Journal of Experimental Psychology:
Learning, Memory, and Cognition
2009, Vol. 35, No. 4, 1089–1096

0278-7393/09/$12.00 DOI: 10.1037/a0015730

Complex Span Versus Updating Tasks of Working Memory: The Gap Is Not That Deep

Florian Schmiedek and Andrea Hildebrandt
Max Planck Institute for Human Development
and Humboldt-Universität Berlin

Martin Lövdén
Max Planck Institute for Human Development
and Lund University

Oliver Wilhelm
Humboldt-Universität Berlin

Ulman Lindenberger
Max Planck Institute for Human Development

How to best measure working memory capacity is an issue of ongoing debate. Besides established complex span tasks, which combine short-term memory demands with generally unrelated secondary tasks, there exists a set of paradigms characterized by continuous and simultaneous updating of several items in working memory, such as the *n*-back, memory updating, or alpha span tasks. With a latent variable analysis ($N = 96$) based on content-heterogeneous operationalizations of both task families, the authors found a latent correlation between a complex span factor and an updating factor that was not statistically different from unity ($r = .96$). Moreover, both factors predicted fluid intelligence (reasoning) equally well. The authors conclude that updating tasks measure working memory equally well as complex span tasks. Processes involved in building, maintaining, and updating arbitrary bindings may constitute the common working memory ability underlying performance on reasoning, complex span, and updating tasks.

Keywords: working memory, complex span tasks, *n*-back, fluid intelligence, structural equation modeling

Working memory (WM) is a construct of great importance in general as well as applied cognitive psychology, with theoretical and empirical relevance for a wide range of behaviors such as reading comprehension, skill learning, and complex problem solving (Feldman Barrett, Tugade, & Engle, 2004). In research demonstrating a strong relation between WM capacity and fluid intelligence/reasoning (Kane et al., 2004; Kyllonen & Christal, 1990; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002), it has proven fruitful to combine experimentally informed task design with an individual differences perspective. Measures of individual differences, however, presume reliable and valid indicators of a construct, which leads to the question of what defines a good WM task (Conway et al., 2005; Oberauer, 2005).

Complex span tasks (CSTs; Daneman & Carpenter, 1980) are well-established measures of WM capacity. Compared to simple span tasks, in which participants have to recall a list of stimuli after a brief retention interval, CSTs require the additional accomplishment of an (often unrelated) secondary task, such as evaluating equations. This combination of short-term storage and processing requirements implements the basic definition of WM as simultaneous storage and processing (Baddeley, 2007). Because of potential trade-offs between the two components, CSTs come with the problem that performance on the secondary task cannot easily be disregarded when determining WM span on the basis of number of items recalled (Conway et al., 2005).

Another prominent paradigm to measure WM performance is the *n-back task* (Cohen et al., 1997). In this paradigm, participants have to evaluate each stimulus presented in a sequence as to whether it matches another stimulus presented earlier in the sequence, with a certain lag. In a verbal 3-back task, for example, participants see a series of letters and have to decide for each whether it matches the one seen three time steps back. This task is frequently used in the cognitive neurosciences (Owen, McMillan, Laird, & Bullmore, 2005), schizophrenia research (Glahn et al., 2005), and cognitive aging research (e.g., Dobbs & Rule, 1989; Kirchner, 1958; Schmiedek, Li, & Lindenberger, 2009).

Recently, Kane, Conway, Miura, and Colflesh (2007) attempted to elucidate the relation between CSTs and *n*-back. They correlated performance on one CST, *operation span* (Turner & Engle, 1989), with different measures of a letter *n*-back task and found only

Florian Schmiedek and Andrea Hildebrandt, Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany; and Institute of Psychology, Humboldt-Universität Berlin, Germany. Martin Lövdén, Center for Lifespan Psychology, Max Planck Institute for Human Development; and Department of Psychology, Lund University, Lund, Sweden. Oliver Wilhelm, Institute of Psychology, Humboldt-Universität Berlin; and Ulman Lindenberger, Center for Lifespan Psychology, Max Planck Institute for Human Development.

The present study was supported by the Max Planck Society, including a grant from the innovation fund of the Max Planck Society (M.FE.A.BILD0005); the Sofja Kovalevskaja Award (to Martin Lövdén) of the Alexander von Humboldt Foundation donated by the German Federal Ministry for Education and Research (BMBF), the German Research Foundation (DFG; KFG 163), and the German Federal Ministry for Education and Research (BMBF; CAI).

