Available online at www.sciencedirect.com
SCIENCE DIRECT
Intelligence 33 (2005) 67–81
INTELLIGENCE

Working memory capacity and fluid abilities: Examining the correlation between Operation Span and Raven

Nash Unsworth*, Randall W. Engle

School of Psychology, 654 Cherry Street, Georgia Institute of Technology, Atlanta, GA 30332-0170, United States

Received 23 December 2003; received in revised form 19 August 2004; accepted 26 August 2004
Available online 6 October 2004

Abstract

The correlation between a measure of working memory capacity (WMC) (Operation Span) and a measure of fluid abilities (Raven Advanced Progressive Matrices) was examined. Specifically, performance on Raven problems was decomposed by difficulty, memory load, and rule type. The results suggest that the relation between Operation Span and Raven is fairly constant across levels of difficulty, memory load, and rule type. Thus, it appears something other than the number of things that can be held memory is important for the shared variance between these two tasks. The results are discussed in terms of the importance of attention control as a possible link between working memory capacity and fluid abilities.

Keywords: Operation Span; Raven; Working memory capacity

A large body of evidence has accumulated over the last decade supporting a substantial relationship between working memory capacity (WMC) and general fluid intelligence (gF; Ackerman, Beier, & Boyle, 2002; Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999; Kyllonen & Christal, 1990). The exact cause of this relationship, however, remains a mystery. The goal of the present study was to better determine the relationship between these two constructs by attempting to identify what variables are important for the relationship. In this regard, we utilized a post-hoc decomposition of the Raven Advanced Progressive Matrices (Raven; Raven, Raven,

* Corresponding author.
E-mail addresses: gtg039d@prism.gatech.edu (N. Unsworth), randall.engle@psych.gatech.edu (R.W. Engle).

0160-2896/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2004.08.003

& Court, 1998) and examined the relation between these decomposed variables with Operation Span (Ospan; Turner & Engle, 1989).

1. Working memory capacity and fluid abilities

Although we realize that no task is a pure reflection of the construct of interest, we examined the relation between Ospan and Raven for several reasons. Across several studies, the correlation between working memory (WM) span measures, such as reading and counting span, with Raven is typically around 0.30 (Conway et al., 2002; Engle et al., 1999; Kane et al., 2004). In these same studies, Ospan tends to correlate with Raven at about 0.34 (i.e., 12% shared variance). Furthermore, factor analyses demonstrate that Ospan loads highly on a WM factor and Raven loads highly on a gF factor, with the path coefficient between the two hovering around 0.60 (Conway et al., 2002; Engle et al., 1999; Kane et al., 2004). Thus, although a moderate first order correlation exists between the two measures, a substantial amount of variance seems to be shared between the two constructs. We hoped to shed light on this shared variance by examining what variables are important for the relationship between putative measures of each construct (i.e., Ospan and Raven).

With this goal in mind, several different research strategies can be used. One strategy is to manipulate a theoretically important aspect of one of the tasks (e.g., the difficulty of processing on the WM span task) and see how that manipulation affects the correlation between the two. Finding that equating participants on the processing component does not affect the correlation between WM span and higher-order cognition would suggest that processing efficiency does not account for the shared variance (Conway & Engle, 1996).

Another strategy that has become popular in determining the relationship between WMC and gF is the use of CFA and SEM techniques. Here, a set of latent variables are defined by a set of tasks thought to reflect those constructs. After the latent variables are defined, several theoretically plausible models are tested to see which model best fits the data. Here, researchers can test the role of short-term memory in the WM–gF relationship (Engle et al., 1999) or test for the possible role of processing speed in the relationship (e.g., Ackerman et al., 2002).

A third strategy, and the one employed in the current paper, is to examine the simple correlation between two tasks by examining different aspects of performance on one of the tasks and seeing how the correlation changes. For instance, Salthouse (1993, Experiment 1) examined the correlation between each item on Raven with both age and WMC. Salthouse (1993) found that the correlation between solution accuracy for each problem and a composite measure of WM was fairly constant across all problems. The same pattern of results held true for the correlations involving age. However, once the WM composite was partialled out of the analysis, the age correlations dropped to near zero. These findings are striking, particularly in light of the fact that Raven problems are arranged systematically such that the easiest items (highest average solution accuracy) are presented first and the most difficult items (lowest average solution accuracy) are presented last. Based on this evidence, it would seem that item variation in terms of difficulty is not a major factor in the WM-Raven correlation. Although, Salthouse (2000) has subsequently found that more difficult items do share some unique variance with age. In this study, Salthouse grouped items into quartiles based on solution accuracy and examined the effects of each quartile after controlling for the earlier quartiles. Each successive quartile accounted for a

small, but significant portion of the variance with age. These studies suggest that item variation in difficulty contributes little to individual differences in WMC and age and that some other factor accounts for most of the variance.

