N-BACK AS A MEASURE OF WORKING MEMORY CAPACITY

A Dissertation
Presented to
The Academic Faculty

by

Tyler L. Harrison

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology

Georgia Institute of Technology
August 2017

N-BACK AS A MEASURE OF WORKING MEMORY CAPACITY

Approved by:

Dr. Randall Engle, Advisor School of Psychology *Georgia Institute of Technology*	Dr. Mark Wheeler School of Psychology *Georgia Institute of Technology*
Dr. Eric Schumacher School of Psychology *Georgia Institute of Technology*	Dr. David Washburn School of Psychology *Georgia State Psychology*
Dr. Christopher Hertzog School of Psychology *Georgia Institute of Technology*	

Date Approved: July 27, 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

One of the most important findings in cognitive psychology, is the relationship of working memory capacity (WMC) to a host of important cognitive activities, the manner in which WMC interacts with many different cognitive variables, and the consequences for the individual when WMC is reduced due to interventions such as sleep deprivation and psychopathology. However, one often over-looked problem is that researchers use different cognitive tasks to measure and study working memory capacity: differential studies have historically used complex span tasks to assess WMC. However, n-back tasks are often used in neuroimaging studies because the task lends itself to the requirements of fMRI studies. The implicit assumption is that both types of tasks measure the same construct. For the present study, both complex span performance and n-back performance was measured, in 328 subject, to see whether they measured the same construct. The size of the stimulus pool for the n-back tasks was manipulated to determine whether n-back tasks with more interference (i.e., with a smaller stimulus pool) were more strongly correlated with cognitive ability. Additionally, the presence of lure trials was manipulated within the n-back tasks to examine whether the most interfering lures were more strongly correlated with complex span performance. From the data, I argue that complex span tasks and n-back tasks measure different sub-processes of WMC and that this causes the two tasks to load onto separate factors, the number of stimuli that an n-back task uses changes its correlation to other measures of cognitive ability, and that the false alarms to lures closest to *n* are most strongly correlated with both complex span tasks and fluid intelligence but only for n-back tasks with a small stimulus pool.

CHAPTER 1. Introduction

One of the most influential findings in the study of cognitive abilities in recent history is the strong relationship between WMC and a huge array of lab and real-world cognitive tasks including measures of fluid intelligence (Gf). Fluid intelligence is the ability to solve novel problems and reason in situations in which one has had little or no experience (Horn & Cattell, 1966). There is a plethora of research exploring the importance of Gf. For instance, individuals with higher intelligence have more successful careers, live longer, and are less likely to be incarcerated (Gottfredson, 1997). Fluid intelligence is the best predictor of job performance in the literature across multiple jobs (Schmidt & Hunter, 1998). Although researchers have a good understanding of the importance of this cognitive ability, they do not have a good understanding of the cognitive mechanisms of Gf and have trouble even defining the construct at the psychological or behavioral level (Neisser et al., 1996).

Working memory is the system of memory and controlled attention that is responsible for maintaining information in memory for brief periods and protecting that information against interference. Working memory capacity is the term used to describe the functioning of the working memory system at the individual differences level. Although, the word "capacity" is used for this ability, I do not mean that WMC is the number of slots that a subject has to hold information in memory (e.g., Cowan, 2001; Miller, 1956). Working memory capacity is much more than the number of items that can be passively maintained in memory. It also includes the ability to deal with interference and distraction such as dealing with proactive interference in a memory task (Kane & Engle, 2000) and maintaining the goal of naming the hue of a color instead of reading the word in the Stroop task (Kane & Engle, 2003). In many ways the "capacity" in WMC is a misnomer but I will still use the term "WMC" to be consistent with previous research.

Unlike for Gf, there are several well-articulated models of working memory and WMC. This has led researchers to see WMC as a key to understanding the mechanisms of intelligence (Engle, Tuholski, Laughlin, & Conway, 1999). This is an example of the cognitive correlates approach to studying individual differences (Sternberg, 1985). For this approach, researchers try to explain a complicated and amorphous construct such as Gf by accounting for individual differences in that construct using simpler and easier to explain cognitive abilities. Although this approach to the study of individual differences was emphasized in the 1980's, it dates back to the inception of psychology (Galton 1883; Cattell, 1886).

Cognitive psychologists have not always agreed on the nature of WMC (e.g., Miyake & Shah, 1999). I review how two major categories of tasks have been used to the study WMC (i.e., complex span tasks and n-back tasks). The present research study was designed to evaluate whether n-back and complex span tasks measure the same construct, account for the same variance in Gf, and to determine what manipulations to the n-back tasks modulate the relationship between the tasks and other cognitive abilities.

1.1 Complex Span Tasks

The initial work on WMC was done by examining performance on complex span tasks (e.g., Daneman & Carpenter, 1980; Turner & Engle, 1989). These tasks required subjects to interleave the performance of an attention-demanding task with an item to be remembered. In the reading span, subjects read a series of sentences and need to recall either the last word of the sentence or a letter, digit, or word that follows each sentence. Subjects receive multiple processing task/to-be-remembered stimulus pairings until the experimenter indicates that the subjects should recall all the stimuli they were instructed to remember in the order that they were presented. Some of the initial findings suggested reading span performance was a good predictor of reading comprehension (Daneman & Carpenter, 1980; Daneman & Merikle, 1996). Some researchers argued that the reading span predicted reading comprehension because the task had a reading component. If subjects were good at reading, these researchers theorized that subjects would have more cognitive resources available to remember words in the reading span task. This hypothesis was refuted when other researchers showed that nature of the processing task has little to no relationship to reading comprehension (Turner & Engle, 1989). Although it is possible to force complex span tasks to load onto two separate factors based on the nature of the to-be-remembered stimuli (verbal vs. visuo-spatial; see Shah & Miyake, 1996), the domain-general components of the complex span tasks account for the bulk of their relation to higher order cognitive abilities (Kane et al., 2004).

In the 1990's researchers embraced both of Cronbach's (1957) two disciplines of psychology (i.e., experimental and differential approaches) to get an understanding of why these tasks were related to reading comprehension and Gf. In a variety of experimental/differential hybrid studies it was shown that neither the difficulty of the processing task (Conway & Engle, 1996), the individual strategies subjects implemented when performing the complex span tasks (Engle, Cantor, & Carullo, 1992), nor word knowledge (Engle, Nations & Cantor, 1990) accounted for why complex span tasks are related to higher order cognitive abilities. This is a point that is particularly important for our discussion: Researchers have manipulated many aspects of the complex span tasks and know how those changes affect both the mean level of performance and the correlations of the complex span tasks.

Since the 2000's research using the complex span tasks as measures of WMC became more common for a variety of important reasons. The complex span tasks became the quintessential measure of WMC in differential cognitive psychology because there were numerous demonstrations that WMC (as measure by complex span tasks) was substantially related to many real-world tasks and also to Gf (e.g., Conway et al., 2002; Kane et al., 2004). Equally important but often overlooked, these studies showed that the complex span tasks had great psychometric reliability and formed a coherent latent factor (see Redick et al., 2012 for the most recent norms). Another development in the study of WMC using complex span tasks was that automated versions of these tasks were created and made available to any researcher who wanted the tasks (Unsworth et al., 2005). Easy access to these tasks had two important consequences: 1.) Psychologists from any discipline could use these tasks in their research 2.) researchers in various labs were all using the same identical tasks.

1.2 N-back Tasks

The n-back task was first used in the late 1950's (Kirchner, 1958) but was not used extensively until the 1990's (Owen, McMillan, Laird, & Bullmore, 2005). For the n-back task, subjects are presented with a series of stimuli. If the stimulus that the subject is presented with is the same stimulus that the subject was shown *n* trials ago, the subject makes a response. For example, if the *n* for a certain task was 3 and the subject was presented with the letters, *A*, *D, G, T, D, T, E*, the subject should make a response only when the second *D* occurs because a *D* was repeated 3 trials ago. The subject should not respond to the second *T* because, although *T* was presented a previous trial, the *T* occurred 2 trials ago and not 3. Researchers call the second *T* a lure trial.

One major difference in the use of n-back tasks compared to complex span tasks is the ultimate goals of the study. Complex span tasks are used primarily in individual differences studies in which the goal is to assess WMC. The n-back tasks have been used to study both working memory (the memory system) and with WMC (the cognitive ability). As for the research into working memory, the n-back task is the commonly used by cognitive neuroscientists (e.g., Awh, Jonides, Smith, Schumacher, Koeppe, & Katz 1996; Braver, Cohen, Nystorm, Jonides, Smith, & Noll, 1997; Nystorm, Braver, Sabb, Delgado, Noll, & Cohen, 2000). The n-back task is used in these neuroimaging studies instead of other working memory tasks for a variety of reasons: subjects only have to make one response (i.e., I saw this stimulus *n* trials ago), it is easy to time-lock every stimulus for each subject (the amount of time it takes subjects to complete the processing component of a complex span task varies), and it is relatively simple to implement a control condition for a neuroimaging study in which subjects look at the stimuli under conditions in which there is no requirement to remember the stimulus. Studies using variations of the n-back task have provided a great deal of evidence for our understanding of what brain regions are involved with working memory (e.g., Jonides et al., 1997). It is clear from these studies that dorsolateral and ventrolateral prefrontal cortex, dorsal areas of the anterior cingulate cortex, medial and lateral areas of the parietal lobes are all important for working memory performance.

