Mem Cogn
DOI 10.3758/s13421-014-0461-7

Shortened complex span tasks can reliably measure working memory capacity

Jeffrey L. Foster · Zach Shipstead · Tyler L. Harrison · Kenny L. Hicks · Thomas S. Redick · Randall W. Engle

Abstract Measures of working memory capacity (WMC), such as complex span tasks (e.g., operation span), have become some of the most frequently used tasks in cognitive psychology. However, due to the length of time it takes to complete these tasks many researchers trying to draw conclusions about WMC forgo properly administering multiple tasks. But can the complex span tasks be shortened to take less administration time? We address this question by splitting the tasks into three blocks of trials, and analyzing each block's contribution to measuring WMC and predicting fluid intelligence (Gf). We found that all three blocks of trials contributed similarly to the tasks' ability to measure WMC and Gf, and the tasks can therefore be substantially shortened without changing what they measure. In addition, we found that cutting the number of trials by 67 % in a battery of these tasks still accounted for 90 % of the variance in their measurement of Gf. We discuss our findings in light of administering the complex span tasks in a method that can maximize their accuracy in measuring WMC, while minimizing the time taken to administer.

Keywords Working memory · Intelligence · Individual differences

J. L. Foster (✉) · T. L. Harrison · K. L. Hicks · R. W. Engle
Georgia Institute of Technology, School of Psychology, 654 Cherry St., Atlanta, GA 30332, USA
e-mail: Jeff.Foster@gatech.edu

Z. Shipstead
Department of Social and Behavioral Sciences, Arizona State University, Tempe, AZ, USA

T. S. Redick
Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA

Published online: 13 September 2014

Introduction

Tasks that measure individual differences in working memory capacity are among some of the most frequently used non-standardized tools in cognitive psychology today (Conway et al., 2005). While these tasks are crucial in understanding the mechanisms and relationships underlying human cognition, they are also consistently used in the wider psychological literature as predictors of a wide range of human abilities in educational, developmental, social, and clinical psychology, to name just a few (e.g., Engle, 2002; Kail 2007; Lee & Park, 2005; Schmader & Johns, 2003).

Why are these tasks used in such a broad range of the literature? Possibly the most important and far-reaching aspect of measuring WMC is that it predicts a number of important characteristics about people. For example, people with higher WMC tend to be better at multitasking, can better comprehend complex language, are better at following directions, have higher SAT scores, and are better at learning new programming languages; in addition, WMC is even a predictor of how well people with schizophrenia will be at managing their medications (Engle, 2002; Engle & Kane, 2004; Heinrichs et al. 2008; Shipstead et al., 2014; see also Barch et al., 2009). Perhaps the most heavily measured relationship with WMC though is its high correlation with the ability to reason and solve novel problems, or general fluid intelligence (Gf; Ackerman et al. 2005; Cowan et al., 2005; Engle, Tuholski, Laughlin, & Conway, 1999; Engle, 2002; Kane et al., 2004; Kyllonen & Christal, 1990).

A variety of tasks have been used as measures of WMC, but some of the most widely used measures within cognitive psychology are the complex span tasks – operation span, symmetry span, and rotation span, amongst others (Conway et al., 2005; Kane et al., 2004; Redick et al., 2012; Unsworth, Heitz, Schrock, & Engle, 2005). To illustrate this point, these complex span tasks are hosted on the Georgia Tech Attention

and Working Memory Lab website (http://englelab.gatech.edu) where, to date, more than 2000 researchers have requested access to these tasks, which have been translated into ten languages. However, while the reliability of these tasks has been repeatedly verified (Redick et al., 2012), there is one consistent concern researchers have expressed with using them: they take a substantial amount of time to administer.

This concern brings about two – mostly related – issues. First, that these tasks are time consuming, and second, that researchers often draw conclusions about WMC using only a single task – perhaps due to time constraints when administering these tasks. More specifically, a single task – such as a complex span task – measures not only the cognitive ability in question, but also other factors unrelated to that ability, such as speed at solving math problems, reading ability, etc. (Loehlin, 2004; Wittman, 1988). In other words, a single indicator of WMC cannot be considered a measurement of WMC itself as the task contains variance from both WMC and the task. Therefore, to draw specific conclusions about WMC researchers should use multiple indicators to create either a composite or factor score of the WMC construct that consists of the variance shared between two or more complex span tasks (Conway et al., 2005; Engle et al., 1999; Loehlin, 2004; Shipstead et al., 2012; Wittman, 1988).

To address this timing concern, the present study examines (1) whether the complex span tasks can be shortened without substantially reducing how reliably they measure WMC, and (2) a cost/benefit analysis of the complex span measures in a way that informs and aids researchers in which task and block combinations they can use to gain the most information in the amount of time they have allotted. Given the wide use of complex span tasks, shortening the time it takes to complete the tasks would allow more researchers to measure individual differences within their current lines of research.

