Behavior Research Methods
2005, 37 (3), 498-505

An automated version of the operation span task

NASH UNSWORTH and RICHARD P. HEITZ
Georgia Institute of Technology, Atlanta, Georgia

JOSEF C. SCHROCK
Maryville College, Maryville, Tennessee

and

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

We present an easy-to-administer and automated version of a popular working memory (WM) capacity task (operation span; Ospan) that is mouse driven, scores itself, and requires little intervention on the part of the experimenter. It is shown that this version of Ospan correlates well with other measures of WM capacity and has both good internal consistency (alpha = .78) and test–retest reliability (.83). In addition, the automated version of Ospan (Aospan) was shown to load on the same factor as two other WM measures. This WM capacity factor correlated with a factor composed of fluid abilities measures. The utility of the Aospan was further demonstrated by analyzing response times (RTs) that indicated that RT measures obtained in the task accounted for additional variance in predicting fluid abilities. Our results suggest that Aospan is a reliable and valid indicator of WM capacity that can be applied to a wide array of research domains.

Working memory (WM) span tasks have been shown to predict performance in both higher order and lower order cognitive tasks (e.g., Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Bleckley, Conway, & Engle, 2001). Indeed, beginning with the work of Daneman and Carpenter (1980), WM span tasks have been shown to predict everything from reading comprehension (Daneman & Carpenter, 1980) to performance on the Stroop task (Kane & Engle, 2003). In addition, these same measures have been useful in predicting phenomena in a wide array of other research domains. For instance, WM span tasks have been shown to predict early onset Alzheimer's (Rosen, Bergeson, Putnam, Harwell, & Sunderland, 2002), the ability to deal with life-event stress (Klein & Boals, 2001), and the effects of alcohol consumption (Finn, 2002; see Unsworth, Heitz, & Engle, in press, for a review).

Several WM span tasks have been developed that follow the lead of Daneman and Carpenter's (1980) reading span task, all of which share the requirement that the to-be-remembered items are interspersed with some form of distracting activity. In addition, all these tasks require serial recall of the to-be-remembered items. What varies from task to task is the nature of the distracting task and that of the to-be-remembered items. Differences in the distracting task include reading sentences (reading span; Daneman & Carpenter, 1980), solving math problems (operation span; Turner & Engle, 1989), counting circles in different colors (counting span; Case, Kurland, & Goldberg, 1982), and judging whether or not letters are mirror images (spatial span; Shah & Miyake, 1996). Differences in the to-be-remembered items include digits, letters, words, shapes, and spatial locations, all of which must be remembered in the correct order.

Although many of these measures differ in both the type of distracting task and to-be-remembered items, they have been shown to have good reliability and validity (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, et al., 1999) and likely reflect a common construct. Engle et al. (1999) demonstrated that the most popular WM span tasks load on the same factor and that common measures of both short-term memory and fluid intelligence load on separate factors. Furthermore, all of these WM span tasks have shown high internal consistency estimates and good test–retest reliability (e.g., Engle et al., 1999; Klein & Fiss, 1999; see also Waters & Caplan, 2003). Thus, although there can be large differences in the types of materials used to assess WM span, performance on these tasks have been shown to share a good deal of common variance and to be reliable indicators of a broader WM construct.

Despite their utility in predicting a wide range of cognitive phenomena, WM span tasks require a great deal of

This work was supported by Grant F49620-00-1-131 from the Air Force Office of Scientific Research. We are grateful to Michelle Grant, Josh Holt, Melissa Jensen, Jessica Parsons, Tom Redick, Paul Tran, and Liz Weldon for data collection assistance. Correspondence concerning this article should be addressed to N. Unsworth, School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332-0170 (e-mail: gtg039d@prism.gatech.edu) or to R. W. Engle, School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332-0170 (e-mail: randall.engle@psych.gatech.edu).

Note—This article was accepted by the previous editor, Jonathan Vaughan

experimenter time in terms of both running participants and scoring their data. For instance, in the Turner and Engle (1989) operation span task (Ospan) participants are required to solve a series of math operations while trying to remember a set of unrelated words. The operation–word strings are presented one at a time, and for each trial, participants are to read aloud and solve the math operation and then read a word aloud. Immediately after the participant reads the word, the experimenter presses a key to move on to the next operation–word string. Following each complete set, the participant recalls the words in the order presented. For example, a three-item set might be:

IS (8/2) − 1 = 1? BEAR

IS (6 ∗ 1) + 2 = 8? DRILL

IS (10 ∗ 2) − 5 = 15? JOB

???