The imaging studies using n-back tasks are important to the study of working memory. However, researchers have also used n-back tasks as measures of WMC (e.g., Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Miller, Price, Okun, Montijo, & Bowers, 2009). There are a few important points to make about how using n-back tasks as measures of WMC might be particularly problematic. First, this is a rather large family of tasks. There is not a ‘standard’ form of n-back as there is with the complex span tasks. There are perhaps as many variations of the n-back task as there are researchers who study working memory. Additionally, there has been little research concerning which manipulations to the n-back tasks change about what the tasks measure. These two points are particularly disconcerting because it may be the case that researchers using the different n-back task are actually measuring completely different cognitive constructs. We simply do not know the overlap of what the tasks measure or, generally, the reliability of the measures.

1.3 Comparing Complex Span and N-back Tasks

Because both complex span tasks and n-back tasks are thought to be measures of WMC, it stands to reason that the two different classes of tasks are measuring the same cognitive construct. Only in the past decade have researchers attempted to answer this question. Kane, Conway, Miura, and Colflesh (2007) examined the relationship between a single complex span task (the operation span) and a single n-back task (letter n-back). They found that a measure of Gf (Raven's Advanced Progressive Matrices) was more strongly related to both the complex span task (r = .33) and the n-back task (r = .42) than both tasks were related to each other (r = .22). This finding is problematic because both tasks are thought to reflect the same construct. However, there are a few potential problems with this study. Correlations were only examined at the task level and not at the latent construct level. It could be the case that task-specific variance and unreliability deflated the correlation between the two WMC tasks. Schmiedek and colleagues (2009) tried to correct for this potential problem by examining the relationship between a complex span factor and an updating factor which had a single n-back task as one of the indicators. These researchers found that the latent relationship between these factors approached unity and argued against the findings of Kane et al. A major concern for this conclusion is that the updating factor had a running memory span task as one of its indicators. This task has been shown to predict the same variance in Gf as the complex span tasks (Broadway & Engle, 2010) and it is as strongly related to the complex span tasks as it is to the other updating tasks in the Schmiedek et al. data set.

Redick and Lindsey (2013) conducted a meta-analysis to examine the relationship between the complex span and n-back tasks. They found that the meta-analytic correlation between categories of tasks was .20, a result that is more similar to the results of Kane et al. (2007) than the results of Schmiedek et al. (2009). The results of this meta-analysis clearly demonstrate that complex span tasks and n-back tasks measure separate constructs. However, Redick and Lindsey did not examine complex span and n-back tasks at the latent level. Additionally, it is still unclear why these two categories of tasks are more strongly related to Gf than they are to one another.

One potential solution to this conundrum might involve manipulations to the n-back task. As mentioned earlier in the introduction, the early research using the complex span tasks manipulated various aspects of the task (e.g., processing task difficulty or word frequency of the to-be-remembered stimuli). These results were critical in determining which aspects of the complex span tasks were important in the prediction of Gf. As of yet, little has been done for the n-back tasks. Kane et al. (2007) reported that their 3-back task was more strongly related to cognitive abilities than a 2-back task but this result seemed to be more related to a ceiling effect in the 2-back task than anything else.

One potentially interesting manipulation in this regard is the presence or absence of lure trials in the n-back. As discussed earlier in the introduction, lure trials are trials in which the subject has seen the same stimulus recently but not *n* trials ago (e.g., a stimulus from 2 trials ago when performing a 3-back task). Gray, Chabris, and Braver (2003) found evidence that lure trials in the n-back led to greater neural activity in the prefrontal cortex for high Gf individuals compared low Gf individuals. Kane et al. showed that performance on lure trials (2-back and 4-back lures for their 3-back task) added unique variance to the prediction of Raven's after accounting for performance for only the target trials of the n-back. Both of these findings suggest that the presence of lures might increase an n-back task's relationship with fluid intelligence.

In a recent study, my colleagues and I attempted to explore this hypothesis (Shipstead, Harrison, & Engle, 2016). We created three different 3-back tasks using three different types of stimuli: words, faces, and wingdings. Critically, we manipulated lures across the three n-back tasks. All three tasks contained 2-back, 4-back, 5-back, 7-back, 8-back, and 9-back lures. The hypothesis was that the lures at positions closer to the target position (i.e., 3-back) would be most strongly related to Gf and the complex span tasks. Instead, we found that exact opposite pattern of results. The lures from farther back in time were the most strongly correlated with both complex span performance and Gf. This result was counterintuitive to the previous literature. Previous studies have shown that WMC is particularly important in situations in which there is a large amount of interference and cognitive control is needed (e.g., Kane & Engle 2000; 2003). One difference between the tasks that we used in this study and the tasks used in most of the previous n-back literature deals with the total number of stimuli used in both types of n-back tasks. Most n-back studies use only a small limited pool of stimuli (around 5 to 15), whereas every stimulus that appeared in the Shipstead et al. n-back tasks was only presented once or twice. This is particularly important because the more times a subject has been presented with a stimulus the greater the interference for the subject to remember exactly when the subject saw that stimulus last. It is likely the case that subjects could rely more heavily on familiarity to perform an n-back task in which the subject only sees a particular stimulus once or twice. Considering that subjects were presented with stimuli that they had never seen before on 50% of trials for the n-back tasks used in the Shipstead et al. study, it is likely that subjects could easily rely on familiarity to reject those stimuli as targets. It is possible that previous studies that highlighted the importance of lure trials (e.g., Gray, Chabris, & Braver, 2003; Kane et al., 2007) found these results because relatively few stimuli were used and repeated for their n-back tasks. One of the goals of the proposed study is to answer the question why different studies found different effects pertaining to lures in the n-back task.

There were three major goals for the present study: 1) to determine whether both types of WMC tasks (complex span tasks and n-back tasks) actually measure the same construct. 2) to determine whether the size of the stimulus pool changed the correlation of the n-backs tasks to the complex span measures. 3) to see if the relationship of lure trials in the n-back tasks to measures of Gf changed when the size of the stimulus pool was manipulated.

CHAPTER 2. METHODS

2.1 Subjects

Subjects were 328 volunteers from the Georgia Institute of Technology subject pool, from Georgia State University, or from the greater Atlanta community. Subjects were between the ages of 18 and 35 at the beginning of the study, had never participated in a study with the Attention and Working Memory Lab before, had English as a native language, and had normal or corrected-to-normal vision.

2.2 Procedure

When subjects first arrived to the lab they were given an informed consent form to read over and sign. At this time subjects were encouraged to ask as many questions as they would like about the procedures of the study. After the consent procedure, subjects were escorted to a group running room. Subjects completed the tasks in this room with a research assistant and up to 4 other subjects. Subjects were compensated with a 30 dollar check for each of the 4 sessions at the end of each session. A 10 dollar check was given to subjects at the end of the 4th session as a completion bonus. All of the tasks that subjects completed and the order in which they were performed are presented in Table 1.

Table 1. The order of all the tasks in the study. The tasks relevant to this paper have been bolded.

Session 1	**Session 2**	**Session 3**	**Session 4**
Operation Span	**Symmetry Span**	**Rotation Span**	Line Discrimination
Raven's Advanced Progressive Matrices	**Number Series**	**Letter Sets**	Pitch Discrimination
Word N-back Large Stimulus Pool	**Wingding N-back Large Stimulus Pool**	**Face N-back Small Stimulus Pool**	Circle Discrimination
Antisaccade	Stroop	Deadline Stroop	Loudness Discrimination
Verbal Fluency I – 1	Verbal Fluency E – 2	Verbal Fluency I – 3	Speed Accuracy Tradeoff – Line Discrimination
SynWin	Analogies	Multitask	Mental Roation
Winding N-back Small Stimulus Pool	**Face N-back Large Stimulus Pool**	**Word N-back Small Stimulus Pool**	Nonsense Syllogisms
Arrow Flanker	Verbal Fluency I – 2	Verbal Fluency E – 3	Control Tower
Verbal Fluency E – 1	Continuous Paired Associates - Words	Continuous Paired Associates - Spatial	Sustained Attention to Response

Table 1 (continued).

Immediate Free Recall	Conjunction Fallacy	Wason Card Selection Task	Demographics
Base Rate	Speed Accuracy Tradeoff- Flanker	Proactive Interference Task	
Visual Arrays 4	Paired Associates Tasks	Speed Accuracy Tradeoff – Lexical Decision Task	

2.3 **Complex Span Tasks**

All complex span tasks had a similar format. Subjects were presented with a to-be-remembered stimulus (i.e., the memory task) and then had to complete a simple distraction task (i.e., the processing task). This pairing of to-be-remembered stimulus presentation and distractor task problem was repeated a number of times until a recall screen appears. Once it appeared, subjects attempted to recall all the to-be-remembered stimuli that they saw in the correct serial order in which they were presented (see Figure 1 for examples of all 3 tasks).