In the current study, we asked subjects to complete three complex span tasks and three measures of fluid intelligence. We restructured the complex span tasks for this experiment by creating three blocks of trials that kept the variation in the number of items remembered constant across blocks. For example, three blocks of the symmetry span consisted of one trial each with two, three, four, and five items to remember, which were then presented in random order. This method allowed us to examine the individual contribution of each block to the task's ability to predict both WMC and Gf. We then compared whether the earlier or later blocks of trials were more predictive of WMC and Gf to determine if the tasks can be substantially shortened, and then compared the amount of variance predicted to the length of time each block of trials took to complete. This analysis illuminated new combinations of tasks that can maximize the variance in Gf and WMC that the complex span tasks measure, while minimizing the amount of time it takes to administer multiple complex span tasks.

Method

Subjects

Five hundred and eighty-nine university ($N = 401$) and community ($N = 188$) subjects, aged 18–35 ($M = 22.36$, $SD = 4.50$; 46.52 % female), successfully completed all six tasks in this study as part of a larger, four-session, study. All subjects reported speaking English fluently before the age of five years.

Materials and procedure

Subjects completed three complex span tasks as measures of working memory capacity (WMC) and three measures of fluid intelligence (Gf) as part of a larger cognitive battery completed in our lab.[1]

WMC measures

We used three established complex span tasks to measure WMC. In a complex span task subjects are given a sequence of to-be-remembered items (such as a sequence of letters). However, subjects must also complete a distractor task (such as solving a math problem) between the presentations of each to-be-remembered item in the sequence. The number of to-be-remembered items, and the corresponding distractor tasks, unpredictably varied from one trial to the next. Subjects saw a sequence of anywhere from two to seven to-be-remembered items (depending on the task), and saw each sequence length three times. Typically, these sequence lengths are randomized throughout the complex span task. Importantly though, this randomization was modified for the current study. That is, subjects completed three blocks of trials with one of each sequence length randomly sampled within each block. To the subject there would be nothing to suggest these sequence lengths were blocked, but doing so allowed us to keep each block of trials equal in terms of sequence lengths.

Other than blocking sequence lengths, the three complex span tasks all followed the procedures of Unsworth and colleagues (2005). The primary distinction of this method is that it requires subjects to maintain their response times for the distraction items in each task – an important requirement to reduce people's ability to rehearse the to-be-remembered items during the distraction task. For this purpose, subjects were timed while practicing the distraction tasks during the instruction phase. They were then required to respond within 2.5 standard deviations (SDs) of their average response time to

[1] The results of the larger battery can be found in Shipstead et al. (2014). It may be noted that the sample size for this study is larger than the sample size in Shipstead et al. This difference is due to subjects who did not complete, or accurately complete, tasks within the battery other than those presented here.

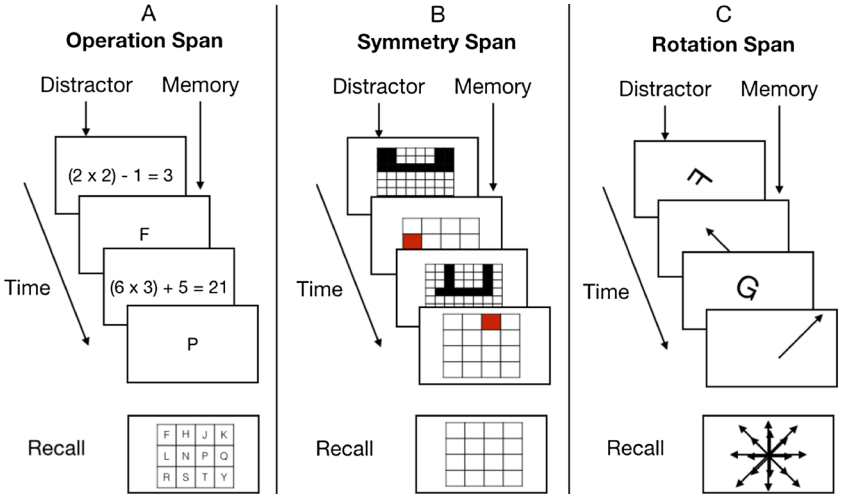

Fig. 1 Examples of Operation Span, Symmetry Span, and Rotation Span

each distraction item. These tasks are available for download from the Attention and Working Memory Lab website (http://englelab.gatech.edu/tasks.html).

Operation span As panel A of Fig. 1 shows, the operation span (OSpan) uses letters as the to-be-remembered items, and simple math problems as the distractor task (Kane et al., 2004; Unsworth et al., 2005). Subjects first solve a math problem, and then see a letter, and then solve another math problem, and see another letter. This math-letter sequence is repeated from three to seven times for each trial with an unpredictable length each time. After each math-letter sequence, subjects are asked to recall, in order, the preceding letters. Scores are calculated by summing the number of letters correctly recalled in the correct order – also known as the partial score (Turner & Engle, 1989).

Symmetry span Panel B of Fig. 1 shows an example item from the symmetry span (SymSpan; Kane et al., 2004; Unsworth, Redick, Heitz, Broadway, & Engle, 2009). The SymSpan task followed a similar method to the OSpan with three key differences. First, the distractor task is judging whether a displayed shape is symmetrical along its vertical axis. Second, the to-be-remembered items are locations of red squares in a 4×4 grid of potential locations. Finally, the number of symmetry-location pairs varied from two to five times per trial.[2] Scores are calculated by summing the number of red square locations correctly recalled in the correct order.