However, Carpenter, Just, and Shell (1990) argued that item difficulty is an important aspect of performance. These authors performed a thorough item analysis of Raven Progressive Matrices and found that an item's difficulty level (as evidenced by its error rate) was due to the number of rule "tokens" that were needed to complete a given problem. Working memory, they argued, is important for maintaining the rule tokens. Specifically, Carpenter et al. argued that the most difficult problems are those that place a heavy burden on WM resources. A problem that requires two rules, or two instantiations of the same rule, should be more difficult than a problem requiring only one rule token. Indeed, based on a classification of the number of rule tokens required for each problem, Carpenter et al. showed that the most difficult problems (again, as evidenced by item error rate) also tended to require the most rule tokens. Therefore, the authors suggested that the amount of information that can be maintained in WM would be an important indicator of reasoning ability.

In order to test this hypothesis, Carpenter et al. constructed two computer simulation models: FAIRAVEN and BETTERAVEN. The authors' suggested that the two models only differed in the fact that BETTERAVEN was better at abstract relations and could hold a larger set of goals in WM. The simulations demonstrated that FAIRAVEN could only solve the first half of the test and hence could only solve the easiest problems, whereas BETTERAVEN solved nearly all of the problems. Accordingly, the authors argued that "One of the main distinctions between higher scoring subjects and lower scoring subjects was the ability of the better subjects to successfully generate and manage their problem-solving goals in working memory" (p. 428). Based on this, it would seem that what is critical for performance on Raven as well as for the shared variance between Raven and measures of WMC is the number of items that can be held in working memory. In our first two analyses, we attempt to shed light on this hypothesis by examining item variations in both difficulty and the number of rule tokens.

Not only can item variations in difficulty and the number of relations, or rule tokens, be examined in this post-hoc manner, but so can other important aspects of performance such as differences in error patterns attributable to differences in rule utilization. Carpenter et al. (1990) identified five different rule types that are involved in solving Raven problems. The five classified rule types, ordered in terms of complexity, are: (1) constant in a row, in which an aspect of the figure stays the same across a row, but changes down a column; (2) quantitative pairwise progression, in which there is a quantitative increment in the figure in adjacent cells; (3) figure addition/subtraction, where parts of the figures are either added or subtracted from one another; (4) distribution of three, in which three categorical attributes of a figure (e.g., all circles) occur within each row; and (5) distribution of two, just like distribution of three except that only two values are distributed throughout a row. By examining the error patterns on these different types of rules, Carpenter et al. (1990) found that most subjects made the most errors on problems involving either distribution of three or two rules. As noted previously, Carpenter et al. (1990) argued that difficult items are difficult because these items require a large number of items to be held in working memory and because these items require more difficult rules.

In a recent post-hoc analysis, Babcock (2002) investigated age related differences on Raven by examining the types of rules required to solve each problem as classified by Carpenter et al. (1990). Babcock (2002) studied whether this rule classification was important to age differences on Raven. That is, are performance differences between older and younger adults restricted to problems requiring a certain type of rule, which upon identification could shed light on the possible mechanisms responsible

for such age differences? However, Babcock found that relatively few subjects attempted problems requiring more complex rules (i.e., distribution of two and three) and thus these problems could not be analyzed. Babcock (2002), also found that high and low ability groups (as determined by their Raven score) differed in their error patterns on problems requiring different rule types. In interpreting these results, Babcock suggested that a possible explanation for the results lay in terms of differential working memory capacities. Specifically, Babcock suggested that older and younger adults may differ in the overall amount of processing resources available to them and thus age differences on Raven are due to quantitative rather than qualitative differences. Once again, this line of reasoning implies that the reason WM measures correlate with Raven is due to the amount of things that can be held in memory.

1.1. Rationale for the present study

The purpose of the present study was to examine the role of individual differences in working memory capacity and fluid intelligence. Specifically, our aim was to examine the hypothesis that the shared variance between working memory span measures and measures of fluid intelligence is due to the number of goals and sub-results that can be held in working memory (Carpenter et al., 1990; Verguts & De Boeck, 2002). Indeed, Verguts and De Boeck (2002) noted that "Persons with a large WM capacity can store more partial results, and, hence, will have a higher probability of solving an item. Therefore, WM capacity and Raven performance are positively correlated" (p. 38). This line of reasoning suggests that the correlation between solution accuracy and a measure of working memory capacity should increase as the number of rules, goals, and/or sub-results on a given problem increases (given that there is enough systematic variability present). That is, items with low memory loads will not exceed even the capacity of low WM span participants and thus most individuals should get these problems right and there should be little systematic variability present. However, as memory load increases so will item discriminability and thus the item-WM span correlations will increase. It is also possible that the pattern of results will resemble those of Salthouse (1993) suggesting that the correlation between working memory span and fluid abilities is rather constant across different types of problems. With this goal in mind, we examined item variations in difficulty, memory load, and rule type in order to better understand the shared variance between these two measures.

2. Method

2.1. Participants

A total of 160 participants were recruited from the subject-pool at Georgia Institute of Technology and from the Atlanta, GA community through newspaper advertisements. Participants were between the ages of 18 and 35 and received either course credit or monetary compensation for their participation. Each participant was tested individually in a laboratory session lasting approximately 1 h.

2.2. Materials and procedure

After signing informed consent, participants completed Ospan (Turner & Engle, 1989) and then Raven Advanced Progressive Matrices, Set II (Raven et al., 1998).

