Learning and Individual Differences 22 (2012) 875–881

Contents lists available at SciVerse ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Computer-paced versus experimenter-paced working memory span tasks: Are they equally reliable and valid?☆

Heather Bailey *

439 Psychology Building, One Brookings Drive, Campus Box 1125, Washington University, Saint Louis, MO 63130, United States

ARTICLE INFO

Article history:
Received 7 November 2011
Received in revised form 3 May 2012
Accepted 2 June 2012

Keywords:
Working memory
Fluid intelligence
Span tasks
Administration
Computer-paced

ABSTRACT

Working memory span tasks are popular measures, in part, because performance on these tasks predicts performance on other measures of cognitive ability. The traditional method of span-task administration is the experimenter-paced version, whose reliability and validity have been repeatedly demonstrated. However, computer-paced span tasks are becoming increasingly more popular. Despite their popularity, no study had systematically compared experimenter-paced and computer-paced versions of the reading span and operation span tasks. Such a comparison is important because research labs in many universities across many countries administer these span tasks with a variety of methods. The purpose of the present study was to evaluate the reliability and validity of computer-paced span tasks and to compare these estimates to those of experimenter-paced span tasks. Results indicated that experimenter-paced and computer-paced span tasks share some overlap, but also measure additional and distinct processes. Computer-paced span tasks were highly reliable measures as well as valid indicators of fluid intelligence (Gf). Thus, computer-paced span tasks may be the optimal type of administration given their methodological advantages over experimenter-paced span tasks.

1. Introduction

Working memory (WM) span tasks have been widely used over the past several decades because performance on these tasks predicts performance on higher-order cognitive tasks (Ackerman, Beier, & Boyle, 2005). One popular verbal span task is called the reading span (RSPAN) task. The original version of the RSPAN task developed by Daneman and Carpenter (1980) involved judging whether sentences (e.g., "Lori left the engine on and parked the grape.") made sense. After a series of sentences, participants were instructed to recall the final word from each sentence (e.g., grape), and word recall on the RSPAN task predicted reading comprehension ability. Another popular verbal span task, the operation span (OSPAN) task (Turner & Engle, 1989), requires participants to solve mathematical equations while remembering words (e.g., "Is (10/2) + 3 = 7?" CHAIR). Importantly, a meta-analysis conducted by Daneman and Merikle (1996) demonstrated that, regardless of whether the processing component involved solving math problems or reading sentences, span tasks with processing components significantly predicted comprehension.

1.1. Span-task administration

Although the processing components have been thoroughly investigated, less research on the administration of span tasks has been conducted. Studies evaluating WM have included a variety of self-paced, experiment-paced, and computer-paced span tasks. With self-paced administration, participants determine how long they spend processing information (e.g., reading sentences on the RSPAN task or solving equations on the OSPAN task) often by pressing a button when ready for the next sentence or equation. By contrast, experimenter-paced administration requires an experimenter to control when the next sentence or equation is presented. Although the experimenter must wait for the participant to complete the current trial (i.e., participants are allowed to read at their own pace), the experimenter ensures that the next trial begins immediately, so that no extra time elapses between trials. Finally, with computer-paced administration, all participants are given a set amount of time (e.g., 4 s) to read the sentence or equation and a set amount of time (e.g., 1 s) to read the to-be-remembered item. After this time has elapsed, the program proceeds to the next trial. Importantly, note that experimenter-paced span tasks must be administered *individually* with an experimenter, whereas both self-paced and computer-paced span tasks can be administered in a group setting.

Given these methodological differences, a systematic comparison of span administration methods is very important because different methods potentially could lead to discrepant findings. For instance, one current debate in the WM literature concerns the extent to which WM and secondary memory predict fluid intelligence (Gf). Secondary memory (SM), a term borrowed from James (1890) similar to long-term memory, involves information no longer in conscious awareness. Some researchers claim that the predictive power of WM span performance is not special and that SM tasks (e.g., paired-

☆ This research was supported by NIH grant F32 AG039162, PI Heather Bailey.
* Tel.: +1 314 935 4138; fax: +1 314 935 7588.
E-mail address: hroth@artsci.wustl.edu.

1041-6080/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.lindif.2012.06.004

associate recall) can serve as better indicators of Gf performance (Mogle, Lovett, Stawski, & Sliwinski, 2008). By contrast, Unsworth and Engle (2007) claim that the WM system is unique because it involves active maintenance of information in primary memory (PM) as well as retrieval of information from SM. Both sides have found evidence in support of their view: Mogle et al. (2008) found that performance on SM, not WM, tasks predicts Gf performance, whereas others have found that span performance accounts for unique variance in Gf performance above and beyond SM performance (Shelton, Elliott, Matthews, Hill, & Gouvier, 2010; Unsworth & Engle, 2006). These conflicting findings have led researchers to ask, why the discrepancy? One important difference between the study conducted by Mogle et al. (2008) and the studies conducted by Shelton et al. (2010) and Unsworth and Engle (2006) is the different types of span administration. The first group used self-paced span tasks whereas the second group used a version of computer-paced span tasks.