Rotation span Panel C of Fig. 1 shows an example trial from the rotation span (RotSpan; Kane et al., 2004, Harrison et al., 2013). The RotSpan follows a method similar to the other two tasks, with three key differences. First, the distractor task is judging whether a rotated letter is presented correctly, or is a mirrored image of the letter. Second, the to-be-remembered items are arrows of either short or long length and pointing in one of eight different directions. Finally, the rotation-arrow sequence is repeated from two to five times per trial. Scores are calculated by summing the number of arrows correctly recalled in the correct order.

Fluid intelligence measures

Raven's advanced progressive matrices (RAPM) The RAPM task consisted of the 18 odd-numbered items from the larger, 36-item RAPM (Raven, Raven, & Court, 1998). In the RAPM, subjects see a 3×3 grid of shapes with the bottom right shape missing. The shapes themselves follow a logical pattern from left to right, and from top to bottom. The subject's task is to choose, from a list of eight possible choices, the shape that logically fits in the missing corner. Subjects were allowed a maximum of 10 minutes to complete all 18 problems. Scores were calculated by summing the number of correct answers.

Letter sets The Letter Sets task consists of 30 items (Ekstrom, French, Harman, & Dermen, 1976). In this task, subjects see five sets of four letters. Four of the sets of letters follow a similar logical sequence, while the fifth set does not follow that logical sequence. For example, four sets of letters may show letters increasing in alphabetical order (DEFG, ABCD, WXYZ, MNOP), while the fifth set may show letters decreasing in alphabetical order (ZYXW). The subject's task is to choose the set of letters that does not logically fit with the rest. Subjects were allowed a maximum of 7 minutes to complete all 30 problems, and their scores were calculated by summing the number of correct answers.

[2] The maximum number of items to recall varies between the three tasks. These differences in the maximum recall number are based on previous studies and were used to maintain consistency with previous research using these tasks.

Table 1 Descriptive statistic and pearson correlations among tasks

	Mean	Standard Deviation	Skew	Kurtosis	1.	2.	3.	4.	5.	6.
1. Number Series	8.54	3.59	-0.21	-0.88	—					
2. RAPM	8.67	3.92	-0.06	-0.89	0.65	—				
3. Letter Sets	15.24	5.35	-0.03	-0.64	0.68	0.61	—			
4. OSpan	54.35	15.46	-0.92	0.31	0.54	0.50	0.45	—		
5. SymSpan	26.62	9.04	-0.43	-0.50	0.56	0.53	0.49	0.53	—	
6. RotSpan	24.56	9.84	-0.44	-0.57	0.56	0.58	0.54	0.52	0.68	—

Note.*RAPM*=Raven's Advanced Progressive Matrices, *OSpan*=Operatiom Span, *SymSpan*=Symmetry Span, *RotSpan*=Rotation Span

Number series The Number Series task consists of 15 items (Thurstone, 1938). In this task, subjects see a sequence of numbers that follow a logical pattern (1, 2, 3, 5, 8, 13, 21). The subject's task is to choose, from five available options, the next number in the sequence (34). Subjects were allowed a maximum of 5 minutes to complete all 15 problems, and their scores were calculated by summing the number of correct answers.

All tasks were part of a larger study conducted across four two-hour sessions. The complex span tasks were always the first task on the first, second, and third day for OSpan, SymSpan, and RotSpan, respectively. The measures of Gf were always the third task on each of the first, second, and third days for Number Series, RAPM, and Letter Sets, respectively.

Results and discussion

Before addressing our primary research questions, we first sought to verify that the measures of fluid intelligence and the complex span tasks formed distinct, but related, constructs using (1) zero-order correlations, and (2) factor analysis.

Table 1 outlines the correlations among the measures of fluid intelligence and working memory. Consistent with the wider literature, all of the tasks show relatively high correlations with each other (Range: .45~.68, all $ps < .01$). Importantly though each Gf task is highly correlated with each other Gf task (.61~.68), and each complex span task is highly correlated with each other complex span task (.52~.67). The generally high correlation within each construct – relative to the lower correlations between each construct (.45~58) – suggests that these tasks did measure two related, but distinguishable, abilities.

The results of the correlations suggest these tasks are highly correlated. As such, we further analyzed this pattern using a two-solution factor analysis to further test their distinguishability.[3] Because of the well established high relationship between these constructs, we used principal-axis factoring with a Promax rotation ($K = 4$) to allow the factors to correlate. The pattern loadings of this factor analysis appear in Table 2. Importantly, the loadings of each task on the two factors confirm that the Gf tasks form the first distinct factor, and the

[3] Using a maximum-likelihood estimation, we found that χ^2 was significantly reduced using a two-factor solution as opposed to a one-factor solution, $\Delta\chi^2(5) = 94.61$, $p < .01$. Based on previous research (Engle, 2002; Engle et al., 1999; Kane et al., 2004) and this analysis, we used the two-factor solution.

