THE UNIVERSITY OF CHICAGO

SOLVING PROBLEMS WITH INSIGHTFUL THINKING

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

BY

PETER TA-KANG HU

CHICAGO, ILLINOIS

JUNE 2018

ABSTRACT

Insightful problem-solving is a well-studied behavior, but one that is not necessarily well-understood. There is a common sense of what insight is, typically relayed through stories like Isaac Newton getting bopped on the head with a falling apple or Archimedes jumping out his bathtub with a shout of "Eureka!" But the current understanding of how we solve problems through insight has several shortcomings. In this dissertation, I use a series of measures to both determine the cognitive underpinnings of insightful thinking and to extend current understanding of insightful thinking to perceptual processes. In the first set of studies, I examine whether insight is better thought of as a single ability or as an emergent property based on a collection of cognitive processes. In the second and third sets of studies, I put forth a novel visuoperceptual task and audio-perceptual task respectively, as possible measures for insight at a perceptual level. In the fourth set of studies, I ask whether we can identify an autonomic component of insightful problem solving by way of the manipulation of emotional valence. In the final set of studies, I examine the role of insight priming in the context of sleep-dependent memory consolidation processes. Together, these studies better inform our current understanding of insightful problem-solving by expanding it to include perceptual processing and deepen the understanding of cognitive processes that underlie this way of thinking.

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my committee chair Professor Howard C. Nusbaum. Without his guidance this dissertation would not have been possible.

I would like to thank my committee members, Professor David A. Gallo, Professor Marc G. Berman and Professor Leslie M. Kay, whose input have been invaluable to this work.

In addition, a thank you to Professor Jonathan D. Cohen of Princeton University, who welcomed me into his laboratory and introduced me to the field of Cognitive Neuroscience. I thank Professor Robert "Bob" Stickgold of Harvard Medical School and Professor Matthew P. Walker of the University of California, Berkeley, for introducing me to the field of Sleep Medicine and the study of sleep-dependent memory consolidation – these are ideas that continue informing my thinking today.

The members of the Nusbaum Lab who have let me bounce ideas off them deserve my gratitude: Shannon Heald, Stephen Van Hedger, Serena Klos, Sophia Uddin, Alison Trude, Patrick Williams, Jean Boulware, Greg Poljacik, Emma Wyatt, and Cassie Kozyrkov.

Thank you to the undergraduate lab members who helped with collecting and scoring the data that appears in this dissertation: Michael Newman, Evan Weingarten, Michelle Kim, and Sophie Holtzmann.

Finally, I would like to thank my parents, Kuang-hao Howard & Mei-yun Hu.

TABLE OF CONTENTS

TABLES viii

FIGURES xi

CHAPTER ONE: INTRODUCTION 1

Specific Aims 8

Specific Aim #1 8

Specific Aim #2 9

Specific Aim #3 10

CHAPTER TWO: GENERAL MECHANISMS OF INSIGHTFUL PROBLEM SOLVING 12

Statistical Methods 18

Experimental Methods 20

Results 28

Discussion 54

CHAPTER THREE: VISUOPERCEPTUAL PROCESSES OF INSIGHT 58

Introduction 59

Study #3 Methods 68

Study #3 Results 71

Study #4 Methods 76

Study #4 Results 78

Study #5 Methods 79

Study #5 Results 81

Discussion 83

Conclusion 86

CHAPTER FOUR: AUDITORY PERCEPTION AND TRANSFER OF SKILL 88

Study #6 Results: Correct Identification 103

Study #6 Results: Signal Detection 111

Discussion 118

CHAPTER FIVE: EMOTIONAL PROCESSING AND VISUOPERCEPTUAL INSIGHT MECHANISMS 123

Study #7 Methods 131

Study #7 Results 137

Discussion 145

CHAPTER SIX: PRIMING INSIGHT THROUGH SLEEP-DEPENDENT MEMORY CONSOLIDATION 148

Study #8 Methods 152

Study #8 Results 154

Discussion 157

CHAPTER SEVEN: GENERAL DISCUSSION 160

REFERENCES 164

APPENDICES 191

Appendix A: Demographics Questionnaire 191

Appendix B: Stanford Sleepiness Scale 193

Appendix C: Sleep Log 195

Appendix D: Participant Questionnaire 197

TABLES

Table 1: Solution Rates for Easy and Hard RAT problems sampled from Bowden and Jung-Beeman (2003) for Study #1. Participants in Bowden and Jung-Beeman responded to stimuli categorized here as “Hard” at a far lower accuracy rate than stimuli categorized as “Easy”. 22

Table 2: Solution Rates for Easy and Hard RAT problems sampled from Bowden and Jung-Beeman (2003. 26

Table 3: Spearman's rank-order correlation coefficients in Study #1. Boldface values in matrix are significant at the 0.05 level. 34

Table 4: Principal Component loadings for task measures in Study #1 36

Table 5: Study #1 Principal Component loadings based on a subset of eight representative factors identified in the omnibus test 38

Table 6: Mean accuracy rates for the seven Functional Fixedness problems in Study #2. 44

Table 7: Response accuracy rates for High, Medium, and Low Valence categories in the Camouflage Perceptual Fluency task. 45

Table 8: Response accuracy rates for High, Medium, and Low Arousal image sets in the Camouflage Perceptual Fluency task 46

Table 9: Matrix of Spearman’s rank order correlation coefficients. Boldface coefficients are significant at the 0.05 level. 48

Table 10: Clusters of factors identified in first four components of a Principal Components Analysis. Values are component loadings. 49

Table 11: Principal Component loadings for accuracy performance in the Functional Fixedness, Number Reduction, Remote Associates, and Camouflage Perceptual Fluency tasks. Boldface values are items of interest. 51

Table 12: Spearman's rank-order correlation coefficients between d-prime for detecting an S in noise and other measures of perceptual experiences. p-values are not corrected for multiple comparisons. 82

Table 13: Participants are sorted into five Conditions that differ on the stimulus type presented at each block of the experiment. 100

Table 14: Spearman's r for correlations between correct response counts and R-SPAN Total scores. Items in bold are significant at the 0.05 level. Values are not corrected for multiple comparisons. 117

Table 15: Spearman's r for correlations between correct response counts and R-SPAN Total scores, Posttest stimuli divided by original presentation block. Items in bold are significant at the 0.05 level. Values are not corrected for multiple comparisons. 118

Table 16: IAPS Valence score means and standard deviations for three stimuli sets. Image sets are generated to represent three different Valence levels. 134

Table 17: IAPS Arousal score means and standard deviations for three stimuli sets used in the experiment..... 134

Table 18: Accurate identification rates by stimuli IAPS Valence category. 137

Table 19: SDNN for camouflage images categorized by IAPS Valence..... 138

Table 20: Heart rate variability (SDNN) at Posttest, categorized by whether participants had been shown the source IAPS image at Pretest or not. 139

Table 21: Condition protocols. AUT-FF is Chrysikou's (2005) modification of the Alternative Uses Task (AUT) with items from the Functional Fixedness problems substituted in. . 152

Table 22: Pretest Elaboration Scores by Experiment Condition 155

Table 23: Pearson's Rank-Order Correlations between Alternative Uses Elaboration Scores at Pretest and Functional Fixedness Scores at Posttest. 157

Table 24: Pearson's Rank-Order Correlations between Alternative Uses Category Scores at Pretest and Functional Fixedness Scores at Posttest. 157

FIGURES

Figure 1: NRT mean response time histograms. Left Panel: All insightful participants. Center Panel: slowest two insightful participants excluded. Right panel: all non-insightful participants. 29

Figure 2: Distribution of R-SPAN Scores in Study #1 32

Figure 3: Scree plot (Study #1). The first four principal components represent over 80% of variance. 37

Figure 4: Correlation Circle (Variable Chart) for principal components identified in Table 5. ... 39

Figure 5: Frequency of accuracy rates by Number Reduction Task sequence in Study #2..... 41

Figure 6: Frequency of accuracy rates by Number Reduction Task participant in Study #2. 42

Figure 7: Accuracy rates by RAT problem 43

Figure 8: Accuracy rates by image Valence in the Camouflage Perceptual Fluency task. 46

Figure 9: Accuracy rates by image arousal in the Camouflage Perceptual Fluency task. 47

Figure 10: Scree plot; proportion of variance explained by 8 principal components. The first three components account for 78.1% of variance. 50

Figure 11: Correlation Circle (Variable Chart) for Principal Components identified in Table 11. 52

Figure 12: Scree plot for four task accuracy measures. The first component accounts for 41.01% of variance and represents performance in Functional Fixedness problems, Remote Associates Task, and Camouflage Perceptual Fluency..... 53

Figure 13: Examples of visual stimuli. From left, 50x50 black and white prototype S, not shown to participants; 50 x 50 black and white noise sample (image 21 of 100), higher correlation with the visual target S; 50 x 50 black and white noise sample (image 95 of 100), no correlation with the visual target S; 50 x 50 grayscale summation of 100 noise images with the highest correlation with the visual target S; 50 x 50 summation of 100 noise images with no measurable correlation to the visual target. 69
Figure 14: Distribution of d-prime scores across participants demonstrates on average an ability to discriminate between the two sets of noise samples. 73
Figure 15: Distribution of betas across participants suggests that the d-prime values are not due to a general bias in responding yes in all stimuli. 73
Figure 16: Distributions of the probability that images are reported to contain the target S. Red bar represents 50% chance. 74
Figure 17: d-prime statistics for a Bag of Features image classifier when trained on S-similar and S-unsimilar noisy-image categories (Condition #1) or on noisy-image categories representing numerals 1 through 9 and the upper-case alphabet (Condition #2). The green dotted line represents chance (d-prime = 0). 79
Figure 18: Response Accuracy by Experiment Condition at Posttest #1; numbers are out of 100 total trials. 104
Figure 19: Response Accuracy by Experiment Condition at Posttest #2. Correct Response values are out of 100 total trials. 105
Figure 20: Accuracy Rates at Pretest and Posttest #1, by Condition. 107
Figure 21: Accuracy Rates at Pretest and Posttest #2, by Condition. 110
Figure 22: Accuracy Rates at Posttest #1 114

Figure 23: Accuracy Rates at Posttest #2 115

Figure 24: Pretest to Posttest 2 (Seen at Pretest) Performance Change..... 116

Figure 25: Example IAPS stimulus image 1610 scored positive in Valence, low in Arousal. Less than 100% (but more than 0%) of independent raters recognized the contents of this image during normative testing. 133

Figure 26: Age Group comparison of heart rate variability (SDNN) at Sessions #1 and #2..... 140

Figure 27: Gender comparison of heart rate variability (SDNN) at Sessions #1 and #2..... 141

Figure 28: Gender comparison of heart rate variability (SDNN) by camouflage image Valence score 143

Figure 29: Gender Group comparison of heart rate variability (SDNN) by Age Group 144

CHAPTER ONE: INTRODUCTION

It is easy to come up with examples of insightful thinking from history, from apocryphal story of Isaac Newton and his apple to Archimedes jumping out his bathtub with a shout of "Eureka!" Moving from historic examples to a singular working definition by which to study the phenomenon is not simple. Examples of insight like Archimedes' might be considered to be a form of problem solving. After all, Archimedes' insight was about figuring out how to measure how much gold was in the king's crown. But must all insights come from explicit problem solving? When looking at a visually ambiguous or noisy image, sometimes there is an "aha" moment of insight in which the booming buzzing confusion clears and the solution recognized: we see the image as representing a scene instead of as being just a collection of ink blotches. Is this a form of "perceptual insight"? Topolinski and Reber (2010) give a definition of insight comprised of four parts: *suddenness*, where the solution pops into mind; *ease*, where problem-related processing is now faster; *positive affect*; and *confidence* that the solution is correct. Yet, research in the field suggests this definition might be too simplistic.

Take, for example, the argument that insight must be *sudden*. In addition to Topolinski and Reber (2010), this assertion is also made elsewhere in the literature (Gick & Lockhart, 1995; Metcalfe & Wiebe, 1987). In the Number Reduction Task (NRT; Thurstone & Thurstone, 1941; Yordanova, Verleger, Wagner, & Kolev, 2010), each stimulus consists of a sequence of numbers, e.g., "1 1 9 4 9". Participants are taught by the experimenter a multi-step algorithm for

"solving" this sequence to arrive at a pattern of responses; participants are specifically informed that the last response in the response pattern is taken as the final answer. What participants are not told, however, is that a secret "shortcut" rule exists that allows participants to skip most of the steps in the algorithm and respond almost immediately with the correct final answer. Participants in the NRT are judged to have achieved successful insight by whether they found the hidden rule or not and how many trials were necessary before a given participant started to reliably use the shortcut. Recent work suggests that there are measurable neural precursors (e.g., changes in event-related potentials) of this "Aha!" moment in the NRT (Yordanova, Verleger, Wagner, & Kolev, 2010; Lang et al., 2006; Darsaud et al., 2011). In addition to neural precursors, there are subtle but robust behavioral cues that the "aha!" moment is imminent. For example, individuals sometimes show reduced response times as they progress through the problems, demonstrate partial solutions, or apply a correct solution on one trial but an error on the next trial (Yordanova et al., 2010). These precursor behaviors suggest that the *sudden awareness* of insight, as identified by Topolinski and Reber, might not strictly be a sufficient and necessary definition.

Even though the NRT might not demonstrate the suddenness aspect of insight, the task appears to demonstrate Topolinski and Reber's *ease* component[1], that problem-related processing becomes faster upon insight. Once the shortcut rule is identified, subsequent trials are measurably faster (Yordanova, Verleger, Wagner, & Kolev, 2010). However, this faster problem-related processing is not found in all tasks that have been used to study insight. One

[1] The "ease" and "sudden" components not unique to Topolinski & Reber (2010); the two components are further discussed by Auble, Franks, and Soraci (1979), Gick and Lockhart (1995), and Metcalfe and Wiebe (1987) as being necessary and sufficient together to define the phenomenon of insight.

example task in which increased speed is not identified is in the Remote Associates Task (RAT; Mednick, 1968). In this task, participants are given three words (for example, “crab”, “pine,” and “sauce”) and are required to identify a root word that associates the three (in this case, “apple”). In the RAT, participants tend to report that the answers just “come” to them, and that the experience in finding the insightful solution was not one of explicit search (Kounios & Beeman, 2009). However, this does not necessarily mean that processing is easier or faster: participants may sometimes try to do a trial-and-error search for a common root word (e.g., “cake” to form “crabcake”, but fails to work for “pinecake” or “cakesauce” before landing upon the correct solution “apple”). Furthermore, it is clear from the normative data provided by Bowden and Jung-Beeman (2003) that certain problems are solved successfully at a higher rate, and that some problems take more time to solve, but Bowden and Jung-Beeman do not provide evidence that increased speed might be found toward the end of a set of RAT problems.

It is evident from these two examples that Topolinski and Reber’s definition is insufficient to satisfactorily describe the phenomenon of insight. Yet, both the NRT and the RAT have both been used, albeit separately, to explore how insightful thinking works. That insight is complicated to define and that a multitude of different tasks have been developed to study the phenomenon of insight, is it reasonable to assume that insight is a singular cognitive factor? Take for example Spearman’s measure of general intelligence (g; Spearman, 1904): a construct based on comparing results on a number of tasks ranging from language (verbal) to mathematics. In particular, Spearman found that performance on one of his measures was predictive of performance on his other measures. From this, Spearman identified a set of positive correlations across a variety of cognitive tasks which he termed the “g factor”. If insight is similarly a single

factor, like g, an "insight factor" could be identified in the same manner through identifying a set of positive correlations among different laboratory measures of insight behavior. The work presented in this dissertation will test this assumption by examining whether performance on one measure of insightful thinking predicts performance on other measures of insight.

If insight can be described as a singular cognitive factor, what is that factor? Alternatively, what cognitive processes might support insightful thinking? According to Thorndike (1898), problem-solving consists of a trial-and-error process in which a series of solutions are applied until the correct one is found. While this explanation may describe the processes of insightful problem solving, it does not describe the selection process used to identify possible solutions. Thorndike's explanation leaves an unanswered question: how is a solution identified as "correct" by the individual thinker?

One possible explanation of how solutions are identified and selected is through spreading activation models. Bowden, Jung-Beeman, Fleck, and Kounious (2005), for example, use the RAT as the basis for a theoretical explanation for how spreading activation can lead to eventual successful retrieval in problem solving. When giving a stimulus consisting of the words "cream", "cube", and "rink", the activation of weaker remote associate words can occur, resulting in the retrieval of "hockey" that works with "rink", but not "cream" or "cube". However, the integration of the three stimulus concepts eventually means one remote associate word will become the strongest-activated path in the spreading-activation network, and be returned as the answer. In this case, that means the word "ice" is retrieved, and the problem is solved.

In a spreading activation model, memories are taken to be stored independently as interconnected concepts where activating one concept will result in closely-connected concepts to be activated as well (Ratcliff & McKoon, 1988). This account of insightful thinking also explains why incorrect solutions, or even why no solutions, are sometimes retrieved: Schwartz and Smith (1997) argue that retrieval failure involves interference of related memories with the "correct" item being recalled. For example, a "tip-of-the-tongue" retrieval failure could be due to a given solution's activation being high enough to signal presence of a solution yet just below threshold for retrieval (Schwartz, 1999). However, it is also possible that solutions are blocked by the retrieval of incorrect solutions (Jones, 1989; Brown & McNeill, 1966; Brown, 1991). In a spreading activation model, activation would spread from related concepts to other related concepts besides the correct answer (e.g., retrieving "Regis Philbin" but not "Kathie Lee" when discussing television talk show hosts). But activating a related concept (e.g., "Donahue") even when it is recognized as the wrong solution might block retrieval of the correct solution (Schwartz, 1999). Emerging from this theoretical account is the idea that insight might be due to competition for selection from activated memories that are related, but not helpful to solving the problem. This competition of related concepts can, at times, ultimately lead to failure (Collins & Loftus, 1975; Anderson, 1983).

Spreading activation can also help explain the "suddenness" component of insight. While a solution or response can be described as having been suddenly identified, this description does not necessarily indicate that solutions or responses come instantaneously. That is, a time delay can occur before the solution or response is suddenly identified. Work by Yaniv and Meyer

(1987) attribute this length of time between the problem presentation and the "aha!" moment, an "incubation period" (during which the thinker does not have an answer) to the process of spreading activation. This incubation period, Yaniv and Meyer argue, is the very time period during which when related concepts are retrieved from memory, a solution selected from among the activated concepts, and returned as a solution (which may be objectively correct or not).

Though these accounts describe what concepts from memory might be activated (correctly or incorrectly), and when it happens, these accounts do not necessarily describe how the *correct* solution for a specific *problem* is identified, particularly in cases where the exact solution might be novel to the individual and thus not specifically stored in memory. Gick and Holyoak (1983) argue that a set of analogs may result in the formation of a schema to ultimately result in analogical transfer, where the solution for one problem is recognized to be the solution of a similar problem. But not all retrieved analogs ultimately result in a solution. Such analogs might be retrieved via spreading activation, as suggested by Holyoak and Thagard (1987). In this view, spreading activation helps us recognize potentially useful analogs to our problem by highlighting similarities.

For both the RAT and NRT, behavioral measures (such as accuracy and reaction time) can be taken as proxies for insightful thinking and serve as a basis for comparing differences across individuals. In such a correlation approach, the tasks traditionally used to probe insight (including the NRT and the RAT) may have more in common with each other than with perceptually focused tasks. This presents a problem, because the literature on perceptual insight is sparse. What is "perceptual insight"? Is recognizing a fuzzy picture or understand garbled

speech (as defined by Rubin, Nakayama, and Shapley, 2002) more stimulus-dependent than other kinds of insight, such as identifying hidden rules in a game? This leads to an important set of questions: is the kind of perceptual insight that occurs when recognizing the contents of a photograph a different kind of insight than the insights identified in more explicit problem-solving? Do perceptual insights share the same common psychological processes as found other measures of insightful problem solving (such as the Number Reduction Task or the Remote Associates Task)? While previous research suggests that perceptual insights can occur when individuals are presented with degraded visual stimuli (Rubin, Nakayama, & Shapley, 2002; Ludmer, Dudai, & Rubin, 2011), is this idiosyncratic to visual processing or do these findings reflect a more general form of insight? The relationship of perceptual insight to cognitive insight remains an open question.

This dissertation will attempt to better clarify the current understanding insight and outline the similarities between "problem solving" insight and "perceptual" insight through a series of Specific Aims. The goal of the research presented here is to develop a framework by which to understand and further study insightful problem solving. The intention of this work is to take a more systematic approach to investigating insight, in contrast to most previous work that draws general insight-related conclusions based on only one task. This work will place the various measures of insight, in the form of differing laboratory tasks, into an organized framework to make better sense of the mechanisms, processes, and other factors that may be involved in producing insight.

Specific Aims

The studies reported here attempt to understand insight in terms of psychological processes that can be inferred from of relationships between insight and other factors such as the role of sleep, working memory, pattern recognition, arousal. It is important to understand how the different characteristics of a task (e.g., the relationship between trials, the way responses are produced and measured, the demands on working memory) reflect a single process underlying insight or a complex assemblage of processes. Thus, the present research examines a set of similarities and differences among tasks.

Specific Aim #1

Do insights arise from a single process or a dynamic assemblage of different mechanisms that only share the ex post affective response in common? Another way to think about this is whether there is one kind of insight process responsible for insights in any situation or problem or whether there are different kinds of insights like the putative difference between spatial and verbal intelligence. The descriptions of insight above do not specifically indict a single process that accounts for insights of all kinds but is really just a collection of reactions to solving a problem and some of those aspects arise in other situations as well.

If performance on very different tasks shares substantial common variance, this would suggest underlying common processing. To the extent that there is a similar correlation across

very different cognitive tasks with different insight measures, this would support the hypothesis that insight across different kinds of tasks arises from a common mechanism. By contrast, substantial differences in performance that result in different patterns of performance across different insight tasks would suggest that there are different kinds of insight manifest in different situations. And, if there are different kinds of insight, there may well be different processes or combinations of mechanisms underlying each kind.

By examining the relationships of performance on working memory with different insight measures used in other studies, it is possible to go beyond the question of the relationship among types of insight tasks to a relationship with one cognitive system—working memory. Insight tasks more associated with working memory performance are likely to depend on this system as part of problem solving. If some insight tasks are correlated with working memory performance and others are not, but there is an intercorrelation among all the insight tasks, this would suggest that there may be similarities and differences in the underlying processing for insight.

Specific Aim #2

Furthermore, some tasks call for explicit, logico-deductive processing such as solving a concrete problem or finding words. But perceptual recognition may be much less an explicit cognizing form of insight. To what extent is perceptual insight related to more explicit problem solving or memory retrieval insight? Perceptual insights have seem more direct and without a lot of explicit problem solving. There are examples of perceptual exposure to an auditory or visual

stimulus and the feeling that there is just no way to understand what is being seen or heard and then, as with other kinds of insight, the confusion resolves into a coherent recognition.

In particular, the research addressing this aim is designed to extend previous research in the visual domain and explore insight in auditory perception. By comparing research in the visual and auditory domains, the studies here allow further examination of whether these two domains share common psychological processes, whether these shared processes represent a distinct form of perceptual insight, whether this perceptual form of insight represents a complex of processes or different sets of processes as described in Specific Aim #1. This will in turn inform whether a framework for insight research will represent perceptual processes of insight separate from cognitive processes of insight, and suggest whether such perceptual processes require further investigation.

Specific Aim #3

The descriptions of insight have a strong affective component associated with the response to solving a problem. To what extent does an antecedent emotion affect achieving insight? For example reducing stress can improve insight (Creswell, Dutcher, Klein, Harris, & Levine, 2013) from a positive self-affirmation exercise. Thus, improving affect (e.g. reducing stress in chronically stressed individuals) may improve insight (cf. Isen, Daubman, & Nowicki, 1987). However, little research has directly tested how positive mood affects insight processes in problem-solving. Given the relationship of emotion to autonomic system arousal (Levenson,

1992), the performance improvement could result from arousal. If the positive emotion effect found by Isen, et al. (1987) is due to increased arousal, it is possible to test this by manipulating arousal and measuring autonomic arousal. Of course to test this it would be important to differentiate between the effects of arousal and emotion which is not done here.

Specific Aim #1, which considers whether there is just one type of insight suggesting the possibility of a single cognitive process for all cases of insight or as a collection of different psychological processes all called insight, will be addressed in Chapter Two. Chapter Two compares different insight tasks that have been used previously to study insight. Specific Aim #2 examines the role of perceptual insight in comparison with more explicit cognitive insight is first addressed in Chapter Two and then subsequently elaborated.

CHAPTER TWO: GENERAL MECHANISMS OF INSIGHTFUL PROBLEM SOLVING

Insight has been studied using a number of different laboratory experiments, each of which examine a specific aspect of cognition surrounding the "aha!" moment. A wide assortment of laboratory tasks have been labeled and studied as measures of insight, but it is not clear whether these tasks as a group test the same underlying phenomenon. This is a particular problem when attempting to understand insight through a research framework, but it is a problem that leads to a testable hypothesis: do these tasks test the same insight ability? More specifically, does performance on one insight task predict performance on another insight task?

A comparison between tasks of insight is certainly not novel in and of itself. Cunningham, MacGregor, Gibb, and Haar (2009) reported an experiment in which participants completed a set of insight tasks, comparing Rebuses (MacGregor & Cunningham, 2008), the Remote Associates Task (Mednick, 1962), Functional Fixedness problems (Duncker, 1945), and verbal analogy problems (cf. Gentner, 1983). The results, in the form of a correlation matrix, illustrate relationships they found between the tasks. These relationships led them to conclude that the solutions could be broken down into a set of six forms of mental restructuring, forms of which includes items as "abstract and non-visualized goals" and "misdirection". Cunningham et al. (2009), however, does not address the issue that each of these tasks appears to depend on different strategies, which in turn may be reliant upon different cognitive mechanisms.

Schooler and Melcher (1995) similarly tested a set of tasks commonly used to assess insight; they found that insight problem solving was not only different from analytic problem solving, but likely dependent on verbal working memory. Schooler and Melcher's protocol included Duncker (1945) and Maier's (1970) Functional Fixedness problems to test insight and analytic problem solving, a task that involved perceiving and recognizing out-of-focus pictures (Bruner & Potter, 1964), an embedded-figures task (Witkin, Oltman, Raskin, & Karp, 1971), a categorization task, anagram-solving, and the Remote Associates Task. Schooler and Melcher contrasted these insight tasks with a set of tasks that are not prima facie insight tasks; these included a set of vocabulary problems, mental rotation problems, and SAT questions which they purported test general abilities. While their resulting correlation matrix of task performance measures allowed Schooler and Melcher to conclude that verbal working memory was likely a key mechanism in insightful behavior, it is not clear if the same conclusion can be made from a larger set of insight tasks. It is also not possible to draw from the limited sample of tasks whether insight might be a singular process or multiple processes, as described by Specific Aim #1.

In a sense, the current work discussed here represents an extension of Schooler and Melcher's protocol, but with a larger set of experimental tasks and with different participant pools. Each participant in the current study of general insight was presented with five experimental tasks of insight: the Number Reduction Task, the Remote Associates Task, Functional Fixedness, and Camouflage Perceptual Fluency task. These tasks were selected because they represent a cross-section of current insight research, allowing for an examination of underlying mechanism in service of Specific Aim #1; each task has been previously used to

study certain aspects of insightful thinking. The Number Reduction Task (NRT) was selected because it requires participants to pay attention to patterns embedded in the task itself. Successful insight in the Number Reduction Task (NRT) is judged by whether the individual found the hidden rule or not, and how many trials were necessary before the participant started reliably using the shortcut. Recent work suggests that there are measurable neural precursors (e.g., changes in event-related potentials) of this "Aha!" moment in the NRT (Yordanova, Verleger, Wagner, & Kolev, 2010; Lang et al., 2006; Darsaud et al., 2011). In addition to neural precursors, there are subtle but robust behavioral cues that the "aha!" moment is imminent. For example, individuals sometimes show reduced response times as they progress through the problems, demonstrate partial solutions, or apply a correct solution on one trial but an error on the next trial.

The Remote Associates Task (RAT) was selected because it appears to depend on semantic memory in a way that the NRT does not. In this task, participants are given three words and are prompted to identify a root word that associates the given three words. In the RAT, participants tend to report that the answers just "come" to them, and that the experience in finding the insightful solution was not one of explicit search despite the requirement that the root word must be retrieved from memory (Kounios & Beeman, 2009). Unlike the NRT, the insight comes from successful memory retrieval, not observing a pattern. The measures of insight between these two tasks are also different: the NRT is measured in terms of time or number of trials until insight is achieved; the RAT is measured in terms of reaction time for each set of word associates or a response accuracy rate.

