9	75.6	14.0	70.7	13.5	73.0	15.2
Backward II	83.7	8.9	82.8	10.7	75.8	13.9	73.4	12.8	70.2	15.9
Subtract 2 I	78.0	9.8	76.2	11.3	72.3	11.7	70.0	14.2	71.9	11.7
Subtract 2 II	81.7	8.4	78.7	12.3	73.8	12.2	71.6	11.0	70.7	11.6
Running item I	62.5	11.0	60.7	10.3	58.1	8.7	56.0	10.2	52.6	13.9
Running item II	63.2	10.8	63.0	11.2	60.4	11.7	58.1	10.0	55.5	11.7

the diagonal in Table 7 are Pearson product–moment correlation coefficients with the effect of age partialled out. As can be seen in the table, there were significant correlations between all of the tasks other than the missing digit task, even when the effect of age was partialled out.

Table 8 shows the squared multiple *R*s (on the diagonal) and the partial correlations between the WM span measures for the combined Phase I and Phase II measures. The partial correlation between any two variables provides information about the variation that is common to the two variables but is not common to any other variables in the matrix. The squared multiple *R* for a variable represents the proportion of variance for that variable that is common with all other variables in the matrix. There are two important features of this analysis. First, only 4 out of 21 correlations remained significant when the variance associated with the other variables was partialed out. Second, the missing digit span task was the only task that did not share a significant amount of variance with the other variables.

Before using confirmatory factor analysis (CFA) to test the hypothesis that these WM measures all test a single construct, an exploratory factor analysis (EFA) was carried out to assess how well the missing digit task associated with the other WM measures. For the purpose of this analysis, as well as for the CFA described below, the two sentence span measures were averaged for each subject, since these measures were essentially the same, with the only difference being the syntactic complexity of the sentences used as stimuli. Because of its obvious lack of association with other measures seen in the EFA, the missing digit task was dropped from all further analyses.

Table 7
Correlations Among Span Measures

Measure	Alph	Bckwd Digit	Miss Digit	Subtr 2	Run Item	Sent Simp	Sent Comp
Alph	**1**	.47*	.23*	.42*	.45*	.43*	.32*
Bckwd Digit	.55*	**1**	.22*	.71*	.58*	.61*	.51*
Miss Digit	.27*	.26*	**1**	.26*	.14	.19*	.14
Subtr 2	.51*	.74*	.29*	**1**	.33*	.54*	.46*
Run Item	.49*	.61*	.17	.56*	**1**	.49*	.38*
Sent Simp	.57*	.67*	.23*	.61*	.52*	**1**	.75*
Sent Comp	.43*	.57*	.18	.52*	.42*	.77*	**1**

Note—Correlations are based on the average of the Phase I and Phase II data. Values above the diagonal are correlations with the effect of age partialled out. Alph, alphabet span; Bckwd Digit, backward digit span; Miss Digit, missing digit span; Subtr 2, subtract 2 span; Run Item, running item span; Sent Simp, syntactically simple sentences; Sent Comp, syntactically complex sentences. *$p < .05$.

CFA was carried out in EQS (Bentler, 1989) in order to investigate the question of whether the five WM span measures (the alphabet, the subtract 2, the running item, and the two sentence span measures) test a single construct. To provide a clear test of whether the five measures tap a single or multiple constructs, separate one-factor and two-factor CFAs were carried out on the five span measures. The two-factor CFA resulted in eigenvalues of 3.3 for the first factor and 0.54 for the second factor. Inspection of the scree plot indicated that the data were best fit by a one-factor solution. Table 9 shows the results of the one-factor CFA. This one-factor solution accounted for about 66% of the variance. These results support the hypothesis that these five tasks measure a common construct.

Test–Retest Reliability

Reliability of the WM span and item scores over time was assessed by correlating the scores from Phase I and Phase II. The top of Table 10 shows the data for the span measures for the group as a whole and for each of the five age groups. The Pearson product–moment correlation coefficients were significant for all of the tasks, other than the missing digit span task, and were in the moderate-to-high range. However, all of these correlations were lower than is desirable for reliability coefficients, which usually fall in the range of .80 to .90 (Anastasi, 1982), and only the sentence span measures met the .70 criterion that some authors have argued is the minimum reliability adequacy (Nunnally, 1978).[1] In addition, although the correlations appear to be higher for some age groups on some tests, it is not the case that test–retest reliability varied systematically across the age groups.