2.2.1. Operation Span

Ospan has demonstrated good reliability and validity (Conway et al., 2002; Engle et al., 1999; Klein & Fiss, 1999). Specifically, previous research has demonstrated that Ospan has good test–retest reliability (e.g., 0.88; Klein & Fiss, 1999) as well as good internal consistency with estimates ranging from 0.61 to 0.83 (Conway et al., 2002; Engle et al., 1999; Klein & Fiss, 1999). Furthermore, as noted previously, the validity of Ospan has been demonstrated in several contexts by showing that it correlates well with other measures of WMC and predicts performance on a number of higher-order cognitive tasks (Conway et al., 2002; Engle et al., 1999).

The Ospan requires participants to solve a series of math operations while trying to remember a set of unrelated words. Participants see one math operation word string at a time, centered on a computer monitor. For each trial, they read aloud and solve the math problem and then read aloud the word. Immediately after the participant reads the word, the next operation-word string is presented. The operation-word strings are presented in sets of two to five items. 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. Three trials of each set size are presented, with the order of set size varying randomly, so that participants cannot predict the number of items. A participant's Ospan score is calculated by adding up the number of items in perfectly recalled trials. For example, a participant who correctly recalled two sets of two-item trials and one set of three-item trials would have a score of seven. Additionally, in order to ensure that participants are not trading off between solving the operations and remembering the words, an 85% accuracy criterion on the math operations is required for all participants.

2.2.2. Raven Advanced Progressive Matrices

As noted previously, Raven is a paper-and-pencil measure of abstract reasoning. The test consists of 36 individual items 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 display of 3×3 matrices of geometric patterns with the bottom right pattern missing. The task for the participant is to select among eight alternatives, the one that correctly completes the overall series of patterns. Participants were allotted 30 min to complete as many items as possible. A participant's score is the total number of correct solutions.

3. Results

As noted in the introduction, the goal of the present investigation was to better understand which aspects of performance on the Raven are important for the association with working memory capacity and in particular, with performance on Ospan. Therefore, here, we present three sets of separate analyses, each examining a different aspect of Raven performance. The first two sets of analyses are based on correct item responses and concern item variations in difficulty and number of rule relations. For these

Table 1
Means, standard deviations, minimum, maximum and correlation for Ospan and Raven

Variable	Mean	Standard Deviation	Min	Max	1	2
1. Ospan	13.25	6.58	2	39	–	
2. Raven	18.93	7.30	0	32	0.335	–

two aspects, we utilized correlation and regression analyses. The last analysis is based on error responses and concerns the different types of rules induced on each problem. For this last aspect, we utilized analysis of variance (ANOVA) methods. Table 1 provides descriptive statistics for both Ospan and Raven. Additionally, the correlation between Ospan and Raven was approximately 0.34, which is similar to correlations previously reported in the literature (e.g., Conway et al., 2002; Engle et al., 1999; Kane et al., 2004).[1]

3.1. Item variations in difficulty

Our first set of analyses concerns item variations in difficulty. Recall that a common account of the correlation between WMC measures and Raven is attributed to differential ability to solve difficult items that tax working memory. Thus, as problems increase in difficulty, so does the demand on working memory. Note that this is a common account of the correlation between these two measures and, thus, although we do not necessarily endorse this hypothesis, we wished to test the veracity of it. By this rationale, then, we should see the point-biserial correlation between a measure of working memory capacity and solution accuracy increase as difficulty increases. Note, that this hypothesis is based on the assumption that there is systematic variability present on the most difficult items. The Raven is constructed such that difficulty increases as problem number increases and as can be seen in Fig. 1, that was the case with these data, namely, proportion correct decreased sharply from early to late items. Therefore, if WMC is more important with greater difficulty (given that there is sufficient systematic variability in performance on an item), plotting the point-biserial correlations between WMC and solution accuracy by problem number should reveal an increase in the correlations that mirrors the function in Fig. 1.

However, as shown in Fig. 2, the correlations between solution accuracy for each item and Ospan, although fluctuating widely, does not appear to increase in any systematic manner as difficulty increases. Indeed, the correlation between Ospan and accuracy on the first problem was as high as with problem 24 (i.e., problem 1 r=0.26, problem 24 r=0.26). These results are strikingly similar to those of Salthouse (1993) who showed roughly the same pattern of correlations between solution accuracy and a WM composite. Both sets of results suggest that there is not a clear relationship between item variations in difficulty on Raven and measures of WM.

Following the lead of Salthouse (2000), we computed quartiles based on the solution accuracy for each Raven problem. The first quartile (quartile 1) represents the nine easiest problems while the last quartile (quartile 4) represents the nine hardest problems (according to accuracy rates). Descriptive statistics for the quartiles and Ospan are presented in Table 2. Performance values for the quartiles are based on proportion correct. Also shown in Table 2 are the correlations between the quartiles and Ospan.

[1] Note that this correlation is nearly identical to a correlation from a much larger sample of participants from our lab who were tested under the exact same testing conditions (e.g., r(1042)=0.349).