The Functional Fixedness (FF) problems were selected because they are expository word problems, unlike the NRT or the RAT; these problems have been used to understand restructuring, a process in which problem state assumptions are reconsidered (Duncker, 1945; Glucksberg & Weisberg, 1966). In one example FF problem, a wrench must be restructured and be thought of as a pendulum weight instead of as a tool used to tighten bolt and nut fasteners. Because reading time and problem-solving time cannot be disentangled, response time is not a useful way to measure insight on these problems. Instead, insight is judged by response accuracy – whether an outside observer agrees that the solution is reasonable and workable.

Finally, the Camouflage Perceptual Fluency (CPF) task was included because the stimuli are pictures and the insight comes from perceptual recognition – with no words present in the stimuli at all. The CPF is based on a protocol described in Ludmer, Dudai, & Rubin (2011), in which participants were shown degraded black-and-white images sequentially. These degraded images are not immediately recognizable upon first look, yet they are instantly recognizable once they have been viewed and recognized. This behavior, in which the insight of image content recognition is not immediately obvious and cannot be "unseen" once successful, is similar to the insight behavior in the Number Reduction Task, in which a secret rule embedded in the task trials is not recognizable on the first encounter, but once recognized becomes instantly and reliably recognizable. Yet the CPF task is also different from the NRT, as the insight in the NRT is one hidden rule that is discovered for the entire experiment (across stimuli), but CPF insight occurs with the recognition of each stimulus' contents. The CPF is also different than the other tasks of insight included here because of the visuoperceptual nature of the stimuli. The Somatic Marker Hypothesis suggests that sensory information are available for use in cognition like

decision-making (Damasio, Tranel, & Damasio, 1991), which is a different type of input than the semantic problems presented in other tasks of insight: the NRT stimuli are arithmetic problems and the RAT stimuli are words, but the CPF stimuli are pictures.

Importantly, little work has been done to demonstrate that each of these tasks test the same "insight" beyond the surface similarity that an "aha!" moment occurs when a solution is found. If a singular insight mechanism or process exists, performance on each of these tasks should be predictive of performance on the other tasks. If a singular insight ability, process, or mechanism exists, an individual who performs poorly on the Number Reduction Task should reasonably be expected to perform poorly on the Remote Associates Task – or any of the other tasks identified in the literature as a measure of insight - as well. If the insight tested across these different tasks is indeed the same insight, this would result in a correlation matrix demonstrating positively-correlated performance relationships between all of the selected tasks (i.e., performance on one of the tasks would predict performance on each of the other tasks). That is, a positive finding of a singular insight mechanism would be identified the same way Spearman's g (general intelligence) was identified: through a correlation matrix demonstrating performance on one task was predictive of performance on other tasks (Spearman, 1904). Such a finding would be in line with the goals of Specific Aim #1, providing evidence that these tasks are measuring the same cognitive mechanism or psychological process

An alternative hypothesis to the singular insight mechanism posited above would be that insight is an emergent property that what appears to be insight is due to other underlying processes. If performance on a given task is found not to be predictive of performance on the

others (e.g., low or negative correlations in group-wise insight rates across tasks), it suggests that insight is not a singular mechanism or process. These processes, in turn, could be correlated more strongly with some insight tasks than others, resulting in a clustered correlation matrix where only certain tasks (via performance measures) would display positive correlative relationships. In this case, it would be useful to understand which tasks demonstrate predictive power with which other tasks. If task performance were found to be predictive of each other only in subset groups of tasks, it is possible that multiple kinds of insight have been found. Alternatively, it is possible that the tasks formed subsets because the insightful behaviors measured by the tasks are emergent properties of other underlying processes or mechanisms. The studies described in this chapter further aims to begin clarifying what processes, such as perceptual processes (Specific Aim #2) might influences might be involved in insightful problem solving.

If the results of this test suggest that the alternative hypothesis is correct, the question remains what factors might underlie insightful problem solving. To this end, a secondary hypothesis can be tested regarding working memory. Several of the insight tasks appear to depend on working memory function more so than the other tasks. The Functional Fixedness problems require the solver to hold an entire written description of the problem state in working memory in order to reframe the problem and realize that the wrench, for example, can be used as a pendulum. The Number Reduction Task also requires the solver to remember enough of the trials as the task progresses in order to identify the hidden rule. Yet, the Remote Associates Task only requires working memory capacity to hold three words as a stimulus and retrieve associate root words from long term memory. Also, the Camouflage Perceptual Fluency task requires a

participant to search and retrieve from long term memory any possible interpretations of the image being viewed. How these tasks of insight might depend upon working memory capacity is an empirical question that is tested here as a secondary hypothesis.

Depending on the alternate hypothesis that insight is dependent on a set of cognitive factors (and is not a singular insight ability), Principal Components Analysis (PCA) is hypothesized to cluster performance measures of Functional Fixedness and Number Reduction Task as being related to a working memory capacity task. Meanwhile, it can be hypothesized that PCA will cluster the Remote Associates Task and the Camouflage Perceptual Fluency task so that are not related to a working memory capacity task.

Statistical Methods

The research presented here were collected as two studies; Study #1 and Study #2 differed primarily in participant recruitment. Participants in Study #1 were recruited from the University of Chicago community for an in-laboratory experiment session. Because the available participant pool on a university campus may not be representative of the general population in their educational attainment, Study #2 was conducted online with a recruitment through the Amazon Mechanical Turk (MTurk) platform. Study #2, which recruited from a participant pool likely to be more representative of the general population, was tested on the Number Reduction Task, the Remote Associates Task, the Functional Fixedness Problems, and the Camouflage

Perceptual Fluency task. In Study #1, the same tasks were assigned but the Camouflage Perceptual Fluency task was replaced with the R-SPAN working memory capacity task.

Statistical analysis of these tasks consists of an omnibus correlation matrix, similar to that used by Spearman to find evidence of general intelligence as a single factor (1904). While a correlation matrix will show relationships, it does not suggest how such tasks might cluster into groups. To determine which task measures might be grouped together (i.e., to begin to determine which component factors might be in common among the insight tasks represented within a given cluster), an omnibus Principal Components Analysis (PCA) was planned with all relevant measures entered. PCA could then be used to determine which measures are less representative of the tasks and a final PCA calculated with a selected set of representative measures.

It is important to note that the purpose of the PCA here is not to reduce dimensonality, but instead to better identify and visualize relationships between tasks and to circumvent problems that can arise when interpreting a correlation matrix. The components here are not used to compare between participants because this study does not contain an experiment manipulation with experimenter-defined conditions. Instead, the function of this study is to determine whether these tasks represent a statistically-significant fully-crossed correlation matrix (e.g., Spearman's general intelligence) or not. The outcome of this question in turn is used to inform the following four chapters.

Finally, common demographic factors are examined to characterize and ultimately rule out effects such as Age and Gender from both Study #1 (participants recruited from the

University of Chicago) and Study #2 (participants recruited online who reside in the United States).

Experimental Methods

In Study #1, participants were recruited through the University of Chicago Department of Psychology's online scheduling system (Sona Systems, Ltd., Tallinn, Estonia) and through advertisements posted on the university campus. 55 adults age 18 or older (M = 22.2 years, SD = 6.4 years; 26 female) were recruited for this experiment. While recruiting materials emphasized normal or corrected-to-normal vision (due to the reading necessary for each question item), visual acuity was not verified beyond self-report. Because of English-language proficiency requirements of the Remote Associates Task and the R-SPAN, recruitment was limited to native speakers of American English.

After informed consent and a basic demographic questionnaire were collected, participants were asked to complete the Stanford Sleepiness Scale (SSS; Hoddes et al., 1973) and the Affect Grid (Russell, Weiss, & Mendelsohn, 1989) to assess changes in wakefulness during the study. Pilot testing of the protocol suggested that the experiment session might become boring, which may result in participants failing to complete the tasks as instructed. The Stanford Sleepiness Scale and Affect Grid were also collected at the end of the experiment session as a check against incompliant task completion. Participants then completed each of the described experimental tasks on a computer in a counterbalanced order except for R-SPAN, presented last

in response to recommendations by pilot-testing participants regarding possible fatigue on that task.

The insight tasks were then presented in randomized order, with participants offered a short break between each task. Stimuli for the Number Reduction Task, Remote Associates Task, and R-SPAN were presented on a standard desktop personal computer monitor using E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA), with participants' responses collected using a standard USB keyboard and mouse. Functional Fixedness stimuli were presented on-screen, but responses were made with pencil-and-paper so that participants could diagram or draw part of their responses as they wished.

The Number Reduction Task (NRT) was based on Thurstone and Thurstone (1941), with sequences lengthened to 8 numbers and presented by computer. Participants were presented on-screen and read aloud by an experimenter instructions based on the procedure given by Wagner, et al. (2004). To maintain similarity with their work, stimuli for the NRT as presented here were sequences of 8 numbers where each number was either a 1, 4, or 9, to match Wagner, Gais, Haider, Verleger, and Born (2004). When solved as instructed, a sequence of 8 numbers resulted in a sequence of 7 responses. Due to the mirroring required to implement the hidden rule, there are 243 mathematically unique sequences in a full complement of 8-number stimuli sequences. The experimental task ended either when participants responded to all 243 unique sequences (i.e., there were no more sequences to present) or when participants started to use the hidden shortcut reliably and accurately. No time limits were imposed during the NRT, but participants were instructed to work as quickly and accurately as possible.

The Remote Associates Task (RAT) consisted of 34 items from Bowden and Jung-Beeman (2003), with half of the items as more difficult than the median and half less difficult (see Table 1). The median split was based on solution success rate at 15 seconds, as given by Bowden and Jung-Beeman. A two-sample t-test found a significant difference in success for the two sets of stimuli ($t(28) = -6.258$, $p < 0.001$). Stimuli were presented on-screen, with a word triplet presented at the top half of the display and a cursor at the bottom for participants to type in their one-word responses. When participants typed in their responses, the next word triplet stimulus was presented. Stimuli were presented in fully randomized order. As with the Number Reduction Task, no time limits were imposed, but participants were instructed to work as quickly and accurately as possible.

Difficulty Condition	Mean	SEM	Min	Max
Easy	51.3%	1.8%	41%	76%
Hard	29.5%	3.0%	17%	39%
All	40.4%	2.7%	17%	76%

Table 1: Solution Rates for Easy and Hard RAT problems sampled from Bowden and Jung-Beeman (2003) for Study #1. Participants in Bowden and Jung-Beeman responded to stimuli categorized here as "Hard" at a far lower accuracy rate than stimuli categorized as "Easy".

Functional Fixedness (FF) problem stimuli consisted of seven word-problem items taken from Duncker (1945) and Meier (1970) that require a written response from the participant (drawings were also permissible). Each problem was presented serially on screen for a maximum of two minutes or until participants pressed the spacebar denoting they had found a solution, whichever came first. The two-minute deadline was chosen because pilot testing showed no

participant taking more than one minute to find an answer, but a time limit was determined to be necessary in case participants failed to remember hitting the spacebar or if they had no solution to the given problem. Although each FF word-problem stimulus was presented by computer via E-Prime software (Psychology Software Tools, Pittsburgh, PA), participants were given a blank sheet of letter size paper for each question and a pen to write down their responses. This allowed participants to draw pictures and diagrams as part of their responses, which would otherwise be difficult to collect with a desktop computer equipped with keyboard and mouse.

In the Camouflage Perceptual Fluency task (CPF; Dallenbach, 1951), participants were shown images taken from IAPS (Lang, et al., 2008), smoothed by Gaussian filter, and flattened from color to black-and-white using a process adapted from Ludmer, Dudai, and Rubin (2011). Difficulty across stimuli was normed by adjusting the smoothing kernel so that more or less image detail is removed from a given image. Participants were shown these degraded black-and-white images sequentially in random order and asked to make recognition responses after each image presentation. Participants were asked to type in their responses into a computer, which were later scored by multiple independent reviewers for accuracy in accordance with existing multiple-judge reliability methods (Lacy & Riffe, 1996). Due to the variability of typed responses and the need for multiple human scorers, participants did not receive immediate accuracy feedback during the task.

As previously noted, performance in the CPF task is based on correct identification of the image contents as determined by independent raters. But because the source images are taken from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 2008), the

images were also normed on Valence (emotional positivity or negativity) and Arousal (strength of emotional content). The images were selected to represent a cross-section of Valence and Arousal scores so that the factors can be examined for their impact on the recognition rates of images. The effect of emotional factors on insight behavior are examined in further depth in Chapter Five. Although the primary reason for including the CPF task in Study #2 are due to questions of visual perception and how it relates to insight, but not emotional content of the stimuli, the IAPS Valence and Arousal measures are assessed and discussed here as well.

The Automated R-SPAN for E-Prime was used in Study #1 to measure working memory capacity (Unsworth, Heitz, Schrock, & Engle, 2005; Broadway & Engle, 2010; Psychology Software Tools, Pittsburgh, PA). Although the Automated R-SPAN is a complete self-contained package with participant instructions presented on-screen, the protocol was modified so that an experimenter read aloud instructions to ensure that the participant understood the mechanics of the task. The reading-aloud of instructions was implemented based on initial task piloting that found several individuals read the instructions too quickly and were unable to comprehend and complete the R-SPAN task.

All experiment tasks were coded specifically for Study #1 except for the Automated R-SPAN; the E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA) for the R-SPAN (Unsworth, Heitz, Schrock, & Engle, 2005) was downloaded and used with permission. Due to time constraints, participants did not complete every task during the experiment session. All 55 participants included in analysis completed the NRT, 51 participants (92.7%) completed the R-SPAN, 52 participants (94.5%) completed the FF, and 53 participants (96.4%) completed the

RAT. This distribution is due to task counterbalancing; participants who could not finish all tasks skipped the last scheduled task.

The NRT, RAT, FF, and R-SPAN tasks were presented to participants in randomized order. After the tasks were completed, participants were re-assessed with the Affect Grid and Stanford Sleepiness Scale questionnaires to determine whether affect, alertness (sleepiness) changed during the course of the experiment session. During debriefing, participants were asked whether they had prior experience with any of the completed tasks, and data from participants who affirmed previous experience with the Number Reduction Task, Remote Associates Task, or specific items from the Functional Fixedness problems. While some participants report familiarity with the idea of the RAT, none were able to recall exposure to specific stimuli. No participants reported familiarity with the NRT. Participants who reported familiarity with FF problems were noted so that those responses could be excluded from analysis.

In Study #2, the Remote Associates Task, Number Reduction Task, Functional Fixedness Problems, and the Camouflage Perceptual Fluency task were implemented in Qualtrics (Qualtrics, Provo, UT) with participants recruited through Amazon's Mechanical Turk service (MTurk; Paolacci, Chandler, & Ipeirotis, 2010). Recent research suggests participants recruited through MTurk are more diverse, are more reflective of the general American population than participants recruited on a university campus, and have been found to make comparably different choices in experimental tasks (Berinsky, Huber, & Lenz, 2012; Henrich, Heine, & Norenzayan., 2010; Ipeirotis, 2010). The Number Reduction Task and the Functional Fixedness Problems implemented in Qualtrics for Experiment #2 consisted of the same stimuli, but due to the online

nature of the tasks Functional Fixedness responses were not deadlined to two minutes and responses to all tasks could only be accepted if typed in. The stimuli for the Remote Associates Task, however, were taken from Bowden and Jung-Beeman (2003) to create a set of 72 hard and 72 easy problems based on Bowden and Jung-Beeman's response rate findings when paced at 2 seconds per problem (Table 2). The shorter response period was chosen as the reference pace because the flat-rate compensation method for participants recruited via the Mechanical Turk service were believed likely to result in quicker responses to stimuli (i.e., completing the experiment session faster yields a higher overall pay rate).

Difficulty Condition	Mean	SEM	Min	Max
Easy	14.0%	1.1%	4%[2]	52%
Hard	1.6%	0.2%	0%	4%
All	7.8%	0.8%	0%	52%

Table 2: Solution Rates for Easy and Hard RAT problems sampled from Bowden and Jung-Beeman (2003.

Participants were 75 adults (35 female) aged 22-62 (M =35.0 years, SD = 9.8 years) recruited to participate remotely through the Amazon Mechanical Turk (MTurk) website. Recruitment was limited to computers who IP addresses were located in the United States using TurkPrime software package (Litman, Robinson, & Abberbock, 2016). The R-SPAN was omitted from Experiment #2 due to technical constraints of the Qualtrics platform. Of the 75 participants, 15 were college-enrolled students, and 8 reported being left-hand dominant.

[2] One Remote Associates problem had a solution rate of 4% at 2 seconds, which would qualify it as a Hard problem, but was sorted into the Easy set due to high solution rates at 5, 15, and 30 seconds.

Participants completed the task on their own time using their own computer equipment. While convenient for participants, the use of an online experiment session like Amazon Mechanical Turk means that the testing environment cannot be directly controlled by the experimenter. However, participants agreed to keep the Qualtrics task window full-screen on their computer and were given opportunities (which consisted of full-screen messages) that asked participants to please clear their work area free of any distractions before the task began.

Pilot testing suggested that the task could be completed within a one hour block of time, but the time limit was set at 6 hours to allow participants to work at their own pace and convenience. Because the task was not completed in a controlled laboratory environment monitored by an experimenter, actual time spent on task is not accurately measurable because it was not possible to guarantee participants did not leave their computer workstations during the experiment session. To minimize this behavior, participants were instructed upon accepting the task to ensure their work environment was free of distractions before proceeding. Participant absence during the task was further minimized because the browser window that displays the experimental tasks could not be closed until the experiment tasks were completed since the check-in procedure used for tracking and payment purposes was located at the end of the experiment. Also, anecdotal evidence suggested that MTurk workers may reserve a task for later when time permitted; allowing a longer time limit would increase the rate of successful task completion.

The study was reviewed and approved by the University of Chicago Independent Review Board (IRB Approval #12-1367). Participants were compensated $6 via the Amazon Payments service for their time spent in participation of the study.

Results

Performance scores in the Number Reduction Task (NRT) were determined by the number of trials that the participant completed before the consistent correct application of the "hidden shortcut". As described in Wagner, Gais, Haider, Verleger, and Born (2004), participants must apply a secret rule that the second response is the same as the final response to be considered as having achieved insight. For this, it is not necessary for participants to verbalize the shortcut, nor to recognize that the hidden shortcut is due to the last three responses of the sequence mirrors the previous three responses. In making sense of participant response data, it is important to note that human participants do not behave ideally - and whether purposefully or accidentally will sometimes skip to the end and by chance happen to enter the correct final response. Assuming a 1/3 random chance of enter a correct response, the point of insight for a given participant is calculated as the first escape-to-end correct response in a sequence of 3 or more trials in which the participant escapes to the end and enters a correct final response. In Study #1 (with participants tested in the laboratory), 30 participants did not achieve insight on the NRT. A one-way ANOVA for response times between participants who did achieve insight ($M = 18125$ ms, $SD = 19488$ ms) and did not achieve insight ($M = 8249$ ms, $SD = 3396$ ms) was significantly different at the 0.05 level ($F(1, 49) = 7.446$, $p = 0.009$). After removing two outliers

who demonstrated insight yet performed slowly (94620 ms and 41495ms respectively), an ANOVA remained significant ($F(1, 47) = 11.59$, $p = 0.0014$), with response times for insightful participants ($M = 12870$ ms, $SD = 6115$ ms) longer than non-insightful. However, the response times for the insightful participants appear to be skewed (breaking assumptions of normality), whereas non-insightful participant response times appear more normally distributed (Figure 1). A Wilcoxon rank sum test was significant ($W = 146.5$, $p = 0.0046$).

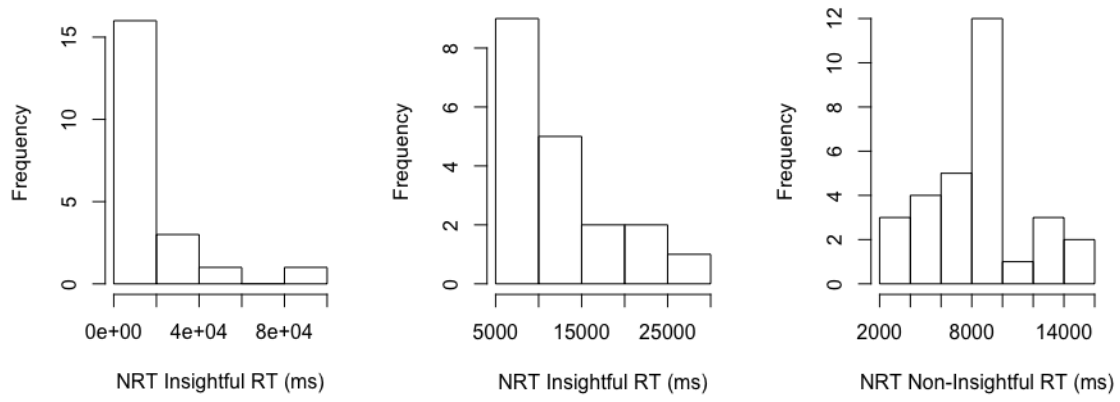

Figure 1: NRT mean response time histograms. Left Panel: All insightful participants. Center Panel: slowest two insightful participants excluded. Right panel: all non-insightful participants.

Responses for the Remote Associates Task were scored on whether the participant typed in the correct linking word or not (e.g., "ice" to form "ice cream", "ice cube", and "ice rink"), then adding up the number of correct responses for a sum total per participant. In the Remote Associates Task, Study #1 participants took a mean time of 14369 ms to respond (or give up and continue to the next problem; $SD = 6994$ ms). When the problems were median-split by difficulty level, participants took a mean time of 11977 ms ($SD = 6297$ ms) on the easy problems, and slower at 16762 ms ($SD = 8756$ ms) on the hard problems, as expected. The difference was significant ($F(1, 94) = 9.654$, $p = 0.002$). When accuracy is taken into account,

participants took only 7792 ms (SD = 3497 ms) when the problems were easy to make a correct response. But when the problems were hard, participants took 10492 ms (SD = 6185 ms) to respond – also as expected. This response time difference was also significant at the 0.05 level ($F(1, 89) = 6.532$, $p = 0.012$). Participants were able to make a mean of 14.04 correct responses (SD = 5.45) of the 34 problems presented. Of the 17 easy problems, participants made a mean of 8.64 correct responses (SD = 3.21). Of the 17 hard problems, participants were significantly less accurate, making a mean of 5.40 correct responses (SD = 2.82), a difference which was significant at the 0.05 level ($F(1, 90) = 24.482$, $p < 0.001$).

Maier (1970) provides desired responses for the Functional Fixedness (FF) problems, but other workable solutions exist. For instance, in the Two-String Problem, the given solution is to use a wrench as a pendulum weight to swing a rope hanging from a ceiling, but swinging other given objects as a pendulum can be considered workable solutions, too. To avoid this problem, independent raters reviewed each solution and determined whether given solutions were "blank (left unanswered)", "incorrect unworkable", "correct workable", or "correct according to Duncker and Maier".

In order to calculate a Normalized Score for a participant's set of responses on the FF, unanswered questions were awarded 0 points, incorrect unworkable responses were awarded 1 point, correct workable responses awarded 2 points, and responses correct according to Duncker and Meier were awarded 3 points. Because participants reported recognizing some problems, scores for recognized Functional Fixedness problems were excluded from further analysis. For each participant, points were summed and normalized (divided by the number of problems

responded without prior experience) to produce the final Normalized Score. Additional scores were calculated by count instead of points: an "Insight Rate" was calculated by counting the number of responses scoring at least 2, and a "correct insight rate" was calculated by counting the number of responses scoring 3. Each participant response to an FF problem was scored by at least two independent scorers, and no set of participant scores varied by more than 2 standard deviations from the mean.

Five participants in Study #1 reported having seen all seven of the presented Functional Fixedness (FF) problems before and are thus excluded from analysis because their responses reflect recall, not insight. An additional five participants reported that they had seen at least one of the problems before; the responses to those previously-seen problems are excluded from analysis as described in Methods. Performance scores were scaled for these participants to allow comparisons with other participants who completed all problems naively.

For each of the word problem stimuli, participants were asked to read the problem prompt on-screen and press the spacebar when they had a solution in mind and were ready to begin writing. The mean time to response was 40504ms (SD = 29600ms). The mean normalized score was 1.979 (SD = 0.128) with a maximum possible score of 3. Participants were able to achieve insight (score a 2 or higher) on average 67.8% of the problems (SD = 19.4%), and identify the correct answer on average, as defined by Duncker and Maier, for 40.7% of the problems (SD = 19.4%).

R-SPAN in Study #1 showed a mean Score of 38.26 (SD = 18.84; Figure 2), with no main effect for Age (median split; $F(1, 38) = 4.054$, $p = 0.051$), no main effect for Gender (male, female, other; $F(2, 38) = 1.98$, $p = 0.101$), and no interaction ($F(1, 38) = 1.958$, $p = 0.169$).

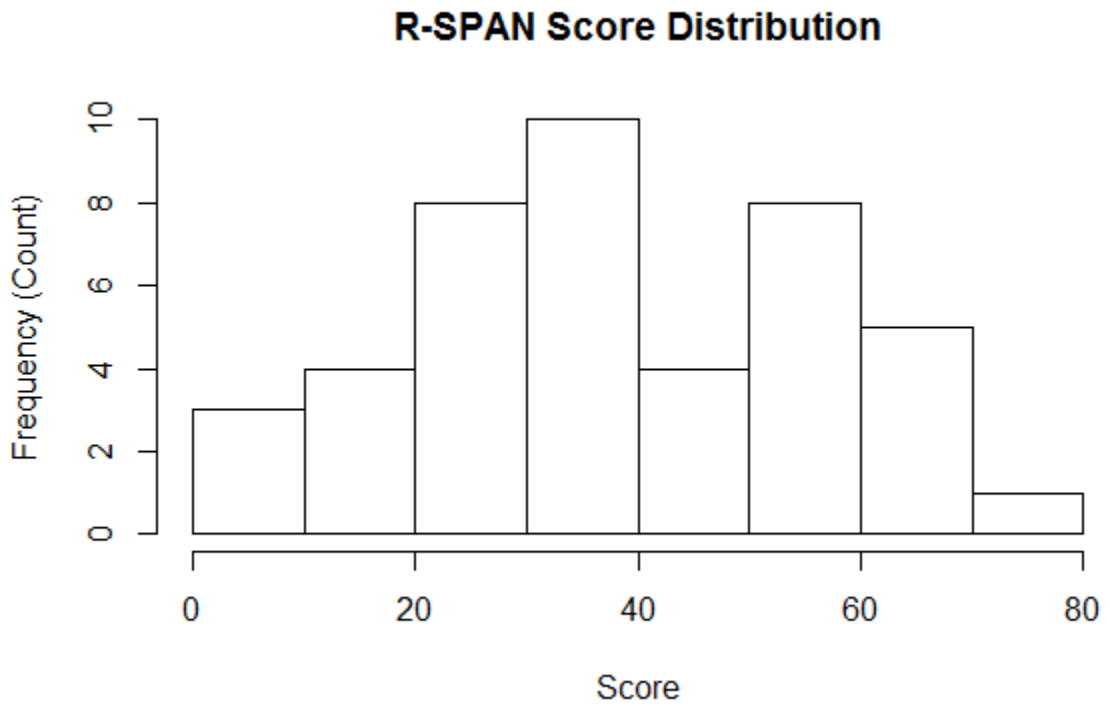

Figure 2: Distribution of R-SPAN Scores in Study #1

To further determine the relationship of Age and Gender to performance in the Study #1 insight tasks, median-split Age (Median = 20 years) was used along with gender as factors in an additional set of two-way ANOVAs. There was no interaction of Age and Gender on the number of trials to insight on the NRT ($F(1, 18) = 0.479$, $p = 0.498$, no main effect of Age (median split; $F(1, 18) = 0.031$, $p = 0.861$), nor Gender ($F(2, 18) = 0.331$, $p = 0.722$). There was no interaction between Age and Gender factors on (correct-response) insight rate on the Functional Fixedness

problems ($F(1, 42) = 0.536$, $p = 0.468$), no main effect of Age ($1, 42) = 0.401$, $p = 0.530$), no main effect of Gender (non-binary categories; $F(3, 42) = 1.425$, $p = 0.249$). A Bonferroni multiple tests correction did not affect these results. Based on these results, Age and Gender are not included as factors in further analyses. The amount of sleep as taken from sleep logs kept by participants was not found to correlate with any insight performance measures used in a later principal components analysis (See Table 3, below).