In order to determine whether test–retest reliability would have been better if the subjects had been tested on several measures and the average of those measures had

Table 8
Partial Correlations Among Span Measures

		Bckwd	Miss	Subtr	Run	Sent	Sent
Measure	Alph	Digit	Digit	2	Item	Simp	Comp
Alph	**.42***						
Bckwd Digit	.12	**.65***					
Miss Digit	.13	.03	**.11**				
Subtr 2	.07	.45*	.13	**.59***			
Run Item	.17	.24*	−.05	.17	**.43***		
Sent Simp	.26*	.19	.02	.09	.08	**.71***	
Sent Comp	−.06	.07	−.02	.06	−.01	.62*	**.60***

Note—Correlations are based on the average of the Phase I and Phase II data. Alph, alphabet span; Bckwd Digit, backward digit span; Miss Digit, missing digit span; Subtr 2, subtract 2 span; Run Item, running item span; Sent Simp, syntactically simple sentences; Sent Comp, syntactically complex sentences. **p* < .05.

been taken as an estimate of WM, we examined test–retest reliability for various composite WM scores. We first examined test–retest reliability for the total group of subjects for a composite score based on the alphabet span and the sentence span (for syntactically complex sentences) measures. The rationale for the inclusion of these two measures was that the two tasks were different and, individually, they had the highest test–retest reliability of any of the measures. Test–retest reliability for these two measures was .82. When we added the subtract 2 span measure, the measure with the next highest test–retest reliability, test–retest reliability was .85. Finally, a composite measure based on the average of all of the tasks, other than the missing digit task, also resulted in a test–retest reliability of .85. Thus, test–retest reliability reaches a more acceptable level when a composite measure based on at least two tasks is used.

The bottom of Table 10 shows the correlation between Phase I and Phase II when the percentage of items correctly recalled measure is used rather than the span measure. As can be seen by comparing the top and the bottom of Table 10, test–retest reliability was somewhat better when item, rather than span, scores were used. Once again, although the correlations appear to be higher for some age groups on some tests, it is not the case that test–retest reliability varied systematically across the age groups.

Internal Consistency

To estimate internal consistency, we calculated the alpha coefficients for each administration of all of the WM span tasks. Alpha coefficients are based on the average correlation of the scores on any size operation set with the total score. The results of the analysis are shown in Table 11. Correlations for the span two set size with the total score were low, due to the high recall accuracy for this set size. The adjusted alpha was, therefore, calculated for set sizes of three to eight, omitting the span two set size. Both the unadjusted and the adjusted values indicate that all of the measures have adequate internal consistency.

Table 9
Factor Loadings From the One-Factor Confirmatory Factor Analysis

Measure	Initial Unrotated
Alphabet span	.747
Backward digit span	.882
Subtract 2 span	.845
Running item span	.775
Sentence span	.810
% of variance	66%

Note—Analysis is based on the average of the Phase I and Phase II data.

Stability of Subject Classification

We examined the stability of subjects' classification into WM span groups across Phase I and Phase II of the experiment and across the seven WM span measures within a phase when different methods were used to classify the subjects. Given that test–retest reliability did not differ substantially across the age groups, for the purpose of this analysis, the subjects were not divided into separate age groups.

Subject Classification Across Time

High-, medium-, and low-span groups. To assess stability of subject classification across time, we classified the subjects into high-, medium-, and low-span groups at each phase and for each task by ranking their scores for each task and assigning them to a group based on the third of the population they fell into, with the proviso that the same span score could not be assigned to more than one group. When a score crossed the boundary between two groups, the cutoff was chosen to be the score that allowed the entire population of subjects to be most closely divided into thirds. Using this method, 35% of the subjects changed their classification from Phase I to Phase II for the alphabet span task, 36% for the backward digit span task, 61% for the missing digit span task, 42% for the subtract 2 span task, 43% for the running item span task, 47% for the simple sentence span task, and 41% for the complex sentence span task. Chi-square analyses showed that a significant number of the subjects changed classifications from Phase I to Phase II for all the tasks (alphabet span, χ^2 = 63.2; backward digit span, χ^2 = 66.6; missing digit span, χ^2 = 13.0; subtract 2 span, χ^2 = 48.9; running item span, χ^2 = 47.1; simple sentence span, χ^2 = 59.2; and complex sentence span, χ^2 = 59.2). Thus, subject classification into WM span groups was not stable when this method was used.