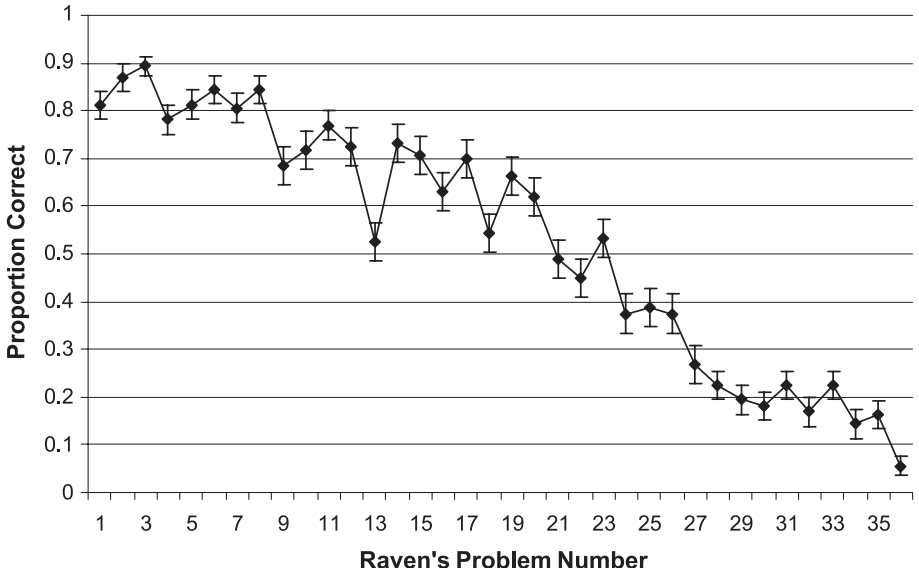

Fig. 1. Mean proportion correct for individual Raven problems. Error bars represent one standard error of the mean.

What is particularly notable about the correlations is that across quartiles 1–3, the correlations are very similar (i.e., approximately 0.30). However, quartile 4, which represents the hardest problems shows a non-significant correlation with Ospan. Although there seems to be adequate variability for quartile 4, this low correlation is probably due to the fact that not as many subjects attempted these problems. Indeed, 80% of participants attempted the first 27 problems, but only 47% of participants finished the test. Thus, only quartiles 1–3 should be interpreted. With this in mind, the results demonstrate that the

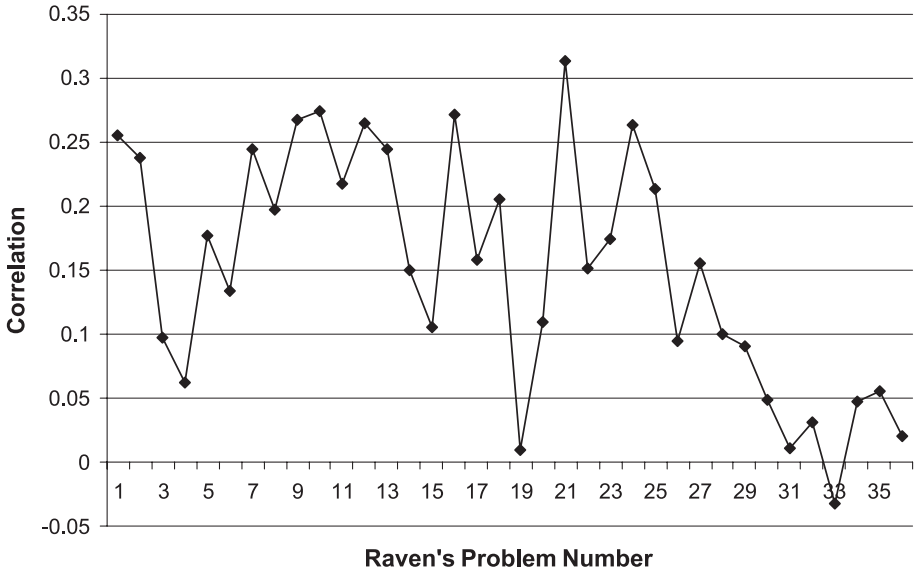

Fig. 2. Point-biserial correlations of solution accuracy with Operation Span for individual Raven problems.

Table 2
Means and standard deviations for accuracy by quartiles and correlations with Ospan

Variable	Mean	Standard Deviation	1	2	3	4	5
1. Ospan	–	–	–				
2. Quartile 1	.83	.21	.324**	–			
3. Quartile 2	.69	.28	.294**	.774**	–		
4. Quartile 3	.44	.30	.331**	.571**	.718**	–	
5. Quartile 4	.18	.20	.079	.262**	.354**	.439**	–

** $p<.01$.

correlation between solution accuracy and Ospan does not increase as difficulty increases but instead remains fairly constant across increasing levels of difficulty.

3.2. Item variations in memory load

The previous set of analyses suggested that the correlation between Raven and Ospan remains fairly constant across levels of difficulty. According to Carpenter et al. (1990), variations in item difficulty are one of the main determinants of performance differences between individuals. Specifically, Carpenter et al. argued that difficult items are difficult because they tax working memory more than easier items. In particular, the authors argued that difficulty increased because the number of rules and rule relations within a given problem increased. Thus, those who perform well on Raven would be those who have larger working memory capacities. Recall, that as a test of this claim, Carpenter et al. (1990) devised two computer simulation models: one of which could solve the first half of the test (FAIRAVEN) and one that could solve practically all items on the test (BETTERAVEN). The second model differed from the first primarily in terms of its ability to hold a larger number of goals in memory.

