Running Head: TEST-RETEST WORKING MEMORY

Repeated Testing of Working Memory Capacity

Laura Carter

Advisor: Dr. Randall Engle

2nd Thesis Reader: Dr. Eric Schumacher

Graduate Mentor: James Broadway

April 25th, 2008

KeyWords: Working Memory, Memory Span, Working Memory Capacity, Repeated Testing, Running Span, Test-Retest

Abstract

Working memory capacity is measured by a variety of memory span tasks and can account for about 40% of inter-individual variation in fluid intelligence (Broadway & Engle, in preparation). In the present study, ten participants performed a widely accepted valid test of WMC, the Running Memory Span task (Pollack, Johnson, & Knaff, 1959), twenty-five times over five sessions to assess test-retest reliability and the extent of practice effects. Results confirmed expectations that memory performance would improve but that the rank ordering of individuals on performance would remain consistent over repeated testing.

Repeated Testing of Working Memory Capacity

Working memory reflects the ability to keep items active in memory in the focus of attention, where measures of working memory reflect the storage and attentional components of working memory capacity (WMC). WMC is the ability to control attention, or the extent of how much information can be temporarily stored and manipulated simultaneously (Engle, Tuholski, Laughlin, & Conway, 1999). WMC is not about a limited number of items in some storage area but about limitations in "the ability to use controlled processing to maintain [unattended information] in an active, quickly retrievable state" (Engle, 2001). By using working memory, a person is able to shift attention from a current task to a distractor task, and then back to the original task without losing concurrent relevant information. Applications of working memory can often be seen in everyday life situations, such as in driving and paying attention to traffic or multitasking at work (Engle, Cantor, & Carullo, 1992).

Case (1974) argued that as mental operations become faster and more efficient, there is more storage space available for information. It is not that processing space increases, but that more efficient mechanisms and processing strategies to encode information are developed. Formation of more efficient mechanisms and strategies allow one to limit the amount of fixed mental resources used and to ultimately leave remaining resources for storage (Engle, Cantor, & Carullo, 1992). According to the *general capacity hypothesis* (i.e., Turner & Engle, 1989; Engle, Cantor, & Carullo, 1992), individual differences in working memory reflect a stable characteristic of people over time. One way to address the role of cognitive efficiency in determining individual differences in

WMC is to examine effects of extensive practice on performance in a working memory task.

Working memory capacity is widely measured using complex span tasks such as Reading Span (Daneman & Carpenter, 1980) and Operation Span (Turner & Engle, 1989). In these tasks, participants must perform a distractor task while also completing a task involving encoding items that must be remembered. For example, in Operation Span a person must solve multiple math problems, and in between these problems, the person is shown a letter to remember. After all of the math problems, the person is then prompted to recall the letters in the order they were presented. Complex working memory span tasks have been shown to reliably predict individual differences in higher order complex abilities like reading comprehension (Daneman & Carpenter, 1980), reasoning (Kyllonen & Christal, 1990), and complex learning (Shute, 1991). Recently, Broadway and Engle (in preparation) showed that the running memory span task (Pollack, Johnson, & Knaff, 1959) can also reliably predict individual differences in higher order cognition, and account for much of the same variance that complex span tasks do.

In a running memory span task, subjects must drop old items from memory and continually add new items to memory over an unpredictable list length (Pollack, Johnson, & Knaff, 1959). Running span tasks can sometimes be called information monitoring or updating tasks since items are continually dropped and added to memory. In comparison to digit span tasks that are generally used in intelligence testing, running span tasks are important because they have been found to have higher correlations to intellectual aptitude (Cowen et al., 2005). In running span tasks, list length is unpredictable and unknown to the participant and makes an updating mechanism necessary. It is expected that the last three to four list items can be remembered without much difficulty, and rehearsal and practice can increase this number of items (Bunting, Cowan, & Saults, 2006). However, there has not been much research devoted to assessing the extent of learning effects in relation to running memory span tasks and how these learning effects can be used over time to predict fluid intelligence. In order to explore these learning effects, the correlations to intelligence and the rank ordering between participants on performance must be explored over repeated testing.

