SPEED-ACCURACY TRADEOFF AND ITS RELATIONSHIP TO HIGHER-ORDER COGNITION

A Thesis
Presented to
The Academic Faculty

by

Christopher Draheim

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Psychology

Georgia Institute of Technology
August, 2017

SPEED-ACCURACY TRADEOFF AND ITS RELATIONSHIP TO HIGHER-ORDER COGNITION

Approved by:

Dr. Randall W. Engle, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Daniel H. Spieler
School of Psychology
Georgia Institute of Technology

Dr. Rick P. Thomas
School of Psychology
Georgia Institute of Technology

Date Approved: May 05, 2017

ACKNOWLEDGEMENTS

I would like to thank Neesia Smith, who was instrumental in managing the lab and data collection for this study. I would also like to thank my committee for their time and comments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

It is a simple idea that there is an adversarial relationship between how quickly one performs an action and how well that action is performed. This phenomenon, known as the speed-accuracy tradeoff (SAT), has received some attention in the literature, notably through modeling work beginning in the 1960's. However, it has not been measured as a cognitive construct using latent variable analysis, as is common with other constructs such as working memory capacity (WMC), fluid intelligence (Gf), attentional control, task switching, memory updating, and so on. The goal of the present study is to address this gap in the literature. Specifically, I propose that the ability to appropriately implement speed and accuracy across different tasks is an important executive function strongly related to higher-order cognition. I tested this hypothesis by implementing tasks of SAT in a large-scale correlational study involving measures of other constructs, namely WMC and Gf. Results are mixed, there is evidence that SAT can be measured at the latent level and that this construct relates to higher-order cognition. However, the magnitude of this relationship is small, and trial-by-trial analyses suggest that lower ability individuals are also capable of adjusting performance to meet task demands.

CHAPTER 1. INTRODUCTION

From Luce (1986):

> …we face a very common problem in psychology: the existence of a tradeoff between dependent variables, in this case false alarms and reaction time. The only sensible long-term strategy is, in my opinion, to study the tradeoff… and to devise some summary statistic to describe it. (p. 56-57).

In the 19th century, Hermann von Helmholtz connected an electrode to a frog leg and demonstrated that the speed of nerve conduction could be quantified. This seemingly pedestrian finding was a pivotal step in showing that reaction time (RT) can serve as a dependent variable to measure neuronal processing. RT has since become an important measure in psychology, even more so with the rise of information-processing (now cognitive) psychology. To that end, RT is the primary dependent variable used to measure cognitive processing today.

The use of RT has its advantages, however, there is one major issue that is well known but often ignored. And that is the relationship between how quickly one performs an action against how well one executes said action. This relationship between speed and accuracy manifests itself in numerous ways. Across individuals, those who are more accurate also tend to be faster due to expertise. Similarly, within an individual, both speed and accuracy increase across extended periods of practice or training. But within an individual and at a particular period of time, emphasizing either speed or accuracy will result in a deficit in the other. That is, emphasizing accuracy will lead to the individual slowing down, and emphasizing speed will result in the individual committing more errors. This sacrifice of either speed or accuracy for the other is known as the speed-accuracy tradeoff (SAT), and it is perhaps the most ubiquitous and pervasive phenomenon we encounter in daily life.

CHAPTER 2. BACKGROUND

2.1 Brief History

Although it is well known that humans balance speed and accuracy, the SAT had not been systematically studied in psychology up until the mid-late 1960's (Fitts, 1966; Ollman, 1966; Pachella and Pew, 1968; Scouten and Bekker, 1967). It was around this time that mathematical decision models were first applied to speed and accuracy data from psychological tasks (Fitts, 1966). These models, known as random walk (Pearson, 1905), assume that speed and accuracy are related stochastic processes and that information for a particular decision accumulates over time until a response is made. The random walk models soon proliferated into other nuanced models, one example being the Ratcliff diffusion model (Ratcliff, 1978). These models are incredibly useful in that different parameters can be estimated that represent bias for one response over the other, drift rate (i.e., rate of information accumulation), indecision time, and threshold (i.e., amount of information that needs to be accumulated before a response is made). The threshold parameter represents speed-accuracy tendencies. Despite their widespread usage, these models are not without their drawbacks. First, most of them are only suitable for measuring performance on very simple two-choice tasks in which decisions are made quickly and there are few stages involved in the decision process. Second, the models require hundreds or even thousands of trials in order to reliably estimate the parameters.

In terms of RT research, SATs are problematic because the assumption is made that all differences in performance will manifest through RT and that all individuals emphasize accuracy to the same extent. This is a tenuous assumption that can lead to faulty conclusions when violated. As an example, a subject taking an average of 5 seconds to respond to a trial is said to have performed worse than a subject taking 4 seconds. But what if the first subject committed no errors whereas the second subject was barely above chance performance? It would seem that the first subject, despite being slower, performed the task better[1]. As such, any results from cognitive tasks using RT ought to be qualified with respect to accuracy as well. And while it is true that researchers often attempt to account for SATs in RT-based research by asking the subject to, "Respond as fast and accurately as possible," this instruction is likely not sufficient. Subjects may not heed the instruction in the first place, or, even more likely, there will be individual differences in what subjects deem to be fast and accurate. Additionally, asking someone to respond, "As fast and accurate as possible," requests two things that are inherently contradictory, analogous to asking someone to go as "northward as southward as possible."

2.2 How to Deal with SATs

Given that SATs are a persistent issue in RT research, what recourse do researchers have?[2] One option is to equate individuals on SAT via task design or instruction. Instructions to respond as quickly and accurately as possible are typically employed in RT research as a quick and easy method to account for the SAT, but again this instruction is likely insufficient. Researchers can, however, create or modify tasks in such ways that makes SATs less of an issue. One good example from our lab is the antisaccade task in which subjects see a cue on one side of the screen immediately followed by a target letter on the opposite side. The target letter is presented briefly and quickly masked such that if

[1] This is, of course, dependent upon task demands and assumes that subjects were given the typical instructions to be as fast as possible without sacrificing accuracy. If subjects were instead instructed to perform the task as fast as possible without worrying about making errors, then the faster subject did indeed perform the task better.
[2] SATs are also problematic for research using accuracy as the dependent variable, however many tasks that rely on accuracy scores do so in a manner that equates or controls for RT in some manner. As such, in the literature it seems that SATs particularly plague RT research.

the subject's attention is captured by the anti-informative cue, he or she will miss the target. Because RT is unimportant in this task (the subject either sees the target or does not) and accuracy is the dependent variable of interest, I would argue that this task is unaffected by subject-level SAT tendencies. For this reason our lab is looking into designing and modifying more attention control tasks that operate like the antisaccade. Another option is to give subjects feedback on their performance level, which is designed to reduce the overall amount of errors thus holding error rates to a minimum such that RT can be looked at in isolation. The idea here being that if error rates are low, then subjects have been effectively equated in terms of one dependent variable and differences in RT can be trusted insomuch as the differences are not being affected by the SAT.

Another approach for accounting for SATs is to analyze the SAT rather than treating it as a nuisance, similarly to how differential research investigates variance among individuals instead of treating these inter-individual differences as error as in experimental research. One such method is to conditionalize accuracy by RT. Using what is known as the conditional accuracy function, accuracy rates are plotted within particular range of RTs. This method was used to good effect in a flanker task in Heitz and Engle (2007) to show attentional differences in low and high working memory individuals. This method is particularly well suited to post-hoc investigations of data, but the SAT can be studied more directly through a priori means. As mentioned, this can be achieved through modeling, requiring a lot of resources, but also experimental manipulation. The most common manipulations are response deadlines and payoff matrices. With response deadlines, subjects are given a limited time to respond on each trial. The amount of time can be the same for all subjects and conditions, or adaptive in some manner (e.g., different for each subject based on their RT distribution from previous trials). Payoff matrices are used to reward subjects for responding per instruction. That is, when the instructions emphasize accuracy the subject is rewarded more for accuracy than speed, and vice versa for instructions emphasizing speed. These manipulations are a psychometric improvement over basic RT research, however individual differences are still an issue. Furthermore, how to set the deadlines and matrices requires preplanning and subjective decisions from the researcher. For instance, deadlines that are too long or too short will not be informative, and thus the researcher must have knowledge of the RT distribution beforehand (see Heitz, 2014, for a more thorough review of SAT history and methodology).

Wickelgren (1977) argues that methods of either partitioning or directly manipulating speed and accuracy tradeoffs are so superior to RT methods that cognitive psychologists should use the former in most instances. We have argued similarly in our lab after reanalyzing RT data from task switching paradigms (Draheim, Hicks, and Engle, 2015). In this particular dataset, the initial analysis using RT revealed that individuals of higher cognitive ability were actually worse at task switching than individuals of lower ability. However, this finding was a result of only looking at RTs, specifically RT-based difference scores (switch costs) that suffer from demonstrably low reliability and validity (e.g., Hughes, Linck, Bowls, Koeth, & Bunting, 2014; Paap & Sawi, 2016). When we considered accuracy, it became quite clear that higher ability individuals performed better, but that they had opted to slow down relative to lower ability individuals, particularly after committing an error. Reanalysis of data from Oberauer, Süß, Wilhelm, and Wittmann (2003) revealed a similar pattern, in which the reliability and strength of the relationship of task switching to higher-order cognition was much lower with RT-based scores and indeed improved when accuracy was taken into account.