Table 2 Rotated factor loadings of the six tasks using Promax (K=4) rotation

	Factor 1	Factor 2
1. Number Series	0.800	0.059
2. RAPM	0.636	0.174
3. Letter Sets	0.813	-0.019
4. OSpan	0.269	0.433
5. SymSpan	-0.040	0.869
6. RotSpan	0.126	0.711

Note. RAPM=Raven's Advanced Progressive Matrices, *OSpan*=Operation Span, *SymSpan*=Symmetry Span, *RotSpan*=Rotation Span

complex span tasks form the second distinct factor. Consistent with previous findings, the factor correlation showed these two factors were highly correlated (r = .76; Engle et al., 1999; Kane et al., 2004; Kane, Hambrick, & Conway, 2005; Oberauer et al. 2005). Taken together, the results of the correlation table and the factor analysis suggest that these two sets of tasks did measure related, but distinct, abilities.

Are later blocks of trials in complex span tasks more predictive of fluid intelligence than earlier blocks of trials?

If later blocks of trials in the complex span tasks are more predictive of fluid intelligence than earlier blocks of trial then shortening the complex span tasks would substantially reduce their ability to predict WMC. To address this question, we first created two – unrotated – factor scores using principal-axis factoring: one using the Gf tasks, and the second using the complex span tasks. For the WMC factor, we recalculated three different factors – each omitting one of the three blocks. This omission was to avoid having any given block of trials predict a factor containing that block of trials. Next, we tested seven regression models to see how much of the variance in the Gf factor all three blocks of complex span tasks predicted, and how much of the variance each block of trials predicted in the WMC and Gf factors. In all seven models the variables were entered simultaneously. The results of these regression models appear in Table 3.

Table 3 Predicted variance in Fluid Intelligence (Gf) and Working Memory (WM) factors using the full model, or only trials from block 1, block 2, or block 3

Blocks Included in Model	Predicted Factor	Total Variance Predicted
All blocks from all tasks (Full Model)	Fluid Intelligence	51.5%
	Working Memory	—
Block 1 only from all three tasks	Fluid Intelligence	46.5%
	Working Memory	71.8%
Block 2 only from all three tasks	Fluid Intelligence	44.3%
	Working Memory	75.4%
Block 3 only from all three tasks	Fluid Intelligence	44.0%
	Working Memory	73.3%

Note. For the WM factor, the block used as the predictor variable was removed from the outcome variable. For example, in the Block 2 only analysis, the working memory factor was created using only data from block 1 and block 3)

Three findings in Table 3 are notable. First, when using all three blocks of all three tasks the regression model accounts for 51 % of the variance in the Gf factor. Second, a comparison between the models using Block 1 only and Block 3 only shows that the last block of trials (R^2 = .73) does not explain significantly more variance in WMC than the first block of trials (R^2 = .72), Z = 0.55, p = .58.[4] Third, the last block of trials (R^2 = .44) does not explain significantly more of the variance in the Gf factor than the first block of trials (R^2 = .47), Z = 0.58, p = .56. Taken together, these findings suggest that the three blocks of trials are statistically equivalent in their ability to measure Gf and WMC.[5]

Can the complex span tasks be shortened without substantially reducing their predictive validity?

Given our findings that the later blocks of trials are no more predictive of Gf or WMC than earlier blocks of trials, we next address the question of whether the tasks can be shortened without substantially reducing their ability to measure WMC. We address this question in three ways. First, we ask if reducing the number of blocks – while using all three tasks – substantially reduces the tasks' ability to predict both the WMC and Gf factors. Second, we ask if reducing the number of blocks in any given task substantially reduces the task's ability to predict the WMC and Gf factors. Finally, we ask whether reducing the number of trials in these complex span tasks reduces the reliability of these measures more than would be expected using the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910).

Figure 2 shows the amount of WMC and Gf factor variance predicted by any given block of trials.[6] This figure

[4] Due to the fact that the dependent measures for these regressions were not identical, we used Fisher's r-to-Z transformation for this analysis.

[5] In addition to these regression analyses, we conducted two confirmatory factor analyses (CFA) using the three blocks of each task to form each of three factors (i.e., an OSpan factor). In the first CFA, the three blocks of trials were allowed to load freely on each task's factor. In the other CFA, the factor loadings were set to be equal for each block. Consistent with the findings of the regression analyses, these two CFAs did not significantly differ from one another, $\chi^2(6) = 10.74$, p = .10. In other words, no single block of trials was a better indicator of the task than any other block.

[6] It is important to note that Panel B demonstrates the individual contribution of each task to a factor score that includes that task – a method that is generally problematic for interpretation. As such, 100 % of the variance in "All Three" is predicted by using Blocks 1, 2, and 3 – the variables are identical. However, since we are interpreting the ability of each task to measure the WMC factor we have chosen to include Panel B regardless of this limitation.