We also looked at the stability of subject classification when a composite measure that combined performance across two, three, or six tasks (as outlined above) was used. When a composite measure based on two tasks (alphabet span and sentence span for complex sentences) was used, 29% of the subjects changed classification across the two phases of the experiment. Using the composite based on three tasks (alphabet, subtract 2, and sentence span), 16% of the subjects changed classification, and when a composite based on all the tasks other than missing digit was used, 25% of the subjects changed clas-

Table 10
Test–Retest Reliability: Correlations Between Phase I and Phase II Working Memory Measures

		Age Group (Years)				
Measure	All Subjects	18–30	50–59	60–69	70–79	80+
			Span			
Alphabet span	.68*	.57*	.52*	.62*	.53*	.71*
Backward digit span	.65*	.55*	.63*	.67*	.60*	.61*
Missing digit span	.22	.06	−.11	.13	.36	.33
Subtract 2 span	.67*	.51*	.64*	.85*	.64*	.59*
Running item span	.61*	.72*	.58*	.63*	.44*	.64*
Sentence (simple)	.73*	.45*	.84*	.57*	.73*	.71*
Sentence (complex)	.76*	.81*	.79*	.71*	.62*	.62*
			Item			
Alphabet span	.74*	.67*	.88*	.86*	.59*	.59*
Backward digit span	.71*	.72*	.83*	.71*	.88*	.45*
Subtract 2 span	.84*	.64*	.91*	.92*	.77*	.73*
Running item span	.79*	.82*	.77*	.73*	.82*	.83*

*$p < .05$.

sification across the two phases of the experiment. Thus, a composite measure results in a somewhat more stable classification than do some of the individual measures.

Upper and lower quartiles. In order to examine the stability of subject classification when the subjects were divided into extreme groups, for each task, subjects who fell into the upper quartile of scores were classified as high span, and those who fell into the lower quartile as low span, again with the proviso that a single span score could not appear in more than one quartile. Of the subjects who fell in the lower quartile in Phase I, 48% of the subjects on the alphabet span task, 43% on the backward digit span task, 66% on the missing digit span task, 36% on the subtract 2 span task, 56% on the running item span task, 45% on the reading span task for simple sentences, and 48% on the reading span task for complex sentences changed classification to some other quartile at Phase II testing. The stability of subject classification was somewhat better for the subjects who fell in the high-span group. Of the subjects who fell in the upper quartile in Phase I, 38% of the subjects on the alphabet span task, 36% of the subjects on the backward digit span task, 56% of the subjects on the missing item span task, 38% of the subjects on the subtract 2 span task, 28% of the subjects on the running item span task, 24% of the subjects on the reading span task for simple sentences, and 22% of the subjects on the reading span task for complex sentences changed classification at Phase II.

We also looked at the stability of classification into the lower and the upper quartiles when various composite measures were used. The percentage of the subjects who changed classification in the lower and the upper quartiles, respectively, were 32% and 19% when the composite based on two tasks was used, 34% and 22% when the composite based on three tasks was used, and 24% and 1% when the composite based on six tasks was used. Thus, a composite measure made up of performance on several WM tasks results in a more stable classification of subjects into groups, particularly for those in the high-span or upper quartile group.

Absolute cutoff scores. Finally, since much of the research using the sentence span task divides subjects into groups on the basis of absolute cutoff scores, we examined the stability of the sentence span scores when this method was used to classify the subjects. We divided the subjects into span groups for the two sentence span tasks, using criteria that have commonly been adopted in the literature (cutoff of 2.5 or less for low span, 3.0 and 3.5 for medium span, and 4.0 and above for high span; e.g., MacDonald et al., 1992). By using these criteria, 40.6% of the subjects were classified as high-span subjects, 18.8% as medium-span subjects, and 40.6% as low-span subjects for the reading span task with simple sentences at Phase I. However, in Phase II, 32% of these subjects changed classification on the basis of their performance on the simple sentences. Of these, the performance of 37% declined, and the performance of 63% increased, in Phase II. Moreover, it was not always the case that the subjects shifted by one category. Twenty-five percent of the subjects who shifted did so by more than one category between Phase I and Phase II. On the basis of performance on the complex sentences, 67.9% of the subjects were classified as low span, 11.9% as medium span, and 20.2% as high span. For these sentences, 25%