2.3 Present Study

It has become increasingly clear to me that the ability to make an adjustment in speed and accuracy is a particular strategy that subjects adopt to meet the demands of the task at hand (see Draheim, Hicks, & Engle, 2015). That is, although it is true that some people may, in general, favor one over the other (e.g., elderly individuals opt for accuracy even when instructed to be fast; Forstmann et al., 2011; Starns & Ratcliff, 2010), it is also the case that adjustments in speed and accuracy are made in response to how difficult or demanding the task is, and how the subject feels he or she is performing on the task at the moment. This is because the optimal balance of speed and accuracy is not the same for all tasks, but depends on myriad factors such as the difficulty of the task, instructions on how to perform the task, rewards or consequences of performing a certain way, and so forth. As such, some individuals will more efficiently alter their performance (i.e., speed and accuracy) to meet these demands. Some subjects will be more averse to making errors than others and therefore will react differently following an error. It does not seem like a stretch to hypothesize that differences in performance adjustment (i.e., SAT) will also relate to higher-order cognition, such as WMC, Gf, attention control, etc. As previously stated, some of our work provides support for this view, as does other work demonstrating that individuals of higher Gf show accuracy, but not RT, differences in task switching (e.g., Unsworth & Engle, 2008).

The goals for the present study are therefore to demonstrate that SAT is an executive function that is related to WMC and Gf. The first challenge will be to establish that a single dependent variable of SAT can form a coherent latent variable, and the second is to model the relationship between this variable and higher-order cognition.

CHAPTER 3. METHOD

3.1 Overview

Results from the present study were collected as part of a larger study designed to answer numerous research questions. The experiment at large consisted of four two-hour long sessions completed on separate days and included forty-five tasks of nearly a dozen cognitive constructs. For my purposes, I will limit the focus to three constructs of interest; WMC, Gf, and speed-accuracy optimization (SAO).

3.2 Constructs

3.2.1 Working Memory Capacity

Working memory is the ability to simultaneously maintain, process, and manipulate chunks of goal-relevant information in readily-accessible form (see Baddeley, 1992, and Engle, 2002, for an overview). Working memory has limitations as to how much information can be maintained at any given time. As such, it is often measured in terms of capacity – estimated to be between three and five chunks of information for the typical young adult (Cowan, 2001). Working memory capacity (WMC) is an important construct in psychology as it has been shown to predict a wide range of cognitive abilities and real-world behaviors. For instance, individuals of higher WMC are better at multitasking (Hambrick, Oswald, Darowski, Rench, & Brou, 2010), task-switching (Draheim, Hicks, & Engle, 2016), language learning (e.g., Baddeley, Gathercole, & Papagno, 1998), language comprehension (e.g., Daneman & Merikle, 1996), attention control (e.g., Kane, Bleckley, Conway, & Engle, 2001), following directions (Engle, Carullo, & Collins, 1991), reasoning (e.g., Kyllonen & Christal, 1990), and many others. Crucially, WMC shares a substantial amount of variance with fluid intelligence (Gf; Ackerman, Beier, & Boyle, 2005; Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Hambrick, & Conway, 2005; Oberauer, Schulze, Wilhelm, & Süß, 2005), which is the ability to reason in novel situations. This relationship resulted in a separate line of research dedicated to exploring the efficacy of improving intelligence through working memory training (see Harrison et al., 2013), and indeed some commercial products claim that performing WMC-like tasks can make you smarter[3]. Additionally, WMC is important in clinical psychology as numerous psychopathologies and diseases are linked to deficits in WMC, such as attention deficit hyperactive disorder (Barkley, 1997), schizophrenia (Goldman-Rakic, 1994), depression (Joormann & Gotlib, 2008), and Alzheimer's (Baddeley, 1991).

We measured WMC with three complex span tasks: symmetry span, rotation span, and operation span. These tasks follow the same design, with the principle difference being the nature of the stimuli. Complex span tasks are dual tasks in which the subject answers a true/false "processing" question, and is then presented with a to-be-remembered stimulus. After several such presentations, the subject recalls the to-be-remembered stimuli in the correct serial order. The processing portion of the task thus serves as a distractor to prevent rehearsal, and the subject must maintain an active representation of the to-be-remembered stimuli in the face of this distraction and as proactive interference builds up. The presence of the processing trials is what

[3] I am not endorsing this claim, but am merely pointing it out in support of my argument that WMC is an important construct.

differentiates these measures from traditional simple span tasks, which measure short-term memory.

As an example (see Figure 1), in the operation span the subject first sees a simple arithmetic string (e.g., (2 x 3) – 3 = ?) and is then shown a number along with a "True" or "False" option. The subject indicates if the number accurately solves the arithmetic problem. After doing so, the to-be-remembered stimulus is displayed (a letter in the case of the operation span), and the process repeats some number of times depending on the set size for that particular trial. Once enough arithmetic and letter pairings have been displayed, a recall screen appears in which the subject has to recall all of the letters shown in the correct serial order. A common dependent variable is the partial span score, which is the total number of correct letters recalled in the correct position (Conway et al., 2005). To ensure that subjects are fully attending to the processing portion of the task and not rehearsing the to-be-remembered stimuli instead, the processing trials have a subject-adaptive response deadline equal to 2.5 SD's above mean RT on the practice trials. Additionally, data from subjects who do not perform the processing trials with at least 85% accuracy are typically thrown out.

We administered only two blocks of each complex span task instead of the typical three (see Foster et al. 2015). Additionally, we added longer set sizes for each of the tasks (e.g., set sizes 8 and 9 for the operation span) in order to better discriminate high ability individuals (see Draheim, Harrison, Embretson, & Engle, 2017). Thus maximum scores for the rotation and symmetry span tasks is 54, and the maximum score for the operation span is 84. Figure 1 shows the three complex span tasks that our lab typically uses.

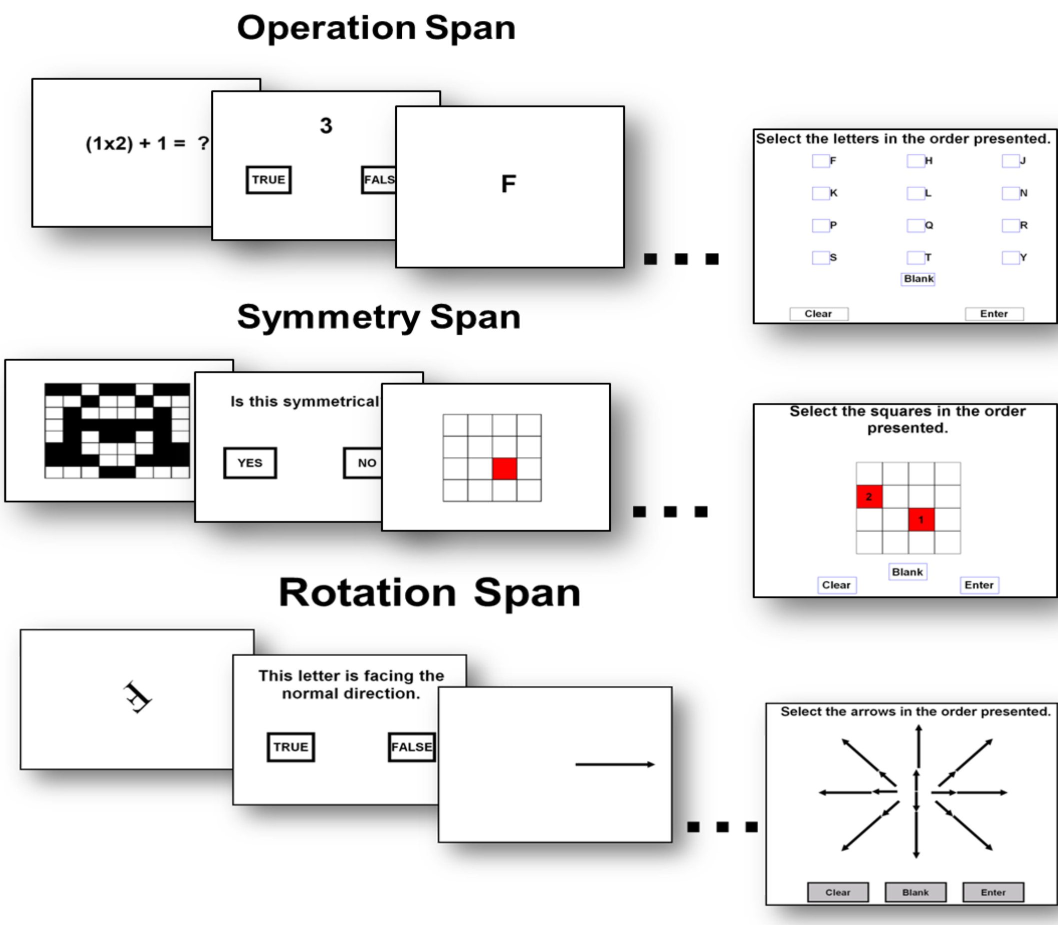

Figure 1 – Illustration of the Complex Span Tasks. Subjects are shown a processing or distractor task and given their mean RT on practice trials + 2.5 x their SD to respond. Then the to-be-remembered stimulus is displayed. This process repeats a number of times until a recall screen appears. In the standard-length tasks, each set size from 3 to 7 is administered three times in the operation span, and each set size from 2 to 5 is administered three times in the symmetry and rotation span. We administered only two blocks and included two larger set sizes for each task.