Using the logic of Carpenter et al., problems that BETTERRAVEN can solve should discriminate individuals better than problems that FAIRAVEN can and thus BETTERAVEN should correlate better with Ospan than FAIRAVEN. That is, as the number of relations increases, so should the correlation with a WMC measure. Therefore, based on the classification provided by Carpenter et al. (1990), we made two composites, one consisting of the problems that both FAIRAVEN and BETTERAVEN could solve and the other based on only those problems that BETTERAVEN could solve. The correlations suggest that those problems that could be solved by both models actually correlated better with Ospan than problems that only BETTERAVEN could solve (i.e., r Both=0.36**, r BETTERAVEN only=0.16*). Note two things about these correlations that could hinder interpretation: (1) as noted previously only 47% of participants finished the test and thus the low correlation for BETTERAVEN could be due to the low number of participants who attempted these problems and (2) it is possible that there is not enough variability present for BETTERAVEN and hence the correlation will be low. In regards to the first point, we performed the same analysis but only with those 76 participants who completed the test and thus attempted these problems.[2] The same pattern of

[2] One reviewer was concerned that only high working memory capacity individuals would finish the test. However, of those participants classified as high working memory (one standard deviation above the mean on Ospan), only 25% of them actually finished the test, whereas 71% of those classified as low working memory (one standard deviation below the mean on Ospan) finished the test. This results in somewhat lower scores for these 76 individuals on the two measures as compared the full sample (i.e. M Ospan=11.12, S.D.=5.90; M Raven=17.50, S.D.=7.59).

Table 3
Means and standard deviations for accuracy by tokens and correlations with Ospan

Variable	Mean	Standard Deviation	1	2	3	4	5
1. Ospan	–	–	–				
2. Token 1	.72	.34	.342**	–			
3. Token 2	.66	.22	.276**	.640**	–		
4. Token 3	.32	.32	.245**	.440**	.533**	–	
5. Token 4	.25	.25	.113	.317**	.507**	.490**	–

** $p<.01$.

correlations emerged (i.e., *r* Both=0.31**, *r* BETTERAVEN only=0.16). Thus, the low correlation does not seem to be due to the fact that only a few participants attempted the last problems. In terms of the second point there does seem to be adequate variability for both composites. Looking only at those 76 participants who attempted the test the standard deviation for Both=0.24 and the standard deviation for BETTERAVEN=0.22. Furthermore, there is some systematic variability present for both because the correlation between the two is 0.67.[3] Thus, the low correlation between Ospan and BETTERAVEN seems to reliable and thus it would seem that what is important for the correlation between Ospan and Raven is due to variability in primarily the fist half of the test.

As a more sensitive test of the hypothesis that item variations in memory load are important for the WM measure-Raven correlation, we grouped items based on the number of "rule tokens" and examined the correlations with these variables and Ospan. Items were grouped based on the appendix provided by Carpenter et al. (1990). Those items with only one rule token were grouped together and formed token 1, while items with two tokens made up token 2 and so on up to problems with five tokens. However, only one problem was classified as having five rule tokens by Carpenter et al. and hence was not analyzed in the current data set. Once again, we would expect that as memory load increased so would the correlation with Ospan (as long as there is enough systematic variability present). That is, if the number of things that can be held in memory is what is important for the relationship between Ospan and Raven, then we should see that problems with highest number of rule tokens correlates the best with Ospan. However, as shown in Table 3, the correlation between the number of tokens and Ospan actually decreased as the number of rule tokens increased. The correlation went from a significant 0.34 for token 1 to a non-significant 0.11 for token 4 (Table 3).

However, once again, we must be cautious in interpreting this result because although 80% of participants attempted problems with 1–3 rule tokens, fewer participants attempted problems making up token 4. Therefore, as before, we examined the correlation between token 4 and Ospan only for those subjects who finished the test. The resulting correlation was 0.14. Additionally, there does seem to be some systematic variability for these problems because the average correlation between the tokens 1–3 with token 4 was 0.52.

Thus, the results suggest that problems with only one rule token are more important for the relationship with Ospan than problems with higher memory loads. Indeed, as shown in Table 4, entering all four token types into a simultaneous regression reveals the fact that only problems with one rule token predict unique variance in Ospan (i.e., 4%). This would suggest that, of the roughly 13% of

[3] The same pattern of results was found when examining data from the full sample of participants.

Table 4
Simultaneous regression predicting Ospan

Variable	*B*	*t*	sr^2	R^2	*F*
Token 1	0.260	2.63**	0.04		
Token 2	0.080	0.723	0.003		
Token 4	0.122	1.30	0.009		
Token 5	−0.070	−0.765	0.003	0.133	5.94**

** $p<0.01$.

variance shared between Ospan and Raven, problems with only one rule token uniquely predict 30% of that variance and 70% of the variance is shared by all rule tokens. Similar to the analysis of difficulty, it seems that the easiest and lowest memory load problems account for a large amount of both shared and unique variance with Ospan (Table 4).