Broadway & Engle (in preparation) tested participants' abilities to remember the last four, five, or six letters from variable-length lists in several different versions of the running span task. They also measured WMC using complex span tasks, and fluid intelligence using two standard tests, Ravens Progressive Matrices and Shipleys Abstractions (Raven, Raven, & Court, 1998; Zachary, 1986). The rate of presentation (1 item/ 250 ms, 1 item/ 1000ms, 1 item/ 2500ms) and sensory modality of stimuli (auditory and visual) varied across tasks, yet much of the same individual differences in higher order intellectual abilities and WMC. The task that involved auditory letter lists presented at a rate of one item per second was found to account for about 20% of variance in fluid intelligence composite, made from z-score averages on the two intelligence tests. The present study used this same auditory running span task, and the sample of participants was taken as a subset from the sample in Broadway and Engle so that the effects of practice could be examined against known rates of performance and relationships to higher order cognition.

The present study addressed three main questions. The first research question sought to explore if participants' performance on the running memory span task would improve after extensive practice. The second question was if the task would be reliable over repeated testing, keeping test-retest reliability. The last question was if the relationship between performance on the task and a test of fluid intelligence would be stable across extensive practice. Participants were expected to improve relative to their initial scores on the task when they had been in the study by Broadway and Engle, but rank ordering of individuals was expected to remain consistent even after practice, indicating good test-retest reliability for the running span task. Participants were selected from the Broadway and Engle sample so that the initial correlation between performance in the task and a composite variable of two tests of fluid intelligence was high. With this relationship between performance on the running span task and intelligence tests known, the present study could address the question of whether relationships to this criterion measure would remain stable after repeated testing and practice on the running span task.

Method

Participants

Participants (N = 15) were recruited from the sample of sixty-one participants in a previous study by Broadway and Engle (in preparation). Participants were selected from this sample so that the full range of WMC was represented, and that the Pearson correlation between these individuals' initial scores on the running span task used in the present study and the intelligence composite was high ($r = .85$, $p < .01$). The auditory running span from Broadway and Engle used in the present study had significantly correlated with the intelligence composite in the earlier sample ($r = .445$, $p < .01$). Of the original fifteen participants that were contacted for the present study, only ten participants were able to complete all five sessions of the present task, but these participants still covered a broad range of initial memory performance. Of the maximum score of 90 letters to report correctly in order, the initial mean running span performance of the present sample was 69.10 letters (*SD* = 11.24; range 41 – 80). The mean for the present sample (N= 10) was not statistically different from the mean (*M* = 61.69, *SD* = 13.42) obtained from the Broadway and Engle sample, t (9) = 2.085, $p = .067$. The correlation between these ten participants' initial scores on the running span task and the composite intelligence variable was only moderate and not significant ($r = .59$, $p > .05$).

Participants were between the ages of 18 and 35 (*M* = 21.7, *SD* = 3.53) and were compensated for their participation with payment of $20 for each session. Upon completion of the fifth session, participants received an extra bonus of $15 for completing all of the sessions.

Materials and Stimuli

Tasks were programmed in E-prime software (Schneider, Eschman, & Zuccolotto, 2002) and administered on a personal computer. Stimuli were auditory vocalizations of letters derived from random lists from the set F, H, J, K, L, N, P, Q, R, S, T, and Y. Letters were presented at a rate of one item per second through head-phones. The vocalizations of letters were compressed into digital sound files and were prepared using Audacity software by individuals with training in sound engineering and diction. The head-phone volume was at the discretion of the participant and during practice trials, the experimenter made sure the participant could adequately hear and distinguish the auditory vocalizations of the letters

Procedure

Each participant completed five one-hour sessions and the average time to complete all five sessions was 16.8 days (*SD* = 3.42). The task was preceded by instructions and practice trials. There were five blocks of trials that occurred within each session, with each block having 18 total trials. The participant was prompted to recall the last four, five, or six items from each list at the end of the presentation phase, and there were six trials for each of these prompted numbers. For a single trial, participants were shown n, n+1, or n+2 items, where "n" refers to the number of items participants were asked to recall, and participants saw each of these presentations (n, n+1, n+2) twice for each prompted number. For example, if a participant was prompted to recall the last five letters, there would have been two trials of seeing five, six, and seven letters, totaling the six trials for each prompted number. Randomization varied the order of how many items were required to be recalled and also varied the order in which these blocks of trials appeared within each session. For each block of trials within a session, only the first block began with instructions; the other four blocks went straight into the task.