Correspondence concerning this article should be addressed to Florian Schmiedek, Institute of Psychology, Humboldt-Universität Berlin, Unter den Linden 6, 10099 Berlin, Germany. E-mail: florian.schmiedek@psychologie.hu-berlin.de

weak correlations in the range of .20. Furthermore, when using the two tasks to simultaneously predict performance on a marker task of reasoning ability (Raven matrices), both accounted for independent portions of variance. These findings led Kane et al. to conclude that the two paradigms "do not appear to be measures of the same construct" (p. 620). Clearly, further research needs to address this surprisingly small overlap of individual differences between the two kinds of tasks.

The results of the study by Kane et al. (2007) were based on single tasks of complex span and *n*-back, which limits their conclusiveness in two ways. First, correlations may be reduced by different content domains contributing nonshared variance to the tasks. Specifically, their *n*-back was based on letters as stimuli, operation span combined words and numerical equations, and the Raven is a figural–spatial reasoning task. Not surprisingly then, Shamosh et al. (2008) reported a much higher correlation ($r = .55$) than Kane et al. between composites of four CSTs and two *n*-back tasks, in which task-specific variance was reduced by means of aggregation. Second, when reliabilities are less than perfect, analyses at the level of observed variables underestimate true correlations because they confound construct variance with content-specific variance, task-specific variance, and measurement error. This is also the case for composites of several variables, as aggregation can only reduce but not eliminate the influence of these factors. Both kinds of problems can be alleviated, however, by using a latent variable approach with multiple task operationalizations from different content domains. This approach greatly reduces the influence of error and content-specific variance, thereby presenting a more accurate (i.e., unbiased) picture of the construct overlap between the two approaches to measuring WM.

This argument can be pushed even further. We argue for using operationalizations that are more diverse with respect to construct-irrelevant task attributes for both classes of paradigms (cf. Little, Lindenberger, & Nesselroade, 1999). Just as for the complex span paradigm tasks as different as *operation span*, *reading span* (Daneman & Carpenter, 1980), *counting span* (Case, Kurland, & Goldberg, 1982), or *rotation span* (Shah & Miyake, 1996) have been employed, one can also put *n*-back on a more general theoretical basis, allowing for a broader range of possible operationalizations. Specifically, a defining characteristic of the *n*-back is that it requires a continuous process of building, maintaining, updating, and releasing arbitrary bindings between items and temporal order positions (cf. Friedman et al., 2006). According to Oberauer, Süß, Wilhelm, and Sander (2007), WM capacity, and its shared variance with reasoning ability, reflects a limited capacity for binding arbitrary component representations. In the case of *n*-back, this would be a limited capacity to reliably bind items to order positions and to unbind and update those at each step in the sequence. The same requirement of building, maintaining, updating, and releasing arbitrary bindings between stimuli and certain representational structures can be found for several other updating WM tasks. For example, in the *memory updating* (MU) paradigm (Salthouse, Babcock, & Shaw, 1991), participants have to memorize several numbers, associated with a set of frames, and update those numbers independently according to a series of arithmetic operations appearing in the frames. Here, the updating requirement applies to numbers and spatial positions. In the study to be presented, we use a spatial version of the *n*-back, a numerical MU task, and an adapted version of the *alpha span* task (Craik, 1986) as operationalizations of updating WM tasks.

To summarize, we propose that an evaluation of different approaches to measure WM should be conducted in a larger theoretical and methodological context, comparing a set of CSTs with a set of updating tasks with content-heterogeneous selections of tasks from verbal, numerical, and figural–spatial domains for both paradigms. Similarly, constructs on the criterion side, as reasoning, should also be based on a content-heterogeneous sampling of tasks. In the present study, we achieved all this by comparing a set of three CSTs with three updating WM tasks and relating both at a latent construct level to a broad factor of reasoning tasks.

We addressed a further issue in this study, also related to the validity of the updating tasks, by using varying presentation times (PTs) for these tasks. PT obviously is a crucial determinant of task difficulty in tasks with paced continuous presentation of a sequence of items or operations, like all updating tasks used in the present study. Therefore the question arises as to how PT affects the psychometric properties of these tasks. A speed manipulation with one fast and one slow PT for each task allowed us to investigate whether this central task parameter of all updating WM paradigms had an effect not only on mean performance but also on the relations to CSTs and reasoning. For the validity of updating WM tasks, it would be desirable if manipulating PT had little effect on these relations. If this was the case, one could use varying PTs to adjust task difficulty for a given measurement purpose and target population and the task would remain a valid indicator of WM.