3.3. Error responses as a function of rule type

The final set of analyses concerned the extent to which subjects rated high, medium, or low on the Operation Span differed on problems requiring different rules. As classified by Carpenter et al. (1990), there are five basic types of rules ordered in terms of complexity (see previous discussion). One might suspect that the different span groups perform similarly on problems requiring certain rules, while performance diverges on rules of another type. If for instance, high and low WM spans differ on problems requiring distribution of two, but not on problems requiring pair-wise progression, we would be able to make some inference about the underlying differences between the span groups.

Participants were classified as being either a high, medium, or low span based on their performance on Ospan. Those participants scoring one standard deviation above the mean were deemed high spans, whereas those scoring one standard deviation below the mean were considered low spans. All other participants were considered mid spans. This resulted in 28 low spans with a mean Ospan score of 5.25 (S.D.=1.67, range 2–7), 108 mid spans with a mean Ospan score of 12.81 (S.D.=3.44, range 8–19), and 24 high spans with a mean Ospan score of 24.54 (S.D.=5.52, range 20–39). Only problems that were attempted and incorrect were considered to see if the span groups differed in error rate for each of the different types of rules.

As shown in Fig. 3, the results suggest that low spans seem to make more errors than do either mid or high spans. Additionally, the proportion of errors is greater for some rules than for others. Specifically, it seems that rules involving either distribution of two or three have a higher proportion of errors than the other rule types. Crucially, however, the proportion of errors for the different rule types is not a function of WM span. Thus, the span groups do not differ on one rule type more than another.

These impressions were confirmed by a 3 (span)×5 (rule type) repeated measures ANOVA with rule type as the within subjects variable. The ANOVA yielded reliable main effects of both span, $F(2,157)=11.16$, $p<0.01$, partial $\eta^2=0.13$, and rule type, $F(4,628)=39.75$, $p<0.01$, partial $\eta^2=0.20$. Post-hoc Bonferroni corrected contrasts revealed that low spans had a higher proportion of errors ($M=0.55$, S.E.=0.04) than both mid and high spans ($p<0.01$). High and mid spans did not differ in proportion errors ($M=0.29$, S.E.=0.05 and $M=0.33$, S.E.=0.02, respectively). Additionally, Bonferroni corrected contrasts in terms of rule type indicated that the proportion of errors was significantly different for all rule types ($p<0.01$) except for constant and add/subtract rules, which did not differ from one another.

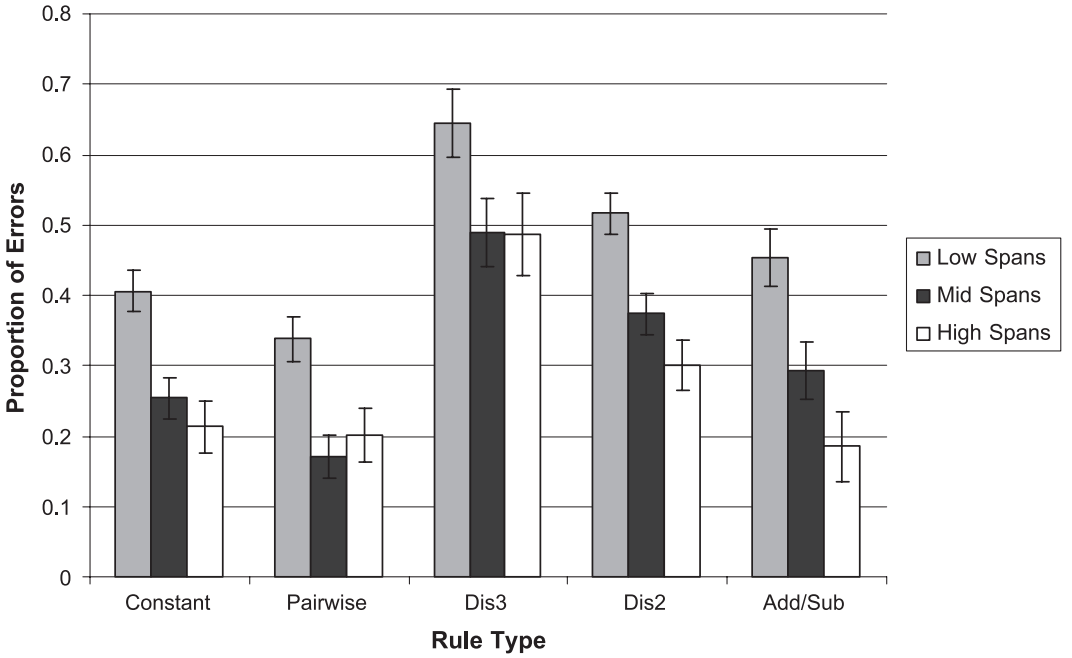

Fig. 3. Mean proportion of errors as a function of rule type and working memory span. Note: constant=constant in a row, pairwise=quantitative pairwise progression, Dis3=distribution of three, Dis2=distribution of 2 and Add/Sub=addition/subtraction. Error bars represent one standard error of the mean.