Given that different administration methods produced different findings, evaluating the reliability and validity of these types of task could shed light on whether span administration influences (1) span performance and (2) span–Gf relationships.

Previous research has evaluated whether methodological diferences between self-paced and experimenter-paced span tasks influenced span performance. Friedman and Miyake (2004) were first to directly compare these administration methods using the RSPAN task. They discovered that participants spent more time on the self-paced trials (which they referred to as "participant-administered") than on the experimenter-paced trials, and they speculated that increased reading times were due to the implementation of strategies (e.g., rehearsal). In fact as time spent on the RSPAN task increased, performance increased indicating that participants used the extra time on self-paced tasks to be strategic. Most important, experimenter-paced, but not self-paced, span performance was related to comprehension performance. However, after controlling for reading times, the correlation between self-paced span performance and comprehension increased. A study conducted by St. Clair-Thompson (2007) replicated and extended these findings to visuospatial span tasks (e.g., the Corsi span task): Performance on experimenter-paced tasks was more highly related to comprehension and mathematical performance than was performance on self-paced tasks. Finally, Lépine, Bernardin, and Barrouillet (2005) found that comprehension and mathematical abilities were more highly predicted by performance on computer-paced span tasks than by self-paced span tasks.

Taken together, these results indicate that both experimenter-paced and computer-paced span tasks are better measures of WM than are self-paced span tasks. In fact, some researchers have speculated that self-paced span tasks may not measure WM; rather, they are more similar to measures of STM because participants are given extra time to rehearse (Conway et al., 2005). However, given that researchers use experimenter-paced (Bunting, 2006; Kane, Poole, Tuholski, & Engle, 2006; La Pointe & Engle, 1990; McCabe, 2008) and computer-paced span tasks (Babcock & Salthouse, 1990; Hambrick & Oswald, 2005; Oberauer, 2005; Pardo-Vazquez & Fernandez-Rey, 2008) interchangeably in the literature (sometimes with little or no justification), these types of span administration should be systematically compared.

1.1.1. Experimenter-paced versus computer-paced administration

As mentioned above, experimenter-paced administration requires a researcher to be with the participant as they complete the span task. The advantage of this type of administration is that the researcher can monitor how well the participant attends to the processing component (e.g., reading sentences or solving equations). Monitoring participants' attention could be important because span performance is based upon memory for the words (not memory for the sentences or equations), and some participants may become aware of this and devote more resources to remembering the words. Having an experimenter present with a participant provides further control during experimenter-paced span tasks. However, the disadvantage of experimenter-paced span tasks is that participants must be run individually. Individual administration of span tasks can be problematic because many WM studies involve a correlational design, which requires large sample sizes.

By contrast, computer-paced span tasks have the advantage of being administered in a group setting. Another advantage of computer-paced administration is the possibility of using span tasks in web-based studies because the presence of an experimenter is not required. Furthermore, computer-paced span tasks can be shared easily between collaborators at different research sites. However, one disadvantage of computer-paced administration is less control over participants' attention. Because an experimenter is not working individually with each participant, no one is monitoring whether participants are attending to the processing component.

Although experimenter-paced and computer-paced span tasks share many similarities (e.g., content, difficulty, processing and storage components, etc.), differences among these tasks may produce systematic changes in their convergent validity — the degree to which performance on these tasks correlates with performance on other tasks thought to measure the same construct. Such differences also may influence their criterion validity, or how well performance on the span tasks predicts other cognitive abilities. Given that these span tasks are presumed to measure the same construct, the goal of the present study was to compare experimenter-paced and computer-paced span tasks on performance, reliability, and ability to predict performance on measures of higher-order cognitive tasks.

Although little attention has been devoted to this issue, two relevant studies have been published. In both, new computer-paced versions of the OSPAN task were introduced and evaluated. In the first of these studies, De Neys, d'Ydewalle, Schaeken, and Vos (2002) introduced a computerized, group-administered version of Turner and Engle's (1989) OSPAN task, which they named the GOSPAN task. Participants completed both the traditional OSPAN task and the GOSPAN task. They found that the GOSPAN task was reliable (Cronbach's alpha = .74) and that GOSPAN performance was significantly correlated with OSPAN performance ($r = .50$, $r = .70$ after correcting for attenuation). Further, these results were replicated by Pardo-Vazquez and Fernandez-Rey (2008) using a Spanish version of the GOSPAN task. However, this computerized task allowed participants to pace how long they solved equations, akin to a self-paced span task. Moreover, no measures of higher-order abilities were administered, so they could not compare the predictive power of the OSPAN and GOSPAN tasks.