The question marks cue participants to write down the words in the correct order. Because the task is experimenter paced, it requires approximately 20 min of experimenter time to run each individual participant and score the responses. Furthermore, because participants are required to say each operation–word string aloud, in order to attenuate rehearsal of the to-be-remembered items, it is difficult to run participants in a group setting. Thus, the current versions of some of the more popular WM span tasks require a great deal of time in order to collect a single score. Given that many research programs, including our own, utilize WM span tasks as a prescreening tool to select individuals who are in the top and bottom quartiles of a given WM span task distribution, this means that a substantial amount of time is allocated to simply selecting participants. Another factor is that, since much of the task relies on the experimenter, there is considerable room for error and inconsistency.

The aim of the present study was to alleviate some of these disadvantages by developing a version of the Ospan task that was reliable, valid, and automated, and therefore easily administered in field, clinical, or laboratory settings. Given that WM span tasks are an increasingly popular tool in a variety of research domains, it is important to have a measure of WM capacity that is easy to administer and can be done with large groups (see De Neys, D'Ydewalle, Schaeken, & Vos, 2002, for similar arguments and task development). To this end, we developed a version of Ospan that is entirely mouse driven, paced on the basis of each individual's time to complete the operations, that automatically produces a score upon completion, and records a variety of response time (RT) measures.

METHOD

Participants

A total of 296 participants were recruited from the subject pool at Georgia Institute of Technology and from the Atlanta, Georgia community through newspaper advertisements. The participants were between the ages of 18 and 35 and received either course credit or monetary compensation for their participation. Seventy-eight of these participants were randomly selected to come back to assess the test–retest reliability of the automated Ospan (Aospan) task. Each participant was tested individually in a laboratory session lasting approximately 1 h.

Materials and Procedure

After giving informed consent, all the participants completed the Turner and Engle (1989) Ospan task, a computer-administered version of the Raven Progressive Matrices (Raven, Raven, & Court, 1998), and the Aospan task. Those 78 individuals in the retest sample completed the Aospan, along with a version of reading span (Rspan; Daneman & Carpenter, 1980) and an abbreviated version of the rotated blocks test from the Air Force Officer Qualifying Test (Berger, Gupta, Berger, & Skinner, 1990). All computer-administered tasks were written in E-Prime Version 1.0 (Schneider, Eschman, & Zuccolotto, 2002).

Ospan. The Turner and Engle (1989) Ospan task requires participants to solve a series of math operations while trying to remember a set of unrelated words. The participants saw one math operation–word string at a time, centered on a computer monitor. For each trial, the participants were required to read aloud and solve the math problem and then read aloud the word. Immediately after the participant read the word, the next operation–word string was presented. The operation–word strings were presented in sets of two to five items. Following each complete set, the participant was instructed to recall the words in the order presented. Three trials of each set size (set sizes 2–5) were presented, with the order of set size varying randomly, so that the participants could not predict the number of items. At recall, the participants were instructed to write the words from the current set in the correct order. In addition, in order to ensure that they were not trading off between solving the operations and remembering the words, an 85% accuracy criterion on the math operations was required for all the participants. The participants received three sets (of set size 2) of practice. For all of the span measures, items were scored if they were correct and in the correct position. The score was thus the total number of correct items in the correct position.[1]

Raven Progressive Matrices. The Raven is a measure of abstract reasoning. This version of the Raven is computer administered and consists of 36 individual items presented in three segments of 12 items each. Within each segment, the items are presented in ascending order of difficulty (i.e., the easiest item is presented first, and the hardest item is presented last). Each item consists of a matrix of geometric patterns with the bottom-right pattern missing. The task for the participant is to select, among either six or eight alternatives, the one that correctly completes the overall series of patterns. Each matrix item appeared separately on the screen along with the response alternatives. The participants used the mouse and simply clicked on the response that they thought completed the pattern. The mouse click registered the response and moved the program on to the next problem. The participants were allotted 5 min to complete each segment. Thus, the task lasted for either 15 min or as long it took to solve all 36 problems. A participant's score was the total number of correct solutions. The participants received two practice problems.