In a single trial, stimuli were presented auditorily in the form of single letters vocalized through headphones at a rate of one letter per second (1/1000 ms). The participant was prompted before each trial about how many how many letters from the end of the series will need to be recalled in the test phase. For example, if the participant was prompted with "Remember the last 5 letters," and the participant heard "P, Q, H, F, R, L, T," the correct response would be "H, F, R, L, T." Participant responded by selecting items in order (by mouse-click) from a grid displaying all the letters that could appear in the task.

Results

A 5 (Session) by 5 (Block) repeated -measures ANOVA was applied to the data to examine the improvement in memory performance on this task after practice and repeated testing. There was a significant main effect of Session, $F(1, 9) = 29.44$ $p < .01$, partial eta squared = .766, indicating that running span performance improved over the five sessions of this study. Participants overall improved by 11.44 letters correctly recalled from Session 1 to Session 5, as can be seen in Figure 1. The final average performance of participants (76.38 items) was not at ceiling, and as Figure 1 shows, there is no plateau in performance. The main effect of Block was not significant, $F\ (1,9) = .909$, $p > .05$, indicating that participants did not improve much *within* each one-hour session. Figure 2 shows the data for each session across the blocks of trials, and five consecutive blocks was one session.

Because the sample is small, a non-parametric statistic, Spearman's *rho,* was used to assess test-retest reliability for the running span task across the five sessions. Table 1 indicates that the rank ordering of individuals was consistent across sessions, which suggests that there is good test-retest reliability on this running span task. Table 1 also shows that correlations are highest between sessions closest to each other in time, and drop as function of temporal distance between sessions. The improvement made by each individual in the study can be seen in Figure 3, where participant's scores in Session 1 of the present study can be compared to their performance in Session 5. Figure 3 suggests that three of the participants made larger improvements compared to other individuals in the sample. To further investigate this, an improvement score was computed for each individual by subtracting Session 1 from Session 5. Linear regression was used to investigate if the improvement score could be predicted from individual scores on a variety of other tests from data that was obtained from Broadway and Engle (in preparation). The intelligence composite, reading span, operation span, and running span score from Broadway and Engle did not significantly predict the improvement score, F (4, 9) = .729, $p > .05$, $R^2 = .368$. This result suggests that there was not differential improvement for people of higher or lower intelligence or WMC.

The third research goal of the present work was to assess the stability of the relationship between running memory span performance and measures of higher-order cognition after repeated testing on this running memory span task. The fifteen original participants were recruited for this reason from the Broadway and Engle (in preparation) sample so that within this new sample for this study, the correlation between the running span task and a composite of two intelligence tests was high. As explained earlier, however, only ten of these participants completed all five sessions of the present study, and the correlation between their initial running span performance and the intelligence measures for this ten person sample was only moderate and not significant. Therefore, the ability to address the stability of predictive validity over the five sessions is limited. However, the question can be addressed based on data from the first two sessions, completed by all of the participants originally recruited.

Table 2 indicates that the *rho* between initial running span performance and the composite of intelligence tests, obtained when the fifteen recruited individuals were participants in Broadway and Engle (in preparation), was .707, $p < .01$. The *rho* obtained with these participants after the first session in the present study was .722, $p < .01$, and the *rho* obtained after the second session in the present study was .595, $p < .05$. Testing for differences among these correlated correlations (Meng, Rosenthal, & Rubin, 1992) did not yield a significant result, $X^2(2) = 1.30$, $p > .05$. These findings indicate that the predictive validity of the running span remains stable even after two hours of practice (10 blocks of trials).

Discussion

Average performance of participants improved over sessions and confirmed Hypothesis 1. From each session to the following session, overall ending and beginning performances were higher than the overall performance of the previous session. This increase from even the first block of trials of each session being higher than the performance on the last block of trials completed in the previous session indicates a learning effect. There was a significant increase in aggregate performance over the five sessions.