Method

Participants

The study consisted of two sessions. The first session was attended by 108 people. Eleven participants did not return to the second session, and 1 person attended the second session only. Thus, our analyses are based on data from 96 participants. The mean age of the sample was 24.9 years ($SD = 2.7$); the percentage of women was 47.2% and of men, 52.8%. The sample was composed of an academically diverse student population (77%), pupils and young adults in professional formation (10%), employees (8%, including 1.7% academics), and unemployed (5%, including 3% academics). Each participant was paid €50 in compensation.

Procedure

Each of the two sessions lasted 3 hr, and the participants attended both sessions within the time span of 1 week. The tasks were completed in small groups of 2 to 5 participants. The WM tasks and Raven's Progressive Matrices were administrated on personal computers (17-in. color monitors, 85 Hz rate, 1280 × 1024 pixels resolution). Additionally, participants completed the paper-and-pencil Berlin Structure of Intelligence (BIS) Test (Jäger, Süß, & Beauducel, 1997).

Updating Tasks

N-back. We administered a figural–spatial 3-back task. Participants had to assess whether the position of a black filled circle that appeared in a 4 × 4 grid was identical to the position of the circle

presented three steps back in the sequence. Participants were asked to press the green key of a button box if the positions were identical or the red key if they were not. Thirty-nine stimuli were presented in each block of trials. The first three stimuli were preparatory, because they had no reference items to be compared with. Each block of trials consisted of 12 targets, 12 nontargets, and 12 lures. Lures are items that match a stimulus presented earlier, but not at the correct temporal distance. In each block, four 2-, 4-, and 5-back, but no 1-back, lures were used. Other than the constraints mentioned, the stimulus sequence was random but fixed across participants. PT for each stimulus was 500 ms. Interstimulus intervals (ISIs) were 2,000 ms for slow and 1,000 ms for fast blocks of trials. Participants had to complete 16 (first eight slow and then eight fast) blocks in total.

Memory updating (MU). Two rows of frames were presented to the participants. The number of frames varied between 2 and 5 per row, defining memory load levels of two to five. At the beginning of each trial block, single-digit numbers were displayed simultaneously for 2,000 ms in each frame of the upper row. Participants were asked to memorize those numbers for each frame. Then, addition and subtraction operations (e.g., "+ 2," "– 4") appeared in the lower row of frames. Numbers in the corresponding frame above had to be updated according to the operations. A sequence of 4 updating operations were included in the 2-frames version, 6 in the 3-frames version, 8 in the 4-frames version, and 10 in the 5-frames version, so that 2 updating operations applied to each frame. Succeeding operations always applied to different frames, in random order. Intermediate and end results for each frame were always in the range of 0 to 9. The final values for each frame had to be entered with the keyboard. A total of 32 blocks of trials were conducted. Participants had to complete 16 blocks (four for each load level) with a PT of the operations of 2,000 ms (ISI was 500 ms), and another set of 16 blocks (four per load level) with a PT of 1,000 ms. The ISI for the slow as well as for the speeded tasks was 500 ms.

Alpha span. Our task differed in some respects from the familiar version by Craik (1986). In the familiar version, after a list of words is presented, participants have to recall the first letters of the listed words in correct alphabetical order. In our version, a sequence of 10 single consonant letters appeared. Together with the presentation of each letter, a number (from 1 to 10) was displayed below the letter. The letters had to continuously be brought into alphabetical order and participants were asked to respond to each letter–number pair whether the number corresponded to the current alphabetical position of the letter among the set of letters that had appeared so far. For example, the items in the sequence could be "B – 1" (order: 1 = B; correct response: yes), "M – 2" (order: 1 = B, 2 = M; correct response: yes), "C – 2" (order: 1 = B, 2 = C, 3 = M; correct response: yes), "K – 4" (order: 1 = B, 2 = C, 3 = K, 4 = M; correct response: no), "Z – 5" (order: 1 = B, 2 = C, 3 = M, 4 = K, 5 = Z; correct response: yes), and so on. Correspondence was confirmed by hitting the green key, and mismatches, the red key. Five out of the 10 stimuli were matches. The participants completed 16 blocks of trials. In the first half of those blocks, stimuli were presented with a PT of 2,000 ms, in the second half the PT was 1,000 ms. In both cases ISI was 2,000 ms.