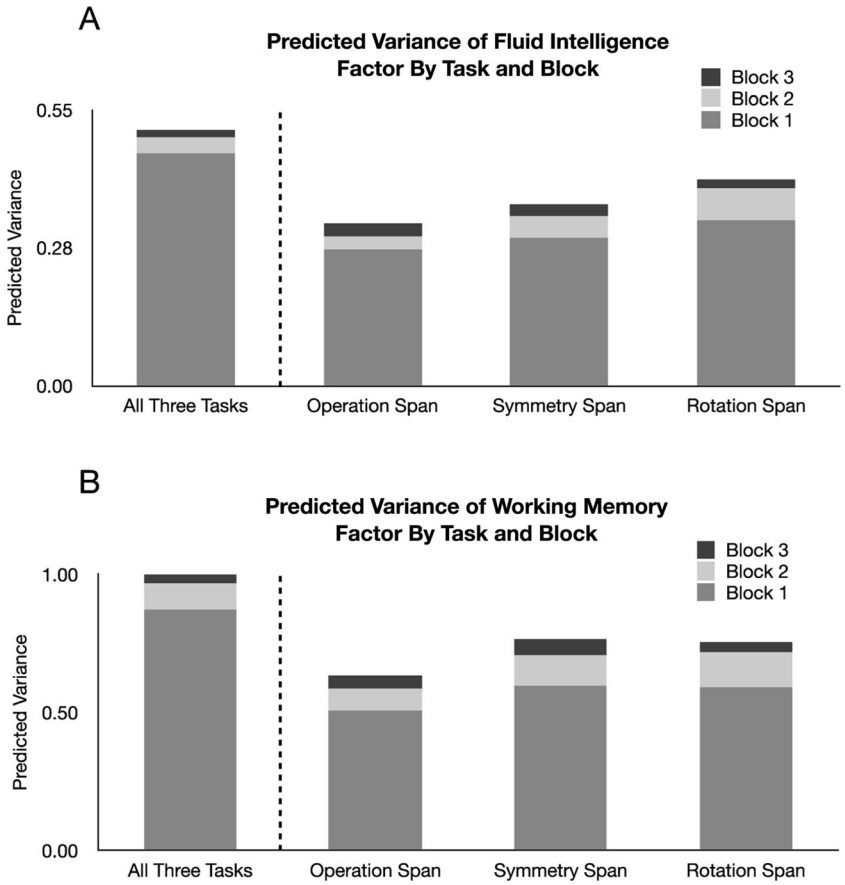

Fig. 2 Predicted variance by task, block and factor predicted

demonstrates the additive variance of each block using hierarchical regression models with Block 1 always entered first, followed by Block 2 and finally Block 3. It shows us four important findings. First, the larger bars on the left side of the dashed line in both Panel A and Panel B show that the majority of both the Gf (R^2 = .47) and WMC (R^2 = .87) factor variance explained by the complex span tasks can be accounted for in the first block of trials in each task, while the third block of trials only accounts for 3.2 % and 1.4 % of the variance in the WMC and Gf factors respectively. Given our relatively large sample size, even a 1.4 % increase in the explained variance is statistically significant ($F(3, 579) = 5.48$, $p < .01$), but in most cases it could be argued that this relatively small increase in explained variance is not a substantial increase. Second, the remaining three bars in Panel A demonstrate that even within a single complex span task, the majority of the Gf variance explained by the task is accounted for in the first block of trials, while the third block of trials only accounts for about 2 % of the variance from any given task. Third, the three right bars of Panel B demonstrate that – similar to Panel A – even within a single complex span task, the majority of the WMC variance explained by the task is accounted for in the first block of trials, while the third block of trials accounts for far less of the variance (4 % ~ 6 %) from any given task. Finally, a comparison between the bar on the left of the dashed line in Panel A, and the three bars on the right shows that even a single block of trials from each of the three tasks (R^2 = .47) is more predictive of the Gf factor than all of the blocks on any given task (R^2s= .32 ~ .41).

Another important question to address when asking if these tasks can be shortened is whether the reduction in the tasks' reliability and correlation to Gf are larger than would be expected based on the reduced number of items in the task. In other words, does reducing the length of the task disproportionally reduce the reliability and predictiveness of the task? To address this question, we next conducted a Spearman-Brown prophecy analysis (Brown, 1910; Spearman, 1910) to compare the expected reduction in task reliability (as measured with Cronbach's alpha) and predictiveness (correlation to Gf) with the actual reduction. These findings appear in Table 4. Table 4 demonstrates two important findings from the Spearman-Brown prophecy analysis. First, for all three tasks, there was no significant decrease from the task's reliability predicted by the Spearman-Brown prophecy to the actual reliability (Ws < 0.91, ps > 0.13; Feldt, 1969). In addition, the actual correlation with Gf was not significantly lower than the predicted correlation with Gf (Zs < 0.05, ps > 0.96). In other words, the reduction in the number

Table 4 Actual and Spearman-Brown prophecy predicted reliability and correlations to the Gf factor using all three blocks of each task, the first two blocks of each task, and the first block only of each task

		All 3 Blocks	Block 1 & 2	Block 1
OSpan	Actual Cronbach's α	0.870	0.817	0.690
	Predicted α	—	0.817	0.690
	Actual correlation to Gf	0.566	0.547	0.523
	Predicted correlation to Gf	—	0.549	0.504
SymSpan	Actual Cronbach's α	0.825	0.753	0.609
	Predicted α	—	0.759	0.611
	Actual correlation to Gf	0.601	0.581	0.543
	Predicted correlation to Gf	—	0.576	0.517
RotSpan	Actual Cronbach's α	0.866	0.808	0.658
	Predicted α	—	0.812	0.683
	Actual correlation to Gf	0.640	0.627	0.576
	Predicted correlation to Gf	—	0.620	0.569

of items matched what would be expected if the later blocks were no more important to measuring WMC than the earlier blocks.