Automated Ospan. This version of Ospan allowed the participant to complete the task independently of the experimenter. The entire task was mouse driven and required the participant to only click the mouse button. The practice session for this task was broken down into three sections. The first practice section was simple letter span. A letter appeared on the screen, and the participants were required to recall the letters in the same order in which they were presented. In all experimental conditions, letters remained onscreen for 800 msec. At recall, the participants saw a 4 × 3 matrix of letters (F, H, J, K, L, N, P, Q, R, S, T, and Y). Letters were used

because previous research has suggested that some of the shared variance between span tasks that use words and a measure of higher order cognition, such as reading comprehension, is due to word knowledge (e.g., Engle, Nations, & Cantor, 1990). Recall consisted of clicking the box next to the appropriate letters (no verbal response was required) in the correct order. The recall phase was untimed. After recall, the computer provided feedback about the number of letters correctly recalled in the current set. Next, the participants practiced the math portion of the task. They first saw a math operation (e.g., (1*2) + 1 = ?). The participants were instructed to solve the operation as quickly as possible and then click the mouse to advance to the next screen. On the next screen a digit (e.g., 3) was presented and the participants were required to click either a "true" or "false" box, depending on their answer. After each operation, the participants were given accuracy feedback. The math practice served to familiarize them with the math portion of the task as well as to calculate how long it would take each person to solve the math operations. Thus, the math practice attempted to account for individual differences in the time required to solve math operations. After the math practice, the program calculated each individual's mean time required to solve the equations. This time (plus 2.5 *SD*) was then used as a time limit for the math portion of the experimental session for that individual. The participants completed 15 math operations in this practice session.

In the final practice session, the participants performed both the letter recall and math portions together, just as they would do in the real block of trials (see Figure 1). As in the Turner and Engle Ospan, the participants first saw the math operation, and after they clicked the mouse button indicating that they had solved it, they saw the letter to be recalled. If the participants took more time to solve the math operations than their average time plus 2.5 *SD*, the program automatically moved on and counted that trial as an error. This served to prevent the participants from rehearsing the letters when they should be solving the operations. The 2.5-*SD* limit was based on extensive piloting. Participants completed three practice trials each of set size 2. After participants completed all of the practice sessions, the program progressed to the real trials, which consisted of three sets of each set size, with the set sizes ranging from 3 to 7. This made for a total of 75 letters and 75 math problems. Note that the order of set sizes was random for each participant. Set sizes