3.2.2 *Fluid Intelligence*

Gf is the ability to reason in novel situations, and is one of the two factors of general intelligence according to the Cattell (1963) model. We used three pattern recognition tasks to measure Gf: Raven's Advanced Progressive Matrices (odd problems; see Figure 2), number series, and letter sets. In the Raven's, subjects see a 3x3 grid of figures with the bottom-right figure missing, and have to choose which completes the figure out of eight possible choices. In the number series, subjects see several numbers and have to fill in the missing number that fits the pattern. In letter sets, subjects see five different four-letter combinations and must choose which one of the five combinations does not follow the pattern of the rest. The link between WMC and Gf is well established, and thus investigating the relationship between these two constructs is not of primary interest for this study (see Ackerman, Beier, & Boyle, 2005; Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Hambrick, & Conway, 2005; Oberauer, Schulze, Wilhelm, & Süß, 2005).

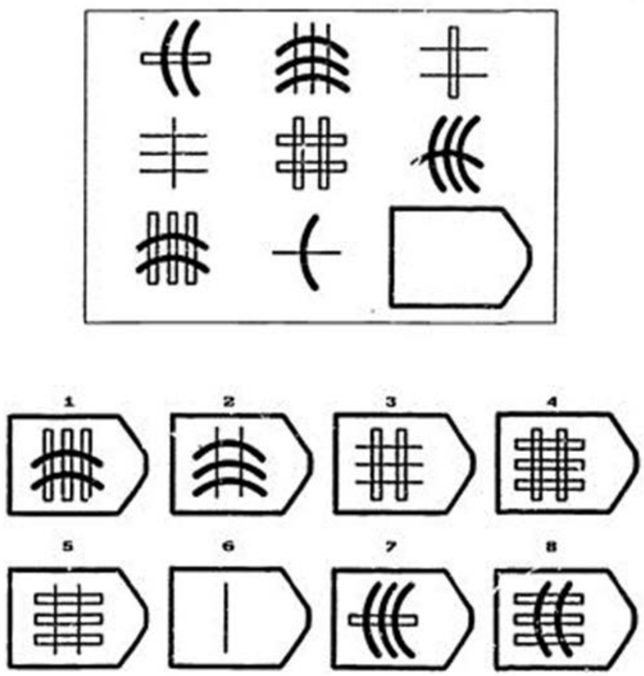

Figure 2 – Example Problem from Raven's Advanced Progressive Matrices. The goal is to choose which of the eight figures completes the pattern. In this example, the correct answer is #5 (I think).

3.2.3 *Speed-accuracy Optimization (SAO)*

The measurement of the speed-accuracy variable is not as straightforward as the other constructs. Here I operationalize speed-accuracy optimization (SAO) as the ability to adjust performance in speed and accuracy in order to meet task demands and perform the task most effectively. I combined response deadline (Pachella and Pew, 1968) and payoff matrix (Fits, 1966) methodologies to create the three SAO tasks. All three followed a similar design shown on Table 1. After performing 48 trials under normal instructions (i.e., baseline trials), the subject's mean RT for those trials was used to create an auditory response deadline for subsequent trials[4]. During those timed trials, the subject received both points and feedback based on their response. If the response was accurate and occurred before the response deadline (the beep), they received two points and saw, "Perfect!" in blue ink. If the response was accurate but after the tone, they lost a point and saw, "Too slow!" in red ink. If the response was inaccurate, but before the tone, they lost a point and saw, "Wrong!" in red ink. If the response was inaccurate and also after the beep, they lost two points and saw, "Wrong and slow!" in red ink[5]. In total, 384 trials in each task included the adaptive response deadline. The points accumulated from these trials served as the dependent variable for the SAO score. The rationale for using the points system was that each subject would have to find an optimal balance between speed

[4] The reason for using an adaptive response deadline is to account for individual differences in RT for these tasks, and thus eliminate variance due to processing speed. That is, if a subject is missing a deadline set from their individual distribution of responses, it is because they responded slowly in terms of their own performance, and not just because they responded slower than other subjects.

[5] Subjects were not shown a cumulative total of their points, but rather only the points they received on each trial immediately following the trial.

and accuracy to maximize their score, and that this optimal balance would be different for each subject depending on their ability and normal speed-accuracy tendencies.

The specific tasks used were arrow flanker, line-length discrimination, and lexical decision task. In the lexical decision task, subjects see a letter string and must decide if the string represents an actual English word. Stimuli were chosen from the English Lexicon Project (Balota et al., 2007) such that all stimuli were four-letters long with a mean accuracy rate of 95-100% and a mean RT of between 600 and 700ms. In the arrow flanker, subjects focus on a cross fixation (+) at the center of the screen until a target arrow appears slightly above the fixation. The target arrow is flanked by two arrows on either side, and the subject must indicate which direction the central arrow is pointing (left or right). On congruent trials, the flanking arrows are in the same direction as the central arrow (→ → → → →); on incongruent trials, the flanking arrows are in the opposite direction (← ← → ← ←); and on neutral trials, the central arrow is flanked by dashes (--- --- → --- ---). An equal number of congruent, incongruent, and neutral trials were presented. In the line-length discrimination task, subjects were simultaneously presented two white bars on a black background and had to decide which one was longer. The bars were jittered such that the subject could not simply line them up to make a decision, and some bar pairings were long whereas others were short. The difference in the length of the bars was set to 35 pixels, which piloting indicated is about twice as long as the average subject's threshold for 75% accuracy.

Table 1 – Trial Design for SAO Tasks.

Amount of Trials	*Purpose*	*Time to Respond*	*Feedback*
8	Practice	Infinite	Accuracy
48	Obtain Baseline RT	5000ms	None
48	SAO	Mean + 1.5 SDs	Time, Accuracy, and Score
48	SAO	Mean + .75 SDs	Time, Accuracy, and Score
48	SAO	Mean	Time, Accuracy, and Score
48	SAO	Mean - .75 SDs	Time, Accuracy, and Score
48	SAO	Mean + 1.5 SDs	Time, Accuracy, and Score
48	SAO	Mean + .75 SDs	Time, Accuracy, and Score
48	SAO	Mean	Time, Accuracy, and Score
48	SAO	Mean - .75 SDs	Time, Accuracy, and Score

Note. The tasks were counterbalanced such that half of the subjects received the trials in the exact order shown, and the other half received the SAO trials in reverse order (i.e., received the quickest response deadline condition first). Subjects had to be at a minimum accuracy of 75% on the practice block (at least 6 correct trials) in order to proceed to the experimental trials. If accuracy was lower, additional instructions appeared and the practice block was repeated. After three such instances, the experimenter was alerted and the subject was either given further instruction from the experimenter or dismissed from the study.

3.3 Subjects

Our screening process required all subjects to be native English speakers aged 18-35 (*M* = 24.5) with normal or corrected-to-normal vision. A total of 351 subjects (179 female) completed all four sessions. Subjects were recruited from Georgia Tech (*n* = 86), Georgia State University (*n* = 36), and the greater Atlanta community (*n* = 229)[6]. Of the 351 subjects, 180 indicated that they were currently attending, or had attended, college. Subjects were compensated at a rate of $10/hour plus a $10 completion bonus after the final session. Georgia Tech students enrolled in introductory psychology courses could choose to receive SONA participation credit instead of financial compensation (1 hour = 1 credit = $10). Institutional review board approval was obtained for the study and there were no protocol deviations to report.

[6] The majority of the subjects were recruited from our existing database, as we ask all subjects if they would like to be contacted for future studies. New subjects were recruited primarily through flyers, newspaper, Craigslist, and SONA.

CHAPTER 4. RESULTS

4.1 Correlational Results

4.1.1 Task-level

Descriptive statistics and zero-order correlations for all tasks in question are shown below (Table 2 and Table 3. As for reliability of the SAO tasks, I obtained internal consistency using an even-odd split-half procedure in which scores were separately calculated for each subject on even trials and odd trials, and then correlated with one another (stepping up the correlation via the Spearman-Brown prophecy formula). The SAO scores were highly reliable in terms of internal consistency estimates, with the lexical decision SAO task having a split-half coefficient of .98, and the other two tasks having a split-half coefficient of .99. Lastly, there was no significant difference in scores based on counterbalance conditions (i.e., if subjects received the trials in ascending or descending order in terms of the response deadline).