Complex Span Tasks

Reading span. We used a version that differed from the original version in that participants did not have to memorize words (see Engle, Tuholski, Laughlin, & Conway, 1999) but rather single letters (see Kane et al., 2004). Several sentences were presented successively on the screen. Below each sentence, a letter was displayed. Participants had to decide whether the sentences were semantically correct (e.g., "The first thing I do in the morning is feed the dog") or incorrect (e.g., "Yesterday in church, the daughter of Till made a terrible plum"). Additionally, they were asked to memorize the letter and, after a sequence of sentence–letter combinations, recall the letters in their order of presentation. We included 12 blocks of trials, three for each load level (of two to five).

Counting span. We implemented the counting span very similarly to Kane et al. (2004). Several displays of blue circles (4–9), green circles (1–5), and blue squares (1–9) were presented. Participants were asked to count the blue circles and hit the green key if their number was even or the red key if the number was odd. The requirement to decide whether the number was even or odd was not included by Kane et al. The number of blue circles had to be memorized for later recall in the order of their presentation. The number of displays ranged from two to six per block of trials. A total of 15 blocks was completed, three per load level.

Rotation span. This task combines recall of a sequence of short and long arrows, radiating from the center of the display, with a letter-rotation task (Kane et al., 2004; Wilhelm & Oberauer, 2006). First, a regular or mirror-reversed letter was displayed on the screen. The letter could be rotated by 0, 45, 90, 135, 180, 225, 270, or 315 degrees. In the processing part of the task, participants were asked to hit the green key when letters were displayed regularly and the red key when they were mirror-reversed. After each processing step (ranging from 2–5 per block of trials), short or long arrows were shown. The arrows pointed in one of the eight directions. At the end of one sequence, participants had to recall the direction and length of the arrows in the order of their presentation and indicate this by clicking on a computer screen layout with the 16 possible positions of the arrow head. There were 12 blocks of trials to complete, three per load level.

Reasoning Tasks

Raven's Advanced Progressive Matrices (RAPM). Fifteen items of the RAPM (Raven, Raven, & Court, 1998) were administrated on the computer. An aggregate of these items was used as one indicator variable of the latent reasoning factor.

BIS test. From the reasoning scale of the BIS test (Jäger et al., 1997; for English descriptions see Carroll, 1993; Süß & Beauducel, 2005; Wilhelm & Schulze, 2002), nine reasoning tasks (three for each content category: verbal, numerical, and figural) were used. Internal consistency of these tasks was high (Cronbach's α = .80). The nine tasks were *z*-standardized and aggregated into three parcels that served as indicator variables for the latent reasoning factor. Each parcel consisted of one verbal, one numerical, and one figural task.

Scoring

For the updating tasks, we defined mean accuracies as dependent variables. For the CSTs, we used a partial-credit unit-scoring

procedure (Conway et al., 2005). The proportions of elements correctly recalled within each block of trials were computed and averaged over blocks, with equal weight given to all memory load levels of each CST. Accuracy on the processing component did not enter into the score. The correlation between the recall score and the processing accuracy was .73 for counting span, .49 for reading span, and .16 for rotation span.

Results

Preliminary Analyses

The data were screened to identify outliers. No individual's score deviated by more than 3.5 *SD*s from the mean of the sample. All the analyses were therefore based on all 96 cases. Descriptive statistics are summarized in Table 1. Internal consistencies were satisfactory to high, except for the RAPM, which might be due to the fact that only 15 of the more difficult items of this test were used in this study. For evidence regarding the satisfactory reliability of the BIS parcels, we refer the reader to the test manual (Jäger et al., 1997).

PT of the stimuli was manipulated in the updating tasks (i.e., the *n*-back, alpha span, and MU tasks). As anticipated, PT had an impact on the means. The fast versions of the *n*-back, alpha span, and MU tasks were more difficult than the slow versions. Table 1 displays the differences of the means (ΔM) and corresponding paired *t* tests, which were all significant and of medium effect size.

A correlation matrix of all tasks is displayed in Table 2. All values were positive. Some of them were not significant. However, this concerned not only several correlations between CSTs and arbitrary updating tasks but also the relation between two different CSTs.

Latent-Variable Analyses

According to our line of argument, the relation between the two sets of WM tasks needs to be examined at the level of latent variables, using a heterogeneous set of tasks. This strategy provides an unbiased picture of construct overlap between the measurement paradigms because influences of unreliability and content heterogeneity are attenuated. To this end, we used confirmatory factor analysis.