There are many possible explanations for this increase in recall ability. First, there may have been strategy learning, meaning participants may have formed better strategies to remember the items as the trials and sessions progressed (for example, silently reciting the items in their head). The forming of new strategies may have made participant's mechanisms for remembering more efficient and this may have improved their performance over sessions. Secondly, there may have been an improvement of the logistics of the task, like learning to press buttons faster. Third, the nature of a repeated testing task is that a participant completes the same task several times, and this repetition results in participants becoming more familiar with the task than they were at the start of the task. If a participant is more familiar with the task and has a better understanding of what is expected of them, there could be a potential increase in performance or ability to recall. Lastly, the participant may have become familiar with the set of letters used, and this familiarity may have made it easier to remember the potential letters presented in the lists. Any of these situations could have provided an improvement in performance, but it is impossible to know without further investigation.

No single participant reached a plateau or had a ceiling effect on the task. The lack of plateau indicates that no person reached the maximum possible score. The number of practice sessions it would take to reach the maximum score on the task is still unknown, and the number of practice sessions it would take for a participant to reach their maximum recall performance is also still unknown since there were no ceiling effects. If the study were to continue with more sessions, the limits of the effects of practice could be further investigated, and it would be possible to explore how much improvement can actually occur. For example, if a low WM ability individual were to keep practicing, is there potential for surpassing a high WM ability individual? Also, is there a point at which more learning and practicing stops helping or actually hinders performance? Further investigating ceiling effects and potential maximum effects could provide answers to these questions.

Hypothesis 2 was also confirmed and rank order of participant performance remained according to Spearman's rho rank order correlation. Performance across sessions was significantly correlated, making this a reliable measure over repeated testing. Lower WM ability individuals remained lower in ability compared to the higher WM ability individuals over the five sessions. Both high and low ability individuals improved to some extent, which shows that practice effects on this test were not completely ability-specific. However, the extent to how much higher and lower ability individuals improved and can improve must be further explored, since only one low span individual (participant 1 in Figure 4) actually completed all five sessions.

Although participants did improve their performance on this working memory test, it does not necessarily mean that they are better at remembering on a different task since transfer was not assessed. There is potential that the results are task-specific, and having participants redo other WM tasks that they had completed before this task, such as Operation Span, Reading Span, or Ravens, could provide information on transfer. If participant performance became significantly better on these tasks and correlations to general intelligence remained, transfer could be examined. Evaluating transfer is the most important future step for this research project.

Even if transfer is cannot be evaluated, this running span task was still found to be reliable over repeated testing. One potential application for this research is in clinical trials. If a particular drug is thought to alter memory in any way, giving this specific task multiple times throughout the clinical trials could provide information on whether the drug is actually altering working memory. If a participant retains the same memory ability over multiple trials, then the practice effects can be taken out, and the actual effect of the drug can be assessed. At the same time, if a person improves in memory, this test would allow for the experimenter to control for practice effects while also assessing the effectiveness of the drug treatment.

There are three main future directions for this research. The first, as stated earlier, deals with evaluating transfer. Evaluating transfer would provide information on if the results of this experiment are task-specific or if practice effects can be generalized further. Another future research direction is to retain and test more low WM ability participants. In the current research, four participants dropped out before session five, and this made their data unusable for evaluating performance across all of the sessions; three out of four of these participants were low WM ability participants. Having more low WM ability participants would provide an opportunity to see how much a low WM ability participant can improve and to analyze overall group correlations to intelligence over the full range of performance. Lastly, increasing the number of sessions for some participants could provide research to see how long it takes an individual to reach maximum performance. With no participant in the current study consistently reaching the maximum possible score or reaching a ceiling, the maximum effects of practice could not be assessed. Because of time constraints of the research, these directions were not explored.

REFERENCES

Bunting, M., Cowan, N., & Saults, J. S. (2006). How does running memory span work? *The Quarterly Journal of Experimental Psychology*, 59 (10), 1691-1700.

Case, R. (1974) Structures and strictures, some functional limitations on the course of cognitive growth. Cognitive Psychology, 6, 544-573.

Cowen, N., Elliot, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, 51, 42-100.

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

Engle, R. W., Cantor, J., & Carullo, J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18,* 972-992.

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.Healy, A. F. (1974). Separating item from order information in short-term memory. *Journal of Verbal Learning & Verbal Behavior,* 13(6), 644-655.

Engle, R. W. (2001). What is working-memory capacity?. In H. L. Roediger III & J. S. Nairne (Eds.), *The Nature of Remembering: Essays in Honor of Robert G. Crowder* (pp. 297-314). Washington, DC: American Psychological Association.