What is of primary interest here is the relationship among the SAO tasks as well as the relationship of SAO to WMC and Gf. The SAO scores correlate fairly well with one another (r = .40 - .55), demonstrating convergent validity. The major concern is the different pattern in how the SAO tasks correlate with the WMC and Gf tasks. The flanker SAO score does not significantly correlate with any of the Gf scores (r = -.01 - .07), and only one WMC score (r = .03 - .16). However, the lexical SAO score correlates moderately and significantly (r = .24 - .39) with all of the WMC and Gf tasks. The line discrimination SAO task is in the middle, as scores correlate weakly (r = .10 - .18, two significant) to WMC scores and not at all with Gf scores (r = .04 - .09, none significant). Another concern is that the line SAO scores are very leptokurtic, with a kurtosis at 4.0. At the composite level (Z-Score average across the tasks). WMC and Gf correlate quite strongly as expected, but more importantly the SAO composite scores correlate strongly with WMC (r = .33) and Gf (r = .29; Table 4). Also included on Table 4 are the Z-score composites for the number of missed deadlines in each of the SAO tasks. That is, how many times in each task the subject responded after the adaptive deadline across all conditions. This variable correlated highly with the overall SAO score unsurprisingly, but less with WMC and Gf (r = .20 and r = .14, respectively). Total accuracy on the SAO trials (i.e., trials in the SAO tasks that included a response deadline) was nearly isomorphic with the SAO score (r = .97), correlating similarly to WMC and Gf than the SAO score (r = .31 and r = .29, respectively), but correlating more weakly with the number of missed deadlines (r = .41). The difference between the correlations of SAO score to missed deadlines versus accuracy on SAO trials and missed deadlines was significant using Steiger's (1980) recommended test of dependent correlations (t = 116.996, p < .001). Lastly, total accuracy for the baseline trials of the SAO tasks (i.e., trials without the deadlines and scoring feedback) correlated strongly with WMC and Gf (r = .32 and r = .34, respectively), very strongly with the SAO score (r = .71), strongly with the number of missed deadlines (r = .39) and very strongly with the accuracy on the SAO trials (r = .72).

Table 2 – Descriptive statistics.

Task	M	SD	Min - Max	Skew	Kurtosis
WMC					
1. OSpan	49.5	18.1	4 - 82	-.35	-.70
2. SymSpan	23.5	10.1	4 - 50	.35	-.23
3. RotSpan	21.5	9.7	3 - 47	.35	-.38
Gf					
4. Raven	8.7	3.7	1 – 17	.14	-.90
5. LetterSet	14.6	4.7	3 – 26	0	-.47
6. NumSeries	8.1	3.2	0 - 15	.26	-.72
SAO					
7. Line	278	66	-48 - 371	-1.5	4.0
8. Lexical	211	75	-102 - 350	-1.1	1.7
9. Flanker	282	54	33 - 364	-1.3	2.5

Note. Ospan = operation span; SymSpan = symmetry span; RotSpan = rotation span. DV for WMC tasks is the partial span score. DV for the Gf tasks is the total number of correct responses. DV for the SAO tasks is the cumulative score for being fast and/or accurate on each trial.

Table 3 – Zero-order correlations of imputed WMC, Gf, and SAO scores.

	WMC			Gf			SAT		
Task	1	2	3	4	5	6	7	8	9
1. OSpan	1.00								
2. SymSpan	.58*	1.00							
3. RotSpan	.55*	.72*	1.00						
4. Raven	.45*	.52*	.58*	1.00					
5. LetterSet	.47*	.45*	.45*	.54*	1.00				
6. NumSeries	.52*	.50*	.50*	.67*	.61*	1.00			
7. Line	.10	.15*	.18*	.08	.09	.04	1.00		
8. Lexical	.24*	.31*	.39*	.36*	.35*	.35*	.46*	1.00	
9. Flanker	.08	.03	.16*	.07	.00	-.01	.55*	.40*	1.00

Note. OSpan = operation span; RotSpan = rotation span; SymSpan = symmetry span.
*$p < .05$.

Table 4 – Zero-order correlations of composite scores.

Measure	1	2	3	4	5	6	7	8
1. WMC	1.00							
2. Gf	.67*	1.00						
3. SA Point Total	.33*	.29*	1.00					
4. SA Missed Deadlines	.20*	.14*	.63*	1.00				
5. SA Accuracy	.31*	.29*	.97*	.41*	1.00			
6. SA RT	.28*	.29*	-.09	.25*	-.19*	1.00		
7. Baseline Trial Accuracy	.32*	.34*	.71*	.39*	.72*	.02	1.00	
8. Baseline Trial RT	.32*	.35*	-.16*	.13*	-.14*	.78*	.05	1.00

WMC = composite Z-score for the three complex span tasks. Gf = composite Z-score for the three reasoning tasks. SAO = composite Z-score of the cumulative score for each of the SAO tasks. MDL = composite Z-score of the cumulative missed response deadlines in each of the SAO tasks. SAcc = composite Z-score of the total accuracy in each of the SAO tasks on trials with response deadlines. BAcc = composite Z-score of the total accuracy in each of the SAO tasks on the baseline trials. Correlations involving MDL were multiplied by (-1) such that a positive correlation indicates that individuals who missed *fewer* deadlines also tended to score higher on the other tasks.
** p < .05.*

4.1.2 Exploratory Factor Analysis

I conducted two different exploratory factor analyses on the data. The first was unrotated and did not specify any factors. From this, a general factor emerged due to positive manifold. Specifically, all task scores loaded .3 or higher onto a general factor, and the SAO task scores loaded .679 - .753 onto their own factor as well. For the second exploratory factor analysis three factors were specified and I used a varimax rotation such that simple structure would emerge. The results are shown below (Table 5). When simple structure was imposed using an orthogonal rotation, the highest loadings for the WMC tasks were on the first factor, the highest loadings for the Gf tasks were on the second factor, and the highest loadings for the SAO tasks were on the third factor. This provides support that the SAO tasks do indeed form their own coherent factor independent of WMC and Gf.

Table 5 – Exploratory Factor Analysis of WMC, Gf, and SAO Tasks.

Task	Factor 1	Factor 2	Factor 3
OSpan	**.758**	.323	.016
SymSpan	**.869**	.264	.062
RotSpan	**.814**	.318	.202
Raven	.381	**.758**	.101
LetterSet	.257	**.791**	.154
NumSeries	.305	**.842**	.055
Line SAO	.091	.271	**.815**
Lexical SAO	.086	.044	**.844**
Flanker SAO	.047	-.003	**.766**

Note. OSpan = operation span; SymSpan = symmetry span; RotSpan = rotation span. Varimax rotation was used to attain simple structure. Boldface indicates the highest factor loading for that variable.

4.1.3 Confirmatory Factor Analysis

I ran a confirmatory factor analysis specifying three factors (WMC, Gf, and SAO), allowing them to freely correlate. Scale of measurement was established by setting one variable per factor to 1, and there were no crossloadings or correlated error terms. The primary area of interest in this model are the factor paths, which indicate the strength of the relationship between the different constructs. It is well established that WMC and Gf share a substantial amount of variance, but what about the SAO tasks? Additionally, will SAO be differentially related to WMC and Gf?

As shown in Figure 2, the model fits well, with a CFI of .96, and the residuals are acceptable (RMSEA = .07). However, the model is significant with $X^2(24) = 51.41$, $p < .01$, indicating that there are some relationships in the covariance matrix that are not adequately reproduced in the model. Ignoring this issue for a moment, it should be noted that the SAO construct shares a substantial and significant amount of variance with both WMC and Gf. The path from SAO to WMC is .44, and the path from SAO to Gf is .36. Again model is not ideal given that it is significant, but these results are encouraging and suggest that the ability to balance speed and accuracy to meet task demands is indeed highly related to higher order functioning. Furthermore, the data suggest a differentiation in the relationship of SAO to WMC and Gf, with SAO being more strongly related to WMC.