A matrix of Spearman's rank-order correlations suggests some relationships between tasks. In Table 3, correlations above the 0.05 significance level are marked in boldface font (values are not corrected for multiple tests). For instance, Number Reduction Task performance is related to performance on the Remote Associates Task. When Spearman's p-values were corrected for multiple comparisons using the Bonferroni method, none were significant at the 0.05 level, however the uncorrected trends are suggestive of relationships; these trends allow the pursuit of a Principal Components Analysis.

	Age	Hours Slept	NRT Trials To Insight	NRT Time To Insight	NRT Mean Trial Time	RAT Acc Easy Prob.	RAT Acc Hard Prob.	RAT Acc All Prob.	RAT RT Easy Prob Correct	RAT RT Hard Prob. Correct	RAT RT All Correct	RAT RT Easy Prob.	RAT RT Hard Prob.	RAT RT All Prob.	R-SPAN Score	R-SPAN Total	FF Insight Rt.	FF Insight Rt. Correct	FF Norm. Score
Hours Slept Mean	0.088																		
NRT Trials to Insight	0.071	-0.018																	
NRT Time To Insight	0.071	0.207	**0.815**																
NRT Mean Trial Time	-0.178	-0.205	-0.724	0.116															
RAT, Accuracy, Easy Problems	0.023	-0.033	-0.092	-0.039	0.097														
RAT, Accuracy, Hard Problems	0.02	-0.051	-0.059	-0.062	0.157	**0.612**													
RAT, Accuracy, All Problems	0.013	-0.054	-0.155	-0.054	0.148	**0.909**	**0.876**												
RAT, RT, Easy Problems Correct Response	-0.001	-0.238	-0.38	**-0.211**	0.218	-0.366	-0.331	-0.409											
RAT, RT, Hard Problems Correct Responses	-0.018	-0.131	-0.269	**-0.354**	0.178	0.057	0.104	0.081	**0.43**										
RAT, RT, All Problems Correct Responses	-0.065	-0.189	-0.387	-0.293	0.224	-0.117	-0.065	-0.116	**0.821**	**0.827**									
RAT, RT, Easy Problems All Trials	-0.141	-0.043	-0.287	-0.188	**0.225**	**-0.099**	-0.108	-0.129	**0.672**	**0.448**	**0.695**								
RAT, RT Hard Problems All Trials	-0.173	-0.192	**-0.399**	-0.232	0.296	0.348	0.066	0.23	**0.366**	**0.617**	**0.667**	**0.673**							
RAT, RT All Trials	-0.15	-0.119	-0.451	-0.263	0.265	0.168	-0.02	0.076	0.533	0.606	**0.741**	0.867	**0.939**						
R-SPAN Score	-0.289	-0.01	0.028	-0.008	0.222	0.246	0.246	0.298	-0.11	0.02	0.013	-0.014	**0.243**	0.146					
R-SPAN Total	-0.212	0.045	0.063	-0.016	0.16	**0.27**	**0.191**	**0.28**	-0.034	0.011	0.071	0.106	**0.305**	**0.233**	**0.935**				
FF Insight Rate	**-0.176**	-0.089	**-0.278**	-0.173	0.228	**0.362**	0.336	**0.384**	-0.038	0.181	0.149	0.206	**0.372**	**0.322**	**0.39**	**0.514**			
FF Insight Rate Correct Responses	-0.299	0.054	-0.443	-0.146	0.271	**0.34**	0.244	**0.33**	-0.028	0.183	0.195	0.157	**0.418**	**0.333**	**0.372**	**0.437**	**0.786**		
FF Normalized Score	-0.165	0.026	-0.335	-0.194	0.181	**0.409**	0.267	**0.384**	-0.014	0.191	0.188	0.14	**0.394**	**0.304**	**0.402**	**0.526**	**0.909**	**0.889**	
FF, RT Mean	0.011	-0.145	-0.195	-0.244	0.041	0.279	0.204	0.25	0.093	0.292	0.316	0.275	0.483	0.463	**0.173**	**0.192**	**0.338**	**0.346**	0.33

Table 3: Spearman's rank-order correlation coefficients in Study #1. Boldface values in matrix are significant at the 0.05 level.

Based on these trending initial results, an omnibus Principal Components Analysis (PCA) was calculated based upon on the 42 participants who completed all tasks (Table 4). This omnibus PCA contains multiple measures (factors) per test in an effort to statistically identify which might be related across tasks. Further analysis based on this omnibus are planned around the outcomes of the current analysis.

The omnibus analysis identified three principal components that accounted for a combined 77.3% of variance, with the remaining components each accounting for no more than 10% of variance.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
Number Reduction Task # Trials To Insight	0.24	0.264	0.064	0.405	-0.342	-0.049	-0.239	-0.439	-0.435	0.213	-0.141	0.079	-0.272	0	0
Number Reduction Task Time To Insight	0.091	0.342	-0.222	0.481	-0.288	0.246	-0.177	0.457	0.19	-0.337	-0.024	-0.07	0.236	0	0
Number Reduction Task Mean Time Per Trial	-0.127	-0.419	-0.173	-0.16	0.084	0.315	-0.694	0.131	-0.37	-0.055	-0.085	0	0.031	0	0
Remote Associates Task Accuracy (easy problems)	-0.261	0.361	-0.274	-0.047	0.222	-0.069	-0.12	0.035	0.038	0.381	-0.174	-0.506	-0.052	-0.467	0
Remote Associates Task Accuracy (hard problems)	-0.231	0.381	-0.018	-0.321	-0.06	0.389	0.141	-0.261	-0.204	-0.325	0.079	0.363	0.122	-0.393	0
Remote Associates Task Accuracy (all problems)	-0.269	0.402	-0.17	-0.187	0.101	0.153	-0.001	-0.109	-0.079	0.064	-0.063	-0.118	0.03	0.792	0
Remote Associates Task RT (easy problems)	-0.263	-0.276	-0.017	-0.034	-0.48	0.179	0.394	0.086	-0.16	-0.036	-0.372	-0.27	-0.127	-0.417	0
Remote Associates Task RT (hard problems)	-0.345	0.005	-0.117	-0.045	-0.342	-0.235	-0.325	-0.143	0.345	0.125	0.44	0.171	-0.038	-0.456	0
Remote Associates Task RT (all problems)	-0.339	-0.144	-0.077	-0.044	-0.453	-0.041	0.021	-0.037	0.116	0.054	0.058	-0.044	-0.089	0.786	0
Functional Fixedness RT	-0.273	0.179	0.2	-0.089	-0.02	-0.68	-0.131	0.272	-0.285	-0.309	-0.296	0.142	0.079	0	0
R-SPAN Score	-0.12	0.117	0.594	-0.047	0.008	0.279	-0.25	0.064	0.443	0.167	-0.433	0.183	-0.167	0	0
R-SPAN Total	-0.131	0.087	0.625	0.05	-0.045	0.122	-0.034	0.118	-0.275	-0.01	0.513	-0.448	0.101	0	0
Functional Fixedness Insight Rate (score<=2)	-0.343	-0.058	-0.01	0.365	0.142	0.096	0.22	0.29	-0.241	0.52	0.068	0.451	0.216	0	0
Functional Fixedness Correct Insight Rate (score=3)	-0.298	-0.212	0.077	0.4	0.188	-0.035	-0.035	-0.543	0.136	-0.211	-0.189	-0.154	0.495	0	0
Functional Fixedness Normalized Score	-0.329	-0.044	-0.055	0.355	0.346	0.06	0.063	0.013	0.015	-0.353	0.135	0.025	-0.699	0	0
Standard Deviations	2.44	1.622	1.459	1.192	1.083	0.845	0.711	0.433	0.345	0.259	0.221	0.164	0.152	0	0
Eigenvalues	5.955	2.63	2.128	1.422	1.172	0.715	0.505	0.188	0.119	0.067	0.049	0.027	0.023	0	0
Proportion of Variance	0.397	0.175	0.142	0.095	0.078	0.048	0.034	0.013	0.008	0.004	0.003	0.002	0.002	0	0
Cumulative Proportion	0.397	0.572	0.714	0.809	0.887	0.935	0.969	0.981	0.989	0.993	0.997	0.999	1	1	1

Table 4: Principal Component loadings for task measures in Study #1

The first principal component appears to represent aspects of the Number Reduction Task, and identifying the Dunker-Maier Functional Fixedness problems at the opposite end of the axis. The second component represents relationship between the Number Reduction Task and the Remote Associates Task. The third component separates out the R-SPAN. The fourth component appears to represent a relationship between the Number Reduction Task and the Functional Fixedness problem, but separates some measures of the Remote Associates Task. These first four components represent over 80% of variance, with the fourth component representing just under 10% of variance (Figure 3).

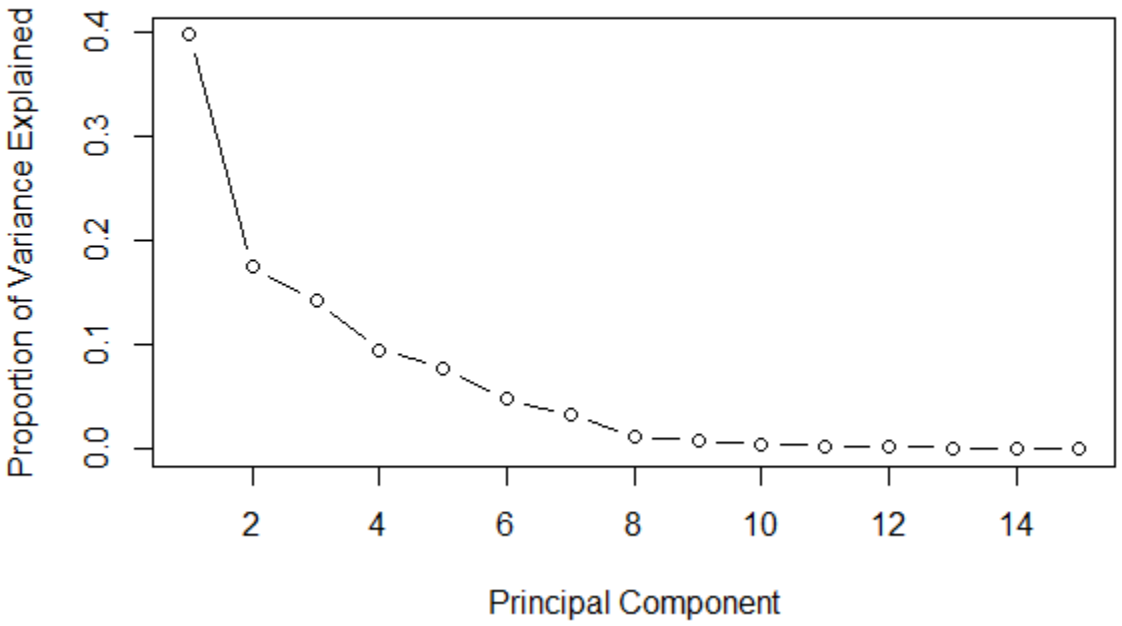

Figure 3: Scree plot (Study #1). The first four principal components represent over 80% of variance.

The question remains whether certain factors, when subset from the omnibus analysis, will show the same relationship between insight tasks. That is, the trends in the omnibus and the

correlation matrix point to which factors may be selected for further principal components analysis. In this analysis, eight factors were selected for inclusion in order for each factor to represent a different aspect of each task (Table 5; see Figure 4 for variable chart). The number of trials to insight in the NRT was selected based on its use in Wagner et al. (2004).

However, the metric does not measure the amount of time the participant has had to think; number of trials only represents the number of exposures to the stimuli. For this reason, time to insight was included, as was mean time per trial. For the Remote Associates Task, accuracy rate was included based on Bowden and Jung-Beeman (2003), as the stimuli were selected based on normative accuracy rate data provided in the work and not on response time measures. The insight rate and response time measures are included for the CPF in order to mirror the selected factors for the NRT. Additionally, R-SPAN Score and Total Score are included based on Conway et al. (2005).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
NRT # Trials To Insight	0.349	-0.514	0.032	-0.159	0.101	-0.378	0.633	-0.187
NRT Time To Insight	0.289	-0.39	-0.429	-0.43	-0.122	-0.139	-0.576	0.17
NRT Mean Time Per Trial	-0.096	0.554	0.09	-0.478	-0.279	-0.595	0.092	0.08
RAT Accuracy (all problems)	-0.291	-0.076	-0.609	0.2	-0.605	0.073	0.346	0.086
R-SPAN Score	-0.411	-0.332	0.356	-0.197	-0.383	0.036	-0.206	-0.604
R-SPAN Total	-0.423	-0.363	0.35	-0.191	0.005	0.048	0.113	0.718
FF Insight Rate (score<=2)	-0.383	0.086	-0.377	-0.541	0.496	0.292	0.201	-0.197
FF Time To Response	-0.453	-0.146	-0.209	0.392	0.371	-0.624	-0.211	-0.063
Standard Deviations	1.664	1.428	1.2	0.865	0.689	0.597	0.34	0.236
Eigenvalues	2.769	2.041	1.439	0.748	0.475	0.356	0.116	0.056
Proportion of Variance	0.346	0.255	0.18	0.094	0.059	0.045	0.014	0.007
Cumulative Proportion	0.346	0.601	0.781	0.875	0.934	0.979	0.993	1.000

Table 5: Study #1 Principal Component loadings based on a subset of eight representative factors identified in the omnibus test

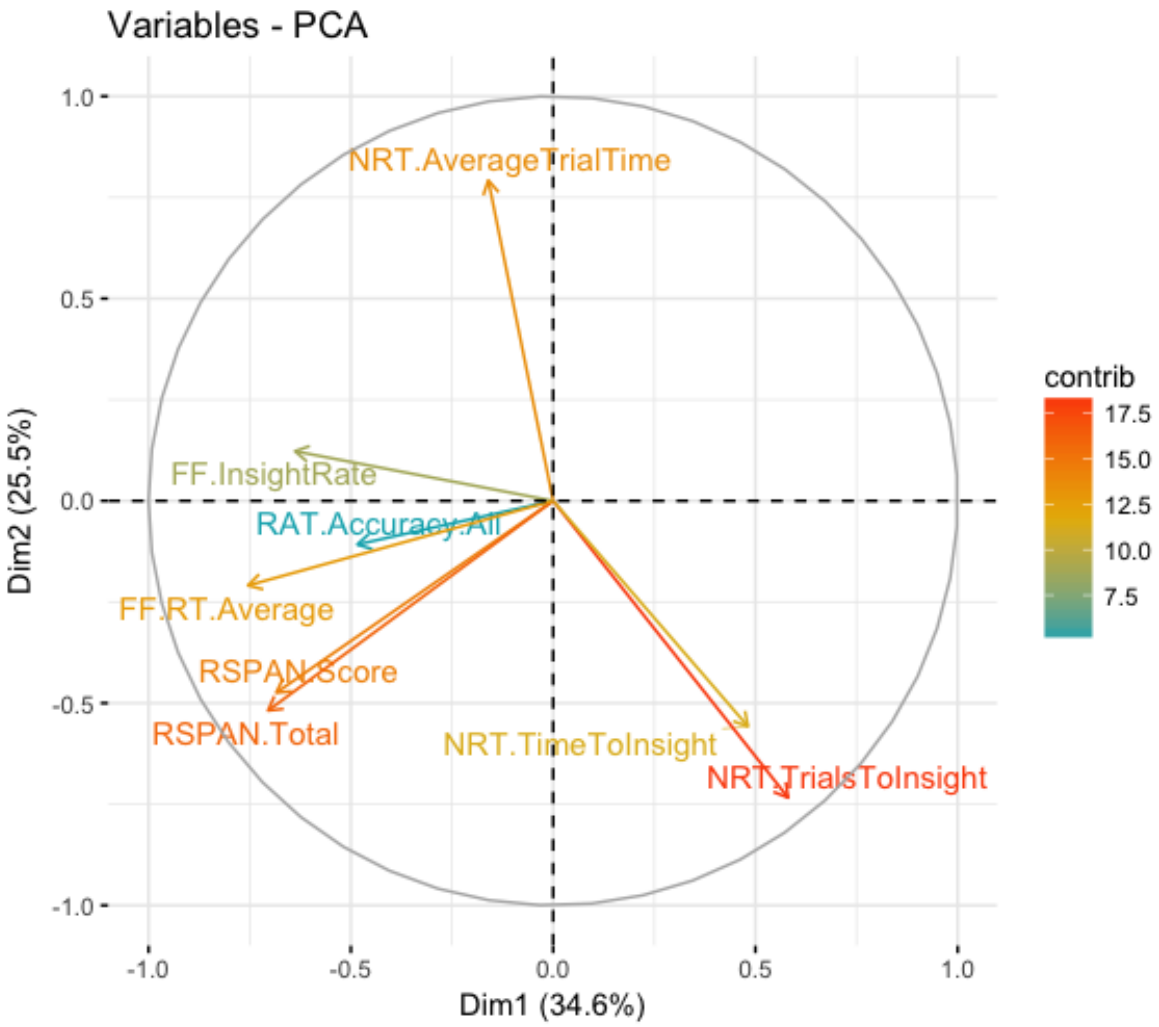

Figure 4: Correlation Circle (Variable Chart) for principal components identified in Table 5.

The first principal component demonstrates the same performance relationship between Number Reduction Task and the seven Functional Fixedness problems, as identified in the omnibus PCA. The second component also appears to represent the Number Reduction Task, like in the omnibus, but here demonstrates a relationship between the Number Reduction Task and the R-SPAN working memory capacity task. The third component appears to strongly represent the Remote Associates Task separately from the others, which can be taken as evidence that it is not closely related with the Number Reduction Task or with working memory capacity.

These three principal components account for a cumulative 78.1% of variance, with each component accounting for no less than 10% of variance, which suggests that the remaining principal components are of lesser statistical importance. However, given that it is unclear how these results apply to the general population (as participants in this study were drawn from the University of Chicago), these results are compared to a participant pool drawn from online in the next study.

Given that Study #1 demonstrated some similarities among tasks albeit drawn from a participant pool that may not reflect the general population, Study #2 recruited from a larger participant population in order to better determine how these tasks grouped together. While the results suggest that insight might not be understood as a single process but instead may be a collection of different properties, replicating Study #1 with a larger and more representative participant pool would provide stronger evidence for the argument presented as Specific Aim #1. Study #2 is not a direct replication of Study #1. The R-SPAN working memory capacity measure was replaced with the Camouflage Perceptual Fluency task in Study #2.

Across participants, Number Reduction Task sequences in Study #2 were accurate (as judged by the final response of the sequence) on average 55.6% the time ($M = 0.556$, $SD = 0.154$; Figure 5). As discussed, accuracy rate is based on a judgment that the participant is applying a hidden rule consistently and correctly, which is different than the 50% odds if accuracy rate were measured based on a two-alternative-forced-choice response.

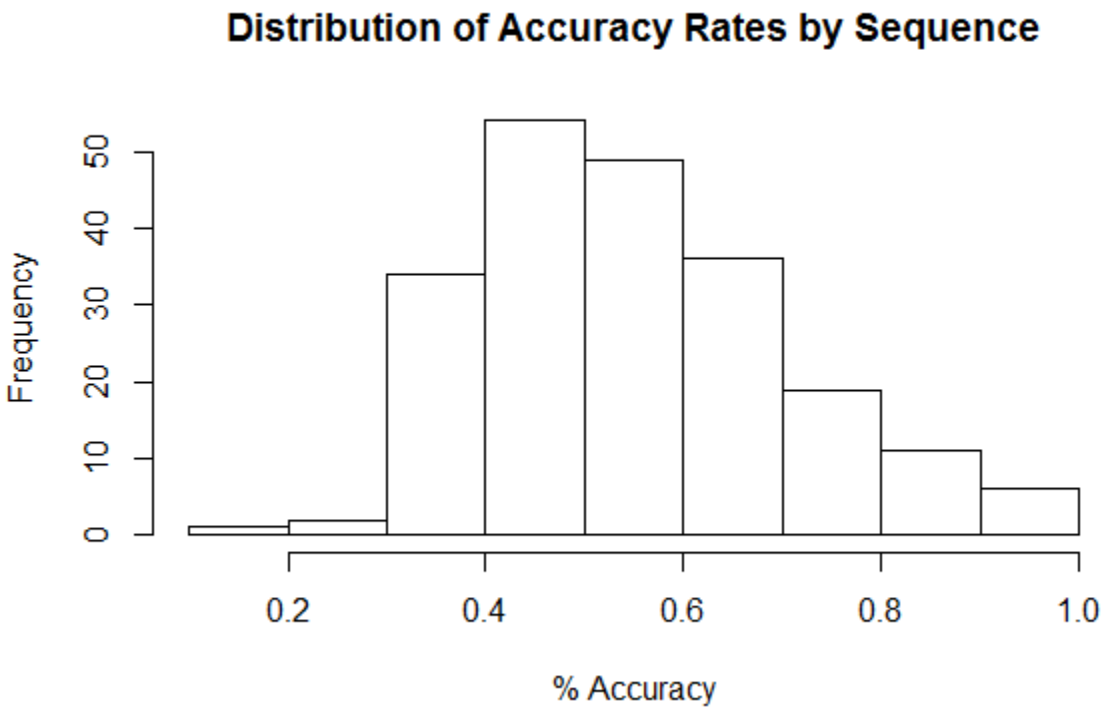

Figure 5: Frequency of accuracy rates by Number Reduction Task sequence in Study #2.

Given that there are only three possible responses for a final sequence response, random guessing would yield an accuracy rate of 33.3%. Yet overall, mean accuracy was 55.5% (M = 0.555, SD = 0.296). The bimodal pattern in Figure 6 (below) suggests that while some participants are completing the task, some are making random responses.

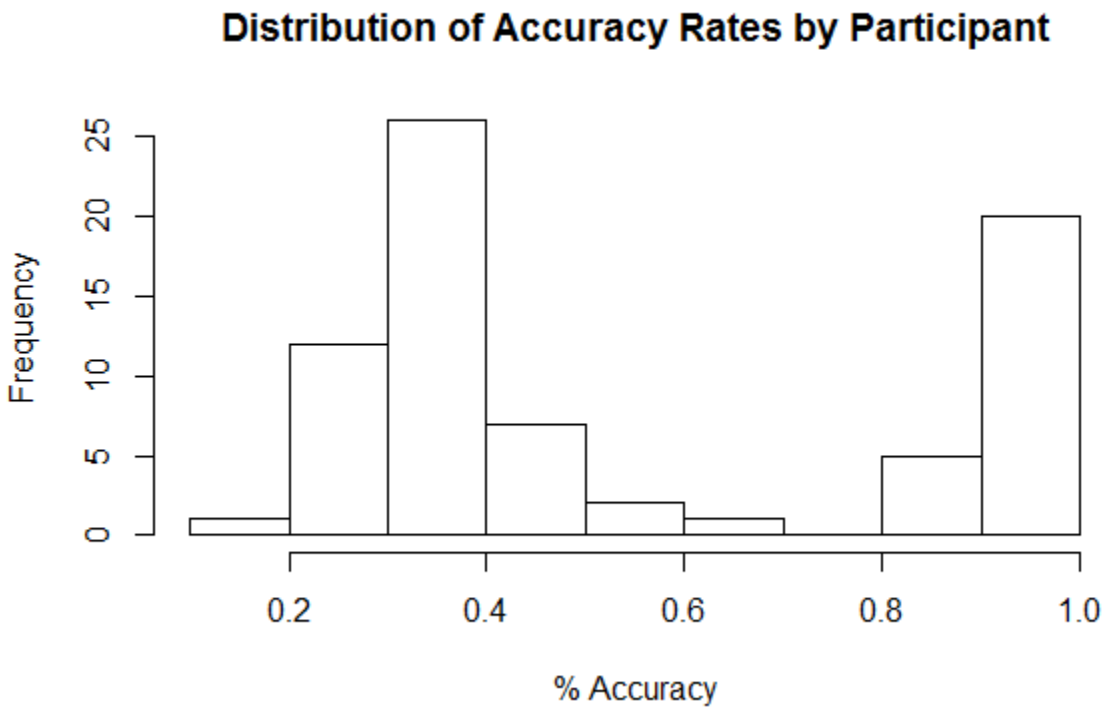

Figure 6: Frequency of accuracy rates by Number Reduction Task participant in Study #2.

Accurate shortcut use on the Number Reduction Task was low among MTurk participants in Study #2 (rate-of-use M = 0.128, SD = 0.061), as expected. The accurate insight rate was higher in Study #1 with University of Chicago participants, which reflects either a difference in participant pools or experimenter demand characteristics. Participants in Study #1 completed the task in a laboratory setting, but participants in Study #2 completed the task online where the setting could not be directly controlled by experimenters. A Pearson's product-moment correlation found a significant relationship between shortcut use and accuracy, also as expected ($r = 0.158$, $t(210) = 2.313$, $p = 0.022$).

Age and Gender factors were also examined on Number Reduction Task performance. When the participant sample was median-split by age, the interaction between Age and Gender was not significant on accurate shortcut use (insight rate; $F(1, 71) = 0.127$, $p = 0.7230$). The main effect of Age ($F(1, 71) = 0.0$, $p = 0.9972$) was not significant. Gender was significant ($F(1, 71) = 4.181$, $p = 0.0446$), but was not significant when a Bonferroni correction was applied ($n = 3$, $p = 0.1338$). Based on these results, Age and Gender factors are excluded from further analysis of the Number Reduction Task.

In the Remote Associates Task, participants in Study #2 solved the Easy problems at a mean rate of 47.6% ($M = 0.467$, $SD = 0.119$) and Hard problems at a mean rate of 34.2% ($M = 0.342$, $SD = 0.092$), as shown in Figure 7, below. A Welch's t-test found this difference to be significant ($t(139.88) = 4.243$, $p < 0.001$).

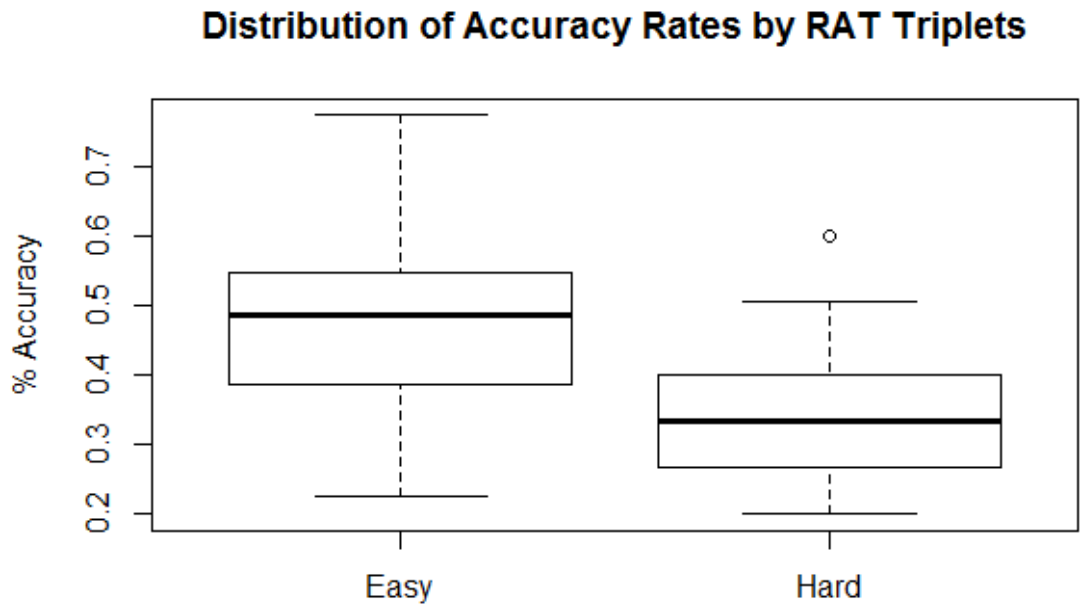

Figure 7: Accuracy rates by RAT problem

In testing for the effects of Age and Gender factors on RAT accuracy rates, the main effect of Gender was not significant ($F(1, 146) = 1.115$, $p = 0.2927$), but the main effect of Age was significant ($F(1, 146) = 7.339$, $p = 0.0076$. Older participants scored higher accuracy rates ($M = 0.4587$, $SD = 0.1749$) than younger participants ($M = 0.3681$, $SD = 0.2178$). Note that the median age in Study #2 is 34 years, but only 20 years for Study #1. Older participants are not well-represented in Study #1, where participants were drawn from the University of Chicago campus.