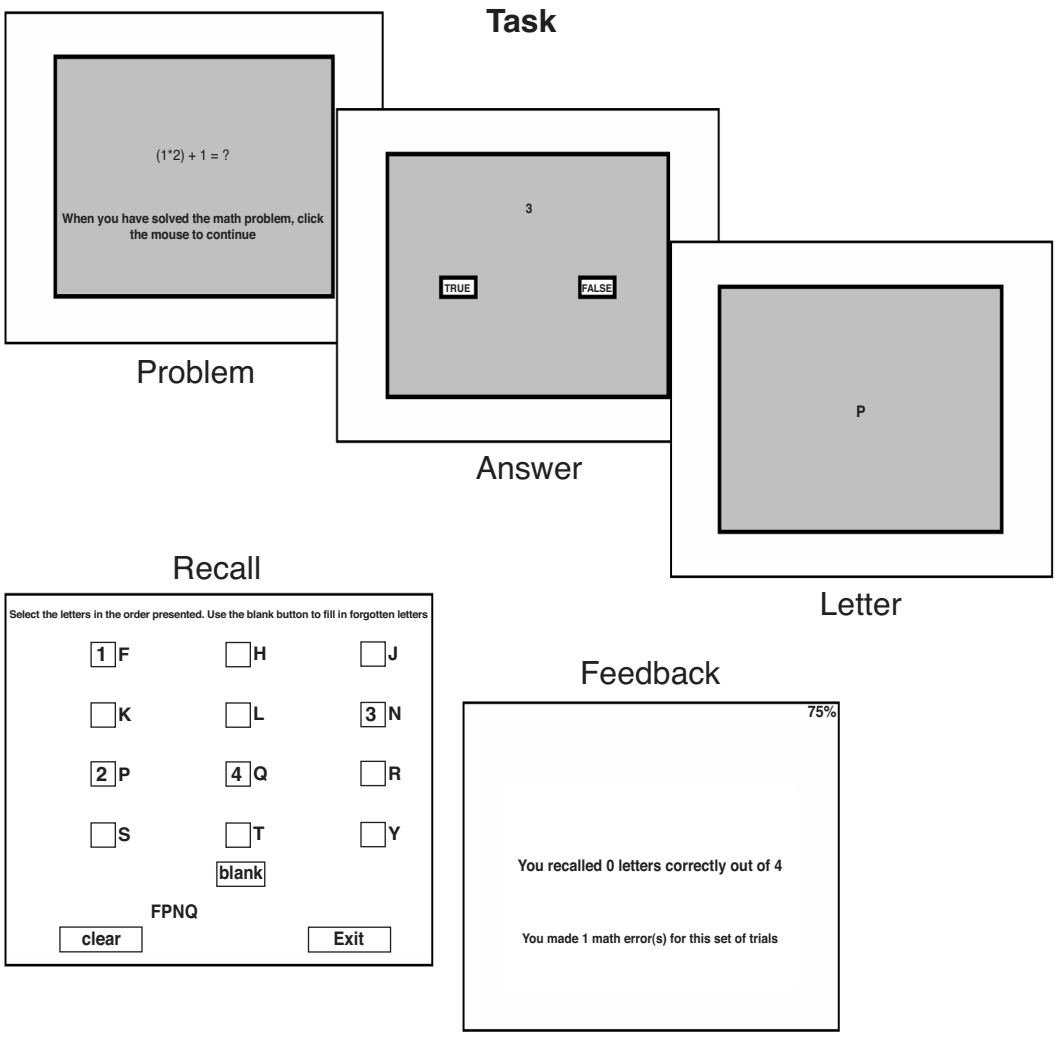

Figure 1. Illustration of the automated operation span task. In the task, first a math operation is presented. After it is solved, participants click the mouse and a digit is presented, which is judged to be either the correct or incorrect answer to the math operation. This is followed by a letter for 800 msec. For recall, the correct letters from the current set are selected in the correct order. After recall, feedback is presented for 2,000 msec.

ranging from 3 to 7 were used because pilot studies showed that these set sizes produced the best distribution of scores (i.e., neither on ceiling nor on floor). Because we wanted to only use those participants who were attempting to solve both the math operations and remember the letters, we imposed an 85% accuracy criterion for all participants. Therefore, they were encouraged to keep their math accuracy at or above 85% at all times. During recall, a percentage in red was presented in the upper right-hand corner of the screen, indicating the percentage of correctly solved math operations.

At the conclusion of the task, the program reported five scores to the experimenter: Ospan score, total number correct, math errors, speed errors, and accuracy errors. The first, Ospan score, used our traditional absolute scoring method. This was the sum of all perfectly recalled sets. So, for example, if an individual correctly recalled 3 letters in a set size of 3, 4 letters in a set size of 4, and 3 letters in a set size of 5, his or her Ospan score would be 7 (3 + 4 + 0). The second score, "total number correct," was the total number of letters recalled in the correct position. Three types of errors were reported: "Math errors" were the total number of task errors, which was then broken down into "speed errors," in which the participant ran out of time in attempting to solve a given math operation, and "accuracy errors," in which the participant solved the math operation incorrectly. The task took approximately 20–25 min to complete.

Rspan. In Rspan, the participants were required to read sentences while trying to remember a set of unrelated letters (B, F, H, J, L, M, Q, R, and X). For this task, the participants read a sentence and determined whether it was sensical or nonsensical (e.g., "The prosecutor's dish was lost because it was not based on fact. ? M"). Half of the sentences were sensical, whereas the other half were nonsensical. Nonsensical sentences were made by simply changing one word (e.g., "dish" from "case") from an otherwise normal sentence. There were 10–15 words in each sentence, and the participants were required to read the sentence aloud and indicate whether it was sensical or nonsensical by saying either "yes" (sensical) or "no" (nonsensical). After the participants gave their response, they said the letter aloud. The experimenter then pressed a key to move on to the next sentence–letter string. At recall, the participants wrote down the letters from the current set in the correct order. There were 3 trials of each set size, with set size ranging from 2 to 5. The same scoring procedure was used as in Ospan.

Rotated Blocks. The rotated blocks task is a paper-and-pencil spatial reasoning task. This version was taken from a study by Kane et al. (2004). For each item, an irregularly shaped 3-dimensional block was presented with some angle of rotation. The participant's task was to select one of the five block alternatives that was the exact same shape of the target when rotated. The participants received 3 practice trials and then were given 8 min to complete 10 real items. The items we employed correspond to problems 332, 333, 334, 336, 337, 338, 340, 341, 342, and 344 from the full Air Force Officer Qualifying Test.

RESULTS

Full Sample

The results are for 252 participants in the full sample (*M* age = 22.51 years, *SD* = 4.75) and 78 participants in the test–retest sample (*M* age = 22.08 years, *SD* = 4.23). Forty-four (15% of 296) individuals were excluded from data analysis because they failed to maintain the 85% accuracy criterion on the math operations for the Aospan.[2] As shown in Table 1, the results suggest that the Aospan is significantly related to the Turner and Engle (1989) Ospan ($r = .45$, $p < .01$). This correlation is similar to other correlations between WM span measures observed in the past. For instance, in the Engle et al. (1999) study, the average correlation between the three WM span measures was .43, and the average correlation between the WM measures in the Conway et al. (2002) study was .51. In addition, Aospan shows a similar magnitude of correlation with a measure of fluid abilities, as does the Turner and Engle Ospan (.38 and .42, respectively). At a surface level, this suggests that the Aospan is a valid indicator of WM capacity.