Table 11
Item–Total Correlations and Cronbach's Alpha

Span Size	Alph I	Alph II	Bckwd Dig I	Bckwd Dig II	Subtr 2 I	Subtr 2 II	Run Item I	Run Item II
Two	.29	.33	.05	–*	.18	.21	.44	.17
Three	.31	.43	.30	.46	.14	.14	.59	.57
Four	.59	.63	.52	.49	.50	.46	.66	.58
Five	.71	.75	.61	.64	.66	.71	.69	.69
Six	.77	.79	.76	.79	.76	.71	.63	.62
Seven	.72	.73	.75	.78	.76	.82	.61	.60
Eight	.62	.66	.75	.74	.71	.76	.61	.53
Alpha	.824	.844	.813	.825	.800	.804	.842	.796
Adjusted alpha†	.837	.856	.835	.849	.817	.821	.842	.814

Note—Alph, alphabet span; Bckwd Dig, backward digit span; Subtr 2, subtract 2 span; Run Item, running item span; I, Phase I data; II, Phase II data. *All the subjects achieved a perfect score. †Span Size 2 omitted.

of the subjects changed classification in Phase II. Similar numbers of subjects improved (54.5%) and declined (45%). It was not the case that the subjects simply shifted by one category. Twenty-one percent of the subjects who shifted did so by more than one category between Phase I and Phase II. These data illustrate that when absolute cutoff scores are used, group membership is not very stable over time and is heavily influenced by the difficulty of the stimulus materials.

Subject Classification Across Tasks

We also investigated the stability of subject classification across the seven WM tasks. We first examined stability of subject classification across the seven tasks when the subjects were divided into high-, medium-, and low-span groups separately for each task. With this method, subject classification across the seven WM span tasks within a phase was not stable. Only 9 out of 139 subjects (6.5%) tested at Phase I were assigned to the same classification group on all seven WM span tasks. Of these, 6 subjects fell in the high-span group on all the tasks, and 3 fell in the low-span group. An additional 19 subjects (13.6%) were classified consistently across six of the seven tasks. Over half of the group (52%) received all three subject group classifications across the seven tasks. A similar pattern was seen at Phase II. Only 14 out of 135 subjects (10.3%) maintained the same classification across the seven tasks. Twelve of these fell in the high-span group, 1 in the medium-span group, and 1 in the low-span group. An additional 10 subjects (7%) were classified consistently across six of the seven tasks. Once again, a large proportion of the group (44%) received all three subject group classifications across the seven tasks.

Subject classification across the seven tasks within a phase was not better when the subjects were divided into extreme groups on the basis of the quartile their scores fell into. Only 1 subject in the lower quartile and 1 in the upper quartile remained in the same quartile across all seven tasks in Phase I. An additional 3 subjects were in the lower quartile on all but one task, and 6 subjects were in the upper quartile on all but one task. In Phase II, 1 subject in the lower quartile and 5 in the upper quartile remained in the same quartile across all seven tasks. An additional 2 subjects were in the lower quartile on all but one task, and 5 subjects were in the upper quartile on all but one task.

DISCUSSION

The results of this study are consistent with many others in the aging literature in showing that older individuals perform more poorly than younger individuals on many WM tasks. Although the effects of age were only moderate, it is important to keep in mind that the elderly subjects in this study were all well educated and cognitively intact and that greater age effects have been found on these tasks in less highly cognitively functioning elderly individuals (Waters & Caplan, 2002). In addition, it should be kept in mind that although the overall sample size was quite large, the sample size in each age group was less than 30, making conclusions about differences across the age groups more tentative than those for the group as a whole.

The effects of age were not significant on the missing digit and running item span tasks. It is not clear why the running item span task did not result in significant age differences, since previous research has shown effects of age on this task (Parkinson, 1980; Talland, 1968). However, many of the subjects indicated that they had used an unforeseen strategy on the missing digit span task and that this strategy actually made the task easier at larger span sizes. On this task, the subjects were read a string of digits. The experimenter then reread the string in a different random order, with one item omitted. The subjects were required to report the missing item. This meant that at the largest span size, they simply had to figure out which of the digits from 1 to 8 the experimenter had not read on the second reading. However, at the lower span sizes, they were required to remember which of the digits from 1 to 8 were originally presented and then were required to figure out which item was missing on the reread. As a result, the scores of all the subjects were near ceiling on this task. This likely accounts for the failure to find an effect of age with this measure.