The second study was conducted by Unsworth, Heitz, Schrock, and Engle (2005). They created a different computerized version of the OSPAN task, which they named the automated OSPAN (Aospan). On the Aospan task, participants completed practice trials consisting of mathematical operations similar to those on the experimental trials. Practice trials allowed the authors to calculate an average time that it took each participant to solve the equations. For the experimental trials, the participant had a set time to solve the equations equal to their average practice time plus 2.5 standard deviations. Note that the Aospan task is a computer-paced span task with an individualized rather than a group fixed rate as in those computer-paced span tasks discussed above. Using individualized processing rates, Unsworth et al. (2005) found that performance on the Aospan task was significantly correlated with performance on the traditional OSPAN task ($r = .45$) and was internally consistent (Cronbach's alpha = .78). Moreover, Aospan performance significantly predicted performance on Raven Progressive Matrices ($r = .38$), a measure of Gf, which was similar to the magnitude of the observed correlation between OSPAN and Raven performance ($r = .42$). These results provided support for the use of a computer-paced OSPAN task; however, the same results have yet to be demonstrated for computer-paced RSPAN tasks.

Although Unsworth and colleagues have reported data concerning the construct validity of an automated version of the RSPAN task (Unsworth, Redick, Heitz, Broadway, & Engle, 2009), the automated and traditional versions have not been as thoroughly examined as the OSPAN tasks.

1.2. The present study

The goals of the present study were twofold. Because little evidence exists for the RSPAN task, the first goal was to evaluate the reliability and validity of computer-paced and experimenter-paced versions of the RSPAN task. The reliability and validity estimates for the RSPAN task also will be compared to those for the OSPAN task. In the present study, the computer-paced span tasks involved a group-fixed rate for the processing component (4 s for all participants) given that numerous recent studies have used this version (e.g., Barrouillet, Gavens, Vergauwe, Gaillard, & Camos, 2009; Colom, Rebollo, Abad, & Shih, 2006; Hambrick & Oswald, 2005; Rowe, Hasher, & Turcotte, 2008; Swets, Desmet, Hambrick, & Ferreira, 2007; Zeintl & Kliegel, 2007).

Most important, the second goal is to evaluate whether both types of span administration account for the *same* variance in Gf performance. Previous research has demonstrated that computer-paced and experimenter-paced versions of the OSPAN tasks accounted for a similar amount of variance in Gf performance, but no evidence speaks to whether they account for shared or unique variance, and the present study will do so.

2. Method

2.1. Participants

One hundred and twenty-five undergraduates (93 females) from introductory psychology courses at Kent State University participated as a course requirement. Mean age was 20.3 (SD = 5.4) years.

2.2. Materials

2.2.1. WM span tasks

All four span tasks were presented on a computer screen. The experimenter went through detailed instructions and examples, and then participants were given two practice trials. Each practice trial consisted of either two operation-word pairs (OSPAN) or two sentence-word pairs (RSPAN). Following the practice trials, participants read summarized instructions and then began the experimental trials. Each span task consisted of 15 experimental trials whose set sizes ranged from three to seven words. Following the final word of each trial, participants were instructed to recall the words in serial order by typing their responses into a text field onscreen, and they had unlimited time to recall the words. The words and the order of set sizes were initially randomized and that order was used for all participants.

2.2.1.1. Experimenter-paced operation span task (EP OSPAN). Participants read a mathematical operation aloud (e.g., "Is $(3\times2)+5=10$?"), and reported whether the answer provided was correct or not. The experimenter recorded their response by pressing "Y" for correct and "N" for incorrect. When the experimenter pressed a key, the to-be-remembered word (e.g., ROCK) appeared onscreen and the participant read the word aloud. Immediately thereafter, the experimenter pressed a key to present the next equation on-screen. Once the word was read aloud, the next operation appeared onscreen.

2.2.1.2. Experimenter-paced reading span task (EP RSPAN). Participants read a sentence aloud (e.g., "Mr. Owens left the lawnmower in the lemon.") and reported whether it was logical or illogical. Once the experimenter recorded the response, the to-be-remembered word (e.g., EAGLE) appeared onscreen and the participant read the word aloud. After the word was read aloud, the next sentence appeared onscreen.

2.2.1.3. Computer-paced operation span task (CP OSPAN). The materials used for the CP OSPAN task were the same as those used for the EP OSPAN task with two exceptions: (1) the mathematical equation appeared onscreen (e.g., "Is $(9\div3)-2=2$") with two buttons ("CORRECT" and "INCORRECT") and (2) participants had 4 s to indicate whether or not the answer provided was correct by clicking on one of the two buttons. After the 4 s had elapsed, a to-be-remembered word was presented for 1 s and then the next equation was presented onscreen.

2.2.1.4. Computer-paced reading span task (CP RSPAN). The CP RSPAN task involved a sentence presented onscreen for 4 s. In the time allotted, the participants pressed the "CORRECT" button if the sentence made sense and pressed the "INCORRECT" button if it did not make sense. After the 4 s had elapsed, a to-be-remembered word was presented for 1 s followed by another sentence for 4 s.