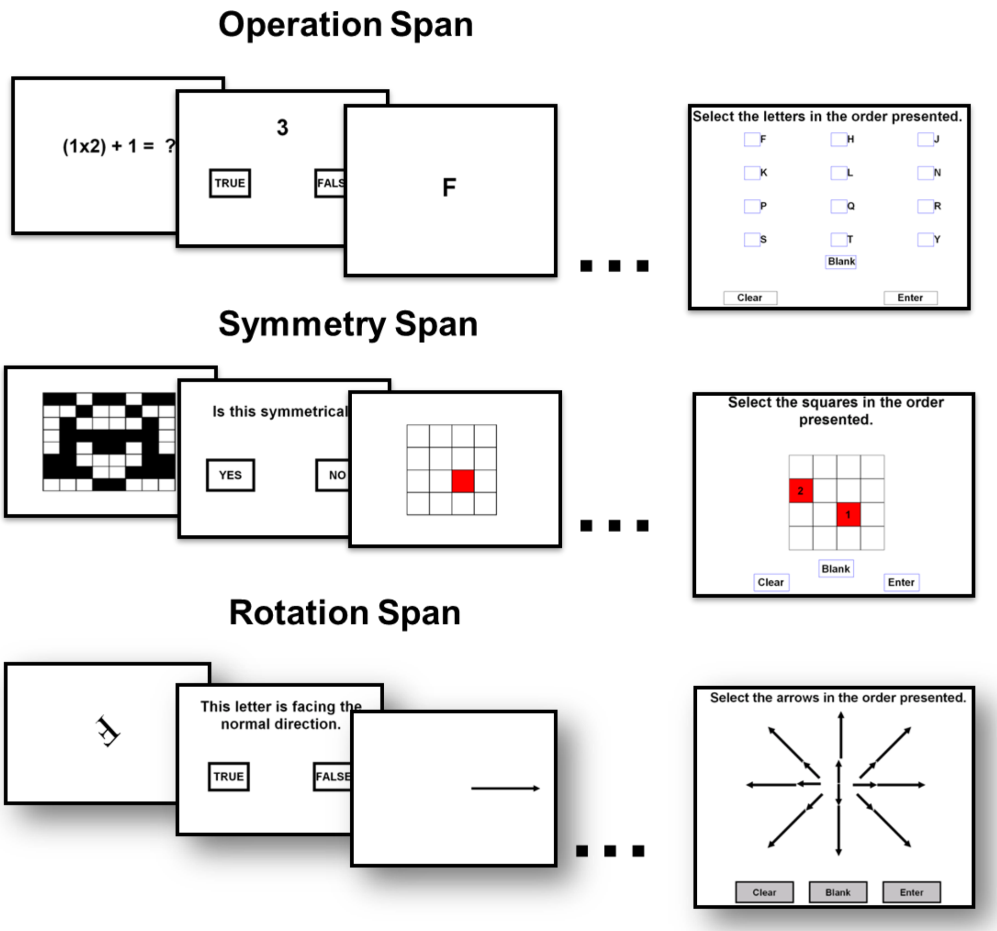
Operation Span
(1x2) + 1 = ?
3
TRUE
FALS
F
Select the letters in the order presented.
F
H
J
K
L
N
P
Q
R
S
T
Y
Blank
Clear
Enter
Symmetry Span
Is this symmetrical
YES
NO
Select the squares in the order presented.
2
1
Blank
Clear
Enter
Rotation Span
This letter is facing the normal direction.
TRUE
FALSE
Select the arrows in the order presented.
Clear
Blank
Enter

Figure 1. Examples of each of the complex span tasks.

2.3.1 Automated Operation Span Task (OSpan; Unsworth, Heitz, Schrock, & Engle, 2005).

For the operation span task subjects had to remember letters and solved simple math equations for the processing task. The to-be-remembered items consisted of 12 phonologically dissimilar letters and the math equations required subjects to perform two simple math operations. There were 14 trials in total and the set-size of each trial ranged from 3-9 letters (two trials of each set size). The dependent variable of interest was the number of letters subjects remembered in correct serial position across the entire task (i.e., the partial-load score, Conway et al., 2005).

2.3.2 Automated Symmetry Span Task (SymSpan; Kane et al., 2004).

For this task, subjects had to remember matrix locations on a 4x4 matrix while they made symmetry judgments. Subjects were shown an 8x8 grid with black and white squares (as shown in Figure 1). They made a judgment about whether the grid was symmetrical about its vertical axis. Afterwards they were presented with the to-be-remembered item, a 4x4 matrix with one of its elements highlighted in red. This process continued until a recall screen appeared. There were 14 trials in total and the set-size of each trial ranged from 2-8 matrix locations (two trials of each set size). Like the operation span tasks, the dependent variable of interest was the partial score.

2.3.3 Automated Rotation Span Task (RotSpan; Kane et al., 2004).

For this task, the processing component consisted of subjects making judgments about whether letters, when rotated to an upright position, were facing the correct direction or were mirror-reversed (see Figure 1). The memory component involved subjects remembering arrows of two different sizes in one of 8 positions (for a total of 16 possible arrows). There were 14 trials in total and the set-size of each trial ranged from 2-8 arrows (two trials of each set size). Like the other two complex span tasks, the dependent variable of interest was the partial score.

2.4 Gf Tasks

2.4.1 Raven's Advanced Progressive Matrices (RAPM; Raven, Raven, & Court, 1998).

For this task, subjects were presented with a 3x3 figure with the bottom-right hand items of the figure missing. The items in the figure were arranged in such a way to follow a logical pattern. Subjects had to determine what this logical pattern is and select the item that completed the figure out of 8 answer choices. Subjects completed the 18 odd problems of the Raven's Advanced Progressive Matrices task and were given 10 minutes to all the problems. The dependent variable of interest was the number of correctly completed problems.

2.4.2 Letter Sets (Ekstorm, French, Harman, & Dermen, 1976).

Subjects were presented with five sets of letters. Each set of letters consisted of 4 letter. Four out of the five sets of letters followed a specific rule (e.g., all sets were in alphabetical order or all sets contained the letter "H"). Subjects had to first discover what the rule was and then select the set of letters that followed this rule. Subjects had 7 minutes to complete 30 problems. The dependent variable of interest was the number of correctly completed problems.

2.4.3 Number Series (Thurstone, 1938).

For each item of this tasks, subjects were presented with a series of numbers. These series followed a specific rule (e.g., each new number required that you add the two previous numbers together or each new number required that you add two to the previous number). Subjects had to determine the specific rule that governed each problem and then selected the next number that correctly completed the sequence out of 5 answer choices. Subjects had 5 minutes to complete 15 problems. The dependent variable of interest was the number of correctly completed problems.

2.5 N-back Tasks

There were a total of 6 n-back tasks. Half of the tasks had a small stimulus pool (10 stimuli) similar to most of the n-back tasks in the previous literature. The other half of the tasks had a large stimulus pool (63 stimuli) similar to the Shipstead, Harrison, and Engle (2016) study. Additionally, like our previous study, there were three different types of to-be-remembered-items (words, wingdings, and faces). Thus, there are two n-back tasks (small stimulus pool and large stimulus pool) for each of the three stimulus types. Each of the n-back tasks consisted of 120 trials with 10% of the trials being targets. For the large stimulus pool n-back tasks, a particular stimulus only appeared 1 or 2 times throughout the entire task. For the small stimulus pool n-back tasks, each stimulus occurred approximately 12 times. In each of the n-back tasks we included 2, 4, 5, 6, 7, and 8-back lures. Each of these lure types appeared approximately 8 times in each of the n-back task. Differing from the Shipstead et al. study, subjects were required to make a response (target present/target absent) for every trial instead of just making a target present response (e.g., go/no-go task). This was changed to prevent subjects from not responding during the entire task. There were two critical dependent measures for the n-back tasks. The first was the overall *d'* score for each of the tasks. This dependent measure was used for all of the latent variable analyses. The second was the number of false alarms that subjects make to each particular lure type.

CHAPTER 3. RESULTS

The descriptive statistics for all of our tasks are presented in Table 2.

Table 2. Descriptive statistics. The dependent measure for the n-back tasks was *d'*. RAPM = Raven's Advanced Progressive Matrices.