Crucially, the span×rule type interaction did not reach significance, $F(8,628)=0.953$, $p>0.20$, partial $\eta^2=0.01$. Together, the results suggested that, although there are differences between the span groups in terms of proportion of errors as well as differences between the different rule types, these two factors did not interact.

4. Discussion

The goal of the present study was to investigate possible determinants of the correlation between Operation Span and Raven Advanced Progressive Matrices. The results demonstrated that item variations in difficulty do not account for the shared variance based on the standard view that more difficult problems are more important for the correlation than easier problems. In fact, the results suggest that the correlation holds fairly constant across the first three quartiles of difficulty and then drops substantially for the hardest problems, although, this latter effect is possibly due to the low number of subjects who actually attempted these problems as well as possibly low systematic variability present for these problems. Regardless, these results are consistent with the work of Salthouse (1993) who demonstrated a similar pattern of correlations with a composite measure of working memory capacity and performance on the Raven, suggesting that the correlation between working memory capacity and solution accuracy was constant across differing levels of difficulty.

The results also demonstrated that item variations in memory load do not account for the shared variance between the two tasks. As memory load increased, the correlations between Ospan and solution

accuracy actually decreased somewhat, resulting in a non-significant correlation for problems with the largest memory loads. These results are contrary to what would be predicted if the shared variance between the two tasks was due to differences in the number of items that could be held in memory. By such a view, one would expect an increase in the correlations. However, the results suggest that, for the most part, the correlations are fairly constant and do not vary systematically with variations in memory load. These results are compatible with the work of Verguts and De Boeck (2002, their Experiment 1) who demonstrated that the correlation between a working memory task and a modified version of Raven occurred even when all of the items had a low memory load.

Finally, the ANOVA results suggest that item variations in rule type also do not seem to account for working memory capacity individual differences and performance on the Raven. That is, the results demonstrated that although low working memory span individuals made more errors overall, and some rule types resulted in more errors than others, these two factors did not interact. Thus, individual differences in working memory capacity did not result in differential performance on certain rules more than others.

It is important to point out that only one measure of working memory capacity was used in the current study and thus the generalizability of the findings are somewhat suspect. However, we note that the Operation Span task is a widely used measure of working memory capacity that has been shown to load highly on a common working memory factor (e.g., Conway et al., 2002; Engle et al., 1999; Kane et al., 2004), and as noted in the introduction the zero order correlation between Operation Span and Raven is similar to those of other working memory tasks to Raven. Additionally, the mean Ospan score of 13.25 is similar to other Ospan scores that have been reported previously (e.g., Engle et al., 1999) and is virtually identical to the mean Ospan score from a much larger distribution in our laboratory (e.g., M=13.72, N=2256). Thus, although only one putative measure of working memory capacity was used in the current study, we feel that the results can be generalized to other working memory span tasks to some degree.

4.1. Working memory capacity and fluid abilities revisited

Taken together, the results of the present study strongly suggest that the number of goals or sub-results that can be held in memory does not account for the shared variance between working memory span measures and fluid intelligence. Thus, the results do not support the hypothesis advanced by Carpenter et al. (1990) that the link between individual differences in working memory capacity and intelligence is due to differences in the ability to hold a certain number of items in working memory. Note that we are not arguing that item variations in difficulty and memory load are unimportant in determining performance on the Raven as suggested by Carpenter et al., but rather we suggest that the reason working memory span tasks are consistently good predictors of fluid ability is due to something else.

Indeed, we have argued elsewhere (Engle et al., 1999; Heitz, Unsworth, & Engle, in press) that the shared variance between working memory capacity and fluid abilities is due to the ability to control attention. This framework suggests that those individuals who score high on a working memory capacity measure are those individuals who are better able to control attention especially in conditions of distraction and interference. This notion is similar to the theory of goal neglect and fluid intelligence put forth by Duncan, Emslie, and Williams (1996).

By our view, it is the central executive component of the working memory system that is important on both working memory span tasks and tasks of fluid abilities. Other researchers have also suggested that individual differences in fluid abilities are due to differences in a general control processor (Embretson, 1995; Marshalek, Lohman, & Snow, 1983; Sternberg, 1985). Embretson (1995) examined the roles of both

general control processing and working memory capacity on the performance of Raven type problems. Similar to the analyses performed here, Embretson utilized the framework outlined by Carpenter et al. (1990) to classify problems based on their memory load and then examined how performance was related to both general control processing and memory load. Instead of examining Raven problems directly, Embretson developed a test bank of items based on Carpenter et al.'s classification scheme (see also Embretson, 1998). In order to examine the roles of both working memory capacity and general control processes, Embretson (1995) argued for the standard view of working memory capacity; that is, the amount that can be held in memory. Thus, the need for working memory capacity should vary systematically with item variations in memory load. For general control processes, however, Embretson argued that this would be needed equally on all problems and thus would be a constant in terms of item variations. Specifically, Embretson argued that, "maintaining an effective solution strategy, depends on control processes, such as selecting an effective strategy, monitoring solution processes, and allocating resources to processing. Such processes are postulated to be involved equally on all items within an item type" (p. 170). Impressively, Embretson demonstrated that together, both constructs accounted for 92% of the variance in reasoning, with general control processes accounting for more variance than working memory capacity. Based on this, Embretson suggested that both are important for fluid abilities, but that control processes were more important.