For the Functional Fixedness Problems in Study #2, mean score was 0.704 ($SD = 0.353$). This score, as discussed, is not a pure accuracy rate; it is a normalized score that accounts for reasonableness and feasibility. A random-effects ANOVA did not identify a significant difference for Problem by participant ($F(6, 485) = 1.02$, $p = 0.411$), which suggests that no particular problem appears to be more or less difficult than another (Table 6).

Problem	Mean	SD
Ten-Coin	0.592	0.426
Candle	0.781	0.289
Fishbowl	0.806	0.349
$100 Pyramid	0.511	0.385
Old Coin	0.744	0.331
Prison Rope	0.790	0.358
Two-String	0.741	0.341

Table 6: Mean accuracy rates for the seven Functional Fixedness problems in Study #2.

Like participants in Study #1, scores in Study #2 did not appear to differ in Age or Gender. A main effect of Age was not significant ($F(1, 67) = 0.235$, $p = 0.630$), nor was a main effect of Gender ($F(1, 67) = 0.603$, $p = 0.440$). The interaction of Age and Gender was also not significant ($F(1, 67) = 0.954$, $p = 0.332$). Subsequent analysis involving the Functional Fixedness problems exclude Age and Gender as factors.

Response accuracy in the Camouflage Perceptual Fluency task was highest for Medium Valence images (i.e., images that were the most emotionally neutral; Table 7). A one-way ANOVA found a significant difference in accuracy rates for the three Valence groups ($F(2, 224) = 71.202$, $p < 0.001$).

Valence	Mean	SD
High	0.213	0.109
Medium	0.463	0.161
Low	0.254	0.138

Table 7: Response accuracy rates for High, Medium, and Low Valence categories in the Camouflage Perceptual Fluency task.

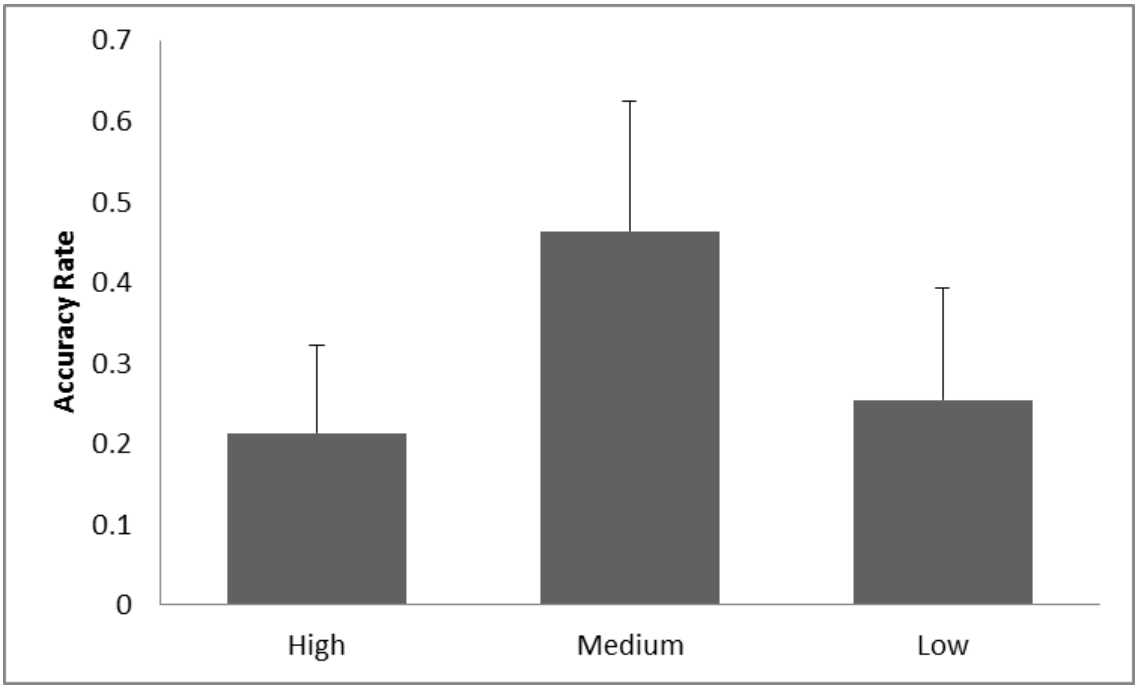

Figure 8: Accuracy rates by image Valence in the Camouflage Perceptual Fluency task.

When images were divided by IAPS Arousal into High, Medium, and Low categories, response accuracy was highest for Low Arousal images (Table 8). A one-way ANOVA found this relationship to be significant ($F(2, 224) = 66.319$, $p < 0.001$; Figure 9).

Arousal	Mean	SD
High	0.187	0.122
Medium	0.298	0.151
Low	0.445	0.139

Table 8: Response accuracy rates for High, Medium, and Low Arousal image sets in the Camouflage Perceptual Fluency task

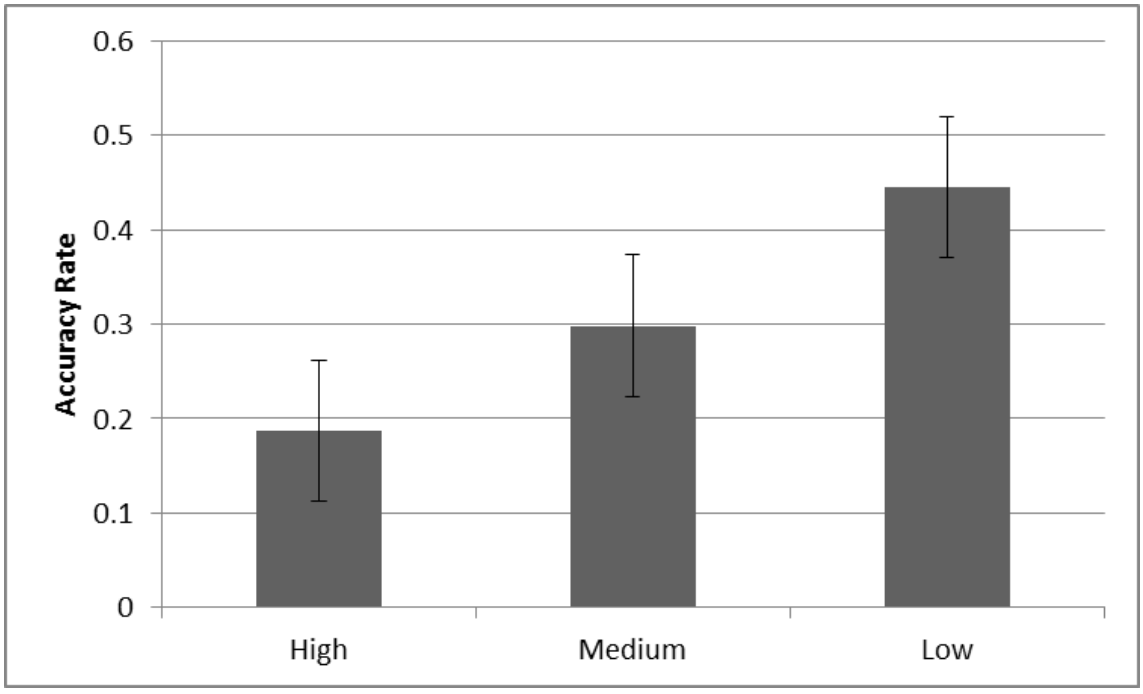

Figure 9: Accuracy rates by image arousal in the Camouflage Perceptual Fluency task.

In assessing the effect of Age and Gender on correct response rate in the Camouflage Perceptual Fluency task, the main effect of Gender was not significant ($F(1, 71) = 0.649$, $p = 0.423$), nor was the main effect of Age (when median split; $F(1, 71) = 0.622$, $p = 0.433$), nor the interaction of Age and Gender on correct response rate ($F(1, 71) = 0.247$, $p = 0.621$). Subsequent analysis involving the Camouflage Perceptual Fluency task in this study thus exclude Age and Gender as factors.

A Spearman's correlation matrix suggests that a set of relationships does exist between each of the tasks (Table 9). The accuracy rate for the Number Reduction Task suggests a positive relationship with multiple Camouflage Perceptual Fluency measures, particularly with strong (high Arousal) and negative (low Valence) items. These results provide a reason for further PCA,

narrowing down the four measures for CPF, two for NRT and RAT. In the next PCA, the measures that appear most characteristic per task are analyzed.

	Age	DM Accuracy Rate	NRT Accuracy Rate	NRT Shortcut Use Rate	RAT Accuracy Rate Hard Problems	RAT Accuracy Rate Easy Problems	CPF Accuracy Rate	CPF Accuracy Rate Low Arousal Images	CPF Accuracy Rate High Arousal Images	CPF Accuracy Rate Low Valence Images
DM Accuracy Rate	0.19									
NRT Accuracy Rate	-0.03	0.12								
NRT Shortcut Use Rate	**0.2**	**-0.05**	-0.3							
RAT Accuracy Rate Hard Problems	0.31	**0.35**	0.13	0.03						
RAT Accuracy Rate Easy Problems	0.21	0.35	0.15	-0.08	**0.89**					
CPF Accuracy Rate	0.08	**0.15**	0.08	-0.23	0.33	0.32				
CPF Accuracy Rate, Low Arousal Images	0.03	**0.33**	0.11	-0.07	0.12	0.09	0.2			
CPF Accuracy Rate, High Arousal Images	0.05	0.25	0.01	-0.15	0.12	0.08	0.19	**0.66**		
CPF Accuracy Rate Low Valence Images	**-0.01**	**0.33**	**0.05**	**-0.16**	**0.09**	**0.07**	**0.26**	**0.9**	**0.78**	
CPF Accuracy Rate High Valence Images	0.11	**0.21**	0.04	-0.06	0.18	0.1	**0.13**	**0.69**	**0.88**	**0.62**

Table 9: Matrix of Spearman's rank order correlation coefficients. Boldface coefficients are significant at the 0.05 level.

In order to further explore these relationships, data were entered into a PCA similar to that in Experiment 1 (Table 10). The PCA identified three main components which together accounted for 72.4% of variance, with each of the remaining components accounting for no more than 10%.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
FF (Accuracy)	**-0.27**	0.033	-0.197	-0.547	**0.646**	0.282	-0.287	-0.084	0.043	0	0
NRT (Accuracy)	-0.129	0.082	**0.626**	-0.389	0.006	-0.618	-0.163	0.099	0.123	0	0
NRT (Shortcut Use)	0.107	-0.042	**-0.693**	-0.002	0.111	-0.671	-0.089	0.191	0.015	0	0
RAT (Accuracy)	**-0.342**	**0.437**	-0.113	0.004	-0.165	0.003	0.075	0.014	0.047	-0.802	0
RAT (Accuracy, Hard Problems)	**-0.314**	**0.427**	-0.169	0.034	-0.23	-0.005	0.035	-0.255	0.59	0.471	0
RAT (Accuracy, Easy Problems)	**-0.344**	**0.409**	-0.031	-0.035	-0.066	0.012	0.12	0.359	-0.656	0.366	0
CPF (Accuracy)	**-0.234**	0.105	0.193	**0.727**	**0.543**	-0.161	-0.193	-0.104	-0.01	0	0
CPF (Accuracy, Low Arousal Images)	**-0.362**	-0.324	-0.033	-0.073	0.05	-0.188	0.496	-0.437	-0.164	0	0.505
CPF (Accuracy, High Arousal Images)	**-0.355**	-0.341	-0.025	0.115	-0.23	0.148	-0.393	0.488	0.19	0	0.49
CPF(Accuracy, Low Valence Images)	**-0.358**	-0.348	0.017	0.023	0.131	0.019	0.48	0.381	0.253	0	-0.541
CPF(Accuracy, High Valence Images)	**-0.355**	-0.311	-0.084	0.016	-0.344	-0.071	-0.439	-0.408	-0.275	0	-0.46
Standard Deviations	2.013	1.595	1.169	0.923	0.847	0.779	0.74	0.422	0.366	0	0
Eigenvalues	4.053	2.543	1.367	0.853	0.717	0.607	0.548	0.178	0.134	0	0
Proportion of Variance	0.368	0.231	0.124	0.078	0.065	0.055	0.05	0.016	0.012	0	0

Table 10: Clusters of factors identified in first four components of a Principal Components Analysis. Values are component loadings.

The first component, which accounts for 36.9% of variance, suggests a cluster of tasks that are related to each other: the seven Functional Fixedness questions, the Remote Associates Task, and the Camouflage Perceptual Fluency task (Table 9 & Figure 10). The Number Reduction Task appears to be represented in the third principal component, which represents 12.4% of variance.

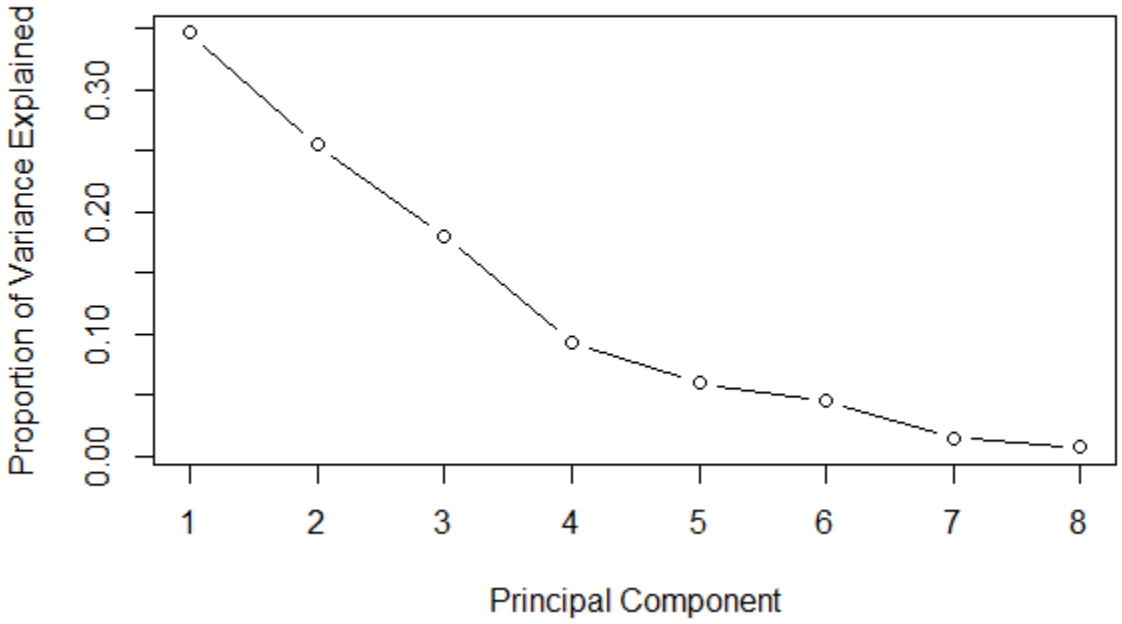

Figure 10: Scree plot; proportion of variance explained by 8 principal components. The first three components account for 78.1% of variance.

Interestingly, the first principal component did not identify particular measures within a task as being particularly important. The only task not represented in the first component, the Number Reduction Task, is represented in the third component. This third principal component also suggests divergence between accuracy rates and use of the shortcut in the Number Reduction Task. The second component appears to represent aspects of the Remote Associates Task, which is also represented in the first component.

Given these results, an abbreviated PCA was completed where each task was represented by one measure. These measures were selected from the previous PCA based on how well they appeared to the experimenter to represent the results of the insight task. To verify which factors

were included, separate PCAs were calculated for different factors within each of the following tasks: the Duncker-Maier Functional Fixedness problems, the Number Reduction Task, the Remote Associates Task, and the Camouflage Perceptual Fluency task. Each of these PCAs identified one primary factor for each task that comprised the majority of variance, represented by the first component, and agreed with the experimenter's selection based on the narrowed omnibus test (above). A PCA based on only these four primary factors (i.e., experimental measures) gathered from the online participants of Study #2 identified four principal components (Table 11).

	PC 1	PC 2	PC 3	PC 4
FF Accuracy	**-0.488**	0.537	-0.494	-0.478
NRT Accuracy	-0.340	**-0.802**	-0.489	-0.048
RAT Accuracy	**-0.627**	0.175	0.072	**0.758**
CPF Accuracy	**-0.507**	-0.195	**0.715**	-0.440
Eigenvalue	1.640	0.944	0.844	0.571
% Total Variance	41.01%	23.61%	21.11%	14.28%

Table 11: Principal Component loadings for accuracy performance in the Functional Fixedness, Number Reduction, Remote Associates, and Camouflage Perceptual Fluency tasks. Boldface values are items of interest.

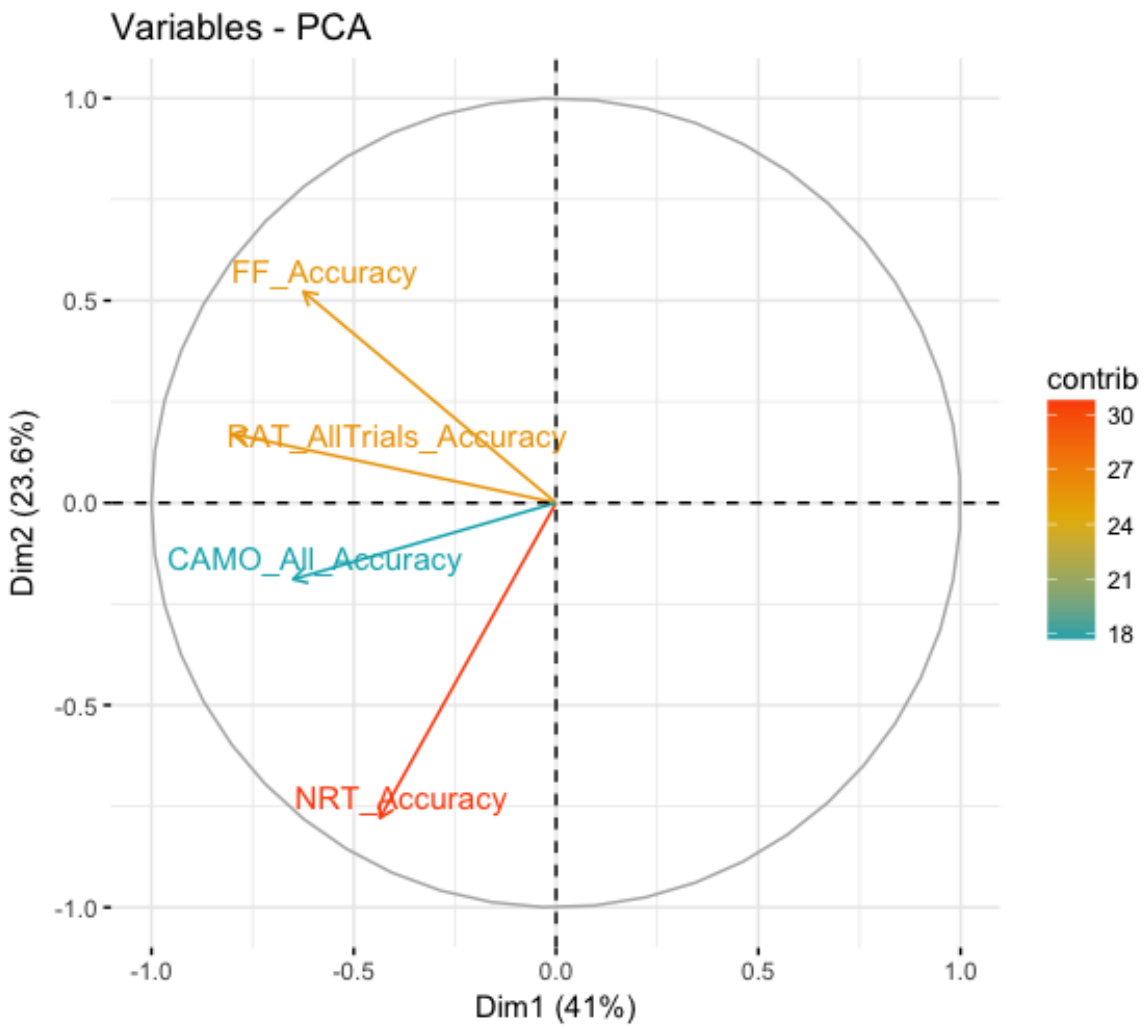

Figure 11: Correlation Circle (Variable Chart) for Principal Components identified in Table 11.

The first component of this PCA accounts for 41.01% of variance (Figure 12) and the second component accounts for an additional 23.61% of variance. When taken together, the first two components suggest a relationship among the Functional Fixedness problems, the Remote Associates Task, and the Camouflage Perceptual Fluency task. Performance in the Number Reduction Task is largely represented in the second component, suggesting performance on the task is not as closely related as performance on the other three insight tasks. The Camouflage Perceptual Fluency task is represented in the third component, suggesting an element of the task

that is unlike the others. This shows a somewhat similar pattern as found the previous analysis, in which the Number Reduction Task and the Functional Fixedness problems were clustered together. However, the PCA for Study #2 added the Camouflage Perceptual Fluency task that was not included in Study #1. Study #1 included the R-SPAN, which is not included in Study #2; The Principal Components Analysis for Study #1 identified a relationship (clustering) of the Number Reduction Task and the R-SPAN, which may suggest an underlying process of insight for these tasks that is related to working memory.

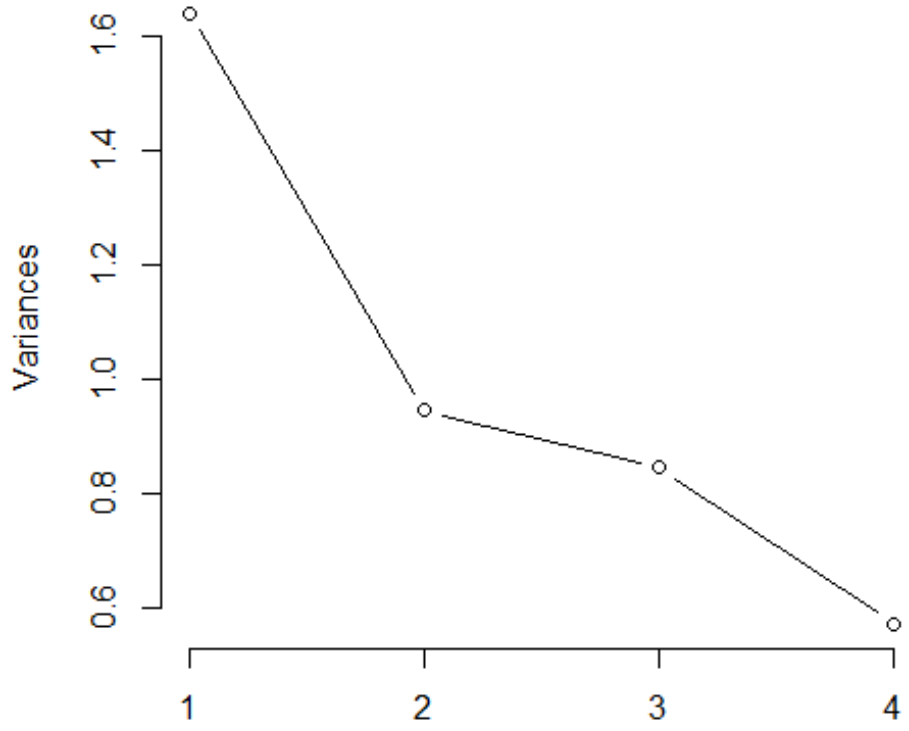

Figure 12: Scree plot for four task accuracy measures. The first component accounts for 41.01% of variance and represents performance in Functional Fixedness problems, Remote Associates Task, and Camouflage Perceptual Fluency.

Discussion

Correlation matrices suggested that performance on the Number Reduction Task and the Remote Associates Task are only slightly correlated, a pattern of results that suggests that perhaps different underlying mechanisms might be responsible for insight with respect to these tasks. At the least, the results suggest that performance on one task does not predict performance on all other tasks of insight (Specific Aim #1). Given that a correlation matrix can only provide a set of relationships but not group them into meaningful clusters, a follow-up Principal Components Analysis provides further evidence that what has been considered "insight" can be broken into two parts.

In one group, Functional Fixedness, Remote Associates Task, and Camouflage Perceptual Fluency task all are represented together, implying a strong shared variance between these three tasks. This cluster was identified in both Study #1 with the Functional Fixedness and Remote Associates tasks through a participant pool drawn from the University of Chicago campus, and in Study #2 with all three of the tasks thorough a participant pool drawn online from Amazon Mechanical Turk. The PCA also clustered the Number Reduction Task separately from the other three tasks. What makes the Functional Fixedness problems, the Remote Associates, and Camouflage Perceptual Fluency so closely related? On the flip side, what makes the Number Reduction Task so different?

One commonality among the tasks in the first group appears to be a relationship with memory retrieval processes. Solutions to the Functional Fixedness problems are based on the premise that a central object must be recognized from memory, and the use for that object must be ignored in favor of a less-common use. The Remote Associates Task requires participants to work retrieve memories that are related to the three given prompts and to recognize which word links all three prompts together. Finally, the Camouflage Perceptual Fluency requires participants to perceive not only that the black-and-white splotches form an image, but be able to recognize what object or scene the image represents. That is, participants must recognize a given stimulus image of "a dog" based on retrieving the correct object category from long term memory stores.

The grouping of the Camouflage Perceptual Fluency task with tasks like the Functional Fixedness Problems is suggestive of a perceptual form of insight (Specific Aim #2). The Camouflage Perceptual Fluency task (CPF) is further represented in multiple principal components. The task is represented in the first principal component likely because it requires long term memory retrieval, but it also appears in lower-ranked components. Perhaps what sets the task apart is in the stimuli: the CPF consists of pictures, unlike the other tasks of insight found in the first principal component. Successful completion of the CPF involves perceptual processes, in particular relying on visual perception, as the CPF consists solely of recognizing what the stimuli pictures represent.

The Number Reduction Task is unlike the other tasks in one important way. Successful insight in the Number Reduction Task is measured both by response accuracy and whether

participants identified the "secret shortcut" that is built into each response sequence. This shortcut, that the last three responses of a given problem sequence mirror the previous three responses, requires encoding and retrieving the pattern of responses across trials. So whereas the other insight tasks included in the battery require long term memory retrieval, solving the Number Reduction Task by way of finding the secret shortcut is heavily dependent on short-term memory encoding and retrieval.

The Remote Associates Task is also represented in lower-ranked principal components, even though the task is already represented in the first component. The Remote Associates Task, as already mentioned, depends heavily on retrieval from long term memory. But what about the task causes it to sort in a way differently than the other tasks found in the first component? Perhaps it is that the types of memories that are retrieved are different from the other tasks represented in the first principal component: the Remote Associates Task is heavily dependent on lexical knowledge (Razumnikova 2007, Gross, Toivonen, Toivanen, & Valitutti, 2012). As pointed out in Kounios and Beeman (2009), this task is language-based in that related items stored in memory must be searched so that this item relates to all three cues given in each trial.

Given these results, this set of insight tasks appear to be divided into two groups based on fluency of long term memory. It is particularly interesting to note that in Experiment #1, R-SPAN appears to be represented in a principal component separate from the other tasks, especially given that R-SPAN is a measure of working memory and not of long-term memory mechanisms. At a minimum, it appears that these tasks are not interchangeable measures of insight, and that insight itself may not be a single factor hypothesized under Specific Aim #1.

In this chapter, the results demonstrate that an insight task that appears to be heavily dependent on visuoperceptual processes is related to other tasks that have been used in the literature to study insight. Another aspect of insight not examined in Chapter Three is visuoperceptual sensitivity. The stimuli used in the Camouflage Perceptual Fluency task are all-or-none when it comes to recognition; I address the issue of perceptual sensitivity in the next chapter.

CHAPTER THREE: VISUOPERCEPTUAL PROCESSES OF INSIGHT

One aspect of perceptual insight not examined in Chapter Two is measuring individual differences in sensitivity. In Chapter Two a link between perceptual insight (as represented in the Principal Components Analysis with the Camouflage Perceptual Fluency task) and several other tasks of insight appearing in the literature (Duncker-Maier Functional Fixedness and Remote Associates Task) was identified. However, a shortfall exists in the literature regarding tasks of perceptual insight. In the Camouflage Perceptual Fluency task, the participants either recognize the contents of the image or not – a judgment that must be made by an independent observer. Some tasks, such as the letter perception task used by Gosselin and Schyns (2003) have idiosyncratic measures that do not allow quantitative comparisons among individuals. The studies described in this chapter demonstrate a sensitivity measure of perceptual insight based on Signal Detection Theory. This task and accompanying measure is first tested on a human participant pool before it is then validated theoretically using a computational classification model.