	Mean	Standard Deviation	Skew	Kurtosis
Operation Span	57.34	14.33	-0.39	1.23
Symmetry Span	27.85	9.21	-0.23	-0.90
Rotation Span	30.42	10.67	-0.78	0.23
RAPM	8.32	3.54	-1.21	-2.12
Letter Sets	14.69	5.09	-0.89	-1.62
Number Series	6.89	2.45	-0.43	0.34
Word N-back Large Pool	1.62	0.43	1.41	2.43
Wingding N-back Large Pool	1.49	0.39	1.87	-1.73
Face N-back Large Pool	1.23	0.59	1.21	0.54
Word N-back Small Pool	1.19	0.32	1.56	-0.92
Wingding N-back Small Pool	1.30	0.34	2.76	-1.61
Face N-back Small Pool	0.94	0.82	1.87	-0.43

3.1 Do Complex Span Tasks and N-back Tasks Measure the Same Construct?

To determine whether the complex span tasks and n-back tasks measured the same latent construct, we first decided to conduct an exploratory factor analysis with all complex span and n-back tasks. To determine the number of factors that should be extracted, we used Kaiser's criterion of extracting the number of factors that have eigenvalues greater than 1.00 in a principal components analysis. Three eigenvalues fit this criterion so I extracted three factors and used a Varimax rotation to try to approximate simple structure (i.e., to make the factors more interpretable). The factor loadings are presented in Table 3.

Table 3. The factor loadings for the exploratory factor analysis. The factors have been rotated by a Varimax rotation and I specified to extract 3 factors because only 3 met Kaiser's criterion.

	1	**2**	**3**
Operation Span	.53	.04	.12
Symmetry Span	.58	.02	.18
Number Series	.65	.08	.14
Word N-back Large Pool	.22	.64	.08
Wingding N-back Large Pool	.14	.45	.10
Face N-back Large Pool	.11	.53	.12
Word N-back Small Pool	.27	.35	.30
Wingding N-back Small Pool	.16	.32	.28
Face N-back Small Pool	.09	.28	.36

There are two critical points to note about the exploratory factor analysis. The first point is that we had to extract 3 factors. If complex span and n-back tasks measured the same construct, a 1 factor solution would be able to account for the bulk of the variance. However, three eigenvalues were greater than 1.00 so 3 factors were needed to account for the majority of the variance in these tasks. A quick glance at the factor loadings shows that Factor 1 seems to be a complex span factor because the 3 complex span tasks have the highest loadings on that factor. Factor 2 seems to be a mix between a general n-back factor and large stimulus pool n-back factor. Factor 3 is a small stimulus pool n-back factor but the small stimulus pool n-back tasks seem to load equally on both Factors 2 and 3. The second major point gleaned from the exploratory factor analysis is that it is somewhat difficult to separate the two categories of n-back tasks (i.e., large stimulus pool and small stimulus pool).

While the exploratory factor analysis provided evidence that we should accept a 3 factor solution for our working memory capacity tasks, there are some limitations with exploratory factor analysis. There are two major limitations concerning our data. First, there is no way to account for the correlations between the n-back tasks that shared common stimuli (e.g., the two n-back tasks with faces). Working memory researchers assume that variance attributed to remembering particular stimuli is not related to a domain-general WMC factor so a statistical model that can attribute this variance to error would be preferable. Second, there is not a perfect criterion for researchers to determine the number of factors to extract. It could be the case that we overfactored and Factor 3 is not important.

To provide converging evidence that n-back and complex span tasks measure different cognitive abilities, three confirmatory factor analyses were conducted[1]. CFA-1 Factor had all of the working memory tasks loaded onto one factor. We also included correlated errors for the n-back tasks that required memory for the same type of stimuli. If this model fit the data well, there would be evidence that both complex span tasks and n-back tasks measured the same latent construct (WMC). This model is presented in Figure 2.

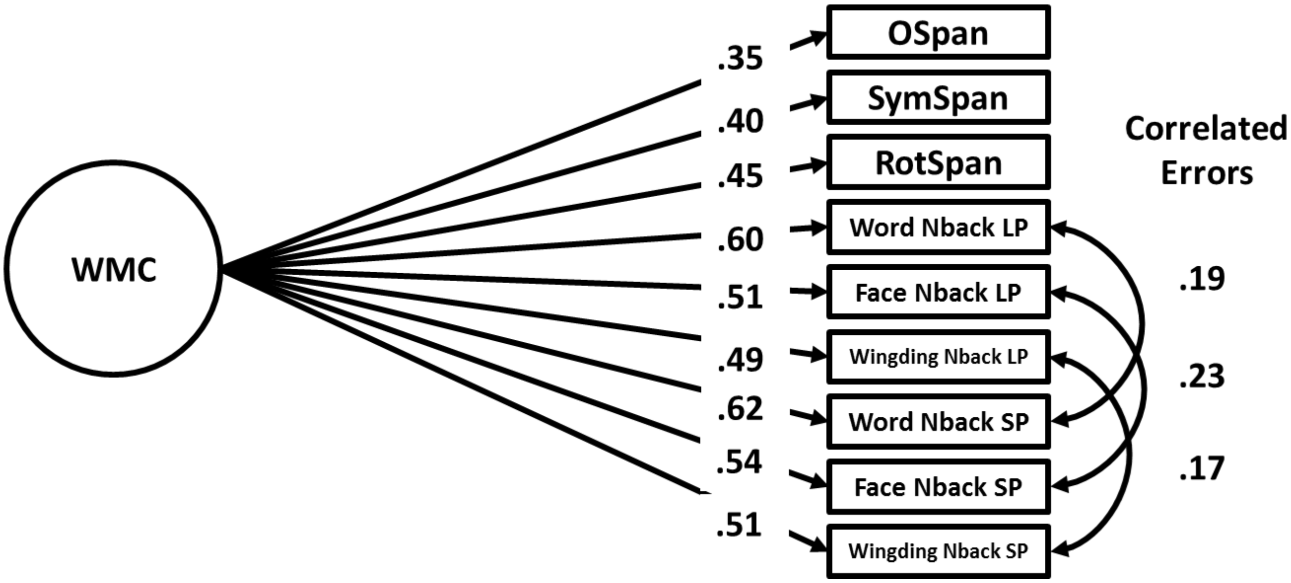

Figure 2. CFA-1 Factor: all complex span tasks and n-back tasks loading onto the same factor. LP = Large stimulus pool, SP = Small stimulus pool. Fit statistics: χ^2 (33) = 44.36, *p* < .05, CFI = .73, RMSEA = .23.

From the fit statistics of this analysis we see that the model fitted the data poorly suggesting the n-back tasks and complex span tasks measure different cognitive abilities. For the next model, CFA-2 Factor, all the n-back tasks were loaded onto one factor and all the complex span tasks were loaded onto another factor. This model tested whether n-back and complex span measured separate factors and the stimulus pool manipulation did not change the construct that n-back tasks measured. The results of this analysis are presented in Figure 3.

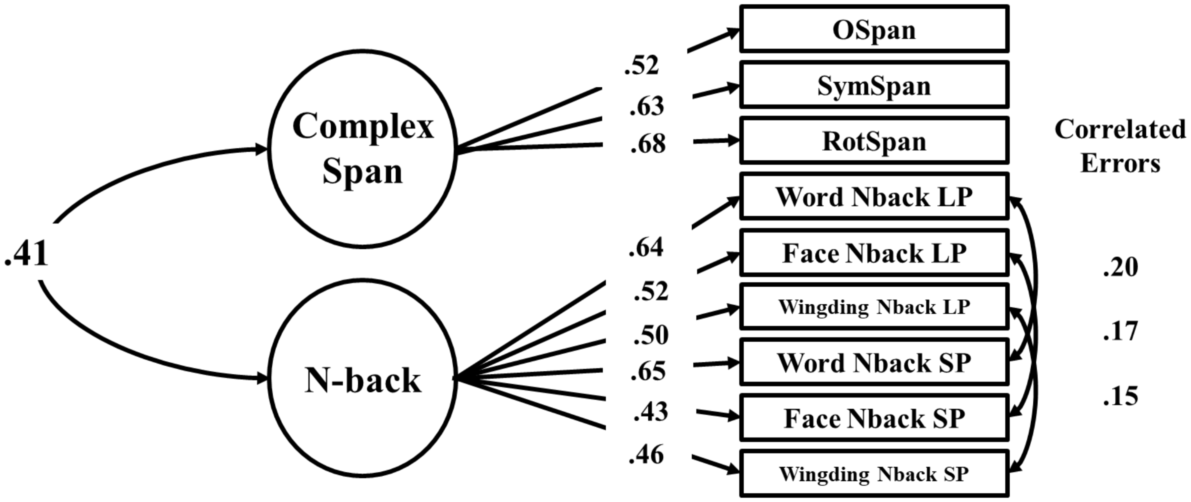

Figure 3. CFA-2 Factor: all complex span tasks loading onto a separate factor than the n-back tasks. LP = Large stimulus pool, SP = Small stimulus pool. Fit statistics: χ^2 (32) = 35.67, *p* < .05, CFI = .81, RMSEA = .12.

Although the fit of the 2 factor model (CFA-2 Factor) was better than the fit for the 1 factor model (CFA-1 Factor), the model still did not fit well (CFI < .90 and RMSEA > .05). The final model we tested was for a 3 factor solution (CFA-3 Factor) with the two different types of n-back tasks (large stimulus pool and small stimulus pool) loading onto two separate factors. The model is presented in Figure 4.