In order to assess reliability, we examined internal consistency for the Aospan. Because there are three presentations of each set size, we combined the first presentation of each set size into one score, the second presentation of each set size into a second score, and the final presentation into a third score. Cronbach's alpha was then computed on the basis of these three subscores. The resulting alpha estimate was .78, which suggests that the Aospan is reliable. When correcting for attenuation, the correlation between the Aospan and the Turner and Engle (1989) Ospan was .57. Together, these initial analyses suggest that the Aospan is both a reliable and valid indicator of WM capacity.

Test–Retest Sample

In order to more thoroughly test both the reliability and validity of this measure, we randomly selected 78 of the original participants to come back to get an estimate of test–retest reliability. These individuals again performed the Aospan, a version of the Rspan task, and a measure of spatial reasoning. The mean lag between the first and second testing was 13 days (median lag = 6 days, ranging from 1 to 173 days). As shown in Table 2, the test–retest sample and the full sample showed remarkable similarity on all measures, suggesting that the test–retest sample was representative of the full sample. Note that Aospan Score 1 is the absolute scoring method given by the program and Aospan Score 2 is the number of correct items in the correct position. Aospan Score 2 was used for all of the analyses. In terms of test–retest reliability, Table 3 shows that Aospan was highly reliable across testing sessions (i.e., .83). In addition, Table 3 shows the correlations between all measures for the test–retest sample. What is notable is that all three WM span measures correlate moderately well with one another, and the two reasoning measures correlate well with one another.

In order to explore these relations more fully, we submitted Ospan, the original testing of Aospan, Rspan, Raven, and the Rotated Blocks test to a confirmatory factor analysis. Here we constrained two factors, one consisting of the three WM measures and one consisting

Table 1
Correlations Between Ospan, Aospan, and Raven

Measure	1	2	3
1. Ospan	–		
2. Aospan	.448**	–	
3. Raven	.423**	.380**	–

Note—$n = 252$. **$p < .01$.

Table 2
Descriptive Statistics for Full and Test–Retest Samples

Measure	*M*	Median	*SD*	Skew	Kurtosis	Lower Q	Upper Q
Ospan							
Full	23.53	24.00	7.92	−.22	−.20	19.00	29.00
Retest	23.23	24.00	8.54	−.19	−.20	17.75	29.25
Aospan Score 1							
Full	39.16	37.50	17.41	−.02	−.49	28.00	51.00
Retest	40.51	42.50	17.86	−.16	−.56	31.00	54.00
Aospan Score 2							
Full	55.25	58.00	13.70	−1.14	1.41	48.25	65.00
Retest	55.97	59.00	13.55	−1.21	1.90	50.75	66.00
Speed errors							
Full	1.15	1.00	1.23	1.49	3.13	.00	2.00
Retest	1.00	1.00	1.09	.99	.03	.00	2.00
Accuracy errors							
Full	4.48	4.00	2.65	.45	−.37	2.00	6.00
Retest	4.28	4.00	2.19	.63	.96	3.00	5.25
Raven							
Full	25.44	26.00	5.36	−.60	.54	22.00	30.00
Retest	25.78	26.00	4.93	−.62	.04	23.00	30.00

Note—Full sample $N = 252$; test–retest sample $n = 78$; Aospan Score 1 = absolute scoring procedure; Aospan Score 2 = correct items in the correct position scoring procedure; Lower Q = lower quartile; Upper Q = upper quartile.

of the two reasoning measures. Our goal was to assess how well the Aospan would load on the WM factor and to replicate a very simple model illustrating the relation between WM capacity and general fluid intelligence (gF) that has been demonstrated in the past (e.g., Engle et al., 1999). Therefore, we predicted that the Aospan would load highly on the WM capacity factor and load as high as the other two WM measures. In addition, we predicted that the correlation between WM capacity and gF would be close to .60 on the basis of the magnitude of correlation that previous studies have shown (see Kane et al., 2004).