2.2.2. Raven Advanced Progressive Matrices (RAPM)

Participants completed 18 trials from the RAPM (Raven, Raven, & Court, 1998). The same trials used by Stanovich and Cunningham (1993) were used in the current experiment. In this task, a display of 3×3 matrices was presented on the computer screen. These matrices consisted of 8 geometric figures with the 9th figure (i.e., the bottom, right-hand figure) missing. Participants were given 8 choices of figures and were instructed to select the one that completed the horizontal and vertical patterns. Participants had 15 min to complete as many of the 18 displays as possible. Performance was based on the proportion of correctly answered items.

2.2.3. Locations test

The second measure of Gf was the locations test from the Kit of Reference Tests for Cognitive Factors (Ekstrom et al., 1976). In this task, participants were shown several problems, each of which consisted of five rows of dashes that were separated into groups. In each of the first four rows, one dash was replaced by an "X"; and in the fifth row, five of the dashes were replaced by numbers (i.e., labeled 1–5). Participants were instructed to find the rule for the location of the "X" in the first four rows of dashes and then to choose the correct location, which was denoted by a number, for an "X" in the fifth row. They completed 2 sets of problems and were given 6 min per set to complete as many problems as possible. They were instructed only to indicate an answer for those sets in which they knew the rule. Again, performance was computed as the proportion of correct answers.

2.3. Procedure

Participants were run individually through tasks. They completed two 1-hour sessions, separated by one week. Each session consisted of one experimenter-paced span task and one computer-paced span task — one of which was an RSPAN task and one was an OSPAN task. In session 1, participants completed an informed consent form, the EP OSPAN task, a demographics questionnaire, RAPM, the CP RSPAN, and the Locations test. In session 2, they completed the EP RSPAN task and the CP OSPAN task. This task order was the same for all participants.

3. Results

The goals of the present study were to (1) evaluate the reliability and validity estimates of computer-paced span tasks and compare

these estimates to those of experimenter-paced span tasks, and (2) evaluate whether both types of span performance account for the same variance in Gf performance. To do so, performance on the computer-paced and experimenter-paced span tasks is presented first followed by the reliabilities of the computer-paced span tasks. Next, the relationships among tasks were examined as well as their relationship with Gf performance.

3.1. Span performance

Performance on the span tasks was computed as the overall mean proportion correct, which was averaged across all trials and then across participants. Recall data were removed from analyses if a participant did not properly attend to the sentences and equations (accuracy<85%), which resulted in a loss of CP RSPAN data from 4 participants and CP OSPAN data from 11 participants. For the remaining data, descriptive statistics for span performance are reported in Table 1. Computer-paced performance ($M=0.66$, $SE=0.02$) was significantly higher than experimenter-paced performance ($M=0.51$, $SE=0.01$, $t(118)=10.30$, $p<.001$, $d=0.86$), which replicated the findings from both De Neys et al. (2002) and Unsworth et al. (2005).

3.2. Reliability

Previous research has established that experimenter-paced span tasks are reliable measures (see Conway et al., 2005), but less data exist for the reliability estimates for computer-paced span task. To assess the reliability (i.e., internal consistency) of each span task, Cronbach's alpha was computed using performance on all 15 trials. Cronbach's alpha for each task is presented in parentheses along the diagonal of Table 2 and indicated that both experimenter-paced and computer-paced span tasks are highly reliable.

3.3. Correlations among span tasks

The correlations among the span tasks are presented in Table 2. As expected, experimenter-paced span tasks were highly related, as were the computer-paced span tasks. Moreover, the RSPAN tasks were significantly correlated, and this correlation increased after correcting for attenuation ($r=.39$); the OSPAN tasks also were significantly correlated (corrected for attenuation $r=.63$). Because all four span tasks are thought to measure WM, these significant correlations provide evidence for the convergent validity of these measures.

3.4. Span–Gf correlations

Next, the predictive validity (i.e., the span–Gf relationship) of each type of measure was compared. Because no evidence exists regarding this comparison for RSPAN performance, these relationships were compared separately for RSPAN and OSPAN performance. To do so, a composite variable for Gf performance was computed by averaging the z-scores on the RAPM and locations test because performance on these two tasks was significantly related ($r=.40$, $p<.01$; $r=.57$ after corrected for attenuation).

Table 1
Mean performance on span and Gf tasks.

Task	Mean	Median	SD	Skewness	Kurtosis
EP OSPAN	.53	.51	.14	.55	.54
EP RSPAN	.48	.46	.13	.53	1.7
CP OSPAN	.75	.80	.16	−1.0	.34
CP RSPAN	.60	.62	.18	−.39	−.46
RAPM	.46	.43	.19	.48	−.37
Locations	.52	.54	.20	.00	−.74

Note. EP = experimenter-paced. CP = computer-paced. SD = standard deviation. RAPM = Raven Advanced Progressive Matrices.

Table 2
Correlations among span and Gf tasks.