In short, these findings suggest that the complex span tasks can, in most cases, be reduced without substantially decreasing their reliability or predictive utility. More specifically, the third block of trials can be removed from all three tasks, and the total loss of explained variance in Gf is only 1.4 %. In many cases, even the total reduction of 4.4 % of the explained variance by removing both the second and third blocks would be considered reasonable. In addition, using multiple tasks – even while reducing the number of trials in each task – results in a far better measure of WMC than using all of the trials in any single task.

Which task and block combinations are the most beneficial?

The question of which tasks and how many blocks of each task should be used will obviously vary based on the research questions being asked, the amount of variance that needs to be predicted, and the length of time that can be allotted to administering the complex span tasks (see Table 5). Several of these factors may need to be taken into account when deciding which tasks and blocks should be used, but some general rules can be found by looking at the results of this study. Before addressing these general rules, it is important to note that this study did not counterbalance the order of the complex span tasks – a decision based on standard methods in individual differences research. While a lack of counterbalancing does not invalidate the data in any way, it does limit some conclusions we can draw about the combination of tasks to use. More specifically, the fact that all of the instructions are similar across the three tasks would likely lead subjects to take longer on the first set of instructions – and on the first task in general – than on the later sets of instructions and tasks. As such, the analyses we present will always include at least the first block of trials in the OSpan before including the SymSpan, and at least one block of trials from both the OSpan and SymSpan before including a block of trials from the RotSpan.

Accounting for this limitation, Table 6 outlines the 39 models with combinations of tasks and blocks that we can draw conclusions about. It is ordered by the amount of time it took 95 % of subjects to complete all of the tasks and blocks

Table 5 Number of minutes it took Subjects (Ss) to complete each section of the task

		Minutes to complete each section			
		Instructions	Block 1	Block 2	Block 3
OSpan	Mean	9.03	3.43	3.20	3.10
	SD	2.90	1.18	1.08	1.03
	Median	8.50	3.06	2.86	2.81
	95% of Ss	13.90	5.69	5.28	5.16
SymSpan	Mean	4.70	2.51	1.56	1.34
	SD	1.52	0.66	0.41	0.40
	Median	4.45	2.36	1.47	1.26
	95% of Ss	7.43	3.58	2.35	2.08
RotSpan	Mean	6.56	1.58	1.62	1.38
	SD	1.82	0.34	0.34	0.29
	Median	6.26	1.50	1.46	1.21
	95% of Ss	10.06	2.26	2.34	1.91

Note 1. 95% of Ss is used as a general measure of how much time to allot for Ss to complete that section

Note 2. OSpan = Operation Span, *SymSpan*=Symmetry Span, *RotSpan*= Rotation Span, *SD*= Standard Deviation

Note 3. Included in the time for the instructions are the actual instructions, three sets of practice trials, and the 15 distractor task practices used for calculating subjects' mean response times to distractor tasks

Table 6 Fluid Intelligence (Gf) and Working Memory (WM) factor variance accounted for by the 39 possible models

Model	OSpan Blocks	SymSpan Blocks	RotSpan Blocks	Minutes for 95% of Ss to complete	Gf Variance Predicted	Proportion of Gf Variance Predicted
1	1	-	-	19.59	27.3%	53.4%
2	12	-	-	24.87	30.0%	58.7%
3	123	-	-	30.04	32.5%	63.6%
4	1	1	-	30.60	40.1%	78.5%
5	1	12	-	32.96	42.4%	83.0%
6	1	123	-	35.04	43.6%	85.3%
7	12	1	-	35.88	41.1%	8/0.4%
8	12	12	-	38.24	43.2%	84.5%
9	12	123	-	40.32	44.2%	86.5%
10	123	1	-	41.05	42.2%	82.6%
11	1	1	1	42.92	46.5%	91.0%
12	123	12	-	43.40	44.0%	86.1%
13	1	1	12	45.26	48.6%	95.1%
14	1	12	1	45.28	47.5%	93.0%
15	123	123	-	45.48	45.0%	88.1%
16	1	1	123	47.17	49.4%	96.7%
17	1	123	1	47.36	48.0%	93.9%
18	1	12	12	47.61	49.3%	96.5%
19	12	1	1	48.20	47.2%	92.4%
20	1	12	123	49.52	50.0%	97.8%
21	1	123	12	49.69	49.6%	97.1%
22	12	1	12	50.54	49.1%	96.1%
23	12	12	1	50.56	48.1%	94.1%
24	1	123	123	51.60	50.2%	98.2%
25	12	1	123	52.45	49.9%	97.7%
26	12	123	1	52.64	48.5%	94.9%
27	12	12	12	52.89	49.7%	97.3%
28	123	1	1	53.37	48.1%	94.1%
29	12	12	123	54.80	50.4%	98.6%
30	12	123	12	54.97	50.0%	97.8%
31	123	1	12	55.70	49.8%	97.5%
32	123	12	1	55.72	48.8%	95.5%
33	12	123	123	56.88	50.6%	99.0%
34	123	1	123	57.61	50.5%	98.8%
35	123	123	1	57.80	49.2%	96.3%
36	123	12	12	58.06	50.3%	98.4%
37	123	12	123	59.97	50.9%	99.6%
38	123	123	12	60.14	50.5%	98.8%
39	123	123	123	62.05	51.1%	100.0%