As shown in Figure 2, this is exactly what we found. The Aospan loaded highly on the WM capacity factor (.68), and this loading was similar to that for the original Ospan and Rspan (.79 and .88, respectively). Furthermore, the correlation between the WM capacity factor and the gF factor was .56, which is clearly consistent with previous findings. The model fit was good [$\chi^2(4) = 6.40$, $p > .17$, RMSEA = 0.08, NFI = .96, CFI = .98]. It should be noted that with a sample size of only 78, the power of this analysis was quite low. Nevertheless, we feel that this analysis is still informative. It showed that all three WM measures had high significant loadings on the WM capacity factor and that the correlation between the two constructs was of similar magnitude, as has been previously reported. Thus, in essence we replicated a model that demonstrates the relation between WM capacity and gF, suggesting that the model is reliable despite its low power. Furthermore, all of the fit indices suggested that the model fit was good. Together, these results suggest that the Aospan shares a good deal of variance with other WM measures and that these measures are moderately related to fluid abilities.

Response Time Analysis of Aospan

As a final demonstration of the utility of the Aospan, we present some analyses regarding the RT measures it provides. One problem with the original span tasks was that in many cases, the processing component and the to-be-remembered items were presented onscreen simultaneously. For example, participants may see: ?IS (8/2) − 1 = 1? BEAR on the screen. Thus, if a measure of RT were collected for this screen, it would be difficult to determine what portion of the overall time was due to completing the operation and what portion was due to encoding the word (although see Engle, Cantor, & Carullo, 1992). The Aospan has an advantage over previous versions of WM span tasks in that it collects two separate RT measures for the processing of the operations as well as RT measures for recall. That is, as shown in Figure 1, RT is collected for each math processing screen (problem and answer screens, respectively) as well as for each mouse click during the recall phase. Thus, it allows for a more detailed analysis of RT measures than do previous versions of the WM spans.

Here, we briefly examine the role that these RT measures play in predicting both performance on the span

Table 3
Correlations for the Test–Retest Sample

Measure	1	2	3	4	5	6
1. Ospan	–					
2. Aospan	.499**	–				
3. Retest Aospan	.453**	.831**	–			
4. Rspan	.710**	.603**	.627**	–		
5. Raven	.411**	.363**	.397**	.403**	–	
6. Rotated Blocks	.306**	.224*	.264*	.147	.387**	–

Note—$n = 78$. *$p < .05$. **$p < .01$.

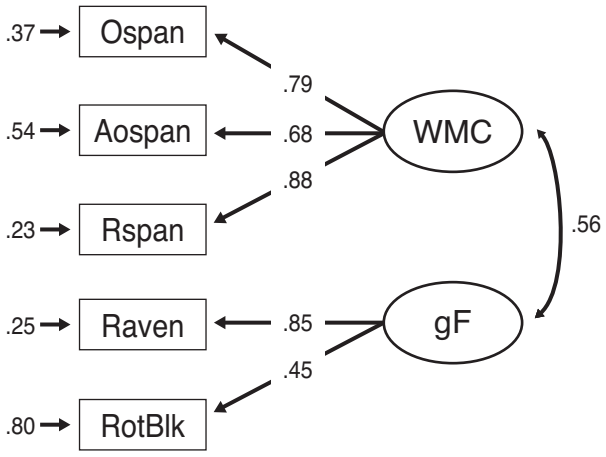

Figure 2. Path model for the structural equation analysis of the relation between working memory (WM) capacity and fluid intelligence. The numbers on the paths leading from the constructs (circles) to the manifest variables (rectangles) are the loadings of each measure on that construct. The number for the double-headed path between the WM capacity factor and the gF factor is the correlation between the two constructs. All paths and loadings are significant at the .05 confidence level. WMC = working memory capacity; gF = general fluid intelligence; Ospan = operation span; Aospan = automated operation span; Rspan = reading span; Raven = Raven Progressive Matrices; RotBlk = Rotated Blocks.

tasks themselves and in predicting measures of higher order cognition. Several recent reports have demonstrated the utility of examining RT for both the processing component of the WM span tasks (e.g., Bayliss, Jarrold, Gunn, & Baddeley, 2003) and recall (e.g., Cowan et al., 2003). In the present analyses, we examined the relation between the two math processing screens and the recall component to WM span accuracy and to performance on the fluid abilities measures.