Task	1	2	3	4	5	6
1 EP OSPAN	(.82)					
2 EP RSPAN	.60	(.87)				
3 CP OSPAN	.53	.37	(.93)			
4 CP RSPAN	.37	.34	.72	(.89)		
5 RAPM	.26	.24	.22	.22	(.68)	
6 Locations	.28	.35	.25	.32	.40	(.80)

Note. All correlations are significant at $p<.05$. EP = experimenter-paced. CP = computer-paced. RAPM = Raven Advanced Progressive Matrices. Cronbach's alpha is reported in parentheses.

As expected, after computing the Gf composite, Gf performance was significantly related to performance on the experimenter-paced span tasks (OSPAN: $r=.29$, $p<.01$; RSPAN: $r=.33$, $p<.01$). Moreover, Gf performance was similarly related to performance on the computer-paced span tasks (OSPAN: $r=.30$, $p<.01$; RSPAN: $r=.32$, $p<.01$). These analyses indicate that both types of span task possess predictive validity and serve as good indicators of fluid intelligence.

3.5. Gf variance accounted for by span tasks

Most important, regression analyses were conducted to evaluate whether experimenter-paced and computer-paced span performance accounted for the *same* variance in Gf performance. One hypothesis is that if span tasks completely measure the same construct regardless of how they are administered, then Gf performance is predicted only by variance shared between computer-paced and experimenter-paced span performance. At the opposite end, if computer-paced and experimenter-paced span tasks differentially measure some construct other than WM, they may account for no shared variance in Gf performance. To address whether Gf performance is predicted by variance shared between computer-paced and experimenter-paced span performance, a series of hierarchical linear regressions was conducted. Again, these analyses were conducted separately for the OSPAN and RSPAN tasks.

3.5.1. OSPAN

EP OSPAN and CP OSPAN performance were entered into the regression analysis as predictors; together span performance accounted for 10% of the variance in Gf performance ($p<.05$). To calculate the amount of variance in Gf performance uniquely accounted for by EP OSPAN performance, CP OSPAN performance was entered as a predictor of Gf in Step 1 of the regression analysis. Then, EP OSPAN performance was entered as a predictor of Gf in Step 2 (see Table 3). These analyses indicated that, after controlling for CP OSPAN performance, EP OSPAN performance no longer predicted Gf performance ($R^2=0.004$). The same analyses were conducted to assess the amount of variance in Gf uniquely accounted for by CP OSPAN performance. After controlling for EP OSPAN performance, CP OSPAN performance accounted for a significant amount of unique variance in Gf performance ($R^2=0.05$, $\beta=.27$, $p<.05$). Finally, the amount of variance in Gf shared between the two types of OSPAN performance was 4.6%, which was calculated by subtracting the amount of unique CP OSPAN variance and the amount of unique EP OSPAN variance from the total amount of variance (i.e., $.10-.004-.05=.046$). Of the variance accounted for in Gf performance, approximately 50% is shared between the two types of OSPAN task.

3.5.2. RSPAN

The same regression analyses were conducted for the RSPAN tasks (see Table 3). Together, CP RSPAN and EP RSPAN performance accounted for 15% of the variance in Gf performance. After controlling for variance due to CP RSPAN performance, EP RSPAN performance significantly predicted Gf performance ($R^2=0.05$, $\beta=.24$, $p<.05$).

Table 3
Summary of hierarchical regression analyses for variables predicting criterion task performance (using composite variables).

Variable	R^2	β	F	Variable	R^2	β	F
		Gf composite		OSPAN		Gf composite	
Step 1				Step 1			
EP OSPAN	.10	.31**	12.76	CP OSPAN	.03	.29*	8.00
Step 2	.10[a]			Step 2	.10[a]		
CP OSPAN	.10	.28*	8.77	EP OSPAN	.05	.21†	3.78
EP OSPAN	.00	.06	0.22	CP OSPAN	.05	.27*	4.95
		Gf composite		RSPAN		Gf composite	
Step 1				Step 1			
EP RSPAN	.10	.32**	11.36	CP RSPAN	.10	.32**	12.07
Step 2	.15[a]			Step 2	.15[a]		
CP RSPAN	.10	.23*	9.36	EP RSPAN	.10	.24*	9.81
EP RSPAN	.05	.24*	5.24	CP RSPAN	.05	.23*	4.81

Note. **$p<.001$, *$p<.05$, †$p<.06$.
[a] Total amount of variance in Gf performance accounted for by EP and CP span performance.

Further, after controlling for variance due to EP RSPAN performance, CP RSPAN performance also accounted for a significant amount of unique variance in Gf performance ($R^2=0.05$, $\beta=.23$, $p<.05$). Finally, the two types of RSPAN performance shared 5% of the variance accounted for in Gf performance (i.e., .15 − .05 − .05 = .05). Approximately one third of the variance in Gf performance is shared between the two types of RSPAN performance.