This chapter consists of material from a manuscript that was previously submitted for publication (Hu, Heald, Malonis, & Nusbaum, 2017) and asks whether insight can be found in perceptual behavior, and what mechanisms might overlap between perceptual and cognitive insight. The insight in this task, as discussed, is one of perception. Participants are asked to identify a target pattern in noisy stimuli, but are not in informed that the stimulus is noise created

with the help of a random number generator. Even so, participants are able to identify features in the noise that does correlate with the target pattern; a reverse correlation reveals that the noise stimuli that participants select do subjectively contain the target pattern. This noisy-picture protocol is similar to the protocol found in the Camouflage Perceptual Fluency task: participants look at a picture and make a judgment about the contents. However, the work presented in this chapter describes the insight behavior subjectively and allows for sensitivity measures that can be used in the type of across-subject comparisons that are missing from Gosselin and Schynns (2003). The studies described are included here because they are another example of perceptual insight and because the task and insights here are related to the Camouflage Perceptual Fluency task described in the previous chapter.

While the work also discusses how the experimental task protocol may be useful in understanding certain psychiatric disorders, this is not a specific aim of this dissertation. However, it is also included because it represents a future direction for research that may be derived from this work that should not be overlooked.

Introduction

When someone sees the image of Jesus in a piece of toast (Liu et al., 2014), does this indicate an overactive visual imagination (Glicksohn & Barrett, 2003), or does it reflect a high degree of sensitivity to expected signal information in noise? Currently the answer to this question leans in the direction of a good imagination (e.g., Bowers, 1979; Lynn & Rhue, 1986;

Parra, 2006). This in part depends on determining what counts as information for a perceiver. The present study tests directly whether there is real expected signal information present in samples that would otherwise be dismissed as noise.

From the perspective of Information Theory (Shannon, 1948), information is a signal that reduces uncertainty regarding which of a set of possible messages was transmitted. But in order for a message to be recognized as a signal, it must be discriminated from noise. We have clear perceptual experiences of what we consider noise and what we consider signal (e.g., the experience of listening to a static-filled radio station), but a formal definition of noise and signal as distinct categories is not at all clear.

Signal Detection Theory provides one framework in which noise and signal are separated: a signal increases sensation above and beyond a background noise level along some sensory dimension (Tanner & Swets, 1954). However, Signal Detection Theory does not formally define the sensory dimension; definitions of signal and noise categories are typically determined in the design of the experiment rather than by formal definitions.

In determining the boundary between what is noise and what is signal, it is useful to consider what noise is. Most definitions of noise are derived from a notion of chance or randomness, such as the "blooming, buzzing confusion" described by James (1893). This definition suggests that noise is a sensory experience of stimulation that does not directly correspond to any known (previously understood or categorized) signal pattern. This suggests that noise can be divided into at least two categories based on source: a psychological source

which is a failure to interpret stimulation regardless of its signal properties, and a physical source that corresponds to a random distribution of spectrally static properties.

Whereas physical sources of noise can be described through physical measurement separate from sensory experiences, such measurements reflect only a snapshot or sample that is aimed at an idealized random population. From Fourier's theorem, any particular sample of noise will have a specific yet randomly sampled spectro-temporal structure. As such, the distribution of spectral properties over time for a given noise sample could, in principle, overlap at least partially with the spectral properties for a given signal.

Gosselin and Schyns (2003) first showed that individuals without any specialized training can classify visual noise samples (generated as matrices of random numbers) when they are instructed to look for the presence or absence of a designated visual target pattern. Even though the target image was explicitly not synthesized into the noise or added to the noise in any way, observers were able to look for the specified target and detect the putative image. In other words, some samples from a large set of visual noise samples can be classified as if observers perceive a visual target in the noise although no target was placed there. Analysis of the samples that received a positive detection response suggested that the choices were not random and were not just due to some kind of bias. Statistical aggregation of the samples that were responded to with a positive detection revealed traces of the image of the designated target albeit different images for different observers. This has been interpreted as evidence that observers can form a stable, clear mental image of the designated target and can use this mental image as the basis for perceptual comparison to specific stimulus noise samples over the course of the experiment.

This means that this mental match-to-visual-noise-sample process can be carried out with some measurable level of reliability over tens of thousands of trials; although given that different observers pick different noise samples yielding different idiosyncratic images, there are clear individual differences.

This finding has now been replicated a number of times in different ways (Liu et al., 2014; Zhang et al., 2008; Rieth, Lee, Lui, Tian, & Huber, 2011; Gosselin, Bacon, & Mamassian, 2004), including an auditory version of the task for vowel detection (Brimijoin, Akeroyd, Tilbury, & Porr, 2013). How is the detection for these "ghost" Fourier traces of signal in noise possible? One view is that sensitivity to very weak signal traces in a high level of noise is aided by top-down perceptual processing in which mental representations of expected or possible patterns are convolved with stimulus information.

Behavioral evidence suggests that top-down linguistic knowledge can aid in comprehension of degraded speech, particularly in the perception of noisy or liminal inputs. For example, Sohoglu, Peelle, Carlyon, and Davis (2013), found that written text improved the reported clarity of noise-vocoded speech, highlighting that knowledge of what is being said helps to guide speech perception. Additionally, multisensory input in the form of lip reading can aid recognition of speech in adverse listening conditions (Ma, Zhou, Ross, Foxe, & Parra, 2009). Similar findings have been identified in recent neuroimaging work using a face detection in noise task (Nestor, Vettel, & Tarr, 2013; Liu et al., 2014). In particular, Liu et al. asked individuals to decide whether pure noise stimuli contained a face or letter, neither of which were physically present. Their results indicated that the fusiform face area (FFA), which is typically active when

faces are present (Kanwisher, McDermott, & Chun, 1997) showed increased activation in the FFA for the noise stimuli that were selected as containing faces compared to noise stimuli that were not selected. Because the faces were illusory, Liu et al. (2014) concluded that top-down knowledge associated with activation in the fusiform gyrus influenced liminal stimulus perception. However, Liu et al. neglected to determine whether there was a detectable face in a given noisy stimulus. While top-down influences could result in FFA activity for select stimuli, it is left unknown whether participants were simply using a bottom-up process for detecting a face in those noisy stimuli that registered activity in the FFA.

The majority of studies investigating the effect of top-down control of perception using "pure-noise" stimuli have inferred the imagined target image by reverse correlation and relating the detected noise samples to physical target properties. Specifically, reverse correlation is used to calculate a "classification image" that shows which points of the noisy stimuli correlate to the same points in other stimuli in the set (e.g., pixel luminance across noisy images), and whether or not the subject detected a signal in the noise (Reith, Lee, Lui, Tian, & Huber, 2011). The resulting classification images not only represent areas of high or low correlation, but also visually appear to show a "superstitious" representation of a top-down prototype (Gosselin & Schyns, 2003). In the reverse correlation paradigm, an observer's perceptual goal of looking for a particular target determines which noise samples are categorized as related to the target and which noise samples are categorized as unrelated. While this results in an idiosyncratic process that is particular to the observer, this need not be the case if there is an objective reality to the presence of signal properties of the target distributed among samples of noise.

It should be possible to select a target ahead of time (e.g., the letter S) and identify noise samples that contain more image content or less image content in terms of target signal properties related to each noise sample (e.g., which noise images would correlate with the target image), yielding a set of putative "true noise" samples (zero similarity to the detection target image) and a set of putative "target signal plus noise" samples using objective statistical methods. This removes the uncertainty of observer-specific sample determination and permits a standard signal detection analysis. With objectively defined signal correlated with noise vs. uncorrelated noise trials, it is possible to measure sensitivity to target trace information relative to measures of signal similarity.

If all samples of noise are effectively perceptually equivalent because they are all produced by independent random number sequences, it should be easy to make clear predictions for classification performance in determining the presence of a visual letter S target. Recognition performance for high-target-correlation noise samples should be identical to recognition performance for low-target-correlation samples. If this were the case, the hit rate and correct rejection rates should each be 50% since the observers would be guessing, yielding a d-prime score near zero. However, recent research (e.g., Gosselin & Schyns, 2003) suggests that different observers might have different mental representations of the target letter, and thus the selection of signal related noises could be idiosyncratic to their individual mental representations. These idiosyncrasies may be detectable using signal detection methods resulting in positive d-prime scores (i.e., higher hit rates for high-target-correlation samples).

To the extent that observers' mental representations overlap and to the extent that they correspond to physical properties in a noise stimulus, we hypothesize that objective selection of a set of noise samples putatively related to a visual target should predict perception (Specific Aim #2). In this way, selecting specific noise samples with higher correlation vs. lower correlation with a prototypical "S" should permit an objective test of the ability of observers to be sensitive to very small amounts of signal information in noise, rather than just a random degree of correspondence with an idiosyncratic image. If this is the case, observers should be able to detect visual targets above chance, resulting in d-primes significantly greater than zero. This allows an objective method for both measuring target signal sensitivity and response bias.

In Gosselin and Schyns (2003), very large numbers of samples of noise were used (2 orders of magnitude greater than the present stimulus set size). The odds of correctly identifying noise samples that work for all observers depends on an additional assumption: rather than imagine that all observers have idiosyncratic mental representations of a designated target, it is possible that observers use a kind of mental representation that guides perception (e.g., Kosslyn, 1981; Bar, 2003).

To the extent that observers use common sensory pattern information across the "target" noise samples and this information is absent from the unrelated distractor noise samples, detection performance should show reliable sensitivity above chance (Specific Aim #2). Study #3 tests the hypotheses that in observers who have a particular target for recognition, that it is possible to measure the relationship of an objectively random noise sample to the target, and that it is possible to predict (on average) recognition performance for the sample.

Study #4 addresses the idiosyncratic observations of the "S" selection among participants in Gosselin and Schyns (2003). The aggregated "S" images are unique to each individual participant, perhaps reflecting variations in the prototype "S" images kept in mind, or perhaps whether participants were strategizing by classifying the stimuli according to what letter of the alphabet it most resembled. Instead of relying on self-report, an image classifier was employed that could be trained to classify stimuli images in these two ways. The Bag of Visual Words classifier (Csurka, Dance, Fan, Willamowski, & Bray, 2004) is a variant of the k-classifiers and requires no prototype image. In this method, the classifier is trained to identify local features that are combined into clusters when given stimuli of known categories. These clusters of features, known as "words", can then be used to classify novel stimuli. This study is a computational verification that participants are completing the task as hypothesized and that they are sensitive to incidental cues that are target-like in the noise stimuli; no human participants were recruited for Study #4.

The Bag of Visual Words classification method does not require a human participant given exposure to an "S" image, as the Bag of Visual Words algorithm works by detecting and classifying features within the noisy stimuli images. However, it is not clear what strategy human participants in Study #3 employed to sort the images. Were participants sorting S versus no-S, or were they identifying any letter in the images and responding affirmatively only when the letter they identified was an S? Study #4 is designed to clarify the strategies participants in Study #3 may have employed.

Despite questions of what strategies individuals might be using to classify these stimuli, a question remains of how the insight described in this task might be related to other experiences of visuo-perceptual insight (Specific Aim #2). Specifically, is it possible that the ability to recognize noise images that are statistically similar to an "S" prototype could be due to perceptual fluency? Do some individuals find the stimuli easier to classify, and is this ability related to other measures of perceptual experiences? Grossberg (2000) has argued that perceptual experiences such as hallucinations are evidence of a normal-top down perception mechanism. Top-down expectations generally balance bottom-up inputs, but Grossberg argues that the dynamics can shift to favor one over another and give rise to "hallucinatory" experiences. If perceptual fluency represents a shift in top-down and bottom-up dynamics, measures of these perceptual experiences may correlate with increased insightful behavior. Here, that behavior is measured as increased sensitivity to "S" in noise.

If the ability to classify images in this task is related to hallucinatory experiences it is useful to compare the ability in this task to self-report measures of hallucinations. The Hearing Voices questionnaire (Posey & Losch, 1983) asks respondents whether they have personally experienced particular common voice-hearing hallucinations. The Launay-Slade Hallucination Scale (Launay & Slade, 1981) is designed to identify individuals who are likely to develop hallucinations. Unlike the Hearing Voices questionnaire, the Launay-Slade is intended to identify other perceptual experiences as well, such as visual hallucinations, intrusive thoughts, and daydreams. The Oxford-Liverpool Inventory of Feelings and Experiences (O-LIFE) scale was originally designed to measure proneness to schizotypy, in particular psychotic characteristics found in normally functioning individuals (Mason, Claridge, & Jackson, 1995; Mason &

Claridge, 2006). In the present study, only the Unusual Experiences scale is used, as it contains questions that specifically pertain to atypical perceptions, magical thinking, and hallucinations. The Tellegen Absorption Scale measures a personality trait called "absorption" that describes periods when an individual's attention is fully engaged to perceptual inputs (Tellegen & Atkinson, 1974). Combined, these questionnaires can be used to assess whether fluency in perceiving an "S" in noise is related to fluency of other perceptual experiences.

Grossberg's (2000) assertion of perceptual experiences such (as hallucinations) are evidence of a normal-top down perception mechanism can be tested in a second way: by comparing performance on sorting noise "S" stimuli with a task of attentional control. If the ability to sort stimuli are related to an ability to shift attentional control in perception, that same ability should result in better performance on another task of attentional control. To test this, participants in Study #5 were additionally asked to complete the Eriksen Flanker Task (Eriksen, 1995).

Study #3 Methods

The letter "S" was chosen for the standard image to maintain consistency with the original study by Gosselin and Schyns (2003). While this is not important to the goals of the study, doing so would strengthen the conclusion that the results of this task are the same form of perceptual insight as identified by Gosselin and Schyns. This standard "S" image, defined by an

image of 50-by-50 pixels, was generated to contain a black “S” over a white background. The letter “S” was sized to take up the full image and did not have serifs (

Figure 13, left-most image). A set of 1,000,000 random-noise black-and-white images of matched dimensions was generated using the Mersenne Twister pseudorandom number generator in MATLAB (2011; Matsumoto & Nishimura, 1998). Generation of the noise stimuli set required 2.5 * 10^9 random numbers; the Mersenne Twister algorithm has a periodicity of about 4.3 * 10^6001, suggesting no repetition in the generated sequence of pseudorandom numbers. Each of the black-and-white images was then transformed into binary vectors and a Pearson’s correlation coefficient was calculated for each noise image relative to the prototype image.

Figure 13: Examples of visual stimuli. From left, 50x50 black and white prototype S, not shown to participants; 50 x 50 black and white noise sample (image 21 of 100), higher correlation with the visual target S; 50 x 50 black and white noise sample (image 95 of 100), no correlation with the visual target S; 50 x 50 grayscale summation of 100 noise images with the highest correlation with the visual target S; 50 x 50 summation of 100 noise images with no measurable correlation to the visual target.

The full set of noise images were rank-ordered by correlation coefficient, after which 100 noise images with the highest similarity to the S target image (Pearson r; M = 0.0782, Range: 0.0729 to 0.0985, SD = 0.051) were selected as the signal plus noise target stimuli. The 100 noise images with the lowest similarity (Pearson r; M = 0.0001, Range: -0.0008 to 0.0008, SD = 0.051) were selected as noise (distractor) stimuli; the remaining noise images were not used.

Measured similarities of images in the target and distractor sets did not overlap with each other. When aggregated, the highest correlating noise images did appear to show an S, similar to Gosselin and Schyns' findings. The lowest correlating noise images, however, did not appear to produce an "S" as would be expected (see

Figure 13 for examples).

The participants were seated so that the stimuli subtended approximately 2 degrees of visual angle (cf. Gosselin & Schyns, 2003). To minimize the idiosyncratic nature of the S reverse correlation images found in Gosselin and Schyns, participants were given instruction in the S to identify in the stimuli. Participants were instructed to examine each stimulus image for a large black "S" on a white background. They were instructed to indicate for each image whether a black "S" was present or not present by pressing one of two buttons associated with each choice on a keyboard. Stimuli were presented on a standard computer workstation via MATLAB (2011) and the Psychophysics Toolbox (Brainard, 1997).

Participants were given two practice trials with explicit accuracy feedback to familiarize the task and image nature. One practice trial presented a target noise sample that was similar to the S target image and the other practice trial presented a dissimilar distractor noise image. These practice images were not part of the 100 similar or 100 dissimilar image sets, but were generated with the same method. Participants were told that one of the practice trials was a target image and the other was distractor noise, but no further guidance was given. At no point in the experiment were participants shown the standard "S" image. After the practice trials, the participants were tested on all 200 stimuli images in a randomized order with a two second delay

between trials; no time limits were enforced. Participants received no accuracy feedback beyond the initial practice trials.

After completing the 200 test trials, participants were debriefed and asked to draw the "S" they thought of while doing the task within the outline of a square (to represent the bounds of the stimuli image) marked on a sheet of plain white office paper. These drawings were then rated by four independent raters on whether they contained a large "S" on white background (i.e., the same instructions given to participants); raters were not shown the prototype "S" image used in creating the stimuli sets. All four raters determined that the drawings of 7 participants were not that of a large S; an additional image drew the disagreement of one rater. Only the 7 participants whose "S" drawings drew complete disagreement among raters were excluded.

Study #3 participants were twenty-eight students at the University of Chicago (10 male; age M = 20.3 years, SD = 1.7 years). Seven participants were excluded because they did not follow experiment instructions with respect to the target being searched for, as determined by independent raters of a post-experiment verification task. Participants were offered course credit or cash compensation at the rate of 0.5 course credit-hours or $5 per half-hour of participation. The experiment was reviewed and approved by the University of Chicago IRB (approval #H09494).

Study #3 Results

Overall, participants correctly responded to 58.1% of stimuli (M = 0.58, SD = 0.06), yielding a classification performance that is significantly above chance ($t(20) = 5.98$, $p < 0.001$). Target signal detection performance was measured using a signal detection analysis and revealed that, for the group, participants' d-prime scores were significantly greater than d-prime = 0 ($t(20) = 5.87$, $p < 0.001$), averaging a d-prime of M = 0.46 (beta = 0.11, criterion c = 0.19). This indicates that on average observers were able to discriminate between the two sets of noise stimuli reliably above chance. In other words, participants as a group did not identify stimuli as containing an "S" by randomly guessing. The results also show a conservative bias against responding that an "S" exists in a given stimulus image.

A histogram of d-prime and beta values for the participants is shown in Figure 14 and Figure 15. Two participants had a d-prime = 0, indicating they were unable to discriminate between the two sets of noise stimuli. But on average, participants were able to discriminate between the two sets of noise samples and determine which samples contained S image properties above chance. The low beta score suggests that the effect is not due to a general bias in responding "yes" to all stimuli.

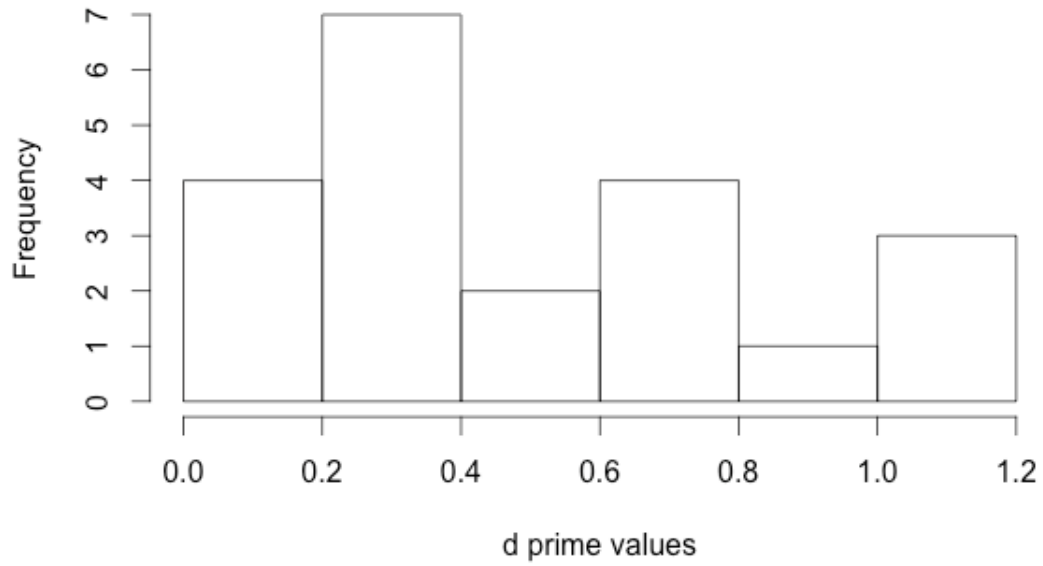

Figure 14: Distribution of d-prime scores across participants demonstrates on average an ability to discriminate between the two sets of noise samples.

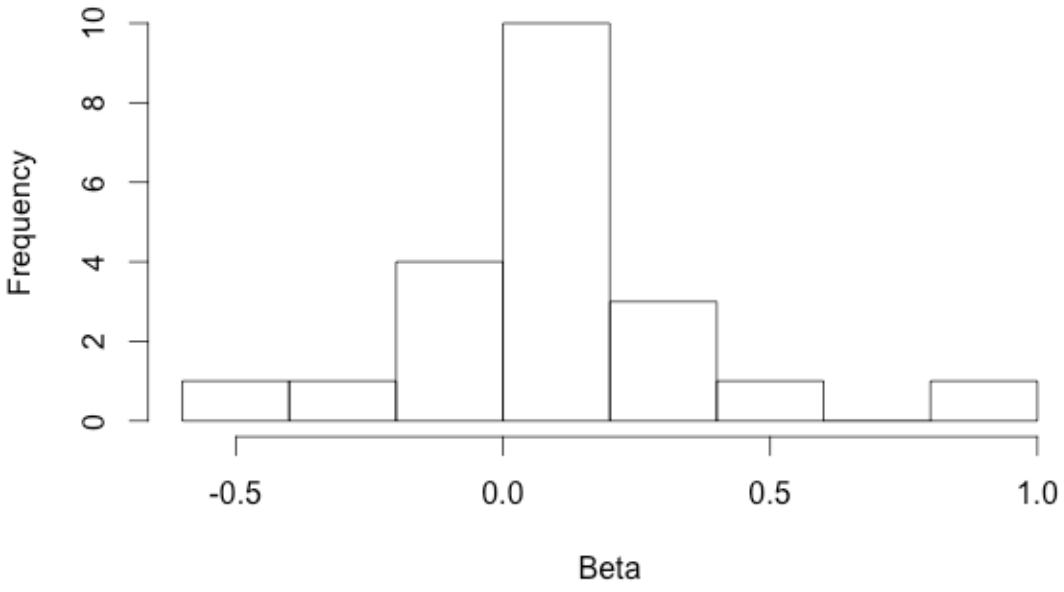

Figure 15: Distribution of betas across participants suggests that the d-prime values are not due to a general bias in responding yes in all stimuli.

A question remains, however, of whether some sample noise images are better representations of the target than others. Instead of describing the data in terms of participant response, the results can also be examined in terms of likelihood across observers to respond to each specific noise stimulus image (Figure 16). Each image in the high correlating "target" noise image set was selected as containing an S on average by 51.4% of participants (M = 0.51, SD = 0.18). Meanwhile, stimuli in the low correlating "distractor" image set were selected as containing an S on average by 35.3% of participants (M = 0.35, SD = 0.15). The difference in selection rates between high and low correlating groups is also significant ($t(193.24) = 6.85$, $p < 0.001$). Together, these results further suggest that participants as a group were more likely to select images from the highly correlating stimuli set as containing an S than to select images from the low correlating stimuli set.

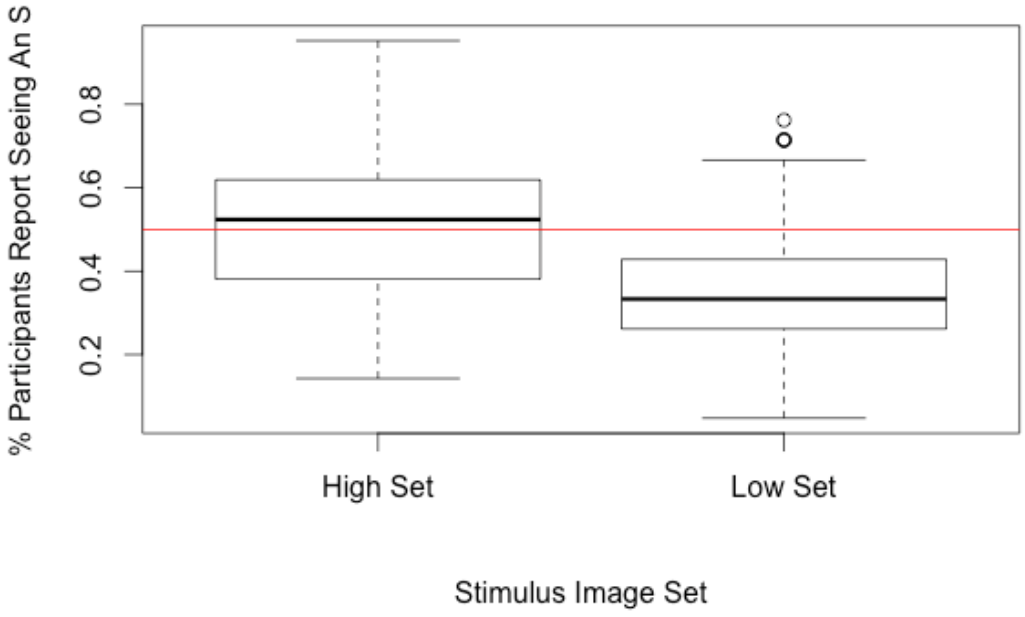

Figure 16: Distributions of the probability that images are reported to contain the target S. Red bar represents 50% chance.

Given a set of 200 randomly-generated visual noise samples and no prior knowledge, it should be impossible for an individual to sort these into two different groups. Independently generated random number sequences are statistically independent by definition. However, for an observer who knows that these samples were selected to correspond to a visual S target or to be unrelated, the extent to which these samples could be sorted on this criterion reflects the ability to relate a mental representation of the visual target to the noise samples. If this determination can be carried out above chance, it indicates the degree to which weak but objectively present signal-related information in noise can be detected by an observer. The present results demonstrate that observers can, on average, discriminate between images that contain a weakly present signal and those that do not, rejecting certain "noise" images over others. It is important to emphasize there is no target in any of these noise samples. Observers are not detecting a visual target mixed with visual noise. Rather in order to perform above chance, they must be sensitive to the weakly present visual features in randomly generated noise consistent with their mental visual representation of a target.

There are two important observations to note. First, previous research using this kind of noise paradigm by Gosselin and Schyns (2003) has been carried out with an assumption that such noise samples that are perceived as target-relevant are idiosyncratic to the observer. This is presumably based largely on the assumption that differences among observers in stimulus perception and mental representations guiding that perception are sufficiently large to make a priori identification of target-similar noise samples impossible. The conclusion that the selection of noises was idiosyncratic to the observer in Gosselin and Schyns' study was made through

post-hoc correlations of individuals' aggregated target-similar noises against various letter prototypes, rather than directly tested. As such, the present results demonstrate for the first time that an objective prior metric of target similarity is sufficiently reliable to be used to sort target-similar and target-dissimilar noises, and that such a sorting may be used to assess detection performance.

The second important observation is that the measured level of similarity between the target-similar noises and the target is extremely low. A correlation coefficient of 0.08 means that less than 1 percent of the noise image variance is sufficient to support reliable target detection. Just as sensory detection of stimulus energy is extremely sensitive even in the general population (Hecht, Shlaer, & Pirenne, 1942), this shows that perceptual sensitivity to signal information is also extremely high in the general population. It takes very little systematic variance between noise and image in order for focused attention to detect the 'ghost' in the noise.

Study #4 Methods

Study #4 did not involve human participants. Instead, a computer algorithm, the Visual Bag of Words (Csurka, Dance, Fan, Willamowski, & Bray, 2004), was employed to test two possible strategies human participants may have used in Study #3. The algorithm was trained in two separate Conditions on stimuli and tested for performance. In Condition #1, the classifier was trained on a subset of the stimuli used in Study #3 that were ranked as S-similar and S-unsimilar, and then tested for accuracy of sorting a subset of images into S-similar or S-unsimilar

categories. In Condition #2, the classifier was trained on subsets of the same one million noise images found in Study #3 categorized as the 100 most statistically similar to each letter A through Z, and each numeral 1-9 (zero was omitted because it was identical to the letter O). The classifier was then tested on a reserved subset of letters that had been classified as being S-similar and a subset that was S-unsimilar.