Table 5 shows the descriptive statistics for the three RT measures. Note that all RT measures are based on the mean of the median for each participant for the initial testing of Aospan. In addition, note that these analyses are based on 72 individuals due to data collection problems for 6 individuals. As shown in Table 4, all three RT measures correlated moderately well with one another. However, as is shown in Table 5, the RT measures correlated less well with the span scores than with the fluid ability measures. For instance, the only significant correlation between the RT measures and the span scores was moderate correlations between problem RT and Aospan and Rspan. Neither answer RT nor recall RT correlated with the span scores. However, all three RT measures did correlate significantly with Raven, and two out of the three RT measures correlated with the rotated blocks task (albeit weakly). This suggests that the RT measures are more related to fluid abilities than to accuracy on the WM span tasks. In order to understand this more clearly, we submitted all of the measures in Table 5 to an exploratory factor analysis with promax rotation. As shown in Table 6, the results suggested a two-factor solution (Eigenvalue for factor 1 = 3.26, Eigenvalue for factor 2 = 1.64), accounting for 49.97% of the variance (note that the scree plot suggested three factors, but when we forced three factors, the solution failed to converge). The first factor shown in Table 6 is clearly a WM span factor, and factor 2 is clearly made up of the RT measures, with the Raven showing some crossloadings between the two factors. In addition, the two factors correlated at −.37. Thus, it seems that there is only a minor relation between accuracy on the WM span tasks and RT on the processing and recall components of the same tasks.

Given that the span scores are typically the only measure of WM performance that is used as a predictor of higher order cognition, we examined the predictive utility of both the span scores and the RT measures in predicting a composite measure of fluid abilities. The gF composite is simply the average *z*-scores of the two fluid ability measures. In order to determine the joint and unique predictive utility of the accuracy and RT measures, we performed two hierarchal regression analyses. As is shown in Table 7, together the span scores and the RT measures accounted for approximately 35% of the variance in the gF composite (e.g., .238 + .114 = .352). Of the 35% of variance accounted for, 11% was uniquely accounted for by the RT measures and 19% was uniquely accounted for by the span measures. Thus, the two types of measures only accounted for approximately 5% of shared variance in the gF composite. If the goal of a study is to predict as much variance in higher order cognition as possible, adding the RT measures into the equation will help boost the predictive power of the span tasks (e.g., see Cowan et al., 2003). However, if the goal of the investigation is to understand why WM span scores

Table 4
Means, Standard Deviations, and Correlations Between the Three RT Measures From Aospan

Measure	*M*	*SD*	1	2	3
1. Problem RT	3,120.78	1,376.78	–		
2. Answer RT	1,035.54	214.88	.419**	–	
3. Recall RT	1,088.37	235.02	.606**	.505**	–

Note—$n = 72$. $^{**}p < .01$.

Table 5
Correlations of RT Measures With All Measures for the Test–Retest Sample

Measure	Problem RT	Answer RT	Recall RT
Ospan	−.111	−.005	−.140
Aospan	−.333**	−.089	−.094
Rspan	−.350**	−.126	.183
Raven	−.401**	−.323**	−.389**
Rotated Blocks	−.204*	−.103	−.255*

Note—$n = 72$. $^{*}p < .05$. $^{**}p < .01$.

Table 6
Exploratory Factor Analysis for All Measures for the Test–Retest Sample

	Factor	
Measure	1	2
Ospan	.868	
Aospan	.673	
Rspan	.804	
Raven	.446	−.377
Rotated Blocks	.318	
Problem RT		.686
Answer RT		.645
Recall RT		.836

Note—$n = 72$. Values less than .30 have been suppressed.

correlate with gF measures, it seems that examining RT for both processing and recall offers little in terms of explaining the relation.

DISCUSSION

In this article, we presented an automated version of the operation span task that is reliable, valid, and easily administered in field, clinical, or laboratory settings. The task is paced for each subject on the basis of the average time it takes that individual to solve the math operations plus 2.5 *SD*. This allows participants to work at their own pace on the operations but restricts them from rehearsing by limiting the amount of time they are allowed to solve the operations. At the end of each set, participants are required to recall the letters in the correct serial order by clicking on the correct box in the correct order. The program provides feedback on the basis of the number of items recalled in each set as well as cumulative accuracy on the math operations. We set our accuracy criterion at 85% in order to ensure that participants were not devoting all of their processing to remembering the letters. At the end of the task, the program provides experimenters with two span scores: absolute span, which is the sum of all correctly recalled set sizes, and total correct, which is the total number of letters recalled in the correct position. The program also reports the number of total math errors, which can be broken down into (1) the number of math errors in which the participant exceeded the time he or she was allowed to solve the math, and (2) the number of accuracy errors in which the participant simply "solved" the math operation incorrectly. One advantage of this task is the fact that it is entirely mouse driven and scores itself—hence, it requires little intervention on the part of experimenter, effectively reducing the amount of time experimenters spend on each participant. The program also records a variety of RT measures for the task, including RT for the two operation screens as well as that for each mouse click during recall.