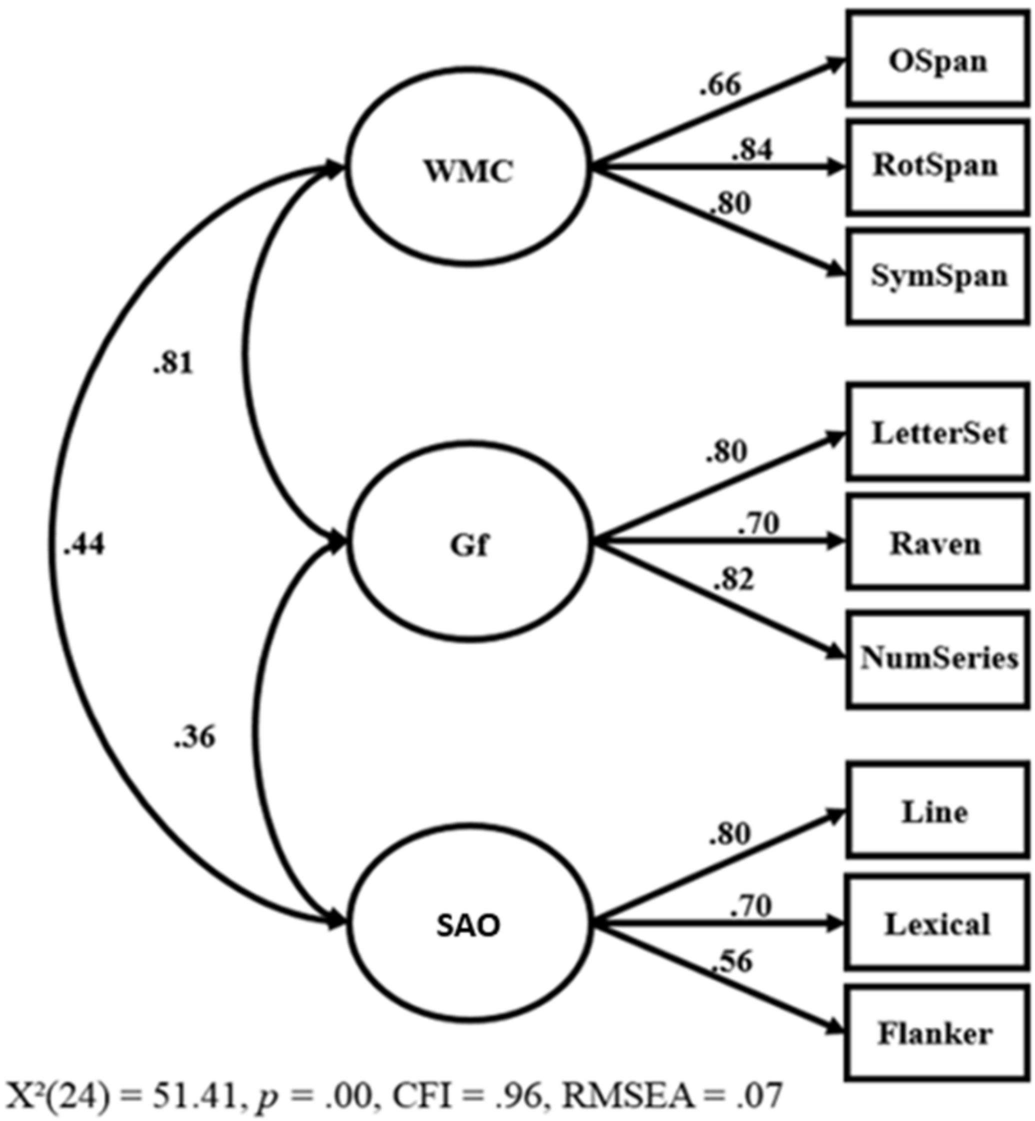

Figure 3 – Confirmatory Factor Analysis – Listwise Deletion. *n* = 213. OSpan = operation span; RotSpan = rotation span; SymSpan = symmetry span. All paths and loadings are significant at the .05 level.

One caveat in respect to the model above (Figure 3) is that listwise deletion was used. Of the 351 subjects from the study, only 213 had acceptable data for all tasks (i.e., non-zero scores within 2.5 SDs of the mean). To test whether this was a problem, I conducted an additional model equivalent to Figure 3 but with all subjects, using data imputation to fill in missing data. This model is shown below (Figure 4). Overall, the model fit, factor paths, and task loadings were very similar between the two models. The major change was the standardized path coefficient between WMC and SAO decreased from .44 to .37 when imputation was used. However, all paths were still significant and the loadings for SAO tasks remained high, thus I can still conclude that the SAO scores form a coherent latent factor that relate to WMC and Gf.

The confirmatory factor analysis shown in Figure 4 is also significant, meaning that there are relationships from the original covariance matrix that the model does not adequately account for. I used the Lagrange Multiplier Test to investigate which relationships would improve the model the most. The top three changes to the model that would most improve model fit were to crossload the lexical SAO task onto the Gf factor, to crossload the rotation span task onto the SAO factor, and to crossload the line SAO score onto the Gf factor. I decided to not make any of these changes to the model as the overall model fit was not an issue and I could not justify making these alterations.

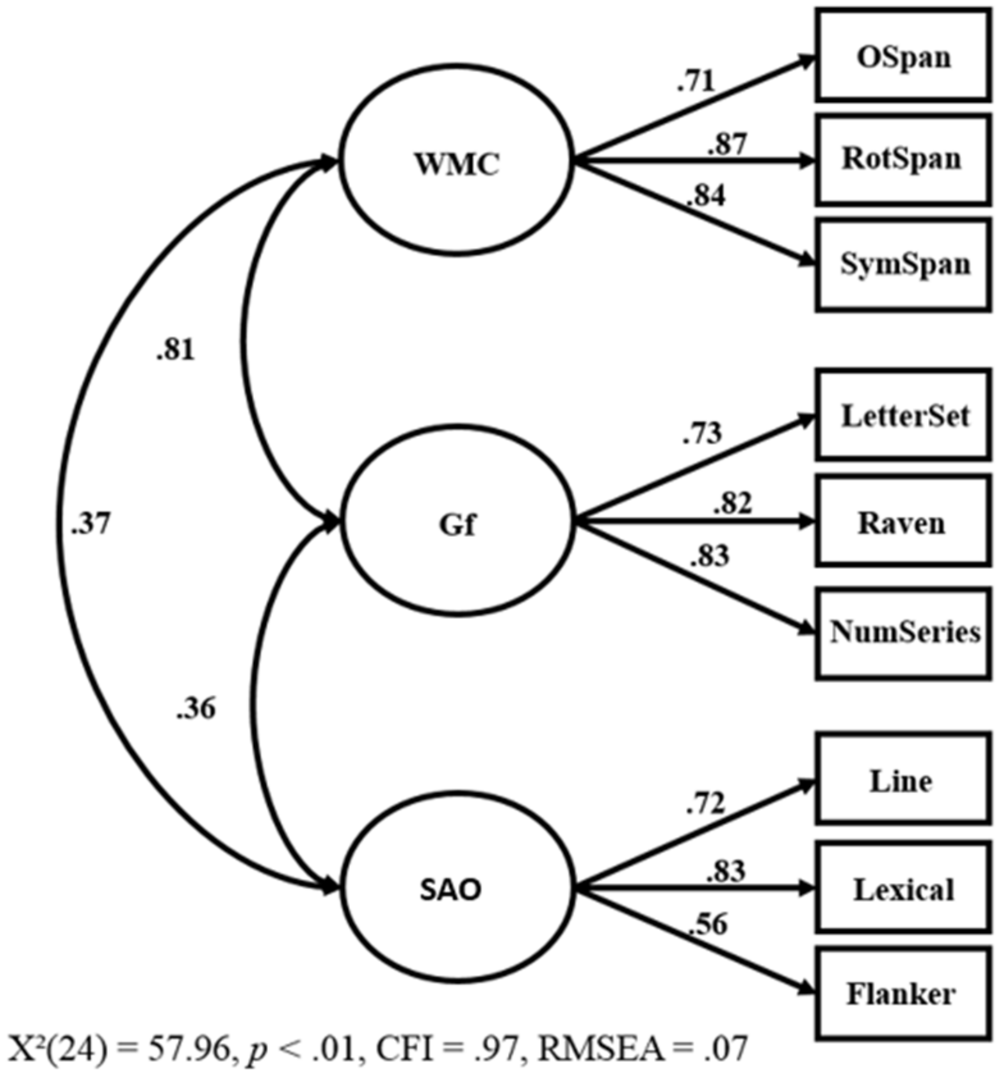

Figure 4 – Confirmatory Factor Analysis – Imputation. OSpan = operation span; RotSpan = rotation span; SymSpan = symmetry span. All paths and loadings are significant at the .05 level.

4.2 Trial-level Analyses

4.2.1 Post-error Slowing

Trial-by-trial analyses can answer questions that mean-level analyses cannot. One example is the phenomenon of post-error slowing, in which subjects tend to be slower on a trial following an error. As discussed, we have shown that post-error slowing during task switching procedures interacted with WMC (Draheim, Hicks, & Engle, 2016). In short, higher ability individuals were much more likely to slow down after an error whereas lower ability individuals did not make this adjustment. As a result, higher ability subjects looked as though they were performing worse on the task when just RT was considered, but when considering accuracy and looking at trial-by-trial data, it became clear that this was not the case.

Post-error slowing for the three SAO tasks are shown below (Table 6). When collapsing across all conditions for each task, there is a difference in mean RT in the flanker task (high spans are faster), but not the line discrimination or lexical decision SAO tasks. Most noteworthy is that high and low subjects alike exhibited post-error slowing in all three tasks and to the same degree, approximately 30ms (significantly different from 0 in all cases). That is, when feedback is provided such that subjects are aware of their errors, high and low cognitive ability individuals both show the same slowing on trials immediately following an error. I conducted the same analysis on accuracy to see how subjects' accuracy rates changed after making an error (Table 7). In short, high spans were more accurate than low spans in all three tasks. However, neither group had differing accuracy rates before and after making an error.

Table 6 – Post-error slowing among high and low spans in SAO tasks.

	Flanker		LineDiscrim		Lexical	
Variable (*ms*)	High	Low	High	Low	High	Low
PreRT	408	469	460	482	510	521
PostRT	434	504	489	503	542	550
Slowing	**26**	**35**	**29**	**31**	**32**	**29**

Note. PreRT = Mean RT on trials preceding an error. PostRT = Mean RT on trials following an error. Highs and lows are defined by tertile split of composite WMC scores. Boldface indicates a statistically significant difference between highs and lows for that variable ($p < .05$).

Table 7 – Post-error accuracy among high and low spans in SAO tasks.

	Flanker		LineDiscrim		Lexical	
Variable (%)	H	L	H	L	H	L
PreAcc	**94**	**89**	**90**	**84**	**88**	**77**
PostAcc	**94**	**89**	**90**	**84**	**88**	**77**
Improvement	0	0	0	0	0	0

Note. PreAcc = Mean accuracy rate on trials preceding an error. PostAcc = Mean accuracy rate on trials following an error Highs and lows are defined by tertile split of composite WMC scores. Boldface indicates a statistically significant difference between highs and lows for that
variable ($p < .05$).