The Bag of Visual Features was run in MATLAB using the Computer Vision System Toolbox (2011). The default 8x8 pixel grid size (used in determining the location of features) was changed to a 1x1 pixel grid size; the default grid size was designed as a means of computational time savings, which was not necessary in the current analysis.

In Condition #1, the same stimuli previously generated for Study #3 were used:1,000,000 images of randomly-generated black-and-white noise that were rank ordered by Pearson's correlation, with the S-similar images being the 100 highest correlating images to the prototype S image described in Figure 13, above, and the S-unsimilar images being the 100 least correlating. Of each 100 images, 30 were randomly selected for training the classifier and the remaining 70 reserved for later testing. This selection for training and testing was repeated 1000 times.

In Condition #2, the same 1,000,000 images from Study #3 were rank ordered by correlation to prototype images of each letter of the alphabet and each numeral, 1-9 (zero was omitted because it was identical to the letter O). For each rank ordering, the 100 images highest-correlating to the given prototype was selected. Of these, 30 images from each ordering were used in training the classifier. The 70 of the 100 noise images that correlated best with the S

prototype and 70 of the 100 noise images that correlated worst with the S prototype were used for testing. This process was repeated 1000 times. Because the test trials included S-similar and S-unsimilar image sets, a signal detection analysis was carried out (following the same methods in Study #3) and d-prime statistics calculated. These results are then considered in terms of Study #3: whether participants are engaging the noise stimuli as hypothesized.

Study #4 Results

In Condition #1, in which the classifier is given stimuli generated and sorted by similarity to a target "S" image, mean d-prime = 0.738 (SD = 0.243) and were statistically significant from zero ($t(999) = 96.234$, $p < 0.001$). This positive d-prime statistic is indicative of discrimination among the noise stimuli and is similar to the positive discrimination found with human participants in Study #3. In Condition #2, where noise images were correlated to a wider range of target characters, mean d-prime = 1.564 (SD = 0.390) and were statistically different from zero ($t(999) = 126.990$, $p < 0.001$); these results are indicative of positive discrimination by the classifier. Two trials of Condition #2 were found to have a d-prime value of zero or less, demonstrating overall good but not perfect discrimination behavior between S-similar and S-unsimilar images (Figure 17).

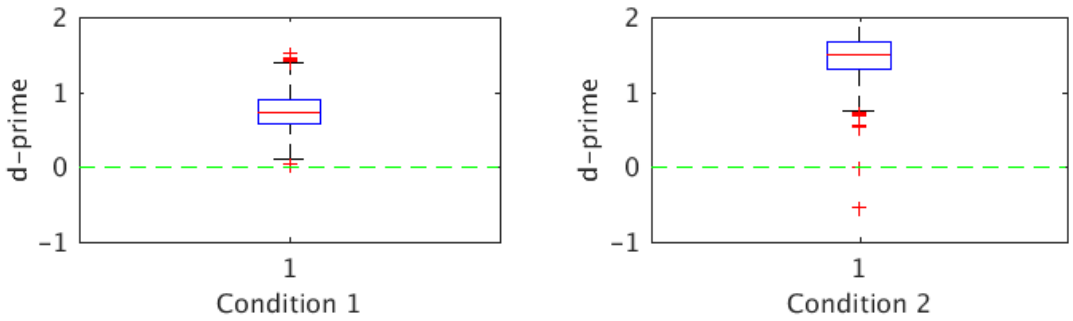

Figure 17: d-prime statistics for a Bag of Features image classifier when trained on S-similar and S-unsimilar noisy-image categories (Condition #1) or on noisy-image categories representing numerals 1 through 9 and the upper-case alphabet (Condition #2). The green dotted line represents chance (d-prime = 0).

Study #5 Methods

Study #5 is a test of attentional processes that may be involved in the ability to perceive and classify these noise stimuli. Participants were 23 members of the University of Chicago community (age M = 23.4 years, SD = 2.2 years; 12 female) who were screened to ensure they had neither participated in the previous S task protocol, nor had they prior knowledge of the task (e.g., by discussing the experiment with previous participants). While recruiting materials emphasized normal or corrected-to-normal vision (due to the reading necessary for each question item), visual acuity was not verified beyond self-report. Participants were compensated with cash or course credit (0.5 course credit-hours or $5 per half-hour completed) in return for their time spent during the study. The experiment was reviewed and approved by the University of Chicago IRB (approval #H09494).

After participants completed informed consent, participants were asked to complete demographic questionnaires before beginning the S image classification task. To maintain consistency across studies, the same 50-by-50 pixel letter "S" was used again for the standard image. The same set of 1,000,000 random-noise black-and-white images of matched dimensions generated using the Mersenne Twister pseudorandom number generator in MATLAB (2011; Matsumoto & Nishimura, 1998) was also used. The same two-alternative-forced-choice paradigm from the previous S classification task was used, as were the same stimuli. Instructions were modified so that instead of being provided with a description of the letter "S" to be searched, participants were instructed to search for a single uppercase letter whose top and bottom edges reached to the top and bottom edge of the noise image. Stimuli consisted of 100 low-correlating and 100 high-correlating images presented in randomized order. These stimuli were presented on a standard computer workstation via MATLAB (2011) and the Psychophysics Toolbox (Brainard, 1997) in the same experiment testing room using the same computer setup.

Following the S classification task, participants were asked to complete the Eriksen Flanker Task. The task consisted of 80 trials presented via E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA) on a Windows-based desktop computer. Participants were asked to judge whether the central character in a given display was "congruent" or "incongruent" with the flanking characters. In the current version of the task, arrows ("<" and ">") were used to maintain similarity with Eriksen (1995). This results in two types of displays that are congruent: "< < < < <" (central arrow and flanking arrows point left) and "> > > > >" (central arrow and flanking arrows point right). This also results in two types of incongruent displays: "< < > < <" (central arrow points left but flanking arrows point left) and "> > < > >" (central arrow points

left but flanking arrows point right). Participants were asked to respond by pressing one of two buttons, one marked “congruent” and one marked “incongruent”, as quickly and as accurately as possible before moving to the next trial. No performance feedback was given to participants in this task.

The questionnaires were presented last to prevent demand characteristics. All questionnaires were distributed on paper and were scored as originally designed (cf. Posey & Losch, 1983; Launay & Slade, 1981; Mason, Claridge, & Jackson, 1995; Tellegen & Atkinson, 1974).

Study #5 Results

Participants as a group were able to sort the noise images (d-prime M = 0.374, SD = 0.0484). A two-tailed one-sample t-test was significant at the 0.05 level ($t(22) = 3.701$, $p = 0.001$); these results are similar to the findings in Study #3. Given that these participants are completing the S visual noise classification task as expected, the question of attentional processes contributing to successful image discrimination can be examined. This is accomplished through the Eriksen Flanker Task that was also administered to the participants.

On the Eriksen Flanker Task, participants as a group were able to respond at a high accuracy rate to congruent stimuli (M = 0.987, SD = 0.018) and incongruent stimuli (M = 0.950,

SD = 0.193). While participants were able to respond faster to congruent stimuli (M = 472.0 ms, SD = 154.0 ms) than incongruent stimuli (M = 586.6 ms, SD = 251.8 ms), a two-sample two-tailed t-test was not significant at the 0.05 level (t(36.469) = 1.4896, p = 0.145).

Spearman's Rank Order correlations between participant S classification d-primes and reaction times for congruent trials of the Eriksen Flanker Task (r(11) = -0.008, p = 0.159) as well as between S Task d-primes and incongruent Eriksen Flanker Task trials (r(11) = 0.101, p = 0.347) were not significant at the 0.05 level.

Interestingly, performance on the S Task was not significantly correlated with any of the given perceptual experiences measures at the 0.05 level (

Table 12). A Bonferroni p-value correction for multiple comparisons did not change these findings.

Questionnaire	Questionnaire Score Mean	Questionnaire Score SD	Spearman's r	Spearman's p-value
Hearing Voices	4.609	2.726	-0.005	0.980
Launay-Slade	35.913	14.283	0.156	0.478
OLIFE Unusual Experiences	9.913	5.607	0.148	0.502
Tellegen Absorption Scale	17.304	6.470	0.041	0.854

Table 12: Spearman's rank-order correlation coefficients between d-prime for detecting an S in noise and other measures of perceptual experiences. p-values are not corrected for multiple comparisons.

Discussion

Fodor (1986) argued against top-down mechanisms in perception by saying that hallucinations are not produced in normal brains and minds, but reflect disordered processes. Evidence for this argument can be found in Bentall and Slade (1985), which concluded that individuals who scored highly on the Launay-Slade Hallucination Scale, whether undergraduate students or hallucinating schizophrenic patients, were "deficient in reality testing and therefore prone to identify imaginary events as real". However, an alternative interpretation is that such individuals may well be identifying traces of real signals in random noise. Others have found that individuals prone to fantasy are more likely to report hearing music, when instructed to do so, if they are deceptively asked to listen to auditory noise (Merckelbach & van de Ven, 2001). Similarly, Barber and Calverley (1964) found that individuals given a hypnotism intervention were more likely to report hearing music in absence of stimuli than with other types of instruction.

Identification of objects in noise also forms the basis of the Snowy Pictures Task (Ekstrom, French, Harman, & Dermen, 1976), as a measure of perceptual ability. Stimuli in the Snowy Pictures Task are a set of black-and-white images in which objects are embedded within a field of noise, with the amount of noise controlled by the experimenter so that the object is either present or entirely hidden (Whitson & Galinsky, 2008). Studies using the Snowy Pictures Task have linked performance to decision-making biases: paranoid schizophrenics and healthy individuals with manipulated dopamine levels display overconfidence in their incorrect

identifications of Snowy Pictures Stimuli (Moritz et al. 2014; Andreou, Bozikas, Luedtke, & Moritz, 2015). Similarly, at-risk-gamblers also respond to more Snowy Pictures stimuli than control non-gamblers (Stea & Hodgins 2012).

In individuals with dementia with Lewy bodies, a disease for which about 80% of patients experience visual hallucinations, Yokoi et al. (2014) found that patients were more likely than others to identify random noise as meaningful. These same individuals identified fewer samples of visual random noise as meaningful after receiving a dose of donepezil, an acetylcholinesterase inhibitor known to help limit hallucinations.

These effects are not limited to patient populations. For example, normal healthy individuals have been reported to make source attribution errors regarding liminal sensory perception (Perky, 1910; Segal, 1972). More specifically, when individuals are asked to describe common objects from their imagination while being presented with liminal images of those objects, participants routinely confused the contents of their imagination with the liminally presented stimuli. Most impressively, participants in those cases were unaware of the liminal inputs' influence. This suggests that everyday individuals' percepts may be undetectably influenced by liminal signals found in random noise, supporting the argument that these behaviors, while not strictly hallucinatory, might reflect a shift in the top-down vs bottom-up dynamic. Perhaps as compelling, Experiment #3 suggests this behavior is not linked to measures of schizophrenic experience, with little or no correlation between S-identifying performance and the four perceptual experience questionnaires.

The Bag of Features classifiers are capable of distinguishing between S-similar and S-unsimilar images, regardless of whether the classifier was trained on a binary choice between S-similar and S-unsimilar, or images similar to S versus images similar to other alphanumeric characters. The results of Experiment 2 suggest that both classification methods provide plausible strategies for human participants in categorizing images as S-similar or S-unsimilar.

As to Study #5, no clear link was established between the ability to identify the letter in noise and other measures of perceptual abilities. Of the four perceptual experience questionnaires administered in the study, none correlated with d-prime values above $r = 0.20$, and none were statistically significant at the 0.05 level. From these results, it cannot be concluded that performance on the task is affirmatively related to perceptual experiences (Specific Aim #2). If a link between the perceptual ability to identify an "S" in noise and these other measures is weak or nonexistent, it is possible that this task represents a particular type of perceptual insight not captured elsewhere – giving further credence for Specific Aim #1 that a single insight ability might not exist.

Even so, these two experiments present a novel method to demonstrate perceptual insight as described by Specific Aim #2. Like other tasks of insight, participants completing this task are able to identify a solution (that is, are able to detect and sort images as containing an "S" or not"), yet are unable to describe the method used to do so. The signal detection aspect of this task is particularly powerful in identifying perceptual insight among individuals by measuring the ability to sort between stimuli previously known to represent a target stimulus (the capitalized sans-serif "S") and other noise images.

The results of the present research, although relevant to understanding the nature of hallucination and perception, raise a very fundamental and important issue about ordinary perception. While Fodor (1986) argued that hallucinations only occur as the result of a disordered brain or mind, the present results suggest that hallucinations may well reflect a very normal yet powerful recognition process. Given that any particular noise sample may have weak similarity to a large number of possible patterns, signal detection (e.g., of the letter "S") may occur as a result of a salient working memory representation that is convolved through attention with a noisy environment. This kind of process, which the current laboratory experiment demonstrates as capable of detecting extremely low levels of target information within noise, may be the basis for the top down enhancement of perception in any noisy situation - thereby demonstrating an important process for supporting the robustness of perception in adverse conditions.

Conclusion

One aspect of perceptual insight not examined in the last chapter is measuring individual differences in sensitivity. The Camouflage Perceptual Fluency task is based on scores assessed by a human judge, which may introduce a degree of subjectivity. The letter recognition task described here provides an important piece of linking evidence. First, it demonstrates a perceptual insight, though classification of noisy images, that is similar to the Camouflage

Perceptual Fluency task. Second, the study protocol removes the need of a human scorer, removing a source of subjectivity.

In an improvement over Gosselin and Schyns (2003), the idiosyncratic measures that do not allow quantitative comparisons among individuals are replaced with a d-prime statistic that reflects individual discriminative perceptual insight. Third, the human performance on the classification task can be modeled with a computational classification model, yielding results that appear similar to human performance and suggest that humans are discriminating among images as hypothesized. Fourth, the ability to discriminate among noise images does not appear to be directly correlated with attentional processes (as tested with the Eriksen Flanker Task), nor does it appear to be directly correlated with unusual perceptual experiences.

The task here demonstrates a new way to assess perceptual insight, a task in which participants must assess and discriminate among noise stimuli in search of a target. This task also introduces the noise discrimination "S" task as a sensitivity measure of perceptual insight. While this task and the Camouflage Perceptual Fluency task demonstrate insight at perception, the two tasks are strictly visuoperceptual. The following chapter extends the understanding of perceptual insight into the auditory domain.

CHAPTER FOUR: AUDITORY PERCEPTION AND TRANSFER OF SKILL

The previous studies have attempted to investigate insight across different types of tasks, including perceptual insight. However, this has focused on the visuoperceptual domain. As noted in Study #1 and Study #2, there appear to be two kinds of response patterns in insight tasks suggesting the possibility of two different types of insight or different ways of achieving it. One form, represented in the RAT and Duncker-Maier Functional Fixedness, is represented in solving discrete separate problems with independent solutions. Solving one problem does not lead to solving others, at least in terms of conceptual content. While there could be a general skill of cued memory retrieval or open monitoring to defeat functional fixedness or strategically find atypical associations across trials, solving one problem does not help solve a second. The second form is seen in the NRT, showing up as different due to PCA results in the first two studies. In the NRT, insight is drawn out over trials, first manifest as implicit insight from the speeding up of responses and then later in trials as a direct reflection of insight in the underlying rule by going right to the final response. Further, the difference in the NRT and other insight tasks is bolstered by the grouping of working memory performance (R-SPAN) in Study #1 showing that working memory fluency is predictive of performance in solving discrete separate insight problems, but less so when insight is drawn out over trials.

Study #2 raised the question of whether perceptual insight, as revealed in camouflaged images is more similar to the NRT than the others. By one view, observers can learn the general masking properties of the camouflage process and thus insight could emerge over trials even though different images are presented. Contrary to that hypothesis, the results indicated that this form of perceptual insight is more similar to the RAT than the NRT in terms of performance variance. Moreover, given that the cognitive single-trial tasks like RAT are similar to working memory performance in Study #1, which is not similar to NRT performance in participants, it is possible that the visual camouflage task also uses working memory. Participants may hold hypotheses about solutions to problems or recognition in mind while testing them cognitively or perceptually against the stimulus (see Nusbaum & Schwab, 1986).

If true, then examining insight in a task that demonstrates learning over time and affects working memory might tap into a perceptual insight that is more similar to the NRT than the camouflage (CPF) task. For example, listening to a series of utterances that are degraded acoustically from a single talker, speech recognition improves showing significant learning (e.g., Schwab, Nusbaum, & Pisoni, 1985). This improvement is talker specific (e.g., Nygaard & Pisoni, 1998) and is demonstrated by improvements in recognition over time, somewhat similar to learning the hidden pattern in the NRT. In this case the perceiver is learning the speaker, whose vocal tract produces the different utterances, rather than learning a single acoustic pattern to memorize. It has been demonstrated that listeners can learn to recognize a talker from sinewave speech—an acoustic transform from the full acoustics of natural speech to be three time-varying sinewaves matched to the center frequencies of the first three formants of speech (Sheffert, Pisoni, Fellowes, & Remez, 2002). While there is generalization from specific

utterances to novel utterances for difficult to understand speech, as if the listener is learning to recognize the talker's vocal patterns (cf. Nygaard & Pisoni, 1998), there is less generalization across source distortions of the speech (Hervais-Adelman, Davis, Taylor, Johnsrude, & Carlyon, 2011). This suggests that listeners may infer how a talker produces speech to some degree but it may be limited by information in the source properties of the speech. The current study (Study #6) was designed to examine perceptual generalization between one acoustic transform of a speech signal (sinewave speech) and another (noise-vocoded speech) between which limited generalization has been demonstrated already. The question is whether this can serve as a basis for investigating auditory perceptual insight as the camouflage images have in vision. The acoustic transforms act in much the same way obscuring spoken word recognition as the camouflage masks image recognition. The difference is that there is prior evidence for learning the transform over time. If there is generalization between transforms, that could serve as the basis for studying whether listeners have an "insight" into the talker. Prior research on sinewave speech has shown there is an insight or aha experience in realizing that sinewave speech is actually speech and not a set of separate nonspeech beeps (Remez, Rubin, Pisoni, & Carrell, 1981).

To make a greater argument that links perceptual processes and insight, it is necessary to extend the argument to a new sensory domain. In this chapter, the link between insight and auditory perception will be argued.

Recent work in auditory perception argues that learning one form of degraded vocoded speech (e.g., Sinewave Speech) will not result in better perception of other forms of degraded

speech (e.g., Noise Vocoded Speech) because of a lack of shared auditory cues. The current work aims to demonstrate that a transfer of learning does occur, questioning the nature of the auditory cues that allow for in-common perception (Specific Aim #2). This is a different test than presented with the noisy picture stimuli of Chapter Three, where insightful thinking was identified through the ability to discriminate target-like and non-target-like noise. Instead, the insight tested here is one of transfer.

In analogical problem solving, individuals are given a problem state to consider. For example, Gick and Holyoak (1980) describe a situation in which a cancerous tumor can be ablated with a beam of radiation, but that beam itself would damage the tissues between it and the tumor. Individuals given this problem were more likely to solve the problem by drawing analogies between the problem state and what appears to be a completely different situation: an army attacking a fortress split up its soldiers across multiple roads so as to approach the fortress from multiple directions, thereby lessening the dangers that come if the soldiers had all attacked the fortress from a single direction. Individuals were expected to transfer similarities between the tumor problem and the fortress problem to come up with the solution: beam the radiation onto the tumor from multiple directions to spread the damage across tissues to lower the radiation absorption by any noncancerous tissues.

If the insight that occurs through analogical transfer can occur at the cognitive level, can transfer occur with insight at the perceptual level? If individuals are given an audio-perceptual "analogy", can it be used to perceive a different audio-perceptual "problem"?

Training through repeated exposure has been demonstrated to facilitate perception and comprehension of degraded vocoded speech, generalizing to new stimuli over time (Schwab, Nusbaum, & Pisoni, 1985; Fenn, Nusbaum, & Margoliash, 2003) and to Noise Vocoded Speech presented with shifted frequencies (Hervais-Adelman, Davis, Johnsrude, Taylor, & Carlyon, 2011). However, what has not been clearly demonstrated is the transfer of auditory perceptual learning. While facilitation has been demonstrated for perception and comprehension within one form of degraded vocoded speech (e.g., Sinewave Speech), exposure of one form has not been observed to facilitate the perception and comprehension of vocoded speech modeled with different acoustic characteristics (e.g., training on Sinewave Speech facilitating performance on Noise Vocoded Speech).

In terms of the Specific Aims, this experiment aims to demonstrate that humans may not be performing a feature extraction on the vocoded speech sounds, and that comprehending "degraded" speech may be a skill that allows transfer of comprehension ability from one form to another. Such results support one hypothesis of Specific Aim #3, which addresses whether or not there is insight at the perceptual level. Sinewave and Noise Vocoded Speech are, by definition, modified speech that are difficult to perceive (comprehend). Transferring perceptual ability (i.e., learning Sinewave Speech but increasing perceptual ability on Noise Vocoded Speech) would lend support for the hypothesis associated with Specific Aim #2.

In natural human speech, sounds are formed by passing air from the lungs through the vocal cords, which induce vibration at certain frequencies, and finally through the throat and nasal passages; vibrating air is further modulated by movement of oral features such as the

tongue, jaw, and lips (Stevens, 2000). These features can be modeled in combination to produce synthesized speech, but these anatomical features can also be used to infer auditory features that are important in the comprehension of human speech.

These models, however, beg the question of what features are necessary in a model for auditory perception and comprehension of language. Noise Vocoded Speech (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995), for example, was developed as a model of cochlear implants and meant to demonstrate how individuals with such implants can hear while missing certain auditory information. In this model, human speech is filtered so as to reduce spectral information while amplitude and temporal cues are preserved: a set of frequency bands are isolated and the temporal envelope is extracted. A low-pass filter, around 20 Hz, is applied before carrier-band noise modulation to smooth the resulting audio. Shannon, Zeng, Kamath, Wygonski, and Ekelid (1995) found that presenting only the amplitude and temporal pattern in as few as three spectral regions was sufficient for listeners to recognize and understand what was spoken.

In Sinewave Speech, human speech is filtered by the generation of a set of temporally dynamic sinusoids; these sinusoids follow formant center frequencies in a moving-window format (Remez, Rubin, Pisoni, & Carrell, 1981). In practice, the first three to five formants are converted into sinusoids, leaving the remaining higher formants unmodeled. According to participants during pilot testing for this experiment, the resulting processed speech sounds like "computers beeping at each other".

With repeated exposure, individuals are able to learn and decipher both Noise Vocoded Speech and Sinewave Speech (Remez, Rubin, Pisoni, & Carrell, 1981; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Sheffert, Pisoni, Fellowes, & Remez, 2001). Furthermore, both normal hearers and individuals with cochlear implants have been found to learn and understand various speech processors (Friesen, Shannon, Baskent, & Wang, 2001). Given such evidence, it is reasonable to ask whether it is possible to transfer skill from one form of degraded speech to another, even though specific auditory cues may be missing in Sinewave or Shannon Speech. It is reasonable to hypothesize that perceptual skill is generally transferrable, and that the ability to do so may be dependent on attentional and working memory mechanisms.

Theoretically, these two forms of degraded vocoded speech present very different information to the listener. Each have been demonstrated to be difficult to understand, but can be understood with learning through repeated exposure. Crucially, it has not been demonstrated whether listeners can transfer what they have learned in perceiving and comprehending one type of degraded vocoded speech to the other type. In this study, Sinewave Speech and Noise Vocoded Speech are used to determine whether transfer of learning can occur between the two.

Based upon a Pretest-Training-Posttest paradigm, the scientific question of whether transfer of learning exists can be tested with two hypotheses. The first hypothesis is that training on Noise Vocoded Speech will result in increased test performance on Sinewave Speech. Experimentally, these participants would be compared against participants who were trained and tested on Sinewave Speech (i.e., no transfer) and against participants who received no training before test on Sinewave Speech. In comparison, individuals who receive Plain (unmodified)

Speech training should perform worse than those trained with Sinewave Speech or Noise Vocoded Speech; this condition would serve as a control for mere-exposure effects.

Insight, in the form of transfer, can reasonably be hypothesized to be found elsewhere as well. A second hypothesis is that training on Sinewave Speech will result in increased performance in perceiving and comprehending Noise Vocoded Speech. While this hypothesis appears to test the same scientific question, it provides evidence that learning is not dependent on the specific form of degraded speech, and that the results from testing the first hypothesis are not due to peculiarities of the chosen stimuli. Thus, if skill can be transferred, participants who are trained on Noise Vocoded Speech and tested on Sinewave Speech would be predicted to perform similarly to participants who are trained and tested both on Sinewave Speech. A prediction can be made that when comparing perception and comprehension on Noise Vocoded Speech, participants who receive Sinewave Speech at both Pretest and Training will perform better than participants who receive no training in any type of degraded vocoded speech.

The research question, as well as the two hypotheses, is meant to test transfer as perceptual insight, in support of testing the hypotheses associated with Specific Aim #2. As discussed, these stimuli are not immediately recognizable as human speech. Yet they are derived from recordings of speech and can be perceived as speech with some instruction or practice. The question remains, however, whether the insight into understanding these forms of degraded vocoded speech is domain specific, or is a generalizable skill or ability (Specific Aim #1).

Study #6 Methods

A male voice actor was recorded in a sound-attenuated room, instructed to use an even unemotional tone while reading aloud a set phonetically-balanced word lists (Egan 1948). The actor was further instructed to pace his speech by a count to five between each word spoken in order to minimize effects of co-articulation. Egan's lists 1-5 were recorded but List 3 was not used due to background noise interruptions during recording; only lists 1, 2, 4, and 5 are used as stimuli in this study. The recorded spoken words were then converted into degraded "Sinewave Speech" using the PRAAT software package (Boersma, 2002; Darwin 2005) using three frequency bands converted to sinusoidal waves as described by Remez, Rubin, Pisoni, and Carrell (1981). Nonsense lure items were created by rank-ordering individual word recordings by length (number of samples) and swapping sinusoidal waves with nearest neighbors in the length-rank-order list. No nonsense-lure item contained more than one sinusoid wave from the same word recording. Two independent raters with experience listening to Sinewave Speech were unable to identify words contained in the nonsense-lure items.

The same word recordings were also converted into Noise Vocoded Speech by separating each word recording into four frequency bands and filtered with the PRAAT software package (Darwin, 2005) using the method described by Shannon, Zeng, Kamath, Wygonski, & Ekelid (1995). Nonsense lure items were generated by rank-ordering the audio files by time length and swapping frequency bands with nearest neighbors in the rank-order list. No lure item contained more than one frequency band from the same word recording. The same two independent raters who rated the Sinewave Speech were also experienced in listening to Noise Vocoded Speech;

they were similarly unable to name the words that were used to generate the nonsense Noise Vocoded Speech lure items.

The experiment consisted of one 1.5 hour experiment session. After agreeing to informed consent, participants were asked to complete a basic demographic form, a set of questions probing sleep quality and times, the Affect Grid (Russell, Weiss, & Mendelsohn, 1989), and the Stanford Sleepiness Scale (Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973). Experiment sessions were scheduled 3-7 days in advance, allowing time for participants to complete a sleep log in the intervening nights. Participants were asked to turn in their sleep logs during the experiment session.

Auditory stimuli were presented electronically using the E-Prime 2.0 software package (Psychology Software Tools, Pittsburgh, PA) on a Windows-based desktop computer equipped with a Sound Blaster Audigy 4 audio card. Audio was presented using Sennheiser HD 280 Pro closed-back headphones in a sound-dampened room. Stimuli were presented at 70dB RMS at the headphones, as measured using a coupler and handheld sound level meter, and adjusted with a Samson S-Amp headphone amplifier. Visual stimuli were presented using a 19" NEC L195GH monitor. Participants were instructed to sit at a comfortable distance from the monitor; chinrests were not used and visual distance measurements were not taken. Participants made their responses by typing on a standard QWERTY USB keyboard.

The experiment protocol consisted of four blocks: Pretest, Training, Posttest 1, and Posttest 2. Participants were allowed short breaks between each block and given instruction before the beginning of each block.

Pretest consisted of 100 items from two of Egan's (1948) phonetically balanced word lists, with recordings presented in randomized order. In each trial, stimulus presentation was preceded by a short break and a fixation cross indicating that a stimulus was about to be played over the headphones. Immediately after the audio stimulus was played, a text entry box was displayed on-screen with a prompt for the participant to type in the word just heard; the participant was also given the option of pressing a labeled button on the computer keyboard indicating that the participant did not perceive a word. No performance feedback was given to participants during Pretest trials (to minimize learning before the Training block), and no nonsense lure stimuli are given (to minimize confusion).