Our view of working memory capacity is similar to that of Embretson's control process. That is, we have suggested that individual differences in working memory capacity are really indicative of differences in a domain-general executive attention component and not indicative of differences in the number of things that may held in memory. Although, some of the shared variance between WM span tasks and measures of fluid abilities such as Raven is probably due to processes other than executive attention. Furthermore, the results of the present study are highly compatible with Embretson's notion that the general control process should be evident equally and equally important across all items. Recall, that we found that the correlation between a measure of working memory capacity and solution accuracy on the Raven was fairly constant across item variations in difficulty and memory load. Additionally, examining error responses on the different types of rules suggested that item variations in rule type were unimportant for the relationship as well. Together, these results suggest that individual differences in working memory capacity did not vary systematically with different item variations, but rather were constant across the different types of problems.

Another possible explanation for the results of the current study is a notion proposed by Verguts and De Boeck (2002) who suggested that it is not only the ability to hold goals and sub-results for a given problem during the solution of that problem, but also it is the ability to successfully hold onto the solution principles of that problem for future use that is important. That is, Verguts and De Boeck suggested that one reason for the shared variance between working memory capacity measures and intelligence is the ability to acquire and reuse correct solution strategies across problems of a similar type. In support of this notion, the authors found that the correlation between working memory capacity and performance on Raven remained even when all of the items had a low memory load. Furthermore, the authors argued that evidence obtained by Carlstedt, Gustafsson, and Ullstadius (2000) was consistent with their view. Specifically, Carlstedt et al. (2000) showed that an intelligence test made up of homogenous items loaded higher on a general intelligence factor than did a similar test made up of heterogeneous items. Verguts and De Boeck (2002) argued that subjects were learning the correct solution strategies and then utilizing them on subsequent problems. Thus, in this view, one reason for the correlation between working memory capacity and intelligence is differential ability to successfully

implement learned solution strategies on problems that require similar solution strategies in the future. It is not the amount that can be held for a given problem, but rather the ability to learn and implement strategies across problems that is important.

However, it is possible that the reason homogenous reasoning tests load more highly on a g-factor than heterogeneous tests is because homogenous tests allow for more proactive interference (PI) to build up during the test and those individuals who are better at combating PI during the test score higher. Thus, the shared variance between these types of intelligence tests and working memory capacity may be due to the fact that susceptibility to PI is an important source of individual differences in both. Indeed, there is abundant evidence suggesting that one aspect of individual differences in working memory capacity is the ability to effectively combat PI (Hasher & Zacks, 1988; Kane & Engle, 2000; Lustig, Hasher, & May, 2001; Rosen & Engle, 1998). This same ability has also been suggested as an important factor in intelligence and cognitive abilities more generally (Dempster, 1991; Dempster & Corkill, 1999). Future research will be needed to test these two theories of the correlation between working memory and intelligence more thoroughly.

Acknowledgement

This work was supported by Grant F49620-00-1-131 from the Air Force Office of Scientific Research.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology. General*, *131*, 567–589.

Babcock, R. L. (2002). Analysis of age differences in types of errors on the Raven's advanced progressive matrices. *Intelligence*, *30*, 485–503.

Carlstedt, B., Gustafsson, J. -E., & Ullstadius, E. (2000). Item sequencing effects on the measurement of fluid intelligence. *Intelligence*, *28*(2), 145–160.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven progressive matrices test. *Psychological Review*, *97*, 404–431.

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.

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory*, *4*, 577–590.

Dempster, F. N. (1991). Inhibitory processes: A neglected dimension of intelligence. *Intelligence*, *15*(2), 157–173.

Dempster, F. N., & Corkill, A. J. (1999). Individual differences in susceptability to interference and general cognitive ability. *Acta Psychologica*, *101*, 395–416.

Duncan, J., Emslie, H., & Williams, P. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology*, *30*, 257–303.

Embretson, S. E. (1995). The role of working memory capacity and general control processes in intelligence. *Intelligence*, *20*, 169–189.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396.

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.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, *vol. 22* (pp. 193–225). San Diego, CA: Academic Press.

Heitz, R. P., Unsworth, N., & Engle, R. W. (in press). Working memory capacity, attention, and fluid intelligence. In O. Wilhelm, & R. W. Engle (Eds.). *Understanding and measuring intelligence*. NY: Sage.

Kane, M. J., & Engle, R. W. (2000). Working memory capacity, proactive interference, and divided attention: Limits on long-term retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*, 33–358.

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

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

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

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*, 107–127.

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

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

Salthouse, T. A. (1993). Influence of working memory on adult age differences in matrix reasoning. *British Journal of Psychology*, *84*, 171–199.

Salthouse, T. A. (2000). Item analyses of age relations on reasoning tests. *Psychology and Aging*, *15*, 3–8.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

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

Verguts, T., & De Boeck, P. (2002). On the correlation between working memory capacity performance on intelligence tests. *Learning and Individual Differences*, *13*, 37–55.