The major focus of this study was on the basic psychometric properties of the WM tasks we administered—in particular, the relationship between span and item measures, test–retest reliability, internal consistency, and classificatory stability.

The pattern of results using the percentage of items correctly recalled was very similar to the results found with the span measures. Correlations between the span and the item scores ranged from .70 to .87 and were virtually identical in the two phases of the study. These correlations are somewhat lower than those we and others have found in previous studies with college students, in which the correlations between span and item measures have ranged from .91 to .95 (Klein & Fiss, 1999; Waters & Caplan, 1996). Furthermore, in this study, the item scores did not seem to be more sensitive than the span scores, since on some tasks differences between age groups were seen with the span, but not with the item, scores. One possible reason for this discrepancy is that the WM tasks used in Waters and Caplan's (1996) and Klein and Fiss's studies were more difficult (as shown by the average span of subjects) than the WM measures for which item scores were obtained in the present study. Contributions to performance from items in lists greater than span would occur more often in harder tasks in which span is lower, possibly increasing the variability of item measures on harder tests.

The correlations between the Phase I and the Phase II span scores were significant for all tasks other than the missing digit span task but were only in the moderate range. As was noted above, some authors have argued that .7 is the criterion for minimum reliability adequacy (Nunnally, 1978); however, reliability coefficients from .80 to .90 are usually considered desirable (Anastasi, 1982). With these criteria, none of the individual WM measures used in this study has adequate reliability when span scores are used. One possible reason for the poor test–retest reliability in this study is that the interval between testing was quite long for some subjects. However, comparison of the subjects who were tested at shorter and longer intervals suggested that this was not the major factor, since test–retest reliability was not better for the subjects who were tested at shorter intervals (see note 1).

Test–retest reliability was considerably better when performance across several span tasks was averaged to yield a composite span score. Test–retest reliability for each of the individual tasks also improved somewhat when the item, rather than the span, scores were used, suggesting that one advantage of item scores is their greater stability over time.

As was outlined in the introduction, there is little available data concerning the test–retest reliability of the measures used in this study. We had previously examined the test–retest reliability of the sentence span tasks used here in a study of 100 college students (Waters & Caplan, 1996). In that study, we found correlations of .65 and .66 for the simple and complex sentence span tasks. These correlations are very similar to the correlations of .73 and .76 found for the group as a whole in this study and are very similar to the correlation of .73 that Tirre and Peña (1992) found for their version of the sentence span task, in which subjects were also required to make a judgment about a sentence. All of these correlations are much higher than the test–retest reliability of .41, which we reported using span scores, and of .52 that MacDonald et al. (2001) reported, using item scores for Daneman and Carpenter's (1980) version of the task, in which subjects simply read a sentence aloud, rather than making a judgment about the sentence (Waters & Caplan, 1996). The present results, in combination with those outlined above, suggest that test–retest reliability for Daneman and Carpenter type sentence span tasks is better when subjects are required to make judgments about the sentences than when they simply read the sentences aloud.

Assessment of the stability of subject classification suggested that the procedure of dividing subjects into discrete WM span groups on the basis of a single administration of a single WM task is an unreliable way of identifying subjects with different WM capacities. For all tasks, chi-square analyses showed that a significant number of subjects changed classification from Phase I to Phase II when they were divided into high-, medium-, and low-span groups on a given task. Division of subjects into upper and lower quartiles was equally unreliable. The data concerning the reliability of subject classification with the sentence span task were very similar to those in our previous study with college students (Waters & Caplan, 1996). Klein and Fiss (1999) reported greater stability of subject classification across three administrations for Turner and Engle's (1989) operation span test. However, even in Klein and Fiss's study, classification of subjects into discrete memory span groups resulted in much less stable performance than did continuous measures. Klein and Fiss attributed the unsatisfactory nature of this index in their study as being due to the small sample size (*n* = 33). This meant that misclassification of a single individual represented a fairly large error in terms of percentage of the total population studied. However, the sample size was much larger in the present study, and yet classificatory stability was extremely low when the subjects were classified into discrete memory span groups. The second, more likely, factor that Klein and Fiss identified was that the distance between the scores of the high- and the low-span groups was relatively small. This is an inherent problem with span measures that have an extremely restricted range of scores but should be somewhat less of a problem when the total percentage of items recalled is used as the measure, since there is a somewhat larger spread in these scores.