4.2.2 Performance Conditional on Score

Related to the post-error slowing analysis, I conducted analyses on how subjects performed in terms of pre- and post-accuracy and RT contingent upon how they scored on the present trial. That is, if a subject is slow but correct (thus losing one point), are they more likely to speed up for the next trial? Similarly, how will subjects respond to being both slow and wrong? The analyses are shown in Tables 8, 9, and 10 below.

Table 8 – Performance in flanker SAO task conditional on score.

	RightSlow		WrongFast		WrongSlow	
Variable	H	L	H	L	H	L
PreRT	**427**	**500**	**403**	**462**	**420**	**488**
PostRT	**414**	**485**	**428**	**488**	**471**	**535**
Slowing	-13	-15	25	26	51	47
PreAcc%	88	87	94	91	**88**	**82**
PostAcc%	92	88	94	90	**95**	**86**
Improvement	4	1	0	-1	7	4

Note. PreRT = Mean RT (in *ms*) on trials preceding the particular score. PostRT = Mean RT (in *ms*) on following trial. PreAcc% = Mean accuracy rate on trials preceding the particular score. PostAcc% = Mean accuracy rate on following trial. RightSlow = subject was correct but did not meet deadline. WrongFast = subject was wrong but responded before deadline. WrongSlow = subject was wrong and did not meet the response deadline. Highs and lows are defined by tertile split of composite WMC scores. The first and last trials for each block were excluded from analysis. Boldface indicates a statistically significant difference between highs and lows for that variable ($p < .05$).

Table 9 – Performance in line discrimination SAO task conditional on score.

Variable	RightSlow H	RightSlow L	WrongFast H	WrongFast L	WrongSlow H	WrongSlow L
PreRT	**498**	**522**	**457**	**476**	**486**	**527**
PostRT	**476**	**492**	**487**	**500**	**511**	**525**
Slowing	-21	-30	30	24	**25**	**-2**
PreAcc%	81	81	**90**	**83**	84	85
PostAcc%	85	86	**90**	**83**	86	84
Improvement	4	5	0	0	2	-1

Note. PreRT = Mean RT (in *ms*) on trials preceding the particular score. PostRT = Mean RT (in *ms*) on following trial. PreAcc% = Mean accuracy rate on trials preceding the particular score. PostAcc% = Mean accuracy rate on following trial. RightSlow = subject was correct but did not meet deadline. WrongFast = subject was wrong but responded before deadline. WrongSlow = subject was wrong and did not meet the response deadline. Highs and lows are defined by tertile split of composite WMC scores. The first and last trials for each block were excluded from analysis. Boldface indicates a statistically significant difference between highs and lows for that variable ($p < .05$).

Table 10 – Performance in lexical decision SAO task conditional on score.

	RightSlow		WrongFast		WrongSlow	
Variable	H	L	H	L	H	L
PreRT	**537**	**569**	508	515	**521**	**579**
PostRT	**524**	**549**	535	539	596	616
Slowing	-13	-20	27	24	**75**	**37**
PreAcc%	**77**	**72**	**89**	**77**	75	73
PostAcc%	**83**	**77**	**88**	**77**	**85**	**74**
Improvement	6	5	-1	0	**10**	**1**

Note. PreRT = Mean RT (in *ms*) on trials preceding the particular score. PostRT = Mean RT (in *ms*) on following trial. PreAcc% = Mean accuracy rate on trials preceding the particular score. PostAcc% = Mean accuracy rate on following trial. RightSlow = subject was correct but did not meet deadline. WrongFast = subject was wrong but responded before deadline. WrongSlow = subject was wrong and did not meet the response deadline. Highs and lows are defined by tertile split of composite WMC scores. The first and last trials for each block were excluded from analysis. Boldface indicates a statistically significant difference between highs and lows for that variable ($p < .05$).

The results from the analyses conditional on score in the flanker SAO task (Table 8) reveal that high spans are quicker than lows in all non-perfect conditions (e.g., trials in which the subject was not both accurate and met the response deadline). However, both groups made similar adjustments to performance after receiving feedback about being slow and/or inaccurate. When correct but not meeting the deadline, both groups slowed down by about 14ms and were slightly (but not statistically significantly) more accurate on subsequent trials. When wrong but meeting the response deadline, both high and low spans slowed by 25-26ms on the subsequent trial and maintained similar accuracy levels (evidence of post-error slowing as in previous analyses). Interestingly, when on trials in which the response was both slow and wrong, both groups slowed by about 50ms and improved accuracy by 4-7% on the subsequent trial. While the 7% improvement for high spans was significantly different from 0, and the 4% improvement for low spans was not, the two were not statistically significantly different from one another. Therefore in all conditions, high and low spans make similar adjustments to performance when given feedback and instruction to either speed up, slow down, or improve accuracy. Also noteworthy is that high and low spans were equal on overall accuracy except on trials preceding or following being both wrong and slow (missing deadline).

In the line discrimination SAO task (Table 9), high spans were faster than low spans in all conditions. However, both groups sped up (21ms for high spans, 30ms for lows) when they were correct and missed the response deadline, and this was also associate with higher accuracy (by 4-5%) on following trials. When subjects were incorrect but met the response deadline, both high and low span subjects exhibited slowing (30ms & 24ms, respectively), but this was not associated with any change in accuracy on the following trial. This was also the only condition with overall accuracy differences, as high spans were 7% more accurate on the trials preceding and following being wrong and fast. Finally, on trials in which subjects were both wrong and missed the response deadline, high spans adjusted by slowing down (25ms) with a slight, but not statistically significant, increase in accuracy (2%). Low spans made no real adjustment on these trials, as their performance was the same before and after being wrong and slow.

In the lexical decision SAO task (Table 10), there were RT differences between highs and lows preceding and following trials in which they were right but missed the response deadline, as well as trials in which they were wrong and missed the deadline. But not when they were wrong and responded before the deadline. When correct and missing the deadline (i.e., slow to respond), both highs and lows sped up (13ms & 20ms, respectively) and this was associated with an increase in accuracy (6% & 5%, respectively). There were differences in overall accuracy between high and low span subjects (by 5-6%). In trials in which subjects were wrong but responded before the deadline, both highs and lows sped up on the following trial to the same degree (27ms & 24ms, respectively), and there was no change in accuracy. Again there were overall accuracy differences between highs and lows here, too. Finally, on trials in which subjects were both wrong and missed the response deadline, both high and low span subjects slowed down, but high spans slowed down much more (75ms) than low span subjects (37ms). Interestingly, this was associated with a massive increase in accuracy on the following trial for high spans (10%), but not for low spans (1%). As a result, high and low spans had different accuracy rates after being both wrong and slow (85% & 74%), but not before (73% & 75%). These results will be discussed further below.

CHAPTER 5. DISCUSSION

The major goals of this study were to demonstrate that SAO ability can be measured with a single dependent variable and that tasks measuring SAO ability form a coherent latent variable related to higher-order cognition. Results favor these hypotheses. Modeling analyses reveal that there are meaningful differences in the amount of points accrued (via accuracy and responding before response deadlines) even when feedback is given. These differences relate moderately (in terms of strength) with high-order cognition (i.e., WMC and Gf). However, the nature of the point system ought to be discussed, as results need to be qualified with respect to the methodology. Each subject received an adaptive deadline that was based on his or her performance on the 48 practice trials for each task. The deadlines were set based on standard deviations from their mean RT on correct trials during practice. While this is an improvement over using non-adaptive response deadlines (i.e., the same deadline for all subjects), this method is not perfect. The extent to which subjects have differing RT distributions will make these deadlines easier or harder for certain subjects. For instance, a subject with a small variance in their RTs (narrow distribution) will encounter deadlines that are all very close to their mean performance, whereas subjects with larger variance will encounter very tough deadlines in the fast conditions but easier deadlines in the slower conditions. I did try to account for this by making it such that each deadline had to be at least 50ms different from the next (e.g., the fastest condition had to be 50ms faster than the second fastest, which had to be 50ms faster than the third fastest) and that there was an absolute limit for all deadlines (i.e., if deadlines were too slow or fast, they were adjusted to some minimum or maximum) based on piloting work, but this is not a foolproof method. Therefore the deadlines might not have affected all of the subjects the same, thus not equating them as I had intended. I still contend that this method is an improvement over static (i.e., non-adaptive) deadlines, however.