During the Training block, participants were similarly presented a stimulus word over the headphones before being asked to type in the word that they heard (or a specially-labeled key indicating the participant did not perceive a word). This is the same procedure participants followed in Pretest, except that additionally after responding, participants were shown the stimulus word printed on screen while the clip was replayed over headphones. This word, printed on-screen, was presented for 1000ms, after which the following trial began. This 1-second timespan was selected based on previous pilot testing. The 100 stimuli presented during the Training block consisted of two phonetically-balanced word lists (each containing 50 words) that had not been presented at Pretest for the given participant.

At Posttest, participants were exposed to a block of 100 Sinewave Speech stimuli randomized with 100 nonsense Sinewave Speech stimuli before being exposed to a set of 100 Noise Vocoded Speech stimuli randomized with 100 nonsense Noise Vocoded Speech. For each trial, the stimulus was played over headphones before participants were asked to type in the word that they heard. The addition of nonsense lures at posttest (Posttest 1 and Posttest 2) is unlike Pretest and Training, when no lures were presented. Posttest 1 contained words from one of Egan's (1948) word lists previously presented at Pretest and one list previously presented at Training, for a total of 100 words. Posttest 2 contained words from the remaining list previously presented at Pretest (not presented at Posttest 1) and words from the remaining list previously presented at Training. Stimuli within a given block were presented in randomized order, mixing presentations between paired word lists. Participants were reminded at each block to press the numeric "1" button at the top-left corner of the keyboard to identify stimuli they believed no word was presented[3]. Participants report failing to use the "No Word" button, electing instead to hit the enter key without typing in a response to denote a no-word response. Several participants reported that they typed in "No Word" as a response for nonsense items. During data analysis, blank entries and "No Word" were considered to be responses for no-word items.

For all blocks, response time on each trial was limited to 60 seconds before the next trial began; no participant reported running out of time when typing responses. Each block consisted of the same trial procedure repeated for a full Egan word list. Word lists were not repeated within-subject, and assignment was counterbalanced across participants.

[3] The phrase "NO WORD" was lettered onto a sticker applied directly to the numeric 1 key in the upper-left corner of the keyboard.

Participants were randomly assigned into one of five Conditions as described in Table 13, below.

	Pretest	Training	Posttest 1	Posttest 2
Condition 1	Sinewave	Noise Vocoded	Sinewave	Noise Vocoded
Condition 2	Sinewave	Sinewave	Sinewave	Noise Vocoded
Condition 3	Sinewave	Plain	Sinewave	Noise Vocoded
Condition 4	Reading	Noise Vocoded	Sinewave	Noise Vocoded
Condition 5	Reading	Plain	Sinewave	Noise Vocoded

Table 13: Participants are sorted into five Conditions that differ on the stimulus type presented at each block of the experiment.

In Condition #1, participants receive a Pretest and Posttest consisting of Sinewave Speech, but critically receive Noise Vocoded Speech at Training. Condition #2 provides a control condition in respect to the first hypothesis and serves as a test condition in respect to the second hypothesis; all stimulus exposures in this condition until Posttest consist only of Sinewave Speech. In Condition #3, the Training block consists of listening to Plain (unmodified) Speech. Participants are expected to perform at ceiling during Training, but the Posttests serve as controls. Conditions #4 and #5 provide further experimental control over the mere exposure effect; participants in these two Conditions are asked to read and respond to words on-screen during the Pretest block. Participants in Conditions #4 and #5 are also asked to wear headphones during blocks when no audio stimuli are given to maintain the same experience across all Conditions and to minimize differences in experimenter demand characteristics.

All participants also completed the R-SPAN working memory capacity task (Daneman & Carpenter, 1980; Unsworth, Heitz, Schrock, & Engle, 2005). In the R-SPAN, subjects read sentences presented on a computer screen and made "sense" or "nonsense" judgements on them. After a sentence and sense judgement, a number was presented on screen. Participants were instructed that after a series of sentences and judgements, they were to recall numbers in the order they were presented by entering them into the computer. A block of sentences consisted of 15 items, 3 each consisting of two, three, four, five, and six sentences that were 13–16 words in length, and presented in ascending length order (Conway et al., 2005).

92 participants were recruited from the University of Chicago community and consisted primarily of undergraduate and graduate students, age 18 or older. Of these, 4 were excluded for experiment equipment issues. Of the remaining 88 participants, mean age was 21.6 years (SD = 4.6 years), 45.5% were male (N = 40), 88.6% were enrolled university students (N = 78). The protocol was reviewed and approved by the University of Chicago Independent Review Board (IRB Approval #H11217).

Study #6 Results

The primary questions this study was designed to address are about transfer in perceptual insight. To answer these questions, outside factors such as emotional state, sleepiness, and amount of sleep must first be addressed.

The Affect Grid measures were highly variable; mean Pleasantness was 0.98 (SD = 1.69), while mean Energy was M = 0.17 (SD = 2.01). Both of these values are within one standard deviation of zero, or "neutral". Participants scored a mean M = 2.72 (SD = 1.08) on the Stanford Sleepiness Scale, which is between "Functioning at high levels, but not at peak; able to concentrate" and "Awake, but relaxed; responsive but not fully alert".

Participants also reported sleeping on average M = 7.9 hours a night (SD = 1.2 hours) on an in-lab questionnaire. This is similar to what participants reported in their sleep logs for hours slept the night before: an average of M = 7.5 hours (SD = 1.9 hours). Combined, the questionnaire responses for sleep and affect suggest that participants were neither excessive in affect (nether positive nor negative), nor did participants appear to lack in sleep.

If transfer of learning exists, we predict that training on Noise Vocoded Speech should result in increased test performance on Sinewave Speech. Therefore, it can be predicted that participants in Condition #1, who receive Sinewave Speech Pretest and Noise Vocoded Speech Training, will perform better than participants in Condition #3, who receive Sinewave Speech Pretest and Plain (unmodified) Speech Training. Participants in Condition #1, who received a Sinewave Speech Pretest and Posttest with an intervening Noise Vocoded Speech Training, would also be predicted to perform similarly to participants in Condition #2, who receive Sinewave Speech Pretest, Training, and Posttest. Secondarily, the protocol design also allows a test of whether training on Sinewave Speech will result in increased performance in perceiving and comprehending Noise Vocoded Speech.

Transfer of learning predicts that when comparing perception and comprehension on Noise Vocoded Speech, participants in Condition #2 who receive Sinewave Speech at both Pretest and Training will perform better than participants in Condition #3 who receive no Training in degraded vocoded speech.

Study #6 Results: Correct Identification

In terms of hit rates where participants correctly identify the Sinewave Speech stimulus word, In a one-way ANOVA, Pretest accuracy did not differ significantly at the 0.05 level for Conditions #1, #2, and #3 ($F(2, 50) = 0.167$, $p = 0.846$). Conditions #4 and #5 were excluded because Pretests in those conditions consisted of retyping words presented on-screen, and accuracy rates were near the 100% accuracy ceiling. Including Conditions #4 and #5 results in an expected significant difference ($F(4, 83) = 313.2$, $p < 0.001$). Given that a significant difference at Pretest was only identified with all five conditions and not with Conditions #1-3, as expected, these results suggest that any findings identified in later analyses will not likely be due to initially different Pretest performance among conditions.

Participants in Conditions #1-3 also received different stimuli during Training. This resulted in a significant difference in performance across the three conditions as expected ($F(2, 50) = 766$, $p < 0.001$). Although participants at Training in Condition #1 and Condition #2 received Sinewave and Noise Vocoded Speech, respectively, participants in Condition #3

received Training on Plain (unmodified) Speech, which should have been a familiar form of speech to all participants.

When comparing across all five Conditions, an ANOVA was significant at the 0.05 level for differences in correct response counts at Posttest #1 (which consisted of Sinewave Speech; $F(4, 83) = 4.828$, $p = 0.002$; Figure 18). The difference appears to be due to the performance of participants in Condition #2, who responded with a mean of 13.5 correct responses (SD = 12.2), while participants in Condition #1 (M = 6.2, SD = 7.4), Condition #3 (M = 5.6, SD = 8.2), Condition #4 (M = 2.4, SD = 3.0), and Condition #5 (M = 2.8, SD = 6.7) did not perform as well.

.

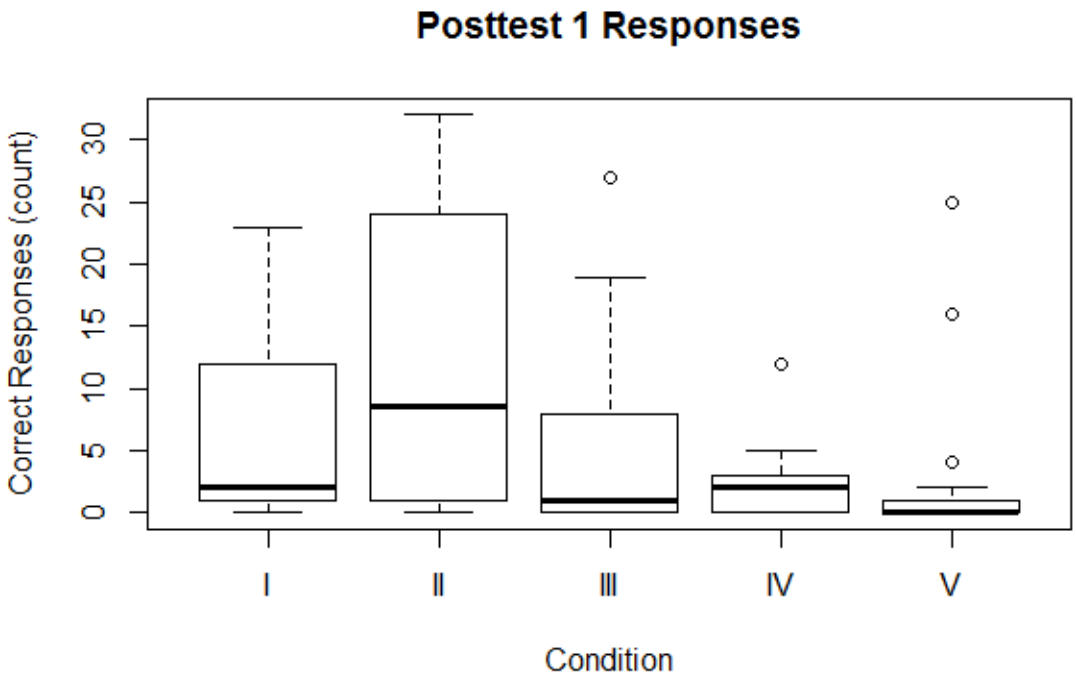

Figure 18: Response Accuracy by Experiment Condition at Posttest #1; numbers are out of 100 total trials.

A d-prime statistic for Posttest #1 results scored by correct word responses was not calculated because Signal Detection Theory does not account for differences in inaccurate versus blank responses.

An ANOVA across all five conditions was also significant for correct responses at the 0.05 level for Posttest #2, which consists of Noise Vocoded Speech stimuli ($F(4, 83) = 9.85$, $p < 0.001$; Figure 19). This appears to be due to performance by participants in Condition #1 ($M = 10.3$, $SD = 6.0$) compared to performance in Condition #2 ($M = 4.6$, $SD = 3.5$), Condition #2 ($M = 4.2$, $SD = 4.3$), Condition #3 ($M = 4.2$, $SD = 4.3$), Condition #4 ($M = 7.7$, $SD = 6.0$), or Condition #5 ($M = 1.4$, $SD = 1.9$).

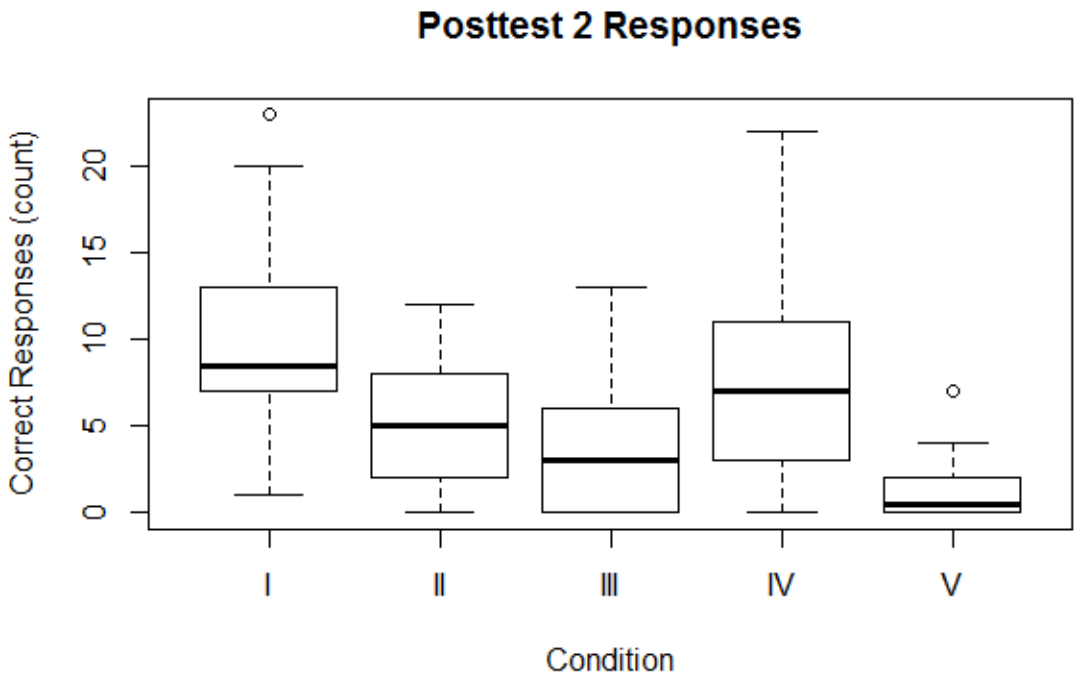

Figure 19: Response Accuracy by Experiment Condition at Posttest #2. Correct Response values are out of 100 total trials.

Given that scores at Posttest #1 and Posttest #2 were significantly different across Conditions, the question remains of what changes might have occurred between Pretest and Posttest. When Pretest and Posttest #1 correct-response scores were compared in a mixed design ANOVA, a significant interaction was identified between Pretest-Posttest #1 (difference score) and Condition ($F(4, 164) = 94.05$, $p < 0.001$), as well as main effects of Test (comparing Pretest and Posttest #1 accuracy, $F(1, 164) = 108.53$, $p < = 0.001$) and Condition ($F(4,164) = 96.42$, $p < 0.001$; all five conditions included). A mixed-design ANOVA for Pretest and Posttest #2 found an interaction with Condition and Test significant at the 0.05 level ($F(4, 164) = 122.3$, $p < 0.001$), as well as significant main effects of Condition ($F(1, 164) = 113.3$, $p < 0.001$; all five conditions included) and Test (comparing Pretest and Posttest #2; $F(4, 164) = 95.4$, $p < 0.001$). These results are as expected given the inclusion of Conditions #4 and #5, where Pretest consisted of reading instead of exposure to auditory stimuli.

Given this, a specific test of Conditions #1-3 is warranted. The first hypothesis about learning transfer predicted that at Posttest 1, participants in Condition #1 would perform better than participants in Condition #3, because Condition #1 participants are able to transfer learning from Noise Vocoded Speech to Sinewave Speech. The correct-response count for Pretest blocks were scaled by 50% in order to make a comparison between the 100 items presented at Pretest and the 50 items previously seen at Pretest that were shown at Posttest #1. A delta change was then calculated based on these scaled values. A significant difference was found for these values by Condition ($F(4, 83) = 353.6$, $p < 0.001$; Figure 20) that remained even when Conditions #4 and #5 were excluded (because their Pretests were not Sinewave Speech; $F(2, 50) = 5.471$, $p = 0.007$). This appears to be due to Condition #2, which participants on average responded to a

(scaled) 3.3 more items correctly (SD = 3.4) than Condition #1 (M = 0.3, SD = 3.4) or Condition #3 (M = 0.0, SD = 1.6).

Change in performance between Pretest and Posttest #1 was significantly different across Conditions ($F(2, 50) = 347.5$, $p < 0.001$; Figure 20). This difference remained even when Conditions #4 and #5 were excluded (because their Pretests were not Sinewave Speech; $F(2, 50) = 8.377$, $p < 0.001$). This appears to be due to Condition #2, which participants on average responded to a (scaled) 9.4 more items correctly (SD = 8.4) than Condition #1 (M = 0.8, SD = 7.4) or Condition #3 (M = 1.3, SD = 3.6).

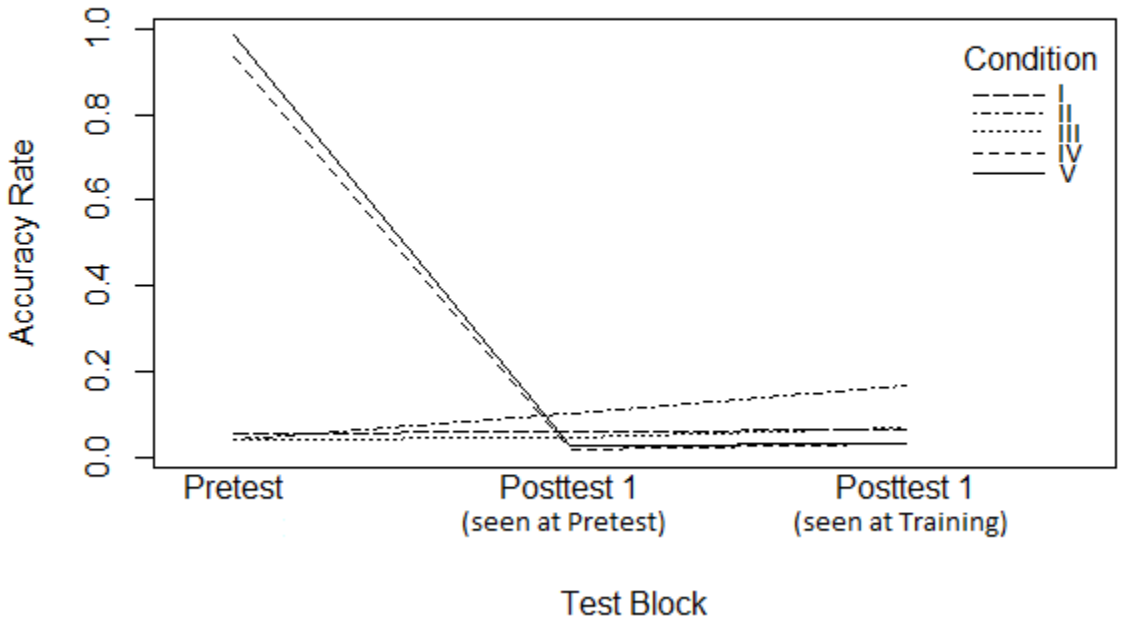

Figure 20: Accuracy Rates at Pretest and Posttest #1, by Condition.

Does Sinewave Speech performance at Posttest #1 demonstrate learning? In other words, was there evidence of transfer between the speech types? When comparing the change in

accuracy between Pretest and Posttest #1 items seen at Training but not in Pretest, a significant difference was found for these values by Condition ($F(4, 83) = 320.1$, $p < 0.001$) that remained even when Conditions #4 and #5 were excluded (because their Trainings were not Sinewave Speech; $F(2, 50) = 9.046$, $p < 0.001$). This again appears to be due to Condition #2, which participants on average responded to a (scaled) 6.4 more items correctly (SD = 5.8) than Condition #1 (M = 0.5, SD = 4.3) or Condition #3 (M = 1.3, SD = 2.4).

The second hypothesis predicted that Condition #2 participants will perform as well as participants in Condition #1 in regard to Posttest #2 performance, and participants in both of those conditions should perform better than those in Condition #3, as learning generalizes or transfers between the vocoded speech types. That is, participants who received Sinewave Speech at Pretest and received Noise Vocoded Speech at Training (Condition #1) should perform as well as participants who received Sinewave Speech at both Pretest and Training (Condition #2); both of these conditions should reflect better performance on Noise Vocoded Speech at Posttest than participants who only heard Plain (unmodified) Speech at Training (Condition #3).

A mixed-design ANOVA did not identify a significant interaction between Conditions #1-3 and Pretest-Posttest #2 ($F(2, 98) = 2.009$, $p = 0.140$). The main effect of Test was not significant ($F(1, 98) = 0.203$, $p = 0.653$), but the main effect of Condition was significant at the 0.05 level ($F(2, 98) = 4.841$, $p = 0.010$). Participants in Condition #2 made on average 4.64 (SD = 3.54) correct responses at Posttest #2, which is lower than Condition #1 (M = 10.28, SD = 5.99) and higher than Condition #3 (M = 4.19, SD = 4.32).

Does performance on Noise Vocoded Speech at Posttest #2 demonstrate insightful transfer? When comparing stimuli at Posttest #2 which were previously presented at Pretest or previously presented at Training, there was no significant interaction between these two categories of stimuli and Condition ($F(4, 164) = 1.054$, $p = 0.381$), and there was no main effect between Posttest stimuli previously presented at Pretest or Training ($F(1, 164) = 1.680$, $p = 0.197$). However, a significant main effect was found for Condition ($F(4, 164) = 14.540$, $p < 0.001$), also as expected.

Change in performance between Pretest and Posttest #2 performance (

Figure 21) was not significantly different across Conditions #1-3 ($F(2, 50) = 2.928$, $p = 0.063$), with Condition #1 participants identifying an average of 4.9 more words ($SD = 6.1$) than Condition 2 ($M = 0.6$, $SD = 4.5$) or Condition 3 ($M = -0.1$, $SD = 8.5$). No significant difference was identified for Posttest #2 stimuli previously seen at Training ($F(2, 50) = 6.138$, $p = 0.004$), nor for stimuli previously seen at Pretest ($F(2, 50) = 2.505$, 0.092).

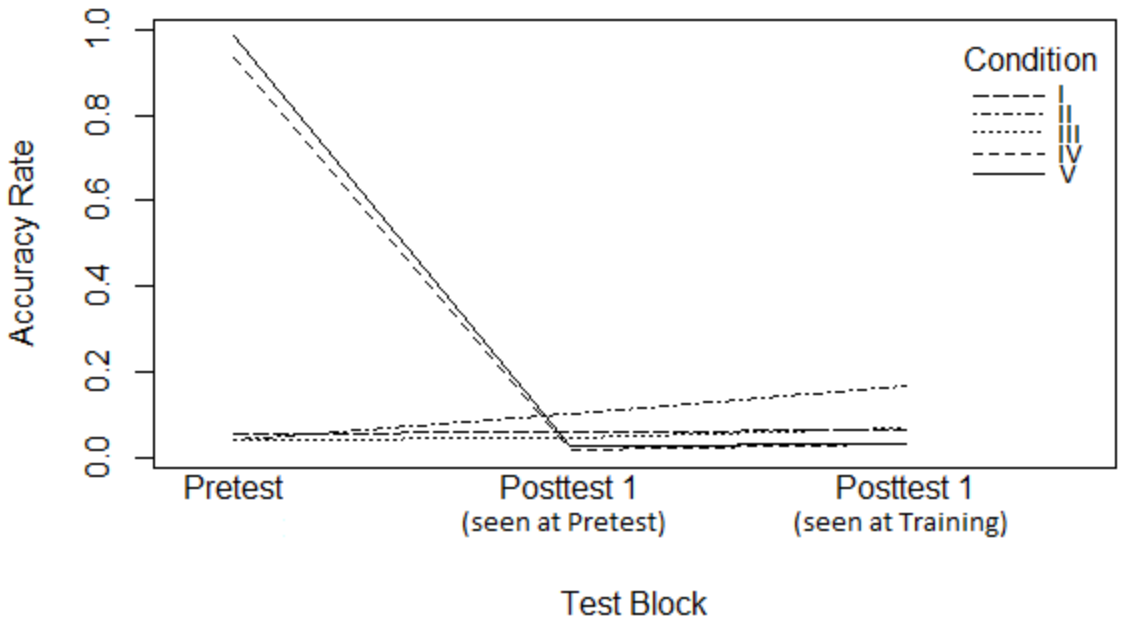

Figure 21: Accuracy Rates at Pretest and Posttest #2, by Condition.

When comparing the change in accuracy between Pretest and Posttest 2 items seen at Pretest, the correct-response count for Pretest blocks were scaled by 50% in order to make a comparison between the 100 items presented at Pretest and the 50 items previously seen at Pretest that were shown at Posttest 2. A delta change was then calculated based on these scaled values. A significant difference was found for these values by Condition ($F(4, 83) = 303.0$, $p < 0.001$) that was not significant when Conditions #4 and #5 were excluded (because their Pretests were not Sinewave Speech; $F(2, 50) = 1.485$, $p = 0.236$).

When comparing the change in accuracy between Pretest and Posttest #2 items seen at Training, a significant difference was found for these values by Condition ($F(4, 83) = 288.3$, $p <$

0.001) that remained even when Conditions #4 and #5 were excluded (because their Trainings were not Sinewave Speech; $F(2, 50) = 3.845$, $p = 0.028$). This again appears to be due to Condition #1, which participants on average responded to a (scaled) 3.4 more items correctly (SD = 3.9) than Condition #2 (M = 0.8, SD = 2.7) or Condition #3 (M = 0.0, SD = 4.6).

Study #6 Results: Signal Detection

In addition to comprehension of speech, the responses can be interpreted in terms of detection of signal from noise regardless of correct identification of the word. Based on whether a stimulus was a word or not, could individuals identify word stimuli from nonword stimuli?

First, did participants randomized into the five experiment conditions perform as expected at Pretest? Differences in Pretest correct responses among the five conditions were significant at the 0.05 level ($F(4, 83) = 7.822$, $p < 0.001$), as expected due to less-than-perfect performance interpreting Sinewave Speech for Conditions #1-3 and performance at-ceiling for Conditions #4 and #5 when participants simply had to type in the words printed on the computer screen,. However, Pretest differences among Conditions #1-3 were not significant at the 0.05 level ($F(2, 50) = 0.362$, $p = 0.698$), also as expected.

Differences in response accuracy at Training between Conditions #1-3 were significantly different ($F(2, 50) = 3.649$, $p = 0.033$), and as expected appears to be due to participants in Condition 3 performing near ceiling (M = 99.2, SD = 0.7), compared to Condition #1 Training

on Noise Vocoded Speech (M = 85.8, SD = 23.3) or Condition #2 Training on Sinewave Speech (M = 82.6, SD = 28.3). Posttest #1 d-prime scores were not significantly different across the 5 conditions ($F(4, 83) = 0.459$, SD = 0.765), nor as were Posttest #2 d-prime scores ($F(2, 83) = 0.378$, $p = 0.824$).

A mixed-model ANOVA found a significant interaction between Condition and Pretest-Posttest #1 ($F(4, 164) = 5.603$, $p < 0.001$), as well as a significant main effect for Condition ($F(4, 164) = 2.922$, $p = 0.023$) and Pretest-Posttest #1 ($F(1, 164) = 4.795$, $p < 0.030$) for correct detections. A mixed-model ANOVA also found a significant interaction between Condition and Pretest-Posttest #2 ($F(4, 164) = 8.780$, $p < 0.001$) and a significant main effect of Pretest-Posttest #2 ($F(1, 164) = 3.989$, $p = 0.047$), but not for the main effect of Condition ($F(4, 164) = 1.673$, $p = 0.159$.

The critical test for the first hypothesis, that training on Noise Vocoded Speech will result in increased test performance on Sinewave Speech, is a mixed-design ANOVA between Pretest-Posttest #1 and Conditions #1-3, was not significant for interaction ($F(2, 98) = 0.307$, $p = 0.737$), main effect of Condition (2, 98) = 0.173, $p = 0.842$), or main effect of Pretest - Posttest #1 ($F(1, 98) = 0.048$, $p = 0.827$).

Did learning occur with Sinewave Speech? When comparing stimuli at Posttest #1 which were previously presented at Pretest or previously presented at Training, there was no significant interaction between these two categories of stimuli and Condition ($F(4, 164) = 0.003$, $p = 1.000$), and there was no main effect between Posttest stimuli previously presented at Pretest or Training

($F(1, 164) = 0.054$, $p = 0.0751$). No significant main effect was found for Condition ($4, 164) = 2.166$, $p = 0.075$).

Posttest #1 performance, divided into stimuli previously presented at Pretest or previously presented at Training, was also compared to Pretest performance. When Conditions #4 and #5 are excluded (because performance at pretest in those Conditions was near-ceiling), no significant interaction was found among Posttest #1 items previously presented at Pretest, Posttest #1 items previously presented at Training, and Pretest ($F(4, 147) = 0.251$, $p = 0.909$). A significant main effect was found among test block ($F(2, 147) = 20.186$, $p < 0.001$), but not found for Condition ($F(2, 147) = 0.233$, $p = 0.793$).