To examine the construct equivalence of the two approaches of measuring WM, we estimated a pair of nested models. In a first model (Model 1A; see Figure 1) CSTs loaded on one latent factor and the slow and fast versions of the updating tasks loaded on a second factor. The errors of the corresponding slow and fast task versions were allowed to correlate. The correlation between the factors was freely estimated. The fit of Model 1A was reasonable, $\chi^2(23) = 36.64$, comparative fit index (CFI) = .97, root-mean-square error of approximation (RMSEA) = .07 (90% confidence interval [CI] = .02, .12), standardized root-mean residual (SRMR) = .05, and the correlation between the latent factors was very high ($r = .96$). In a nested version of Model 1A, the correlation between the latent factors was fixed to 1 (Model 1B; see Figure 1). The following fit indices resulted for Model 1B: $\chi^2(24) = 36.76$, CFI = .97, RMSEA = .07 (90% CI = .01, .11), SRMR = .05.

We evaluated whether the free estimation of the correlation significantly improved the model fit with a chi-square difference test. The chi-square difference between Models 1B and 1A was not significant, $\Delta\chi^2(1) = 0.12$, indicating that fit did not improve significantly when the correlation was freely estimated. Therefore, regarding individual differences at the latent level, CSTs and updating tasks were measuring the same construct.

Table 1 shows that the manipulation of PT in the updating tasks had an impact on the means. If PT played a different role when processing fast than when processing slow versions of the updating tasks, the speed manipulation should also affect the covariances. If this was the case, then separate factors for the more or less speeded versions of the updating tasks would not correlate to 1. We tested this by comparing two models (not shown in figures) in which the slow versions of the three updating tasks loaded on one factor and the fast versions on another factor (task-specific residuals were allowed to correlate). The two factors correlated very highly ($r = .97$; not significantly different from $r = 1.00$), $\Delta\chi^2(1) = 0.59$. As a model with two latent factors correlating perfectly is equivalent to a one-factor model, we can conclude that that the fast and slow versions of the updating tasks were measuring the same construct.

In the next step, we tested a model for reasoning ability (Model 2; not shown in a figure). An aggregate of the 15 RAPM items and the BIS parcels loaded on the latent factor. This model had

Table 1
Descriptive Statistics and Cronbach's Alphas for All Variables

Variable	No. of blocks	*M*	*SD*	Skew	Kurtosis	α	ΔM	*t*	*d*
NB_s	8	.72	.14	−0.64	0.17	.96			
NB_f	8	.67	.14	−0.41	0.38	.95	.05	5.41*	0.50
AS_s	8	.63	.12	−0.58	0.66	.83			
AS_f	8	.59	.14	−0.73	2.21	.94	.04	5.22*	0.50
MU_s	16	.70	.14	−0.41	0.91	.86			
MU_f	16	.63	.14	−0.05	0.84	.80	.07	7.21*	0.66
RS	12	.72	.15	−1.94	4.87	.84			
CS	15	.84	.14	−2.84	11.28	.85			
RoS	12	.67	.15	−0.57	−0.34	.76			
RAPM	15	.54	.18	−0.21	−0.54	.63			

Note. NB_s = *n*-back slow; NB_f = *n*-back fast; AS_s = alpha span slow; AS_f = alpha span fast; MU_s = memory updating slow; MU_f = memory updating fast; RS = reading span; CS = counting span; RoS = rotation span; RAPM = Raven's Advanced Progressive Matrices; α = internal consistencies (Cronbach's alphas); ΔM = difference of means between slow and fast trials; *d* = effect size measure (Cohen's *d*).
* $p < .05$.

Table 2
Correlations Among All Tasks (N = 96)

	NB_s	NB_f	AS_s	AS_f	MU_s	MU_f	RS	CS	RoS	RAPM	K_1	K_2	K_3	K
NB_s	—													
NB_f	.91**	—												
AS_s	.43**	.42**	—											
AS_f	.38**	.40**	.79**	—										
MU_s	.35**	.28**	.46**	.45**	—									
MU_f	.26*	.25*	.49**	.40**	.70**	—								
RS	.18	.18	.21*	.17	.29**	.25*	—							
CS	.17	.17	.17	.16	.27**	.14	.30**	—						
RoS	.51**	.40**	.45**	.38**	.41**	.34**	.16	.28**	—					
RAPM	.31**	.28**	.29**	.36**	.30**	.23*	.20*	.41**	.32**	—				
K_1	.43**	.42**	.44**	.37**	.56**	.39**	.31**	.30**	.46**	.36**	—			
K_2	.31**	.30**	.38**	.31**	.50**	.36**	.33**	.26*	.41**	.31**	.60**	—		
K_3	.37**	.36**	.47**	.42**	.60**	.45**	.25*	.20*	.37**	.42**	.69**	.59**	—	
K	.42**	.42**	.49**	.43**	.64**	.46**	.34**	.29**	.48**	.42**	.88**	.84**	.87**	—