Subject classification across the seven tasks was also extremely unstable, with a very small proportion of the subjects being classified in the same manner across all the tasks. This finding suggests that, if subjects in a study are classified into different WM groups, the effect of WM in a particular study may depend heavily on the task used to classify the subjects into these groups.

The finding that subject classification across the seven WM tasks was unstable touches on the issue of whether all of the tasks used here measured the same construct. Correlational analyses showed that there were significant correlations between performance on all of the span tasks, other than the missing digit task. The magnitude of the correlations found in this study was very similar to those reported by other researchers for subtests of the tasks reported in this study (e.g., Dobbs & Rule, 1989; Light & Anderson, 1985; Park et al., 1996; Salthouse, 1988a). Moreover, the pattern and magnitude of the correlations among the measures was very similar when the effects of age were partialled out, suggesting that the correlations can not be accounted for simply by the fact that all of the tasks are sensitive to age.

Squared multiple *r*s showed that the missing digit span task was the only task that did not share a significant amount of variance with the other variables. Moreover, a CFA showed that all of the tasks, other than missing digit, loaded on a single factor that accounted for 66% of the variance. This finding suggests that these tasks do tap a common latent variable, although there is still a significant amount of variance in the data unaccounted for.

In summary, there are three main results of this study. The first is that WM declines with age on almost all the tasks. The second is that many WM tests have acceptable psychometric properties in elderly subjects. This conclusion must be qualified by the observation that test–retest reliability is acceptable only when performance is measured over span size or items, not when subjects are classified into WM groups on the basis of either relative performance or cutoff scores. The final conclusion is that performance on different WM tests is only moderately correlated. The practical implications of these findings are that researchers should use a composite measure that may capture commonalties among the measures and that is much more likely to result in a reliable and stable characterization of subjects' WM performance. The present study suggests that the use of a composite measure based on two or three tasks (alphabet span, subtract 2 span, and sentence span) results in better test–retest reliability and greater stability of subject classification than does the use of any single measure.

REFERENCES

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

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.

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

Botwinick, J., & Storandt, M. (1974). *Memory, related functions and age.* Springfield, IL: Thomas.

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

Caplan, D., & Hildebrandt, N. (1988). *Disorders of syntactic comprehension.* Cambridge, MA: MIT Press, Bradford Books.

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. 409-421). Amsterdam: North-Holland.

Craik, F. I. M., Morris, R. G., & Gick, M. L. (1990). Adult age differences in working memory. In G. Vallar & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory* (pp. 247-267). Cambridge: Cambridge University Press.

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

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

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.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-Mental State": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, **12**, 189-198.

Fozard, J. L., Nuttall, R. L., & Waugh, N. C. (1972). Age-related differences in mental performance. *Aging & Human Development*, **3**, 19-43.

Gathercole, S., & Baddeley, A. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology*, **81**, 439-454.

Gick, M. L., Craik, F. I. M., & Morris, R. G. (1988). Task complexity and age differences in working memory. *Memory & Cognition*, **16**, 353-361.

Goodglass, H., & Kaplan, E. (1972). *Assessment of aphasia and related disorders.* Philadelphia: Lea & Febiger.

Hartley, J. T. (1986). Reader and text variables as determinants of discourse memory in adulthood. *Psychology & Aging*, **1**, 150-158.

Hayslip, B., & Kennelly, K. J. (1982). Short-term memory and crystallized-fluid intelligence in adulthood. *Research on Aging*, **4**, 314-332.

Hooper, F. H., Hooper, J. O., & Colbert, K. C. (1984). *Personality and memory correlates of intellectual functioning: Young adulthood to old age.* Basel: Karger.

Hultsch, D. F., Hertzog, C., & Dixon, R. A. (1990). Ability correlates of memory performance in adulthood and aging. *Psychology & Aging*, **5**, 356-368.

Hultsch, D. F., Hertzog, C., Small, B. J., & Dixon, R. A. (1999). Use it or lose it: Engaged lifestyle as a buffer of cognitive decline in aging? *Psychology & Aging*, **14**, 245-263.

Hultsch, D. F., Hertzog, C., Small, B. J., McDonald-Miszczak, L., & Dixon, R. A. (1992). Short-term longitudinal change in cognitive performance in later life. *Psychology & Aging*, **7**, 571-584.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, **99**, 122-149.

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.

Light, L. L., & Anderson, P. A. (1985). Working memory capacity, age and memory for discourse. *Journal of Gerontology*, **40**, 737-747.