Note. Block number of each task that was used is listed under the task headings.Table is sorted by the amount of time it takes for 95% of subjects (Ss) to complete the specified blocks/tasks. Underlined Models specify models that predict less Gf variance than a model that takes less time to complete

within each model, and shows the amount of variance in the Gf factor that the model explains. In addition, the far right column shows the percentage of the maximum 51.1 % of the full model (model 39) Gf variance that the complex span tasks can predict. This column allows us to look at a reduction account of the variance by seeing what percentage of the variance prediction is lost by reducing the full model. Importantly, more than half of these models (the 23 underlined models) demonstrate task and block combinations that account for less of the variance in the Gf factor than another combination that takes less time to administer. Models that are not underlined are displayed in Fig. 3, which graphs the

amount of variance in Gf explained by each model by the amount of time it took 95 % of subjects to complete the tasks in the model.

While many conclusions can be drawn by Table 6 and Fig. 3, one key finding seems vitally important for researchers measuring WMC. One of the more commonly used, albeit inappropriate, methods that researchers use to draw conclusions about WMC is to administer only the OSpan (model 3). However, the OSpan alone accounted for less than two-thirds of the variance in the Gf factor that the full model accounted for. While some researchers may be using this single indicator due to a lack of knowledge about the need for multiple indicators, we believe that most do so due to time constraints. That is, the OSpan alone takes just over 30 minutes for 95 % of subjects to complete the task. However, these data suggest an alternative method that takes a similar amount of time: using one block of the OSpan and one block of the SymSpan (model 4). While taking less than a minute longer for 95 % of subjects to complete the tasks, this combination accounts for more than three-quarters of the variance from the full model, and adds a second indicator to more appropriately create a WMC factor. More specifically, for virtually the same amount of time it takes to conduct the OSpan task alone, researchers could use this specific task/block combination to account for an additional 14.9 % of the variance (see Table 6) that the full model predicts in the Gf factor.

Assuming that researchers who are not already using multiple indicators do so due to time constrains, the results outlined in Table 6 demonstrate the best combinations of tasks and blocks to use given a specific amount of time allotted for WMC measures. For example, if a researcher can only allot 35 minutes to completing WMC tasks, then the best combination would be to use the first blocks of the OSpan, and all three blocks of the SymSpan (model 9). At 50 minutes, the ideal combination would be to use model 20.

While these models can guide researchers in choosing the combinations that give them the most explained variance given a specific amount of time, one question still remains: what combination can reduce the time to complete the tasks without a substantial decrease in their predictive utility? While again, specific circumstances can lead to needing more or less variance accounted for, a good - albeit arbitrary - cut-off point is the point at which 90 % of the full model variance is accounted for. Using this cut-off point, the data suggest that a single block of each of the three tasks (model 11) is sufficient for the WMC tasks to predict the Gf factor. At 42.9 minutes to complete, this combination reduces the time it takes to complete the three WMC measures by 28 % [1 - (42.9 / 59.8) = .28], while still accounting for 91 % of the full model variance when predicting the Gf factor.

These findings have important implications for both individual differences research and the wide range of research conducted using measures of WMC. First, they highlight some of the deficiencies in using only a single task to measure WMC, while offering an alternative solution that greatly increases the ability to accurately measure WMC with little effect on the time it takes for subjects to complete the tasks. In addition, they offer the thousands of research laboratories using complex span measures insight into how to reduce the time-cost of using these measures without substantially reducing the amount of Gf factor variance they predict or, in some cases, increasing the predicted variance.

Summary

Overall these findings offer a means to substantially reduce the time it takes to accurately measure WMC using complex span tasks, without substantially reducing their validity. It is important to note that the findings of this study used only a single type of WMC measurement: complex span tasks. There

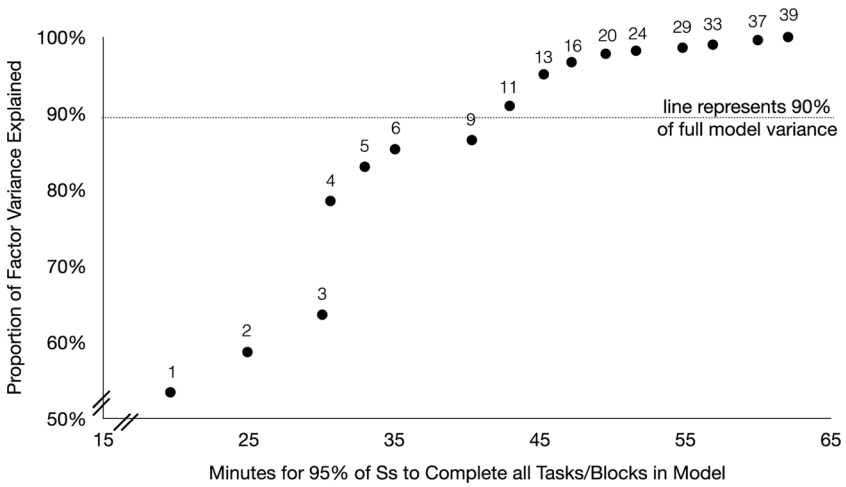

Fig. 3 Proportion of fluid intelligence (Gf) factor variance explained by model (numbers reference models in Table 6)

are many paradigms that are used to measure WMC such as n-back, visual array, and binding tasks (Gevins & Cutillo, 1993; Luck & Vogel, 1997; Wilhelm et al. 2013). Furthermore, just as using a single task to measure a construct like WMC creates problems in differentiating between task and construct variance, there are similar concerns with using a single paradigm (Loehlin, 2004; Wittman, 1988). As such, using multiple paradigms to measure WMC should result in a better measure than multiple tasks all from the same paradigm.