Note. NB_s = *n*-back slow; NB_f = *n*-back fast; AS_s = alpha span slow; AS_f = alpha span fast; MU_s = memory updating slow; MU_f = memory updating fast; RS = reading span; CS = counting span; RoS = rotation span; RAPM = Raven's Advanced Progressive Matrices; K_1 = reasoning, Parcel 1 (Berlin Structure of Intelligence [BIS] test); K_2 = reasoning, Parcel 2 (BIS test); K_3 = reasoning, Parcel 3 (BIS test); K = general reasoning (BIS test).
* $p < .05$. ** $p < .01$.

excellent fit, $\chi^2(2) = 0.81$, CFI = 1.00, RMSEA = .00 (90% CI = .00, .15), SRMR = .01, and factor loadings were satisfactory (.43) to high (.74 to .84).

Our concern was not only to test the internal structure of the WM tasks and show that they are all measuring the same construct. We also compared the two classes of WM paradigms with respect to their relation to reasoning. For this purpose, we computed structural models with three correlated latent variables—one for CSTs, updating, and reasoning tasks, respectively. The structural Models 3A and 3B are shown in Figure 2.

The correlations between the three latent factors were all freely estimated in Model 3A. Fit indices of this model were very good, $\chi^2(59) = 76.47$, CFI = .97, RMSEA = .05 (90% CI = .00, .08), SRMR = .05. The two WM factors correlated to $r = .97$, the CST factor and the reasoning factor to $r = .78$, and the updating factor and reasoning to $r = .84$.

With Model 3B, we tested whether the correlation of the latent updating factor with reasoning was significantly higher than the one of the CST factor. The correlations between updating and reasoning and between span and reasoning were fixed to be equal in Model 3B. Fit indices of Model 3B were also very good, $\chi^2(60) = 76.75$, CFI = .97, RMSEA = .05 (90% CI = .00, .08), SRMR = .05. Both WM factors correlated to $r = .82$ with reasoning and to $r = .99$ with each other. The difference between Models 3A and 3B was not significant, $\Delta\chi^2(1) = 0.28$. Free estimation of the correlations did not lead to significant improvement of the model fit. Thus, at the latent level, associations of the two different WM paradigms to reasoning were high and did not reliably differ from each other.

Discussion

The goal of the study was to examine the relation between updating tasks of WM, such as the *n*-back, and CSTs by means of a latent variable analysis, thereby overcoming the problem that correlations are attenuated by task-specific and content-specific

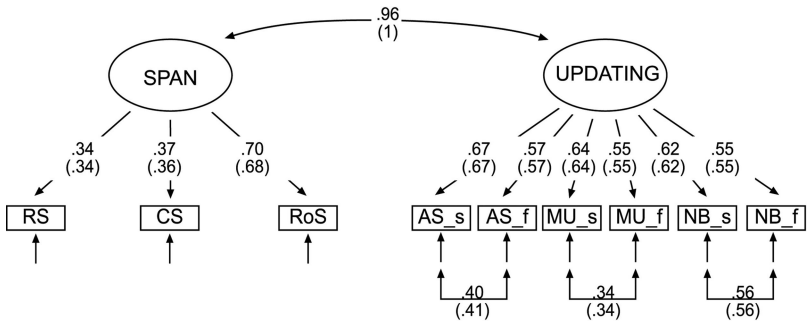

Figure 1. Models 1A and 1B. Ovals represent latent factors. SPAN = Complex Span factor; UPDATING = Updating factor. Rectangles represent manifest variables. RS = reading span; CS = counting span; RoS = rotation span; AS_s = alpha span slow; AS_f = alpha span fast; MU_s = memory updating slow; MU_f = memory updating fast; NB_s = *n*-back slow; NB_f = *n*-back fast. The errors of the slow and fast task versions were allowed to correlate. The values for Model 1B are displayed in parentheses.