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.

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

Miyake, A., Carpenter, P., & Just, M. (1994). A capacity approach to syntactic comprehension disorders: Making normal adults perform like aphasic patients. *Cognitive Neuropsychology*, **11**, 671-717.

Miyake, A., Emerson, M. J., & Friedman, N. P. (1999). Good interactions are hard to find. *Behavioral & Brain Sciences*, **22**, 108-109.

Nelson, M. J., & Denny, E. C. (1960). *The Nelson–Denny Reading Test.* Boston: Houghton Mifflin.

Nunnally, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology & Aging*, **17**, 299-320.

Park, D. C., Smith, A. D., Lautenschlager, G., Earles, J. L., Frieske, D., Zwahr, M., & Gaines, C. L. (1996). Mediators of long-term memory performance across the life span. *Psychology & Aging*, **11**, 621-637.

Parkinson, S. R. (1980). Aging and amnesia: A running span analysis. *Bulletin of the Psychonomic Society*, **15**, 215-217.

Salthouse, T. A. (1988a). Resource-reduction interpretations of cognitive aging. *Developmental Review*, **8**, 238-272.

Salthouse, T. A. (1988b). The role of processing resources in cognitive aging. In M. L. Howe & C. J. Brainerd (Eds.), *Cognitive development in adulthood* (pp. 185-239). New York: Springer-Verlag.

Salthouse, T. A. (1990). Working memory as a processing resource in cognitive aging. *Developmental Review*, **10**, 101-124.

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

Stine, E. A. L., & Wingfield, A. (1987). Process and strategy in memory for speech among younger and older adults. *Psychology & Aging*, **2**, 272-279.

Talland, G. A. (1965). Three estimates of word span and their stability over the adult years. *Quarterly Journal of Experimental Psychology*, **17**, 301-307.

Talland, G. A. (1968). Age and the span of immediate recall. In G. A. Talland (Ed.), *Human aging and behavior* (pp. 93-129). New York: Academic Press.

Tirre, W. C., & Pe a, C. M. (1992). Investigation of functional working memory in the Reading Span Test. *Journal of Educational Psychology*, **84**, 462-472.

Tompkins, C. A., Bloise, C. G., Timko, M. L., & Baumgaertner, A.

(1994). Working memory and inference revision in brain-damaged and normally aging adults. *Journal of Speech & Hearing Research*, **37**, 896-912.

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

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.

Waters, G. S., & Caplan, D. (1997). Working memory and on-line sentence comprehension in patients with Alzheimer's disease. *Journal of Psycholinguistic Research*, **26**, 377-400.

Waters, G. S., & Caplan, D. (2002). Working memory and on-line syntactic processing in Alzheimer's disease: Studies with auditory moving windows presentation. *Journal of Gerontology: Psychological Sciences*, **57B**, 1-14.

Waters, G. S., Caplan, D., & Hildebrandt, N. (1987). Working memory and written sentence comprehension. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 531-555). Hove, U.K.: Erlbaum.

Wechsler, D. (1981). *The Wechsler Adult Intelligence Scale–Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1987). *The Wechsler Memory Scale–Revised*. San Antonio, TX: Psychological Corporation.

Welford, A. T. (1958). *Aging and human skill.* London: Oxford University Press.

Wingfield, A., Stine, E. A. L., Lahar, C. J., & Aberdeen, J. S. (1988). Does the capacity of working memory change with age? *Experimental Aging Research*, **14**, 103-107.

NOTE

[1] One possible reason for the fairly low correlation between Phase I and Phase II scores is that the length of time between Phase I and Phase II varied somewhat across subjects. The average number of days between Phase I and Phase II was 56.4 days, but the range was from 27 to 189 days. Klein and Fiss (1999) argued that the low test–retest reliability found in the previous study by Waters and Caplan (1996) may reflect the benefit of practice for individuals tested after brief intervals and the absence of such effects for those tested at longer intervals. To test this possibility, we divided the subjects into two groups: those who had been tested at an interval of less than 50 days (M = 41, range = 27–50) and those who had been tested at an interval of more than 50 days (M = 70 days, range = 52–189 days). It was not the case that the correlations were substantially or systematically higher for those who had been tested with a shorter interval between Phase I and Phase II.

(Manuscript received July 27, 2001;
revision accepted for publication March 19, 2003.)