Of course, the lack of counterbalancing has limited the conclusions we can draw about the most ideal methods of delivering complex span tasks. Indeed, a close inspection of Fig. 3 would suggest that the first block of the RotSpan and SymSpan may be more beneficial than the first block of the OSpan. While this finding may be driven by the OSpan being the first complex span task, it may alternatively suggest that one block each of the SymSpan and RotSpan may be more beneficial, and quicker, than one block each of the OSpan and the SymSpan. Future research could look at the role of task order and task position in the task's ability to predict Gf. Regardless of this limitation, our findings do reveal evidence of improved combinations that can substantially reduce the time it takes to accurately measure WMC, without substantially reducing their predictive validity.

In addition to the single paradigm and counterbalancing limitations, three other limitations exist with the nature of this experiment. First, we measured WMC using two spatially oriented tasks (SymSpan, RotSpan) and only one verbally oriented task (OSpan). It is important to note that the use of two spatially oriented tasks may place undue weight on spatial abilities in the WMC factor (Kane et al., 2004). This limitation might also explain why the OSpan task showed the lowest relationship to the WMC factor. Second, the tasks were run on separate days within a larger battery of tasks, rather than back-to-back as they would in a single session experiment. It is possible that the importance of later blocks of trials may increase when the tasks are run back-to-back. Finally, a close inspection of Table 5 will show that the instructions of each task – which includes several practice trials – take a substantial amount of time to complete. The question of whether the instructions can be substantially shortened without reducing the tasks' ability to measure WMC is an important question unanswered by this research.

Perhaps the most beneficial finding of this study is that a single block of the OSpan and a single block of the SymSpan predict more variance in the Gf factor than all three blocks of the OSpan. This finding is most important in light of a large number of studies that draw conclusions about WMC ability from a single task, and offer an alternative method with little or no cost in completion time, that drastically improves the validity of measuring WMC. However, this finding should not be taken to justify using only a single block of two tasks rather than using the full battery of tasks. Indeed, the minimum to account for 90 % of the variance in the Gf factor would be recommended, which includes one block each of all three tasks – a combination that reduces the time taken for the full model by 28 %.

Acknowledgments This work was supported by grants from the Office of Naval Research (N00014- 12-1-0406 and N00014-12-1-1011) and the Center for Advanced Study of Language (H98230-07-D-0175 and H98230-07-D-0175) to Randall W. Engle.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131,* 30–60.

Barch, D. M., Berman, M. G., Engle, R., Jones, J. H., Jonides, J., MacDonald, A., … Sponheim, S. R. (2009). CNTRICS final task selection: working memory. *Schizophrenia Bulletin*, *35*, 136-152.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12,* 769–786.

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51,* 42–100.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests (pp. 109-113)*. Princeton, NJ: Educational Testing Service.

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

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 44, pp. 145–199). NY: Elsevier.

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.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45,* 99–105.

Gevins, A., & Cutillo, B. (1993). Spatiotemporal dynamics of component processes in human working memory. *Electroencephalography and Clinical Neurophysiology, 87,* 128–143.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science, 24,* 2409–2419.

Heinrichs, R. W., Goldberg, J. O., Miles, A. A., & McDermid Vaz, S. (2008). Predictors of medication competence in schizophrenia patients. *Psychiatry Research, 157,* 47–52.

Kail, R. V. (2007). Longitudinal evidence that increases in processing speed and working memory enhance children's reasoning. *Psychological Science, 18,* 312–313.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, & Boyle (2005). *Psychological Bulletin, 131,* 66–71.

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 visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189–217.

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

Lee, J., & Park, S. (2005). Working memory impairments in schizophrenia: a meta-analysis. *Journal of Abnormal Psychology, 114,* 599–611.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Chicago: Psychology Press.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. (2005). Working memory and intelligence—their correlation and their relation: Comment on Ackerman, Beier, & Boyle (2005). *Psychological Bulletin, 131,* 61–65.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4: The Advanced Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28,* 164–171.

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

Shipstead, Z., Harrison, T. L., Trani, A. N., Redick, T. S., Sloan, P., Bunting, M. F., … Engle, R. W. (2014). *Working memory capacity and executive functions, Part 1: General fluid intelligence*. Manuscript submitted for publication.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138,* 628–654.

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric monographs*.

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

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory, 17,* 635–654.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37,* 498–505.

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology,* 4.

Wittman, W. W. (1988). Multivariate reliability theory. Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology (2nd ed.). Perspectives on individual differences. (pp. 505-560)*. New York: Plenum.