3.6. Factor analysis

Regression analyses indicated that administration method affected the span tasks' predictive validity. Thus, an exploratory factor analysis was conducted using principle axis extraction with Varimax rotation to assess whether the span tasks fell onto a one-factor model (i.e., WM span) or a two-factor model (i.e., EP span and CP span). Factor loadings shown in Table 4 indicated that a two-factor model accounted for more variance than did a single-factor model (83.78% versus 63.65%). Next a confirmatory factor analysis was conducted to test the goodness of fit for a one-factor and a two-factor model (fit indices are shown in Table 5). Comparative fit index (CFI) values of .95 or higher and root mean squared error of approximation (RMSEA) values of .06 or lower are indicative of good model fit (Hu & Bentler, 1999). Results indicate that two factors – EP and CP span, which were free to correlate – fit the data better than did a single span factor.

3.7. Structural equation modeling

Finally, structural equation modeling was conducted to determine if a one-factor or two-factor model better predicted Gf. The one-factor model predicting Gf did not provide good fit, $\chi^2(13, N=125)=45.05$, $p<.001$, CFI = .89, and RMSEA = .14. However, the two-factor model shown in Fig. 1 did provide a very good fit, $\chi^2(11, N=125)=12.57$, $p>.05$, CFI = .99, and RMSEA = .03, suggesting that administration method affects what a span task measures.

Table 4
Factor loadings for single and two-factor models of span performance.

	Single factor	Two factor	
Task	1	1	2
CP OSPAN	.796	.934	.154
CP RSPAN	.776	.839	.379
EP OSPAN	.865	.175	.839
EP RSPAN	.749	.304	.829
Variance accounted for	63.65%	83.78%	

Table 5
Goodness of fit indicators of models for span performance from confirmatory factor analysis.

Model	χ^2	df	p	CFI	RMSEA
Single factor	19.98	2	<.001	.87	.25
Two factor	0.10	1	.752	1.0	<.0001

Note. χ^2 = Chi-square value; df = degrees of freedom; p = Chi-square p value; CFI = comparative fit index; RMSEA = root means square error of approximation.

4. Discussion

The two goals of the present study were to evaluate (1) the reliability and validity of computer-paced versions of the OSPAN and RSPAN tasks and (2) whether both types of span task account for the same variance in Gf performance. Relevant to the first goal, computer-paced span tasks demonstrated high estimates of reliability (Cronbach's α). Further, these tasks demonstrated convergent validity because performance was significantly related to performance on other WM tasks as well as predictive validity because performance was significantly related to Gf performance. Note here that although the OSPAN-related results replicated previous research (De Neys et al., 2002; Unsworth et al., 2005), the results of the present study were the first to demonstrate that a computer-paced version of the RSPAN task demonstrated similar reliability and validity to that of an experimenter-paced version.

Interestingly, among the four span tasks, performance on the two versions of the RSPAN tasks had the lowest correlation ($r=.34$). Although they were completed during separate sessions on different days, the same was true of the OSPAN tasks and they had a stronger relationship ($r=.53$) than did the RSPAN tasks, $z=2.0$, $p<.05$. Further, in both studies that compared performance on traditional, experimenter-paced span tasks and "computerized" (De Neys et al., 2002) or "automated" (Unsworth et al., 2005) span tasks, only OSPAN performance was evaluated.

One explanation for the lower correlation between the RSPAN tasks is that the processing component (i.e., reading sentences) interfered with the storage component (i.e., remembering words), and hence participants could incorrectly recall words from the sentences rather than the target words. Interference may differentially affect performance depending on administration type: EP RSPAN tasks typically require participants to read the sentences and the target words aloud, whereas CP RSPAN tasks typically require participants to silently read the stimuli. Previous work has shown that span performance is significantly higher on tasks that involve silent versus aloud reading (Beaman, 2004), which may explain why span performance was significantly higher on CP than on EP span tasks. Further, reading the stimuli aloud on the EP RSPAN task may create more interference and explain the lower correlation between the RSPAN tasks.

To evaluate this possibility, recall errors were coded as sentence intrusions if the word had come from one of the sentences within the *same* trial. Of the 75 to-be-remembered words, significantly more recalled words were sentence intrusions on the CP version ($M=2.75$, $SE=.31$) than on the EP version ($M=1.42$, $SE=.18$), $t(66)=5.03$, $p<.001$, $d=0.47$. Further, after controlling for rates of CP (but not EP) sentence intrusions, the partial correlation between performance on the two RSPAN tasks increased ($r=.43$). These results suggest the two versions of the RSPAN task differ on more than pacing, and that sentence intrusions partially explain the lower observed correlation between CP and EP RSPAN performance. However, interference from the processing component occurred more often and influenced performance more so on the CP RSPAN task. Thus, reading the stimuli aloud