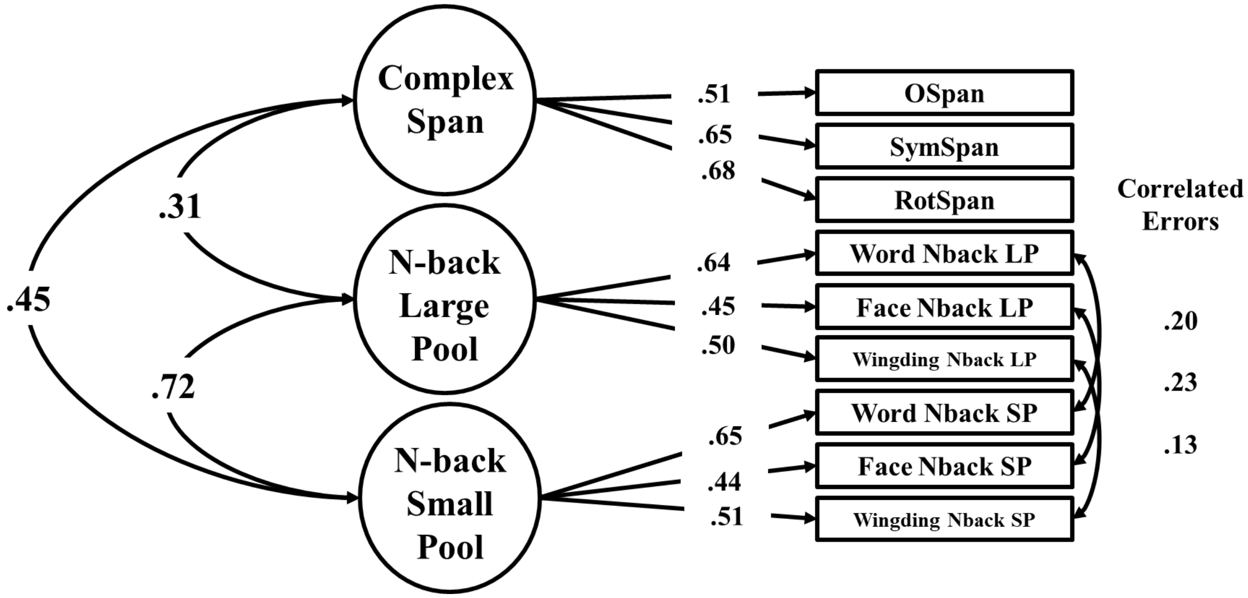

Figure 4. CFA-3 Factor: all complex span tasks and n-back tasks loading onto the same factor. LP = Large stimulus pool, SP = Small stimulus pool. Fit statistics: χ^2 (30) = 30.38, *p* < .05, CFI = .92, RMSEA = .04.

The fit of CFA-3 Factor was good. The results of our CFA models and exploratory factor analysis both converged on a three factor solution. Thus, the data suggest two conclusions 1.) n-back tasks and complex span tasks measure different cognitive constructs and 2.) the manipulation of the stimulus pool size changed what the n-back tasks were measuring. This last point does not necessarily mean that the two different types of n-back measure two separable cognitive abilities. It could be the case that both categories of tasks measure the same construct but that the small pool (or large pool n-back tasks) are measuring an additional cognitive construct (e.g., interference) and this is leading the 3 factor solution to fit the best.

3.2 Does the Manipulation of Stimulus Pool Size Change the Correlations of N-back Tasks to Complex Span Tasks and Measures of Fluid Intelligence?

To test to test whether the n-back tasks with small stimulus pools correlated more strongly with both complex span tasks and Gf than n-back tasks with large stimulus pools a confirmatory factor analysis (CFA-3 Factor Gf) was conducted with 4 factors: complex span tasks, n-back tasks with a large stimulus pool, n-back tasks with a small stimulus pool, and Gf. The results of this analysis are presented in Figure 5.

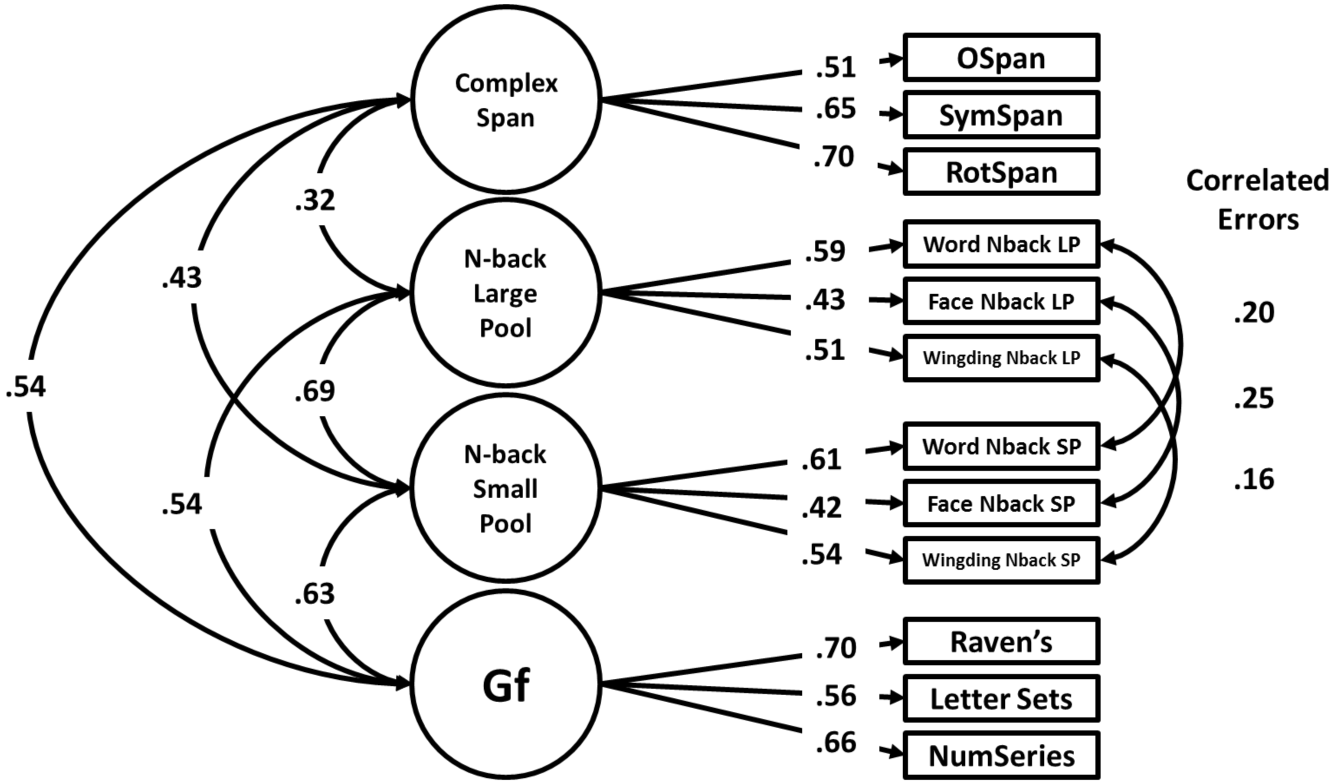

Figure 5. CFA-3 Factor Gf - Confirmatory factor analysis with complex span tasks, n-back tasks with a large pool of stimuli, and n-back tasks with a small pool of stimuli all loading onto the different factors. LP = Large stimulus pool, SP = Small stimulus pool. Fit statistics: χ^2 (57) = 84.05, *p* < .05, CFI = .91, RMSEA = .06.

The first point to note about this model is that the fit of this model is good (RMSEA at .06 or below and CFI above .90; Byrne, 1994). Next, we see that the n-back factor for the tasks with a small stimulus pool correlates more strongly with both the complex span factor and with Gf than with the n-back factor for the large stimulus pool tasks. To test whether these differences were significantly different the factor correlations between the complex span factor and the small pool n-back factor and between the complex span factor and the large pool n-back factor were set to be equal. This significantly hurt model fit, $\Delta\chi^2$ (1) = 6.47, $p < .05$, showing that the small pool n-back factor is more strongly correlated to the complex span factor than the large pool n-back factor is. A similar process was used to test the difference between both n-back factor's correlations to Gf, $\Delta\chi^2$ (1) = 4.90, $p < .05$. The analysis revealed that the small pool n-back factor was more strongly correlated with Gf than the large pool n-back factor is. Overall, it seems that the n-back factor with the small stimulus pool tasks was more strongly correlated with other measures of cognitive ability.

Finally, I sought to examine whether the small pool n-back factor had a stronger relationship with Gf than that of the complex span factor. I set the Gf/small pool n-back and the Gf/complex span correlations to be equal and this significantly hurt model fit, $\Delta\chi^2$ (1) = 5.13, $p < .05$. This analysis shows that Gf is more strongly correlated with an n-back factor than with a complex span factor, at least for some n-back tasks. This finding has potentially important ramifications if researchers are interested in the relationship between WMC and Gf. The answer to the strength of this relationship would depend on which type of working memory task the researchers used. This finding may also enlighten which cognitive processes both tasks are measuring.