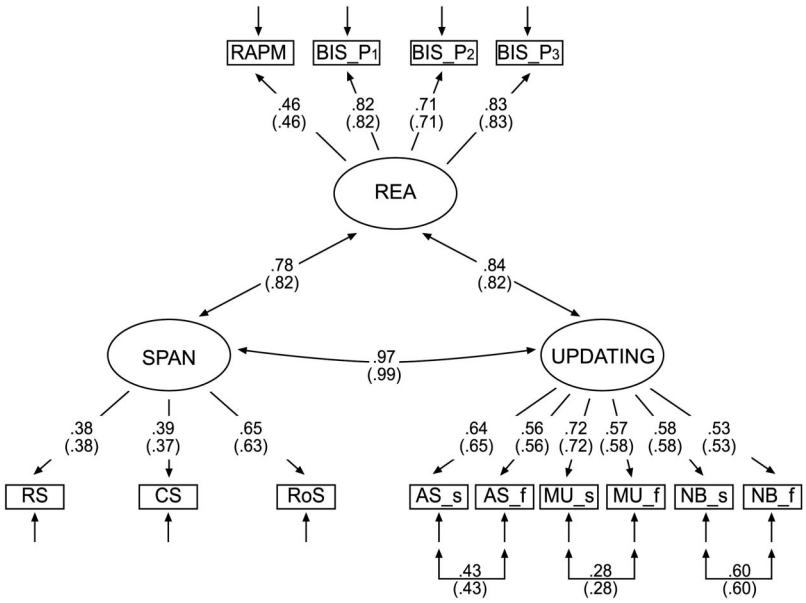

Figure 2. Models 3A and 3B. Ovals represent latent factors. REA = Reasoning factor; SPAN = Complex Span factor; UPDATING = Updating factor. Rectangles represent manifest variables. RAPM = Raven's Advanced Progressive Matrices; BIS_P_1 = Reasoning scale of the Berlin Structure of Intelligence Test (BIS), Parcel 1; BIS_P_2 = BIS Parcel 2; BIS_P_3 = BIS Parcel 3; RS = reading span; CS = counting span; RoS = rotation span; AS_s = alpha span slow; AS_f = alpha span fast; MU_s = memory updating slow; MU_f = memory updating fast; NB_s = *n*-back slow; NB_f = *n*-back fast. The errors of the slow and fast task versions were allowed to correlate. The values for Model 3B are displayed in parentheses.

sources of individual variation as well as measurement error. Results showed that the three updating tasks used here were reliable indicators of a latent factor that was statistically identical to a CST latent factor. Therefore, interindividual differences in WM capacity can be equally well described with both families of tasks. Consequently, the relations of both WM factors with reasoning are of equal magnitude.

What are the common sources of variance that produce the perfect relation of CSTs to updating tasks at the latent level, or in other words, what is common to all CSTs that is also common to all updating tasks investigated here? At least three theoretical accounts come to mind. First, the controlled attention view by Engle and colleagues (e.g., Engle, 2002) proposes that the ability to resist interference from internal and external distraction is central to WM performance. It seems difficult to rule out that individual differences in controlled attention might contribute to task performance in CSTs as well as in updating tasks, as both require a continuous focus on complex processing requirements. It is also difficult, however, to specify how much of the common variance of WM tasks can be explained by individual differences in controlled attention.

Second, Unsworth and Engle (2007a, 2007b; see also Mogle, Lovett, Stawski, & Sliwinski, 2008) have proposed that individual differences in CSTs are partially due to individual differences in the ability to maintain information accessible in primary memory and partially due to individual differences in the ability to retrieve information from secondary memory, with the latter being more strongly responsible for the relation to fluid abilities. To reconcile this view with the present findings, it would be necessary to assume that retrieval processes from secondary memory also play a central role in the updating tasks. The fact that for two of the updating tasks, *n*-back and MU, the total number of items to be maintained in primary memory at any time was within short-term memory limits makes such an assumption difficult to entertain.

Third, it is possible to interpret the commonalities among the WM tasks and to reasoning within the framework of Oberauer et al. (2007). These authors suggested that building, maintaining, and updating arbitrary bindings are the cognitive mechanisms shared between WM and reasoning performance (for similar views and supporting evidence from cognitive neuroscience, see Miller & Cohen, 2001; Smith & Jonides, 1999). These demands are apparent for the updating tasks used in our study but are also plausible determinants of CST performance. Because CSTs require the recall of items in correct serial order, there are some crucial ingredients to successful performance: the reliable establishment of bindings of content (words, numbers, directions) to context (serial positions), its maintenance in the presence of an interfering secondary task, and the necessary release and updating of bindings from one trial to the next. Releasing of bindings might hereby play a similarly important role as the creation of bindings (e.g., Durstewitz, Kelc, & Güntürkün, 1999).