The SAO scores also correlated nearly perfectly with total accuracy on the SAO trials and strongly, but significantly less so, with the number of overall missed deadlines. This suggests that the SAO scores (i.e., points accrued by being quick and/or accurate on the response deadline trials) were heavily influenced by if the subject was accurate or not on each trial, and less so by their ability to respond before the response deadline. Missed deadlines also correlated weaker with WMC and Gf, demonstrating that high ability individuals and low ability individuals were relatively equal in terms of how many response deadlines they missed, but showed more differences in terms of overall accuracy (high ability individuals being more accurate). This is likely a product of the response deadlines being adaptive for each subject based on their baseline RTs. One caveat of the SAO score relating to higher-order cognition is that accuracy on both baseline and experimental (i.e., trials with deadline and scoring feedback) correlated similarly to WMC and Gf as the SAO scores did with WMC and Gf. As such, it can be argued that the SAO scores are merely a function of accuracy. Perhaps the SAO scores are not a reflection of the subject's ability to balance speed and accuracy to meet task demands, but rather a general reflection of if the subject tends to be overall accurate or not. To answer this, I did another regression model predicting WMC from baseline accuracy and then SAO score. If the SAO score is nothing more than accuracy tendencies, it should not be a significant predictor above and beyond baseline accuracy levels. Table 11 shows this regression model. When SAO score is entered after baseline accuracy, it still significantly predicts WMC, though admittedly the overall effect size is small SAO scores account for 2% more variance in WMC than baseline accuracy. Thus I argue that the SAO score is heavily influenced by baseline accuracy levels, but that the scores also reflect a bit more than just these accuracy tendencies, likely the ability to adopt a certain responding criterion in the face of changing task demands and quick deadlines.

Table 11 – Predicting WMC from baseline accuracy and SAO score in the SAO tasks.

	Predictor	R	Adjusted R-Square	SE	*p*
Step 1	Baseline Accuracy	.32	.10	.783	<.001
Step 2	Baseline Accuracy + SAO Score	.35	.12	.776	.027

Note. Baseline accuracy on the non-SAO trials from the flanker, line discrimination, and lexical SAO tasks were entered in step 1. In step 2, the SAO scores from the SAO trials of these tasks were entered. The outcome variable being predicted is the composite WMC score.

Interestingly, despite the latent variable of SAO ability relating to higher-order cognition, trial-by-trial analyses revealed that high and low subjects generally adjust to performance in a similar manner when given feedback. Across the three tasks, there were only two conditions in which high and lows differentially adjusted performance based on feedback – when both wrong and slow (i.e., missing the response deadline) in the lexical decision and line discrimination SAO tasks. In both of these cases, high spans responded to the feedback by slowing down (25ms in line discrimination, 75ms in lexical decision), but this only resulted in improved accuracy for the following trial in the line discrimination task (by 10%). Low spans did not slow down after being both wrong and slow in the line discrimination task, but did slow down in the flanker (by 47ms) and lexical decision SAO tasks (37ms). This slowing was not associated with increased accuracy on the following trials, however. It thus appears that high and low spans typically respond similarly when given feedback about performance, except in some cases in which their responses are both wrong and slow. In this case, it is more difficult to make the decision on how to alter performance since it is not clear if one should be quicker or emphasize more on accuracy. High spans seem to favor slowing down, sometimes associated with increase accuracy. Low spans also slow down, but to a lesser degree, and this does not result in higher accuracy. In conditions in which it is more clear on how to adjust performance (i.e., being slow but correct or quick but incorrect), high and lows adjust similarly *when given feedback*

Given the totality of results from the trial-by-trial analyses, it seems the case that all subjects are capable of altering performance to meet task demands to the same degree, counter to the latent variable analyses. In my opinion, these results indicate that the SAO scores derived from the tasks are not good estimates of actual SAO ability. I conducted analyses looking at if there were individual differences in the number of total deadlines missed, mean RT, or mean accuracy rates, and if so if these differences related to WMC and Gf. In terms of total accuracy, accuracy correlated significantly to WMC (.r = .15 in (flanker, r = .22 in line discrimination, and r = .32 in lexical decision) in the SAO conditions. However, WMC also significantly correlated with baseline accuracy in the practice trials (r = .23, r = 17, and r = .26, respectively) to the same degree. Therefore there was no interaction in terms of high spans being more accurate than lows in the SAO trials than the baseline trials. The pattern of results was the same for RT; there was a significant correlation between WMC and RT in the SAO trials (r = .42, r = .15, and r = .11, respectively) such that high ability individuals were faster. However, correlations were similar between WMC and baseline RT in the practice trials (r = .40, r = .14, and r = .19, respectively), suggesting no interaction. In terms of missed deadlines in the SAO conditions, missed deadlines correlated significantly with WMC in the lexical decision task (r = .18, p = .005) but not the line discrimination (r = .11, p = .079) or flanker (r = .12, p = .072). It therefore appears that there was a weak relationship between WMC and the amount of response deadlines subjects missed in the SAO trials[7]. As such, it appears that higher WMC individuals performed better on the SAO tasks (in terms of the score) due to higher overall accuracy rates, which results in more points earned.

The post-error slowing findings are interesting given prior data from our lab demonstrating that there indeed are differences in post-error slowing among high and low

[7] Relationships between all of these discussed variables and Gf were very similar to the relationships to WMC, with the magnitude of the correlations being slightly smaller or slightly weaker in some cases.

span subjects. In particular, subjects in this study also performed a flanker task (96 trials) in which no feedback was given about their performance. The results from this task show that high spans (again, top third in terms of WMC composite score) slowed down after errors by an average of 55ms (527ms prior to error on average, 582ms on trials following an error on average), whereas low spans only slowed down by 30ms (592ms before error, 622ms after). This 30ms for low spans is consistent with results from previously discussed analyses, but the 55ms is a much larger slowing for high spans. This difference in post-error slowing among highs and lows might be the result of the absence of feedback on every trial that was present in the analyses discussed previously. It therefore appears that the presence or absence of trial-by-trial feedback is a determining factor of how high and low spans differentially alter performance, particularly after making an error. These flanker data do potentially suggest one very key point, however, in that low spans may be aware of making errors even without being given direct feedback about their accuracy levels. More evidence in favor of this position ought to be obtained before making any strong conclusions, however.

CHAPTER 6. CONCLUSION

More broadly, one end goal for this line of research is to establish that SAO is both an ability and an executive function related to higher-order cognition. The evidence from the present study suggest that these hypotheses are viable. A potential application of this is that impulsive individuals (e.g., those with A.D.H.D.) might benefit from programs aimed toward teaching them to favor accuracy over speed when the situation requires it, and to better evaluate whether the task at hand requires a greater emphasis of accuracy or speed. Such training will stand a better chance of being successful than attempting to increase intelligence through WMC training, because the goal is to train a particular strategy that can be applied universally, rather than train an entire construct in hopes it transfers to another. If these training programs were to be implemented, then formerly impulsive individuals would do better in class, score higher on standardized tests, and, in general, have a better quality of life because of their ability to perform tasks more efficiently and effectively.

One useful takeaway from the trial-by-trial analyses is that low ability or otherwise poorly functioning individuals seemingly benefit from direct feedback, as they perform better when given feedback than when no feedback is present. As such, program could be implemented that are directed towards initially providing performance feedback, and then teaching subjects better self-evaluative skills such that direct feedback is no longer required.

Ultimately, I argue that this study provides a starting point and lays the foundation for future researchers to use related methodologies to study the nature of SAO differences and how low performing individuals can be instructed, taught, and/or trained to meet task demands not necessarily in the research lab, but in the real world. In the future, I intend to do conduct studies using a more in-depth modeling approach (e.g., diffusion modeling) to even better understand the nature of individual differences in adopting the optimal SAO to meet task demands.

APPENDIX A. DESCRIPTIVE STATISTICS FOR SAO TASKS

The following tables show more in-depth descriptives for the line, lexical decision, and flanker SAO tasks. Note that due to a coding issue, a couple variables are missing from the tables. The variables are labeled by condition, for instance "LDTVeryFastAcc" is the accuracy for the 96 trials of the lexical decision task that had the quickest response deadline. LexicalVeryFastMissedDL is total number of missed response deadlines for that same condition. LexicalMDL is the total number of missed deadlines across all conditions. "Score" refers to the points accrued in the trials with a response deadline.

Table A1 – Descriptive Statistics for Lexical Decision SAO Task.

Descriptive Statistics

	Mean	Std. Deviation	Skewne...	Kurtosis
	Statistic	Statistic	Statistic	Statistic
LDTBaselineAcc	.92	.092	-2.826	9.807
LDTTotalAcc	79.65	9.978	-1.386	1.711
LDTVeryFastAcc	.78	.115	-.940	.606
LDTFastAcc	.83	.117	-1.303	1.542
LDTSlowAcc	.85	.107	-1.399	1.849
LDTVerySlowAcc	.86	.107	-1.651	2.833
LDTBaselineRT	703.41	165.748	2.051	8.723
LDTTotalRT	531.75	93.063	.155	3.653
LDTVeryFastRT				
LDTFastRT	495.18	95.948	-.128	2.714
LDTSlowRT	546.94	100.013	.305	3.712
LDTVerySlowRT	556.36	102.066	-.030	4.320
LexicalMDL	41.87	25.028	1.028	1.377
LDTVeryFastMissedDL	25.85	14.748	.969	.984
LDTFastMissedDL	9.32	7.662	1.427	2.958
LDTSlowMissedDL	4.37	4.713	1.805	5.122
LDTVerySlowMissedDL	2.33	3.021	2.548	8.740
LexicalScore	210.88	88.004	-1.173	1.570
LDTVeryFastScore	28.67	28.169	-.727	.828
LDTFastScore	53.37	25.116	-1.121	1.562
LDTSlowScore	62.10	22.096	-1.262	1.540
LDTVerySlowScore	66.74	21.003	-1.525	2.405

Table A2 – Descriptive Statistics for Line Discrimination SAO Task.