Even though all 3 factors are correlated with Gf in the previous confirmatory factor analysis, it could be the case that they are all predicting the same variance in Gf. To answer this question, I conducted a structural equation model (SEM) with all 3 working memory factors predicting Gf. With this analysis, I can show which WMC factors uniquely predict Gf. The results of this analysis are presented in Figure 6.

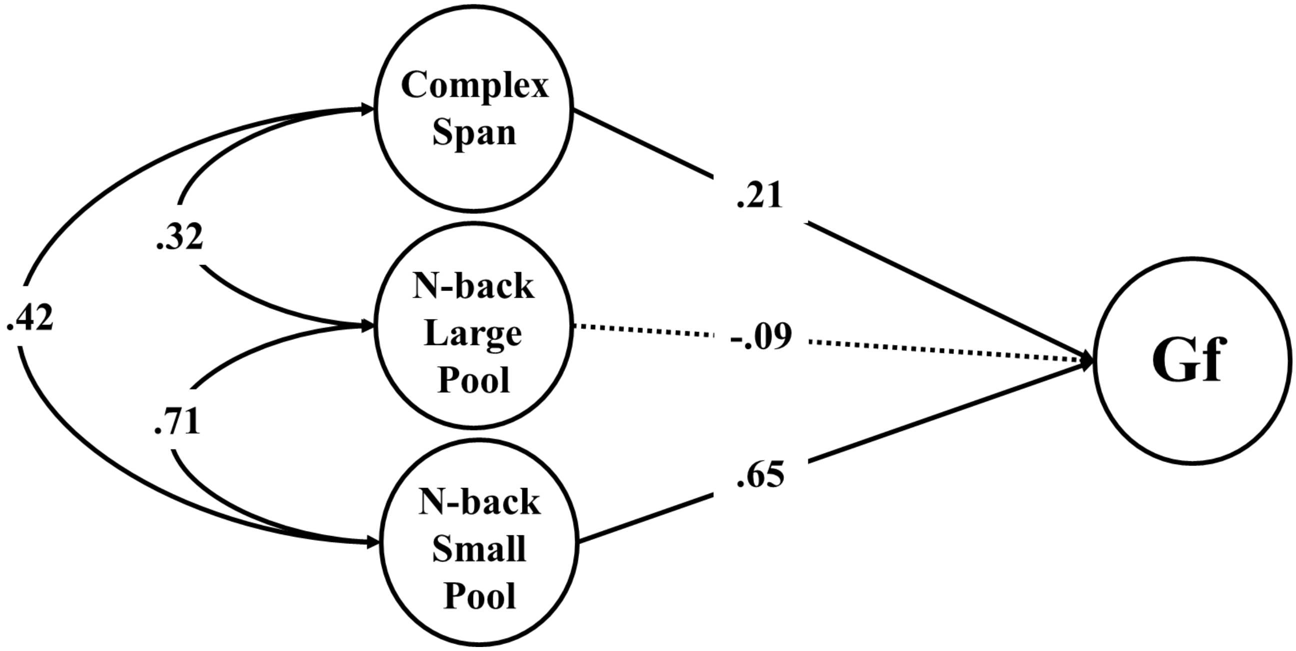

Figure 6. SEM – The structural model of the analysis with each WMC factor predicting Gf. The measurement model was nearly identical with the results of CFA-3 Factor Gf. Fit statistics: χ^2 (57) = 80.35, *p* < .05, CFI = .92, RMSEA = .07.

The model fit of our structural equation model was acceptable (ideally RMSEA should be lower). The critical finding of this model is that only the complex span and the n-back small stimulus pool factor uniquely predict Gf. Although the n-back large stimulus pool factor was significantly correlated with Gf in the previous confirmatory factor analysis (CFA-3 Factor Gf), it does not predict Gf above and beyond the complex span factor and the n-back small stimulus pool factor. If a researcher was looking to predict Gf from a variety of measures of working memory, the researcher should include both complex span tasks and n-back tasks because both tasks are predicting unique variance in Gf. Additionally, if a researcher had to choose a single working memory task to predict Gf, from these data, an n-back task with a small stimulus pool, would do the best job of predicting Gf.

3.3 **Does the Relationship Between Lures in the N-back Task and Cognitive Ability Change with the Stimulus Pool Size of the N-back Task?**

For the lure analyses, the false alarm rate for each lure position (2-back through 8-back) was calculated for each of the 6 n-back tasks. These false alarm rates were collapsed across the n-back tasks with the same stimulus pool size (the pattern of results were similar for the three different types of stimuli). The results of these analysis for the lures correlations to the complex span factor are reported in Figure 7 and the results for the correlations to Gf are reported in Figure 8.

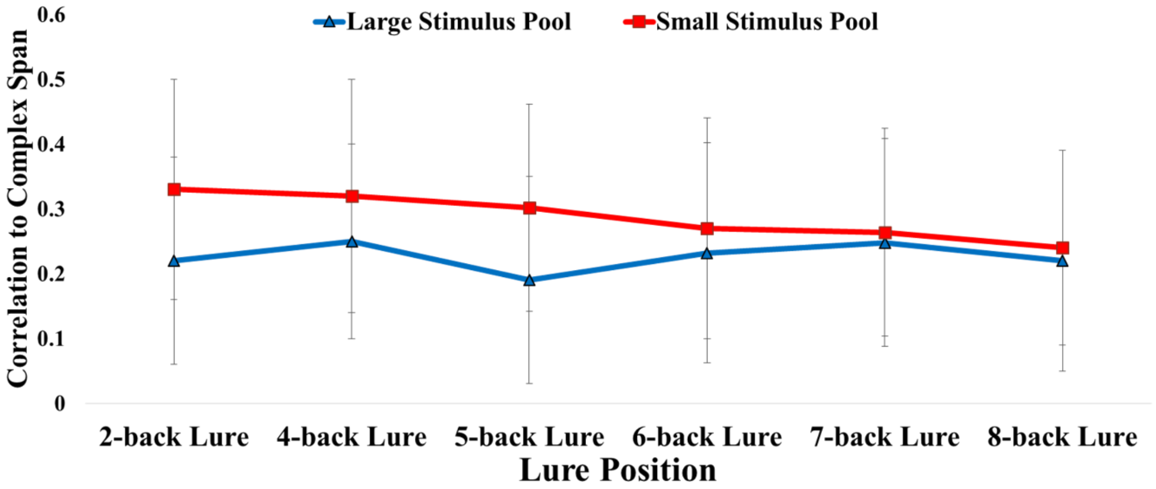

Figure 7. The correlations of false alarms for each lure position to the complex span factor collapsed across stimulus type (word, face, and wingding) and separated by the size of the stimulus pool. We flipped the sign of the correlations (false alarms were negatively correlated with complex span performance) to aid the reader. The error bars represent 95% confidence intervals of the correlations.

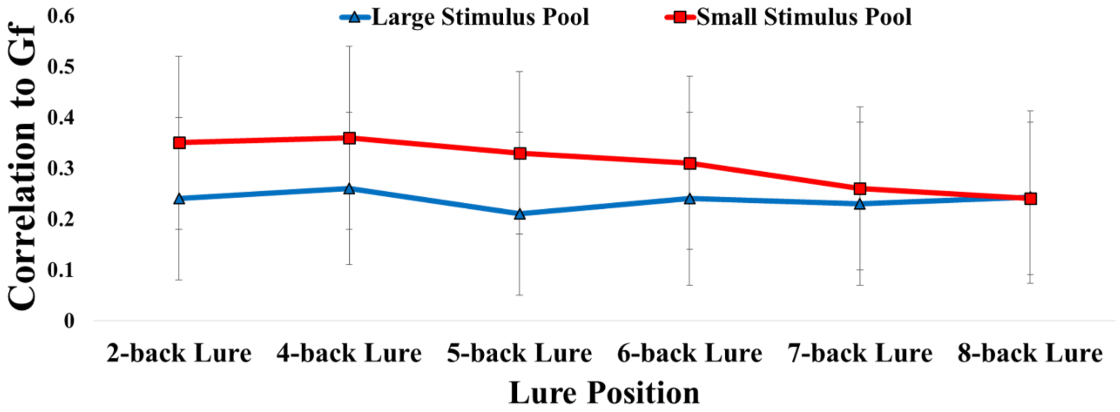

Figure 8. The correlations of false alarms for each lure position to Gf collapsed across stimulus type (word, face, and wingding) and separated by the size of the stimulus pool. We flipped the sign of the correlations (false alarms were negatively correlated with Gf) to aid the reader. The error bars represent 95% confidence intervals of the correlations.