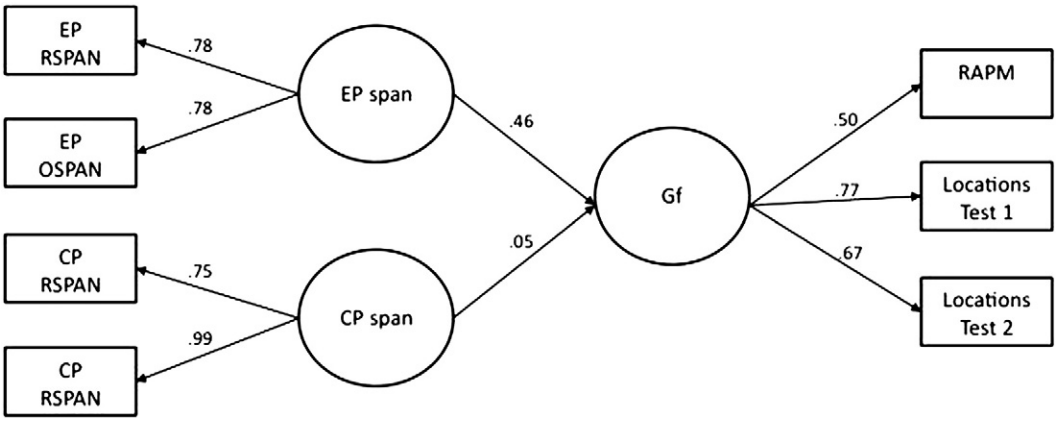

Fig. 1. Structural equation model for a two-factor model – experimenter-paced (EP) and computer-paced (CP) span – predicting fluid intelligence (Gf). EP RSPAN = experimenter-paced reading span; EP OSPAN = experimenter-paced operation span; CP RSPAN = computer-paced reading span; CP OSPAN = computer-paced operation span; RAPM = Raven Advanced Progressive Matrices; Locations test 1 = first sub-test for the Locations test; Locations test 2 = second sub-test for the Locations test.

on the EP RSPAN seemed to help participants distinguish between sentence and target words at recall. Because CP RSPAN tasks involve participants reading silently, they may measure how well individuals inherently distinguish between the sentence and target words.

Regarding the second goal, results indicated that both types of OSPAN and RSPAN tasks predicted shared variance in Gf performance (presumably that due to WM). However, these span tasks also accounted for unique variance, which can be thought of as method-related variance. Interestingly, administration methods were so influential that the span tasks loaded onto two separate factors – CP and EP span – which predicted Gf better than did one WM factor. Why does administration affect predictive validity, particularly for RSPAN? One possible reason is differences in strategy affordance. Because participants had more control over the processing component on the EP RSPAN task, time spent reading sentences was longer and more variable on this task ($M = 5.1$ s, SD = 1.6) as compared to time spent reading on the CP RSPAN task, on which all participants had 4 s. Although the experimenter pushed through the task quickly, participants may have used extra time on the processing task to rehearse the items or to formulate other strategies (Bailey, Dunlosky, & Kane, 2008; Barrouillet, Bernardin, & Camos, 2004; Dunlosky & Kane, 2007). According to the time-based resource-sharing model (Barrouillet et al., 2004) participants have the opportunity to refresh their memory trace of the to-be-recalled information when the processing component is less cognitively demanding or when it allows small amounts of extra time.

If participants were more strategic on the EP RSPAN task, then strategic behavior may be partially responsible for method-related variance predicting Gf performance. To evaluate this explanation, time spent reading the sentences was examined. Although reading time does not directly indicate strategy use, it has been used as a proxy for strategic behavior (Engle, Cantor & Carullo, 1992; Friedman & Miyake, 2004). The extra time spent reading sentences was calculated by subtracting the reading time on the CP RSPAN task (i.e., 4 s) from that on the EP RSPAN task. The extra time spent on EP RSPAN was not correlated with Gf performance ($r = -.01$); in fact, this extra time did not even affect EP RSPAN performance ($r = .11$, $p > .10$).

Strategy affordance does not seem to explain why EP tasks account for unique variance in Gf performance; however, state anxiety may. Moutafi, Furnham, and Tsaousis (2006) found that higher levels of anxiety lowered Gf performance. Given that an experimenter is seated right beside the participants during EP tasks, certain participants may experience anxiety that leads to both lower span and Gf performance. Thus, EP tasks may capture variance due to WM as well as anxiety that are both related to Gf performance.

This method-related variance should be considered in future studies using WM for psychometric purposes or as an indicator of higher-order cognitive abilities. Psychometrically, computer-paced span tasks showed high reliability and validity. As indicators of higher-order abilities, computer-paced span performance predicted Gf as well as performance on traditional (experimenter-paced) span tasks. Results from the present study indicate computer-paced versions of verbal span tasks measure what they were designed to measure (WM), and they do so as well as experimenter-paced span tasks. In fact, computer-paced span tasks may be the better option because they (1) are easier to implement, (2) can be administered in a group setting (for a discussion of the benefits of group testing, see Shelton, Metzger, & Elliott, 2007), and (3) demonstrated stronger convergent validity ($r = .72$) than did the experimenter-paced span tasks ($r = .60$), $z = 1.68$, $p < .05$.

Although computer-paced span tasks have many advantages, they do have two potential disadvantages mentioned in the Introduction: (1) less control over participants' attention on the processing task and (2) equating processing task time for all participants. However, the automated span tasks introduced by Unsworth et al. (2005) sidestep these potential problems by monitoring processing-task accuracy and tailoring time limits according to each individual's ability.