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

Meng, X., Rosenthal, R., & Rubin, D. B. (1992) Comparing correlated correlation coefficients. *Psychological Bulletin*, 11, 172-175.

Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology*, 57 (3), 137-146.

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

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-prime user's guide. Pittsburgh: Psychology Software Tools Inc.

Shute, V. J. (1991). Who is likely to acquire programming skills? *Journal of Educational Computing Research*, 7, 1-24.

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

Zachary, R. A. (1986). *Shipley Institute of Living Scale: Revised Manual*. Los Angeles: Western Psychological Services.

Figure Captions

Figure 1: Significant effect of average total improvement over repeated testing of 5 sessions

Figure 2: Average performance across all blocks of trials (25), where five blocks equaled one session

Figure 3: Individual participant performance from session 1 to session 5

Table 1:

N = 10	T1 Average	T2 Average	T3 Average	T4 Average	T5 Average
T1 Average					
T2 Average	.891**				
T3 Average	.770**	.830**			
T4 Average	.758*	.867*	.964**		
T5 Average	.697*	.733*	.842**	.903**	

Note.
* Significant at the .05 level (2-tailed)
**Significant at the .01 level (2-tailed)
Spearman’s *rho* correlations between sessions (T, or times) for N=10 sample

Table 2:

N = 15	gF composite	T0 Average	T1 Average	T2 Average
gF composite				
T0 Average	.707**			
T1 Average	.722**	.806**		
T2 Average	.595*	.774**	.932**	

Note.

* Significant at the .05 level (2-tailed)

**Significant at the .01 level (2-tailed)

Spearman's *rho* correlations between the composite intelligence variable, performance on the initial session from Broadway and Engle (T0), and performance on the first 2 sessions (T1, T2) for N=15 sample

Figure 1:

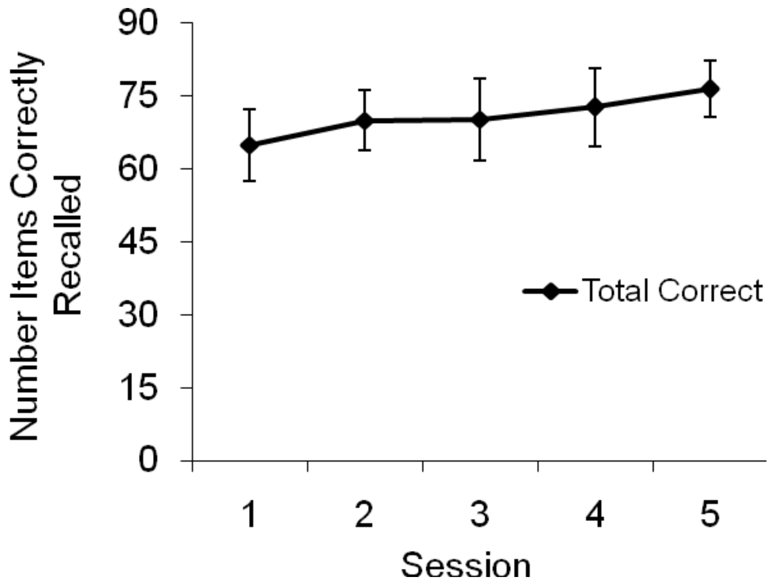

90
75
60
45
30
15
0
Number Items Correctly Recalled
Total Correct
1
2
3
4
5
Session

Figure 2:

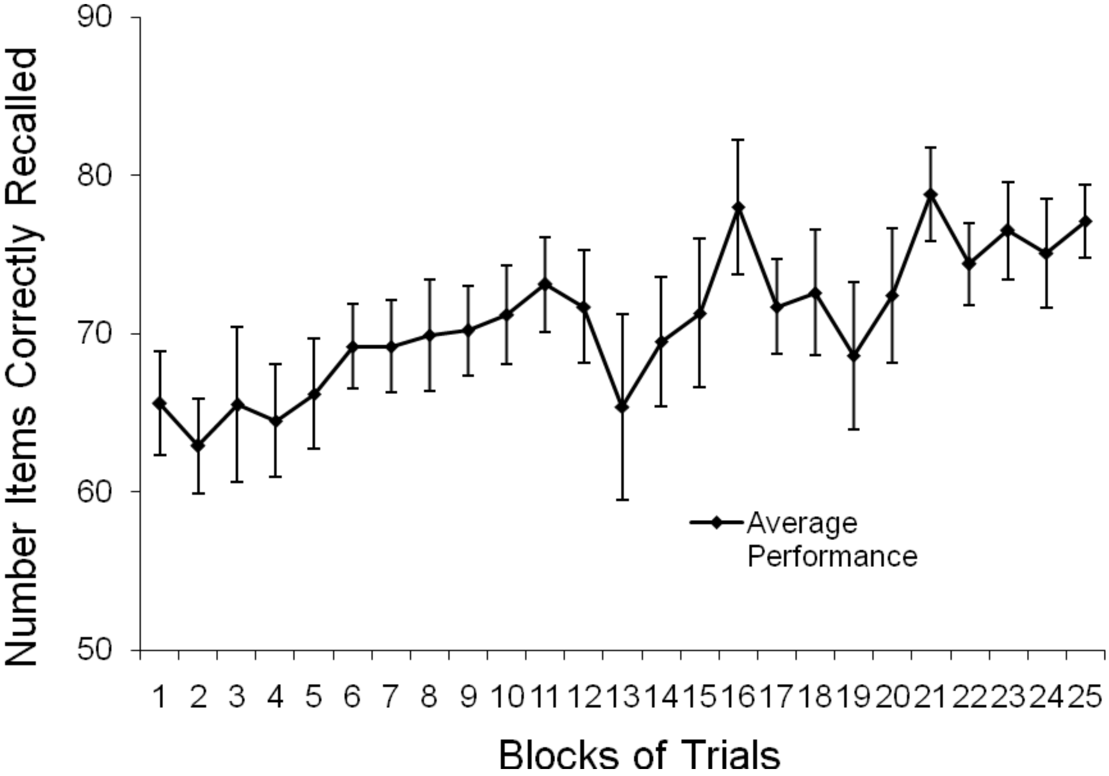
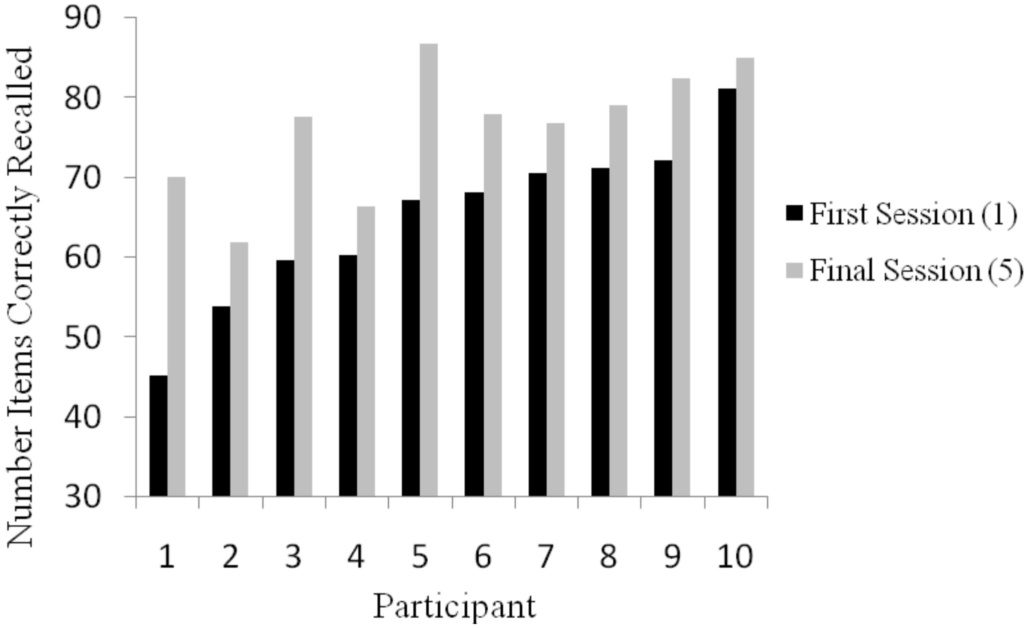

90
80
70
60
50
Number Items Correctly Recalled
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Blocks of Trials
Average Performance

Figure 3:


90
80
70
60
50
40
30
Number Items Correctly Recalled
1
2
3
4
5
6
7
8
9
10
Participant
■ First Session (1)
■ Final Session (5)