The first thing to note about these analyses is that false alarms to lure trials are more strongly correlated to both complex span and Gf in the small stimulus pool n-back tasks compared to the large stimulus pool n-back tasks. To test this, we measured the average number of false alarms subjects made to lure trials for both the large stimulus pool and the small stimulus pool task regardless of lure position and regardless of stimulus type. The correlation between complex span performance and the average number of false alarms subjects made during the small stimulus pool n-back tests (*r* = .32) was significantly greater than the correlation between complex span performance and false alarms during the large stimulus pool tasks (*r* = .25) when we used a repeated-measures test of correlational differences (Cohen, Cohen, Aiken, & West, 2013). The same is true for the correlations between Gf and small pool n-back tasks (*r* = .35) and between Gf and large pool n-back tasks (*r* = .29).

Additionally, false alarms on lure trials closer to the target n (i.e., lure trials 2-back and 4-back) are more strongly correlated to both complex span performance and Gf that lure trials in which the distance of the lure is further back in time (e.g., 7-back or 8-back). However, this finding only holds up for the small stimulus pool n-back tasks and not the large stimulus pool n-back tasks.

CHAPTER 4. DISCUSSION

The present study was designed to answer three questions: 1.) whether both types of WMC tasks (complex span tasks and n-back tasks) measure the same latent ability. 2.) whether the size of the stimulus pool changed the correlation of the n-backs tasks to both complex span tasks and Gf. 3.) whether the relationship of false alarms during lure trials in the n-back tasks to complex span measures and Gf change when the size of the stimulus pool was manipulated.

4.1 What Makes a Task a Working Memory Task?

At the latent level, with multiple measures of both complex span and n-back, we showed that the two different types of tasks measure different cognitive constructs (with the n-back tasks additionally measuring two separate constructs when stimulus pool size is manipulated). This is potentially problematic because researchers (e.g., Jaeggi, Buschkuehl, Jonides, & Perrig, 2008) claim that both of these tasks measure WMC and use findings based of one type of task (e.g., complex span) to inform their findings using a different type of task (e.g., n-back). At this point in time the reader might be asking themselves whether n-back tasks or complex span tasks are the "true" measure of WMC. To adequately answer this question, we must go beyond the results of this research study.

As mentioned in the introduction, working memory is the memory system that is used for maintaining information for brief periods of time in the presence of interference. The differences in the functioning of this memory system is "working memory capacity." Like every hypothetical construct in psychology, we cannot measure WMC directly. There is no instrument that researchers can insert into the brain and get a complete measurement of WMC. Rather, it must be inferred from performance on tasks that we theoretically believe capture this ability, which is in turn complicated by the fact that almost all cognitive tasks require a multitude of abilities, not all of them are critically important to the researcher. For example, the operation span task requires the maintenance for letters in memory for a brief amount of time, the ability to solve math problems, the ability to disengage from previously correct answers, the ability to recognize letters, and many other sources of variation.

How does a researcher isolate the sources of variation that are of interest to them? Experimental and differential researchers have different methodologies for tackling this problem. Experimental psychologists will create a control task that requires all the cognitive processes except the ones of interest (e.g., naming colored X's in the Stroop task; Stroop, 1935) and compare performance with a task that requires all the same processes as the control task plus a cognitive process or processes of interest (e.g., naming color words in which the hue of the word does not match what the word reads in the Stroop task). Neuroimaging studies use a similar line of reasoning when comparing brain activity during an experimental block with activity during a control block (Huettal, Song, & McCarthy, 2008). On the other hand, differential psychologists will measure multiple tasks that are all theorized to measure an ability and look at the common variance among all the tasks. Ideally, all the cognitive processes that are not important to the researcher will end up in the error term of a model and the researcher should be left with a relatively process-pure measurement of the construct of interest.

Both approaches have their advantages and disadvantages in trying to measure an isolated cognitive process or ability. Experimental methodologies are well equipped at determining which manipulations harm or help a given cognitive process for a group of individuals. However, some of the tasks that are particularly useful for experimental psychologists do a poor job at measuring a cognitive ability between individuals. For instance, the attention network task uses response time from different varieties of flanker trials to compute 3 different response time difference scores corresponding to 3 different attention networks (Fan, McCandliss, Sommer, Raz, & Posner, 2002). Although the task is elegant, and, on the surface, it seems like the task is parsing out 3 different cognitive processes there are some major problems with using this task for measuring individual differences in attention. The task relies on the use of difference scores which are known to be psychometrically unreliable (Cronbach & Furby, 1970) and, empirically, the alerting and orienting networks effects of the task are not reliable (Redick & Engle, 2006). In many instances, researchers want to use the tasks that they have used in their experiments for assessment of a particular cognitive ability without having a good idea about what the task actually measures.

Alternately, differential methodologies are well suited at reliably measuring a construct and empirically showing that the construct that they are measuring has real-world applications (i.e., criterion validity). However, differential researchers run the risk of drawing inappropriate conclusions about the structure of cognitive abilities because of the tasks they selected to measure. For instance, my colleagues and I recently conducted a study in which we used only 3 complex span tasks to measure WMC (Harrison, Shipstead, & Engle, 2015). We drew conclusions about why matrix reasoning tasks were related to

WMC. However, it could be argued that our WMC construct was too narrow and did not include all the relevant cognitive sub-processes that are important for WMC. It could also be argued that maybe our results were a product of the task-specific variance related to performance on the complex span tasks. Both of these concerns are valid and all differential studies are vulnerable to them. Another problem that one encounters with differential research, is the tendency for researchers to uncover a factor using factor analysis and reifying that factor to be the same thing as a cognitive ability. It could be the case that the researcher "discovered" a factor that is simply due to the method of stimulus presentation in the task or due to the response collection of the task.

Back to the question, are complex span tasks and n-back tasks both measures of WMC or is one measure "better" than the other in measuring WMC. There are many important sub-processes that underlay the ability of WMC. For instance, the ability to maintain information for a brief period of time, the ability to control the focus of attention, the ability to selectively forget or disengage from previously relevant information, and others (e.g., Duncan & Owen, 2000; Shipstead, Lindsey, Marshall, & Engle, 2014; Unsworth, Fukuda, Awh, & Vogel, 2014). Both the complex span and the n-back tasks measure these processes to a certain extent. However, the n-back tasks require subjects to disengage from information to a greater extent than the complex span tasks. Additionally, the complex span tasks require a more difficult retrieval from working memory compared to the n-back tasks (because subjects have to generate the correct answer and not recognize the correct answer). If researchers were to use only one category of task, they would be missing out on measuring the entirety of WMC and potentially underestimating the role that WMC plays in whatever construct they are studying. The present research shows that relying on a single category of WMC task hinders the prediction of WMC to an important cognitive ability (Gf).

The outcomes of this research dovetail with a recent theory that my colleagues and I have proposed (Shipstead, Harrison, & Engle, 2016). In this recent article, we argued that complex span tasks are particularly good at predicting performance on a cognitive measure when the maintenance of information is required. Additionally, measures of Gf are particularly well-suited at predicting performance when a task requires the disengagement of previously relevant information. For instance, verbal fluency tasks (e.g., name as many animals as you can think of without repeating) requires subjects to search for exemplars of a particular category in memory without recalling exemplars that have been previously generated. We found that Gf predicts verbal fluency tasks above and beyond WMC and a host of other cognitive abilities. This theory might explain why these two different categories of WMC tasks, n-back and complex span, are less related to one another than they are to Gf. The ability to maintain information is particularly important in the complex span tasks because the subject is required to recall the to-be-remembered stimuli in a particular order. For the n-back task, the subject does not need to recall any information but has to recognize whether a particular stimulus is the same one they saw n trials ago. Thus, the difficulty in the n-back tasks is not to maintain information but to disengage from information that is no longer relevant to the task (i.e., stimuli further back than 3 trials ago). I argue that complex span tasks are better are measuring the maintenance of information in WMC while n-back tasks are better at measuring the disengagement of information in WMC.

Complex span and n-back tasks fail to load onto the same factor not because one category of tasks is a "true" measure of WMC and the other is not, but because each task emphasizes a different sub-process of WMC. If a researcher wants to assess WMC to the best of their ability, they should incorporate both the n-back and the complex span tasks.

4.2 Does Manipulating Stimulus Pool Size Change What N-Back Tasks Measure?

We found that the size of the stimulus pool for the n-back tasks changes the correlations between those tasks to both complex span and Gf. The stimulus pool manipulation had such a profound impact on what the n-back tasks measured that we had to construct a model with 2 n-back factors to adequately model the data. The more times a particular stimulus repeats in the n-back task, the more interference there is for the subject in determining whether they saw this particular stimulus *n* trials ago. According to prominent models of WMC (e.g., Engle & Kane, 2004), increasing the interference of a cognitive task should increase that task's correlation to WMC. This pattern of results is exactly what we found in the present research. However, not only were the n-back tasks with a small stimulus pool size more strongly correlated with a complex span factor compared to n-back tasks with a large stimulus pool size, but they were more strongly correlated to Gf as well. To help explain these results it will be fruitful to discuss research from the recognition memory literature.