Table 7
Hierarchical Multiple Regression Analyses of WM Spans and RT Measures on gF Composite

Measure	Variable	ΔR^2
WM span–RT		
Step 1	Ospan, Aospan, Rspan	.238**
Step 2	Problem RT, answer RT, recall RT	.114*
RT–WM span		
Step 1	Problem RT, answer RT, recall RT	.161**
Step 2	Ospan, Aospan, Rspan	.190**

Note—$n = 72$. *$p < .05$. **$p < .01$.

This task was shown to be both reliable and valid. The fact that the task correlated only moderately with the Turner and Engle (1989) Ospan may seem like a cause for concern. However, it is important to point out that these are really two very different tasks, with the main similarity being that the processing component in both involves math operations. The tasks differ in that the to-be-remembered stimuli are words in one and letters in the other, the presentation of the stimuli is different, with one involving only one screen and the other involving three screens, as well as the fact that at recall in one task participants must generate the items, whereas in the other they must select the correct items from a pool of items. Thus, the low correlation between the tasks is to be expected to some extent, given these differences. However, it is not the absolute magnitude of the zero-order correlation that is crucial, but rather that the two tasks show the same pattern of correlations with other tasks (e.g., Bollen, 1989). Thus, an important finding is that the automated version of the Ospan task was shown to load on the same factor as two popular WM measures, the original version of Ospan and Rspan in both a confirmatory and an exploratory factor analysis. Furthermore, in the confirmatory factor analysis, the WM factor was moderately correlated with a factor making up two measures of spatial reasoning, replicating previous findings in the literature (e.g., Kane et al., 2004). In addition, the automated task correlated with other WM span measures with a similar magnitude of correlation that has been reported previously, and the measure had a similar magnitude of correlation with measures of fluid abilities that has been reported previously (e.g., Engle et al., 1999). Based on this, it can be concluded that the Aospan taps the same underlying construct as both the Turner and Engle Ospan and a version of Rspan, and this construct is highly related to fluid abilities.

The utility of the Aospan was further demonstrated by an examination of RT for both the processing and recall components of the task. These analyses demonstrated that although the RT measures were moderately correlated with the span scores, the RT measures predicted unique variance in a composite of fluid abilities, suggesting that an examination of the RT measures in the WM span tasks helps to boost their power in predicting higher order cognition. This task can be obtained from the attention and working memory lab Web site (available at http://psychology.gatech.edu/renglelab).

REFERENCES

Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. D. (2003). The complexities of complex span: Explaining individual differences in working memory in children and adults. *Journal of Experimental Psychology: General*, **132**, 71-92.

Berger, F. R., Gupta, W. B., Berger, R. M., & Skinner, J. (1990). Air Force Officer Qualifying Test (AFOQT) form P: Test Manual (AFHRL-TR-89-56). Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

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.

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

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

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

De Neys, W., D'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica*, **42**, 177-190.

Engle, R. W., Cantor, J., & Carullo, 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., Nations, J. K., & Cantor, J. (1990). Word knowledge and working memory capacity. *Journal of Educational Psychology*, **82**, 799-804.

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.

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. (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.

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.

Raven, J. C., Raven, J. E., & Court, J. H. (1998). *Progressive matrices*. Oxford: Oxford Psychologists Press.

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.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime Version 1.0 [Computer software]. Pittsburgh: Psychology Software Tools Inc.

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.

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

Unsworth, N., Heitz, R. P., & Engle, R. W. (in press). Working memory capacity in hot and cold cognition. In R. W. Engle, G. Sedek, U. von Hecker, & D. N. McIntosh (Eds.), *Cognitive limitations in aging and psychopathology: Attention, working memory, and executive functions*. New York: Cambridge University Press.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, **35**, 550-564.

NOTES

1. Note that the statistical analyses were rerun using the absolute scoring procedure, in which the span score is the sum of all perfectly recalled sets. For both analyses, the results were virtually identical.

2. The majority of these participants' errors were accuracy errors ($M = 17.30$, $SD = 10.08$). In addition, they tended to have very low Ospan and Raven's scores (M Ospan score = 19.07, $SD = 8.03$; M Aospan score = 39.89, $SD = 17.28$; M Raven score = 18.84, $SD = 5.64$). The pattern and magnitude of correlations did not change when analyzing the data with these participants included.

(Manuscript received March 16, 2004;
revision accepted for publication August 14, 2004.)