4.1. Conclusion

Experimenter-paced and computer-paced span tasks both presumably measure WM, but how a span task is administered will influence which additional processes it may capture. Which type of span task should researchers use when assessing WM? Computer-paced span tasks are viable options, not only because they have methodological advantages over experimenter-paced span tasks, but also because they are highly reliable and valid measures that serve as good indicators of fluid intelligence.

Acknowledgments

Thanks to Melissa Bishop, Marc Guerini, JB House, Sarah Papakie, and Nathan Sears for assistance with data collection. Thanks also to Katherine Rawson for help in programming the computer-paced span tasks and to John Dunlosky and Mary Pyc for helpful comments on earlier versions of this paper. A special thanks to Chris Was for his input on the latent variable analyses.

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.

Babcock, R. L., & Salthouse, T. A. (1990). Effects of increased processing demands on age differences in working memory. *Psychology and Aging*, *5*, 421–428.

Bailey, H., Dunlosky, J., & Kane, M. J. (2008). Why does working memory span predict complex cognition? Testing the strategy affordance hypothesis. *Memory & Cognition*, *36*, 1383–1390.
Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology. General*, *133*, 83–100.
Barrouillet, P., Gavens, N., Vergauwe, E., Gaillard, V., & Camos, V. (2009). Two maintenance mechanisms of verbal information in working memory. *Developmental Psychology*, *45*, 477–490.
Beaman, C. P. (2004). The irrelevant sound phenomenon revisited: What role for working memory capacity? *Journal of Experimental Psychology*, *30*, 1106–1118.
Bunting, M. (2006). Proactive interference and item similarity in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 183–196.
Colom, R., Rebollo, R., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition*, *34*, 158–171.
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.
Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.
Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433.
De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica*, *42*, 177–190.
Dunlosky, J., & Kane, M. J. (2007). The contributions of strategy use to working memory span: A comparison of strategy assessment methods. *Quarterly Journal of Experimental Psychology*, *60*, 1227–1245.
Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.
Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 972–992.
Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory and Language*, *51*, 136–158.
Hambrick, D. Z., & Oswald, F. L. (2005). Does domain knowledge moderate involvement of working memory capacity in higher-level cognition? A test of three models. *Journal of Memory and Language*, *52*, 377–397.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.
James, W. (1890). *Principles of psychology.* New York: Holt.
Kane, M. J., Poole, B. J., Tuholski, S. W., & Engle, R. W. (2006). Working memory capacity and the top-down control of visual search: Exploring the boundaries of "executive attention.". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 749–777.
la Pointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1118–1133.
Lépine, R., Bernardin, S., & Barrouillet, P. (2005). Attention switching and working memory spans. *European Journal of Cognitive Psychology*, *17*, 329–345.
McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language*, *58*, 480–494.
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.
Moutafi, J., Furnham, A., & Tsaousis, I. (2006). Is the relationship between intelligence and trait Neuroticism mediated by test anxiety? *Personality and Individual Differences*, *40*, 587–597.
Oberauer, K. (2005). Binding and inhibition in working memory — Individual and age differences in short-term recognition. *Journal of Experimental Psychology. General*, *134*, 368–387.
Pardo-Vazquez, J. L., & Fernandez-Rey, J. (2008). External validation of the computerized, group administrable adaptation of the "operation span task". *Behavioral Research Methods*, *40*, 46–54.
Raven, J. C., Raven, J. E., & Court, J. H. (1998). *Progressive matrices.* Oxford, England: Oxford Psychologists Press.
Rowe, G., Hasher, L., & Turcotte, J. (2008). Age differences in visuospatial working memory. *Psychology and Aging*, *23*, 79–84.
Shelton, J. T., Elliott, E. M., Matthews, R. A., Hill, B. D., & Gouvier, W. D. (2010). The relationships of working memory, secondary memory, and fluid intelligence: Working memory *is* special. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 813–820.
Shelton, J. T., Metzger, R. L., & Elliott, E. M. (2007). A group-administered lag task as a measure of working memory. *Behavior Research Methods*, *39*, 482–493.
St. Clair-Thompson, H. L. (2007). The influence of strategies upon relationships between working memory and cognitive skills. *Memory*, *15*, 353–365.
Stanovich, K. E., & Cunningham, A. E. (1993). Where does knowledge come from? Specific associations between print exposure and information acquisition. *Journal of Educational Psychology*, *85*, 211–229.
Swets, B., Desmet, T., Hambrick, D. Z., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology. General*, *136*, 64–81.
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. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language*, *54*, 68–80.
Unsworth, N., & Engle, R. W. (2007). 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., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.
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.
Zeintl, M., & Kliegel, M. (2007). How do verbal distractors influence age-related operation span performance? A manipulation of inhibitory control demands. *Experimental Aging Research*, *33*, 163–175.