Although the time between encoding and retrieval is short for the n-back tasks, they can still be considered recognition memory tasks. The most prominent theory of recognition memory is the dual-process theory (e.g., Jacoby, 1983; Yonelinas, 2002). Although the specifics of the theory change from researcher to researcher, the crux of the theory remains the same; recognition memory is accomplished by two separable cognitive processes: familiarity and recollection. Familiarity is a quicker cognitive process in which a subject has a vague feeling of remembering a particular item but cannot remember the context in which the item was encoded. Recollection is a slower process in which the subjects remembers the context in which the item was presented and is more confident that the item was previously presented. They are many lines of evidence demonstrating that recollection and familiarity are distinct cognitive processes. Most germane to the present discussion are findings in which older adults have impaired recollection compared to young adults but their familiarity is relatively intact (e.g., Naveh-Benjamin, 2000). Older adults are also known to have impaired WMC compared to young adults (e.g., Hasher & Zacks, 1988).

For n-back tasks with a large stimulus pool size there is not much interference between the to-be-remembered stimuli. Thus, subjects can rely both on a graded sense of familiarity or recollection to successfully complete the task. If only recollection is related to WMC, then familiarity acts as a suppressor effect and lowers the correlation of the n-back task to other measures of WMC and to Gf. However, this is not the case with n-back tasks with a small stimulus pool size. Because familiarity is not helpful in this task (stimuli are repeated multiple times), subjects are more likely to rely on recollection. If recollection is the mechanism behind the tasks' correlation to both complex span tasks and Gf, it stands to reason why the small stimulus pool n-back tasks are more strongly related to other cognitive abilities.

An important point to make here is that the stimulus pool size is not a manipulation of n-back tasks that has ever been systematically studied in an individual differences paradigm. When set size is manipulated, the n-back tasks loaded onto two separate factors. This just goes to show the importance of knowing how every aspect of a task can change what the task will measure.

The final point that I wish to make about the n-back tasks with small stimulus pools is that they had a significantly higher correlation to Gf than the complex span tasks. This is particularly striking because the specific complex span tasks that were used in this study have been modified over the course of 3 decades to maximize their reliability and validity (Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005). The n-back tasks used in this study were created for the first time with little insight into what aspects of the tasks would increase both the reliability and the validity. This suggests that there is a particular sub-process of WMC that is measured by the n-back tasks with small stimulus pool sizes that is strongly related to Gf. I argue that this sub-process is the ability to disengage from previously relevant information. As mentioned before, this ability is theoretically linked with tasks that measure Gf (Shipstead, Harrison, & Engle, 2016).

4.3 **Do Lure Trials Closer to** *n* Correlate More Strongly to Cognitive Ability?

From our previous study, I expected to replicate the finding that false alarms on lure trials farther back in time (e.g., 7-back or 8-back lures) would be most strongly correlated with complex span tasks and Gf compared to lures closer to *n* (e.g., 2-back or 4-back lures) for the n-back tasks with a large stimulus pool size (Shipstead, Harrison, & Engle, 2016). This result was not replicated in the present research study. For the n-back tasks with large stimulus pools lure position did not change the correlation of false alarms to other cognitive abilities. The difference in the results of the two studies could be due to several factors. First, in the previous study, n-back tasks were constructed to either have lures close to *n* (i.e., near lures) or lures farther away from *n* (far lures). Subjects would receive the two different versions of the n-back tasks on two different days, sometimes a week apart. Practice effects might have caused the false alarms to lures farther back in time to be more strongly correlated with cognitive ability (for two of the three pairs of n-back tasks the near lure version occurred first). Additionally, having lure trials until occur at a certain range of positions may have led subjects to set a certain expectation while performing the task. A similar effect is found with the Stroop task (MacLeod, 1991). Increasing the proportion of congruent trials in the Stroop task increases the interference effects of the incongruent trials. Perhaps having no lure trials close to *n* changed how subjects performed the task. For the present research, all lure types were manipulated throughout the task.

Another difference was that the present research required subjects to make a decision for every trial. Subjects had to press a key if they thought the trial was a target or press a different key if they thought the trial was not a target. For the Shipstead et al. study, the task was a go-no task and subjects had to press a key only when they saw a target. This led some subjects to not press a key throughout the entire task and, because we did not know if the subjects understood the instructions, we had to remove a significant number of our subjects from further analyses (approximately 15%). These two differences may have accounted for the discrepancy of the two studies' findings.

On a more important note, we found that lures closer to *n* were more strongly correlated with both complex span performance and Gf for the n-back tasks with small stimulus pools. This finding makes more sense with theories of WMC (e.g., Engle & Kane, 2004). Interference increases the relationship of a task to WMC. Thus, the more interfering lures (i.e., those closer to *n*) should be more strongly correlated with WMC. These findings dovetail with a recent study in which prefrontal activity predicted how well both young and older adults were able to successfully retrieve memory items in the presence of lures (Fandakova, Lindenberger, & Shing, 2014). More frontal activity is needed for tasks with more cognitive interference and this should lead to a larger correlation with Gf.

4.4 Final Remarks

The present research is much more than a study about how to measure WMC. For instance, I showed that the ability to disengage from previously relevant information is measured by the n-back task and that this is why the n-back tasks can be more strongly correlated to Gf. This is important to any researcher interested in the nature of WMC. A researcher will miss a criterial aspect of WMC if only one category of WMC task is measured. Finally, I showed that changing the stimulus pool size and the presence of lure trials changes which subprocesses of WMC are required to perform the n-back task.

REFERENCES

Awh, E., Jonides, J., Smith, E. E., Schumacher, E. H., Koeppe, R. A., & Katz, S. (1996). Dissociation of storage and rehearsal in verbal working memory: Evidence from positron emission tomography. *Psychological Science, 7*, 25-31.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, *5*, 49-62.

Broadway, J. M., & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, 42, 563-570. doi:10.3758/BRM.42.2.563

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. Sage.

Cattell, J. M. (1886). *Psychometrische Untersuchungen*. W. Engelmann.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

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

Cowan, N. (2001). Metatheory of storage capacity limits. *Behavioral and brain sciences*, *24*, 154-176.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist*, *12*, 671.

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": Or should we?. *Psychological bulletin*, *74*, 68-80.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior,* 19, 450-466. doi:10.1016/S0022-5371(80)90312-6

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422-433.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in neurosciences*, *23*, 475-483.

Engle, R. W., Nations, J. K., & Cantor, J. (1990). Is" working memory capacity" just another name for word knowledge?. *Journal of Educational Psychology,82,* 799.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General*, *12,* 309.

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

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of cognitive neuroscience*, *14*, 340-347.

Fandakova, Y., Lindenberger, U., & Shing, Y. L. (2014). Deficits in process-specific prefrontal and hippocampal activations contribute to adult age differences in episodic memory interference. *Cerebral Cortex, 24*, 1832-1844.

Galton, F. (1883). *Inquiries Into the Human Faculty & Its Development*. JM Dent and Company.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*, 79-132.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature neuroscience*, *6,* 316-322.

Harrison, T. L., Shipstead, Z., & Engle, R. W. (2015). Why is working memory capacity related to matrix reasoning tasks?. *Memory & cognition*, *43*, 389-396.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of learning and motivation*, *22,* 193-225.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of educational psychology*, *57*, 253-270.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6,* 1-55.

Jacoby, L. L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior*, *22,* 485-508.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*, 6829-6833.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*, 394-412.

Jonides, J., Schumacher, E. H., Smith, E., Lauber, E. J., Awh, E., Minoshima, S., & Koeppe, R. (1997). Verbal working memory load affects regional brain activation as measured by PET. *Cognitive Neuroscience, Journal of*, *9,* 462-475.

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

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 336.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of experimental psychology: General*, *132*, 47.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*, 189-217. doi:10.1037/0096-3445.133.2.189

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology, 55*, 352.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological bulletin*, *109*, 163.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review, 63*, 81-97.

Miller, K. M., Price, C. C., Okun, M. S., Montijo, H., & Bowers, D. (2009). Is the n-back task a valid neuropsychological measure for assessing working memory?. *Archives of Clinical Neuropsychology*, *24*, 711-717.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive psychology*, *41*, 49-100.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1170.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., & ... Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77-101. doi:10.1037/0003-066X.51.2.77

Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., & Cohen, J. D. (2000). Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. *Neuroimage*, *11*, 424-446.

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

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales.* New York, NY: Psychological Corporation.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2015). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*.

Redick, T. S., & Engle, R. W. (2006). Working memory capacity and attention network test performance. *Applied Cognitive Psychology*, *20*, 713-721.

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic bulletin & review*, *20*, 1102-1113.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, *124*, 262-274.

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: the gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1089.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: an individual differences approach. *Journal of Experimental Psychology: General*, *125*, 4.

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and Disengagement. *Perspectives on Psychological Science*, *11*, 771-799.

Shipstead, Z., Lindsey, D. R., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language*, *72*, 116-141.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. CUP Archive.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*, 643.

Thurstone, L. L. (1938). Primary mental abilities.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent?. *Journal of Memory and Language*, 28, 127-154. doi:10.1016/0749-596X(89)90040-5

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive psychology*, *71*, 1-26.

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.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and language*, *46*, 441-517.