Descriptive Statistics

	Mean	Std. Deviation	Skewne...	Kurtosis
	Statistic	Statistic	Statistic	Statistic
LineBaselineAcc	.9696	.06459	-3.681	16.517
LineTotalAcc	.8433	.08382	-1.801	4.270
LineVeryFastAcc	.8490	.10472	-1.289	1.681
LineFastAcc	.8792	.09119	-1.767	3.963
LineSlowAcc	.8886	.09110	-1.912	5.440
LineVerySlowAcc	.8970	.08922	-1.843	4.012
LineBaselineRT	801.7084	268.65405	1.497	3.342
LineTotalRT	490.2870	105.14222	1.533	5.947
LineVeryFastRT	450.8768	94.67553	.816	3.494
LineFastRT				
LineSlowRT	503.4354	117.00251	2.064	10.516
LineVerySlowRT	522.1035	129.15429	1.892	6.355
LineMDL	18.99	21.316	2.951	12.849
LineVeryFastMissedDL	11.10	11.080	2.315	8.031
LineFastMissedDL	3.89	5.482	2.966	12.487
LineSlowMissedDL	2.34	4.319	4.279	24.869
LineVerySlowMissedDL	1.67	4.055	5.190	33.868
LineScore	271.42	80.172	-1.744	3.753
LineVeryFastScore	55.87	27.223	-1.357	2.166
LineFastScore	68.85	20.453	-1.768	3.615
LineSlowScore	72.18	19.911	-2.001	5.551
LineVerySlowScore	74.51	19.017	-1.914	4.260

Table A3 – Descriptive Statistics for Flanker SAO Task.

Descriptive Statistics

	Mean	Std. Deviation	Skewne...	Kurtosis
	Statistic	Statistic	Statistic	Statistic
FlankBaselineAcc	.9719	.07028	-5.225	31.873
FlankTotalAcc	88.7595	5.91697	-2.275	7.007
FlankVeryFastAcc	.8888	.07871	-1.614	3.953
FlankFastAcc	.9224	.06777	-2.000	5.199
FlankSlowAcc	.9432	.05960	-2.212	5.415
FlankVerySlowAcc	.9440	.06601	-3.026	12.140
FlankBaselineRT	566.4035	119.12914	1.836	5.097
FlankTotalRT	462.0419	75.39359	1.677	4.552
FlankVeryFastRT	419.1584	67.04399	1.295	4.388
FlankSlowRT	482.3012	89.64593	1.873	5.747
FlankVerySlowRT	487.6524	86.17026	1.693	4.601
FlankerMDL	42.8512	22.61503	1.206	2.375
FlankVeryFastMissedDL	23.8062	12.37537	1.195	1.880
FlankFastMissedDL	10.6990	6.93892	1.247	2.523
FlankSlowMissedDL	5.3875	4.69212	1.809	4.424
FlankVerySlowMissedDL	2.9585	3.06611	2.433	9.295
FlankerScore	283.0415	59.52408	-1.597	3.832
FlankVeryFastScore	50.8166	22.45561	-.959	1.303
FlankFastScore	70.3772	16.64143	-1.598	3.534
FlankSlowScore	79.6055	14.21451	-1.990	5.018
FlankVerySlowScore	82.2422	13.97903	-2.659	9.945

APPENDIX B. SCORES BY COUNTERBALANCE CONDITION

Scores for the counterbalance conditions in the SAO tasks are shown below (Figure B1 and B2). There were no statistically significant differences. Note that these descriptives might not like up 100% with the descriptives from Table 2, and this discrepancy is due to using the imputed scores here rather than the raw listwise deletion scores.

Group Statistics

	CounterBalance	N	Mean	Std. Deviation	Std. Error Mean
LineScoreImputed	0	170	277.4376	69.63819	5.34101
	1	167	264.0140	83.25768	6.44267
LexicalScoreImputed	0	170	211.4334	81.62013	6.25998
	1	167	209.9596	88.69548	6.86346
FlankerScoreImputed	0	170	280.4951	56.62422	4.34288
	1	167	284.8826	55.68327	4.30890

Figure B1 – Descriptive Statistics of Counterbalance Scores for Each Task.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
LineScoreImputed	Equal variances assumed	1.427	.233	1.607	335	.109	13.42354	8.35546	-3.01224	29.85931
	Equal variances not assumed			1.604	322.810	.110	13.42354	8.36865	-3.04045	29.88752
LexicalScoreImputed	Equal variances assumed	.615	.433	.159	335	.874	1.47376	9.28260	-16.78578	19.73329
	Equal variances not assumed			.159	331.635	.874	1.47376	9.28948	-16.79998	19.74749
FlankerScoreImputed	Equal variances assumed	.885	.347	-.717	335	.474	-4.38754	6.11870	-16.42345	7.64836
	Equal variances not assumed			-.717	335.000	.474	-4.38754	6.11778	-16.42165	7.64656

Figure B2 – Significance Testing of the Different Counterbalance Conditions.

APPENDIX C. EXPLORATORY FACTOR ANALYSIS WITH NO ROTATION

Table C1 – Exploratory Factor Analysis of WMC, Gf, and SAO Tasks.

Task	Factor 1	Factor 2	Factor 3
OSpan	**.721**	-.275	.290
SymSpan	**.772**	-.252	.411
RotSpan	**.821**	-.116	.343
Raven	**.789**	-.173	-.279
LetterSet	**.748**	-.093	-.385
NumSeries	**.778**	-.210	-.395
Line SAO	.526	**.679**	-.086
Lexical SAO	.385	**.753**	.077
Flanker SAO	.300	**.702**	.079

Note. OSpan = operation span; SymSpan = symmetry span; RotSpan = rotation span. No rotation was used. Boldface indicates the highest factor loading for that variable.

REFERENCES

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131,* 30-60.

Baddeley, A. D., Bressi, S., Della Sala, S., Logie, R., & Spinnler, H. (1991). The decline of working memory in Alzheimer's disease. *Brain, 114* (6), 2521-2542.

Baddeley, A. D. (1992). Working memory. *Science, 255* (5044), 556-559.

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*, 158-173.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*, 445-459.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychological bulletin, 121* (1), 65.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1-22.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12* (5), 769-786.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review, 3* (4), 422-433.

Draheim, C., Harrison, T. L., Embretson, S. E., Engle, R. W. (In press). What Item Response Theory can tell us about the complex span tasks. *Psychological Assessment.*

Draheim, C., Hicks, K. L., and Engle, R. W. (2015). Combining reaction time and accuracy: The relationship between working memory capacity and task switching as a case example. Revision submitted to *Perspectives on Psychological Science*.

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

Engle, R. W., Carullo, J. J., & Collins, K. W. (1991). Individual differences in working memory for comprehension and following directions. *Journal of Educational Research*, *84*, 253-262.

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* (3), 309-331.

Fitts, P. M. (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Psychology, 71*, 849–857.

Forstmann, B. U., Tittgeymeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., and Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: A structural model-based approach. *The Journal of Neuroscience, 31* (47), 17242-17249.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory and Cognition*.

Goldman-Rakic, P. S. (1994). Working memory dysfunction in schizophrenia. *The*

Journal of neuropsychiatry and clinical neurosciences.

Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology, 24,* 1149-1167.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D.Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science, 24* (12), 2409-2419.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8,* 1-19.

Heitz, R. P., & Engle, R. W. (2007). Focusing the spotlight: individual differences in visual attention control. *Journal of Experimental Psychology: General*, *136*(2), 217.

Hughes, M. M., Linck, J. A., Bowls, A. R., Koeth, J. T., Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task switching paradigm: Their reliability and increased validity. *Behavior Research Methods*, *46* (3), 702-721.

Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130*, 169-183.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*, 66-71.

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

Luce, D. (1986). Response Times: Their Role in Inferring Elementary Mental Organization. New York, NY: Oxford University Press.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.M. (2005). Working memory and intelligence—their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *American Psychological Association*, *131*, 61.

Oberauer, K., Süß, H. M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence, 31* (2), 167-193.

Ollman, R. T. (1966). Fast guesses in choice reaction time. *Psychonomic Science*, *6*, 155–156.

Paap, K. R., & Sawi, O. (2016). The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods*, *274*, 81-93.

Pachella, R. G., and Pew, R.W. (1968). Speed-accuracy tradeoff in reaction time: Effect of discrete criterion times. *Journal of Experimental Psychology, 76,* 19–24.

Pearson, K. (1905). The problem of the random walk. *Nature, 72*, 294.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review,* 85, 59–108.

Schouten, J. F., and Bekker, J. A. (1967). Reaction time and accuracy. *Acta Psychologica 27*, 147-153.

Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise:

Boundary optimality in the diffusion model. *Psychology and Aging*, *25*(2), 377.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological*

Bulletin, *87* (2), 245-251.

Unsworth, N. & Engle, R. W. (2008). Speed and accuracy of accessing information in working memory: An individual differences investigation of focus switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 616-630.

Wickelgren, W. A. (1977). Speed-accuracy Tradeoffs and Information Processing Dynamics. A*cta Psychologia, 41*, 67-85.