An important finding from this study was that even though manipulating PT had the expected effect on means, it did not affect the reliability and validity of the tasks in any important way. This is encouraging, given that practical considerations will often require adjusting PTs to reach a certain level of accuracy. Regarding our theorizing about individual differences in WM capacity, this finding demonstrates that no additional source of variation unre-

lated to higher cognition is introduced by decreasing the PTs of updating tasks within a reasonable range; the decrease lowers mean performance while keeping the rank order of interindividual differences intact.

Our results are of particular interest for the large number of researchers using the *n*-back task. Although earlier studies have shown that MU and alpha span are reliable indicators of WM (e.g., Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000), studies of the construct validity of the *n*-back have been few and results have been mixed. Here, we have shown that a spatial 3-back task is a comparably good indicator of the updating WM factor as the other two tasks. Although single-task operationalizations always are problematic on conceptual and psychometric grounds, *n*-back is as good a marker of WM as any of the other more established tasks used here. It should be underscored again, however, that for drawing conclusions at the construct level, a broader operationalization with several tasks from different content domains is preferable to research designs based on individual tasks.

To conclude, we found a high latent correlation (.96) between a CST factor and an updating factor that is not statistically different from unity. Accordingly, both factors predict reasoning equally well. Updating tasks measure WM equally well as CSTs. These results indicate that reasoning, CSTs, and updating tasks share common processing mechanisms. Building, maintaining, and updating arbitrary bindings may constitute these mechanisms, but further research including additional tasks designed to directly assess bindings are needed to elucidate this assumption.

References

Baddeley, A. (2007). *Working memory, thought and action.* Oxford, United Kingdom: Oxford University Press.

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

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

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997, April 10). Temporal dynamics of brain activation during a working memory task. *Nature, 386,* 604–608.

Conway, A. R. A., 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.

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

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

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

Durstewitz, D., Kelc, M., & Güntürkün, O. (1999). A neurocomputational theory of the dopaminergic modulation of working memory functions. *The Journal of Neuroscience, 19,* 2807–2822.

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

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.

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

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science, 17,* 172–179.

Glahn, D. C., Ragland, J. D., Abramoff, A., Barrett, J., Laird, A. R., Bearden, C. E., & Velligan, D. I. (2005). Beyond hypofrontality: A quantitative meta-analysis of functional neuroimaging studies of working memory in schizophrenia. *Human Brain Mapping, 25,* 60–69.

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test, BIS-Test: Form 4. Handanweisung* [The Berlin Intelligence Structure Test, BIS Test: Form 4. Test manual]. Göttingen, Germany: Hogrefe.

Kane, M., Conway, A., Miura, T., & Colflesh, G. (2007). Working memory, attention control, and the *n*-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 615–622.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. E. (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.

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

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

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4,* 192–211.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202.

Mogle, J. A., Lovett, B. J., Stawski, R. S., & Sliwinski, M. J. (2008). What's so special about working memory? An examination of the relationships among working memory, secondary memory, and fluid intelligence. *Psychological Science, 19,* 1071–1077.

Oberauer, K. (2005). The measurement of working memory capacity. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 393–408). Thousand Oaks, CA: Sage.

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

Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 49–75). New York: Oxford University Press.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). *N*-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25,* 46–59.

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.

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

Schmiedek, F., Li, S.-C., & Lindenberger, U. (2009). Interference and facilitation in spatial working memory: Age-associated differences in lure effects in the *n*-back paradigm. *Psychology and Aging, 24,* 203–210.

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.

Shamosh, N. A., DeYoung, C. G., Green, A. E., Reis, D. L., Johnson, M. R., Conway, A. R. A., et al. (2008). Individual differences in delay discounting: Relation to intelligence, working memory, and anterior prefrontal cortex. *Psychological Science, 19,* 904–911.

Smith, E. E., & Jonides, J. (1999, March 12). Storage and executive processes in the frontal lobes. *Science, 283,* 1657–1661.

Süß, H.-M., & Beauducel, A. (2005). Faceted models of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 313–332). Thousand Oaks, CA: Sage.

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

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

Unsworth, N., & Engle, R. W. (2007a). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114,* 104–132.

Unsworth, N., & Engle, R. W. (2007b). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133,* 1038–1066.

Wilhelm, O., & Oberauer, K. (2006). Why are reasoning ability and working memory related to mental speed? An investigation of stimulus–response compatibility in choice reaction time task. *European Journal of Cognitive Psychology, 18,* 18–50.

Wilhelm, O., & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. *Intelligence, 30,* 537–554.

Received June 23, 2008
Revision received February 4, 2009
Accepted February 6, 2009 ■