When comparing the change in accuracy between Pretest and Posttest #1 items seen at Pretest (Figure 22), the correct-response count for Pretest blocks were scaled by 50% in order to make a comparison between the 100 items presented at Pretest and the 50 items previously seen at Pretest that were shown at Posttest #1. A delta change was then calculated based on these scaled values. A significant difference was found for these values by Condition ($F(4, 83) = 12.3$, $p < 0.001$) that was no longer significant at the 0.05 level when Conditions #4 and #5 were excluded (because their Pretests were not Sinewave Speech; $F(2, 50) = 0.575$, $p = 0.566$).

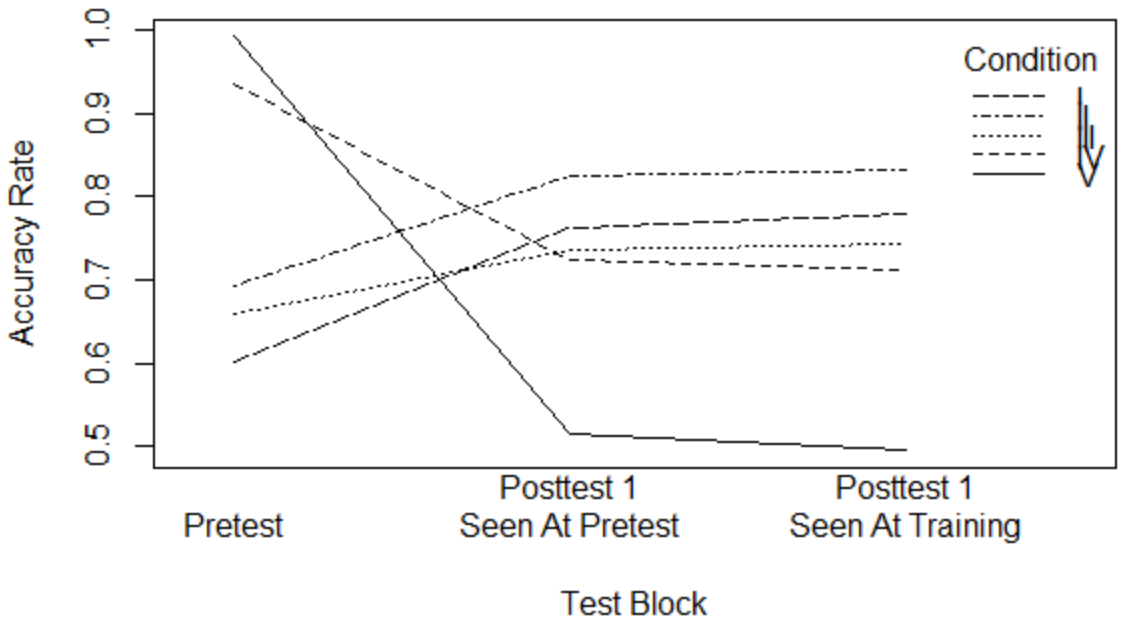

Figure 22: Accuracy Rates at Posttest #1

When comparing the change in accuracy between Pretest and Posttest #1 items seen at Training, a significant difference was found for these values by Condition ($F(4, 83) = 13.66$, $p < 0.001$) that was no longer significant at the 0.05 level when Conditions 4 and 5 were excluded (because their Pretest stimuli were not Sinewave Speech; $F(2, 50) = 0.769$, $p = 0.469$).

The critical test for the second hypothesis, that training on Sinewave Speech will result in increased performance in perceiving and comprehending Noise Vocoded Speech), is a mixed-design ANOVA between Pretest-Posttest #1 and Conditions #1-3, was not significant for interaction ($F(2, 98) = 1.348$, $p = 0.264$), main effect of Condition (2, 98) = 0.216, $p = 0.806$), or main effect of Pretest-Posttest #1 ($F(1, 98) = 0.159$, $p = 0.691$).

Change in performance between Pretest and Posttest #2 was not significantly different across Conditions #1-3 ($F(2, 50) = 1.855$, $p = 0.162$). No significant difference was identified for Posttest #1 stimuli previously seen at Pretest ($F(2, 50) = 1.592$, 0.214), nor for stimuli previously seen at Training ($F(2, 50) = 1.499$, $p = 0.233$).

When comparing the change in accuracy between Pretest and Posttest #2 items seen at Pretest (Figure 23 and Figure 24), the correct-response count for Pretest blocks were scaled by 50% in order to make a comparison between the 100 items presented at Pretest and the 50 items previously seen at Pretest that were shown at Posttest #2. A delta change was then calculated based on these scaled values. A significant difference was found for these values by Condition ($F(4, 83) = 12.150$, $p < 0.001$) that was not significant when Conditions #4 and #5 were excluded (because their Pretest stimuli were not Sinewave Speech; $F(2, 50) = 2.699$, $p = 0.077$).

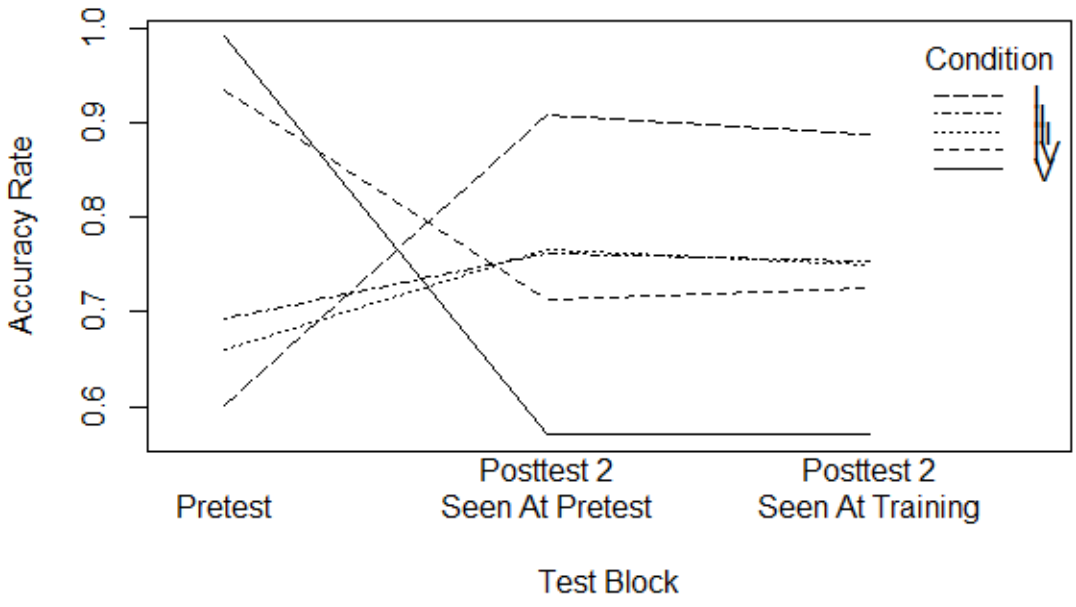

Figure 23: Accuracy Rates at Posttest #2

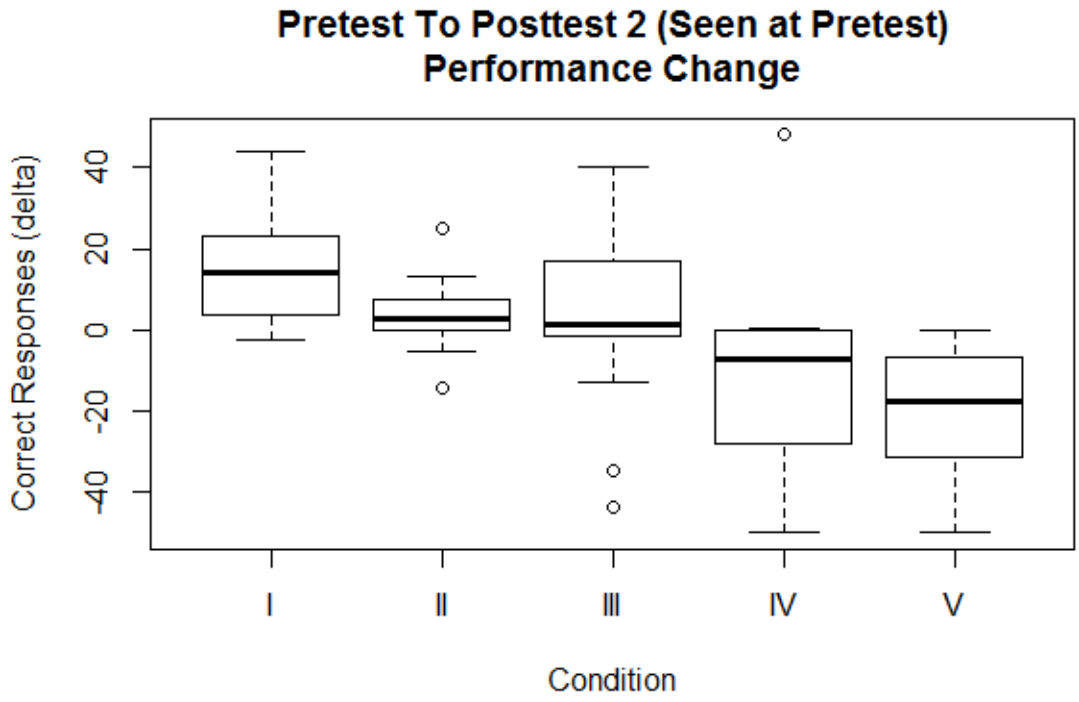

Figure 24: Pretest to Posttest 2 (Seen at Pretest) Performance Change

When comparing the change in accuracy between Pretest and Posttest 2 items seen at Training, a significant difference was found for these values by Condition ($F(4, 83) = 10.940$, $p < 0.001$). This appears to be due to Conditions #4 and #5, in which performance decreased at Posttest #2 ($M = -11.1$, $SD = 21.9$ and $M = -21.1$, $SD = 16.6$, respectively) but increased in Conditions #1-3 ($M = 15.3$, $SD = 14.1$; $M = 3.5$, $SD = 9.1$; and $M = 5.3$, $SD = 20.7$, respectively). When Conditions #4 and #5 were excluded, the difference among Conditions #1-3 was not significant at the 0.05 level ($F(2, 50) = 2.699$, $p = 0.078$).

The secondary question of working memory capacity was tested under a hypothesis of whether performance in identifying the stimuli correlated with working memory capacity. R-

SPAN Span Score did not differ across condition ($F(4, 48) = 1.676$, $p = 0.171$), nor did R-SPAN Total Score ($F(4, 48) = 1.362$, $p = 0.261$), indicating no evidence of significant between-subjects differences. For Pretest and Training, correlations (Spearman's rank-order r) were calculated between R-SPAN Total score and correct response counts. Because the posttest blocks contain lure items, a d-prime score can be calculated for those blocks. For Posttest #1 and Posttest #2, Spearman's rank-order correlations were calculated between d-primes and R-SPAN Total scores. Correlation coefficients were also calculated for participant responses when scored for correct word response accuracy and for detection of word (

Table 14).

	R-SPAN Score: Correct Word				R-SPAN Total Score: Correct Word			
	Pretest	Training	Posttest 1	Posttest 2	Pretest	Training	Posttest 1	Posttest 2
Condition 1	-0.374	**0.417**	-0.124	-0.140	0.372	0.105	0.130	0.090
Condition 2	-0.076	-0.024	-0.024	0.036	0.317	**0.850**	**0.838**	**0.771**
Condition 3	-0.305	0.526	-0.143	0.086	-0.350	0.535	-0.346	0.185
Condition 4	-0.207	0.714	0.525	0.257	0.207	0.714	-0.031	0.257
Condition 5	-0.025	-0.224	-0.055	-0.104	0.027	-0.091	-0.015	-0.022

	R-SPAN Score: Signal Detection				R-SPAN Total Score: Signal Detection			
	Pretest	Training	Posttest 1	Posttest 2	Pretest	Training	Posttest 1	Posttest 2
Condition 1	-0.096	0.231	-0.183	-0.112	**0.661**	0.166	0.099	0.034
Condition 2	0.429	0.171	0.216	0.119	0.690	**0.781**	**0.886**	0.690
Condition 3	0.170	0.535	0.108	0.331	0.270	0.526	-0.014	0.225
Condition 4	0.000	0.371	-0.116	0.143	-0.207	0.543	-0.203	-0.029
Condition 5	-0.086	-0.310	-0.458	-0.373	-0.097	-0.100	**-0.542**	-0.454

Table 14: Spearman's r for correlations between correct response counts and R-SPAN Total scores. Items in bold are significant at the 0.05 level. Values are not corrected for multiple comparisons.

The R-SPAN Total Score relationships at Posttest do not appear to be related to whether the stimuli were originally presented at Pretest or at Training (Table 15).

	R-SPAN Total Score: Correct Word				R-SPAN Total Score: Signal Detection			
	Posttest 1 Seen At Pretest	Posttest 1 Seen At Training	Posttest 2 Seen At Pretest	Posttest 2 Seen At Training	Posttest 1 Seen At Pretest	Posttest 1 Seen At Training	Posttest 2 Seen At Pretest	Posttest 2 Seen At Training
Condition 1	0.562	0.499	0.198	-0.374	**0.624**	**0.637**	**0.636**	0.420
Condition 2	**0.792**	**0.886**	0.602	**0.771**	**0.913**	**0.862**	**0.802**	0.639
Condition 3	-0.135	-0.350	0.129	0.197	0.130	0.000	0.041	0.342
Condition 4	0.123	-0.247	**0.870**	0.203	-0.203	-0.088	0.029	0.086
Condition 5	-0.126	0.043	0.073	-0.237	**-0.528**	-0.504	-0.437	-0.410

Table 15: Spearman's r for correlations between correct response counts and R-SPAN Total scores, Posttest stimuli divided by original presentation block. Items in bold are significant at the 0.05 level. Values are not corrected for multiple comparisons.

Discussion

The results of this study are inconclusive. Several factors stemming from demographics and experiment design did not appear to be the source of statistical variance. Participants appeared to be well-slept and reasonably alert given responses on questionnaires. Yet when performance is examined by whether participants were able to name the correct word utterance or were able to detect when stimuli contained words, the results overall do not appear to provide evidence of skill transfer. Maybe there is not transfer of insight.

At Posttest 1, participants in Condition #2 (who are trained and tested on Sinewave Speech) perform best, though the differences among Conditions #1-3 were not statistically significant; these are null results regarding the main hypothesis. Participants in Condition #2 were exposed only to Sinewave Speech and their learning would not demonstrate insightful

transfer. Participants in Condition #1, who are trained on Noise Vocoded Speech but tested on Sinewave Speech, were hypothesized to demonstrate insightful transfer. However, these participants appear to perform comparably to participants in Condition #3. The participants in Condition 3 received Plain (unmodified spoken) Speech at training but are tested on Sinewave Speech; these participants were not expected to demonstrate any insightful transfer as they did not receive performance feedback, only auditory exposure to the Sinewave Speech stimuli. Interestingly, this pattern holds whether data are examined for correct-word responses or for word-nonword stimulus detection.

At Posttest 2, which consisted of Noise Vocoded Speech stimuli and when results were scored for correct-word accuracy, participants in Condition #1 (who were trained on Noise Vocoded Speech) performed the best when Posttested on Noise Vocoded Speech, again as expected, as Training and Posttest #2 for these participants consisted of Noise Vocoded Speech. Critically, however, participants in Condition #2 (who were trained on Sinewave Speech) performed worse at the Noise Vocoded Speech Posttest. When results were scored for word-nonword detection, there was no significant difference across Conditions.

Similarly, when performance was measured as a difference between Pretest and Posttests, the data did not provide evidence of an insightful transfer of learning between the two forms of degraded speech – whether the stimuli were Sinewave Speech or Noise Vocoded Speech.

While these results do not demonstrate transfer of skill across forms of degraded vocoded speech, correlative results regarding R-SPAN suggest a working memory mechanism at play,

which reflects the findings of Chapter Two. Correlations between R-SPAN Total Score and signal detection at Training, Posttest #1, and Posttest #2 blocks suggest a positive relationship between learning degraded vocoded speech and working memory capacity. Participants who do best in these blocks are those with the highest R-SPAN Total Scores. The lack of a statistically significant relationship at Pretest also makes sense; participants are not given feedback at this point of the experiment, and are responding without any guidance. Interestingly, working memory capacity seems to make a difference for learning specifically, not for how good someone naturally is at this task. Interestingly, in Condition #2, the positive relationship extends into Posttest 2. Participants in Posttest 2, Condition #2, who did best at naming the Noise Vocoded stimuli had the best R-SPAN Total Scores.

When Posttest stimuli were divided by when they were first presented to participants, the same trend continued: Condition #2 participant performance correlated with R-SPAN Total Scores. However, the relationships were not limited to novel (presented without feedback at Pretest) or repeated (presented with feedback at Training) words; stimuli in both groups reflect a positive correlation with R-SPAN Total Score across participants. While insightful transfer (as tested in this chapter) was not tested with items like the Remote Associates Task, Number Reduction Task, Camouflage Perceptual Fluency task, or Functional Fixedness, a hypothesis can be made that this skill transfer may be related to the Number Reduction Task because that task was also identified to be related to working memory capacity (cf. Specific Aim #1).

Although this study did not clearly demonstrate transfer of learning in the domain of degraded vocoded speech, the results do suggest a cognitive mechanism that is involved in

learning this type of auditory stimuli, which Specific Aim #2 asks about perceptual behavior. The experiment, as designed, has several shortcomings that can be addressed in a subsequent study. For example, the stimuli consisted of single word utterances. These stimuli are short, and their presentation – even with a fixation cross presented on-screen to alert participants to a new trial – may be too short. One possibility may be to allow participants to replay the stimuli. Another is to use longer stimuli, such as sentences. A second issue with the stimuli is that they are from Egan's (1948) phonetically balanced word lists. These words are designed primarily to address the problem of phonetic frequency, but leave open the problem of the distribution of these words in modern common usage. That is, whether some of these words are used more frequently than others was not accounted for in this design.

The working memory capacity component of this study (R-SPAN) suffers from low participant numbers, as this component of the study was added after the initial round of data had been collected. If a Bonferroni correction for multiple tests is applied for 48 independent correlations, none of the correlations significant at the 0.05 level would pass the new significance threshold. However, the results do suggest the importance of working memory capacity in successful feedback training and subsequent performance.

Given the difficulties in creating balanced unbiased word stimuli lists in this experiment, it is reasonable to question how well Egan (1948) normed the word lists that were used here. Egan designed the stimuli to be phonetically balanced, and the Pretest-Training-Posttest design of this experiment was designed on the assumption that participants were listening to entire words without cueing on specific auditory cues present in the modified vocoded speech. But

what about emotional balance? Some of the words in the list might be interpreted on average as happy and positive, while others might be interpreted as sad and negative. To the extent that affective valence may change the arousal of the listener, this could affect insight. As noted earlier, positive affective states can improve problem solving. How does arousal affect the probability of insight?

CHAPTER FIVE: EMOTIONAL PROCESSING AND VISUOPERCEPTUAL INSIGHT MECHANISMS

The description of insight includes an increase in positive affect—people feel good from achieving an insight. Furthermore, positive mood improves performance on problem solving. This linkage suggests there may be a very general affective influence on problem solving. However, does this same affective influence affect the probability of insights? From the individual's perspective, a flash of insight is a positive, arousing experience; having a flash of insight means a solution was found for a vexing problem, after all. We might experience a quickened heartbeat due to excitement (Wulfert, Roland, Hartley, Wang, & Franco, 2005) or because we are happy or angry (Schwartz, Weinberger, & Singer, 1981). Such claims of emotional and physiological reactions have been grounded in recent research, as emotion and mood have been found to facilitate insight in various laboratory protocols. Study #7 was designed to explore the possibility that positive affective arousal, as contrasted with negative arousal, may benefit insights.

Subramaniam, Kounios, Parrish, and Jung-Beeman (2008), for instance, found that a region of the anterior cingulate cortex normally associated with conflict detection and decision-making was sensitive to positive moods and biased individuals to find insightful novel solutions over other kinds of solutions. Positive affect, induced by watching a few minutes of a comedy film, has also been demonstrated to increase the rate of insightful solutions on Functional

Fixedness problems (Isen, Daubman, & Nowicki, 1987). Positive emotions have also been demonstrated to affect attentional processes, changing the scope of responses on a figure-similarity rating task (Fredrickson & Branigan, 2005).

Positive mood in particular has been shown to increase performance in analogical reasoning problems (Jausovec, 1989), in an Alternative Uses task (Zenasni & Lubart, 2002), and on other similar verbal tests of creativity (Abele-Brehm, 1992; Vosburg, 1998). Participants in a positive mood have also been found to complete more Remote Associates Task problems than those who were in other moods (Estrada, Isen, & Young, 1994; Rowe, Hirsch, & Anderson, 2007). Similarly, participants who watched a positive emotional film clip were subsequently found to solve more Functional Fixedness problems than those who watched emotionally negative or neutral film clips (Isen, Daubman, & Nowicki, 1987). While existing mood has been demonstrated to affect performance on these tasks, it is not clear from the existing evidence that emotional content endogenous to the sensory stimuli input might affect insightful problem solving (a question specified in Specific Aim #3).

Emotional content has been found to support memory recall mechanisms on which insight may depend. For instance, several studies suggest that emotionally salient materials are remembered better, whether they are humorous in nature (Chambers & Payne, 2014), emotionally positive (Boucher & Osgood, 1969; Lang, Dhillon, & Dong, 1995), negative (Payne, Stickgold, Swanberg, & Kensinger, 2008; recall after a night of sleep), or emotionally arousing whether positive or negative (Hu et al., 2006; recall after a night of sleep). Similarly, a body of

research exists for speeded recognition of emotional stimuli (Leppänen & Hietanen, 2004; Tracy & Robins, 2008; Kuperman, Estes, Brysbaert, & Warriner, 2014).

Importantly, the effects of mood and insight do not seem to necessarily operate on the immediate timescale; Olton and Johnson (1976) argue that the incubation stage where problem solvers are "stuck" is a defining feature of insight, during which some offline processing occurs to help the individual find a solution. Offline processing has been shown to increase rate of finding inferences (Ellenbogen, Hu, Payne, Titone, & Walker, 2007), and longer delay periods that include sleep have been demonstrated to help consolidate emotional memories (Hu, Stylos-Allan, & Walker, 2006).

Despite this evidence, whether a given individual experiences a flash of insight does not appear to be contingent on an existing positive mood. While positive mood has been found to make insight more likely, existing evidence does not preclude insight from occurring in the absence of a positive mood. While the story of Archimedes' bathwater is anecdotal, some evidence suggests that positive mood and related physiological changes do occur with such moments of insight. For instance, mood is already known to affect memory retrieval. Teasdale and Fogarty (1979) noticed that there were significant latency differences to retrieve pleasant or unpleasant memories. Heart rate changes have been noted in recall of positive versus negative images (Palomba, Angrilli, & Mini, 1997; Vrana, Cuthbert, & Lang, 1989). Increased skin conductance activity has been previously correlated with problem solving engagement (Pecchinenda & Smith, 1996).

While positive mood can underlie changes in insightful behavior, it is not clear from previous work whether emotional state at the "aha!" moment compounds the insight-related heart rate and galvanic skin response increases (i.e., provides contributory input), or if emotional mechanisms do not contribute to insight mechanisms. If an emotional contribution to the insightful "aha!" moment exists, will the physiological size of the "aha!" change with emotion? A null hypothesis would suggest that there is no emotional contribution, that there is no change in physiological magnitude of the "aha!" effect with emotional contribution.

Various effects of the emotional content of stimuli, and fear-relevant stimuli in particular, have been well-studied. Fear-inducing stimuli, for instance, capture attention in the form of facilitated search without explicit attentional control (Öhman, Flykt, & Esteves, 2001). Besides cases of fear-inducing stimuli and visual search, emotional stimuli have also been found to increase Gabor patch contrast sensitivity. Individuals whose attention was cued with a location prime performed better when the prime was that of a frightened face, but not a neutral face or an upside-down face (Phelps, Ling, & Carrasco, 2006) - perhaps suggesting that emotion can facilitate and benefit attention in the context of early visual perception.

Critically, none of these studies directly probe the interaction of insight, perception (Specific Aim #2), and emotion (Specific Aim #3). While evidence suggests that positive mood of a given individual might facilitate insight and contribute to early visual perception processes, it is not clear that positive moods can contribute to insightfulness at the perceptual level. This experiment aims to determine whether emotional content can benefit or otherwise modulate perceptual insight. If emotional content does contribute to insight, then the physiological changes

(i.e., measurable factors such as heart rate and galvanic skin response) due to emotion and insight should be additive. If they are not contributory, or otherwise unrelated, the addition of emotional arousal should not affect the increased arousal found with insight and result in no physiological difference between neutral and arousing instances of insight.

The idea that the autonomic nervous system might affect (or even guide) behavior is not new; the Somatic Marker Hypothesis makes the assertion that certain inputs influence decision-making (Bechara & Damasio, 2005). These markers are theorized to provide a "gut reaction" that can shift attention to change cortical processing resulting in changed problem-solving behavior (Damasio, Tranel, & Damasio, 1991). Emotion has been identified as one signal arising from the body that that can be used to inform decision-making (Damasio, 1996; Hänsel & von Känel, 2008); this will be tested in terms of insightful problem solving by incorporating emotional content into visuoperceptual stimuli. But how can this emotional signal be measured and tested?

One such "somatic marker" is galvanic skin response (GSR) as measured by electrodermal skin conductance (SCR), which Bechara, Damasio, Tranel, and Damasio (1997) identify as producing a measurable and reliable signal of decision-making well before the given individual is consciously aware of the decision. However, skin conductance is a slow response with a response latency on the order of seconds (Boucsein et al., 2012), which may not be sensitive enough to capture an emotional signal. Instead, heart rate variability may be a better measure of autonomic emotional response in an experimental setting.

Heart rate variability (HRV) is the measurable small changes in the heart rate that can be observed and identified as perturbations relative to a steady beat. These perturbations have been identified as one marker of emotion–related arousal in humans (Berntson et al., 1997; Vaschillo et al., 2008). Heart rate is controlled by a balance between sympathetic and parasympathetic influences, with changes in either system resulting in a change in heart rate (Berntson et al., 1997; Thayer & Lane, 2000). The parasympathetic system's influence is exerted during periods of calm and low arousal. Parasympathetic activation exerts effects on heart rate more slowly than sympathetic, with peak effects measured on the order of seconds (Appelhans & Lueckin, 2006). The sympathetic nervous system's influence on heart rate occurs much faster, with a peak effect and return to baseline within 1 second (Berntson et al., 1997; Pumprla, Howorka, Groves, Chester, & Nolan, 2002).

HRV, in turn, is theorized to be the result of oscillations resulting from the two different rates imposed by the sympathetic and parasympathetic nervous systems. The perturbations from a steady beat, observed as oscillations in heart rate over time, can then be measured by electrocardiography using a moving window across time to identify how much the recorded heartbeats vary from the average moving rate (Appelhans & Lueckin, 2006).

Emotional state is reflected in the autonomic nervous system, where emotions can change the balance between the sympathetic and parasympathetic systems (Levenson, 1992), and thereby perturb the heart rate. In an experimental context, this can result in a set of separable states that can be observed by measuring HRV: a higher-emotional state in which the parasympathetic system is exerting control resulting in higher measured HRV, and a lower-

emotional state where the sympathetic system is exerting control, resulting in lower HRV (cf. Porges, 1995, 2001). Thus, the emotional state during the pre-insight period may be measured by using HRV as a proxy (cf. McCraty, Atkinson, Tiller, Rein, & Watkins, 1995; Park, Oh, Noh, Kim, & Kim, 2018). The collected heart rate data may then in turn be examined for evidence of emotional inputs (via autonomic processes) that affect insightful thinking (Specific Aim #3). In practice, HRV is calculated from heart rate data that is gathered using a heart rate monitor. This is convenient for research purposes because it allows emotional state to be measured without directly asking the individual to identify their emotions, and without need to interrupt any emotional experience.

Compared to neutral stimuli, strongly negative or positive camouflage images (i.e., high IAPS Arousal) are hypothesized to elicit an additive physiological effect in heart rate variability if emotional assessment contributes to perceptual insight processing. As a control, unmanipulated novel images would be presented to elicit an emotion-related physiological reaction without insight. Any additive effects of emotion-plus-insight are expected to result in greater heart rate variability than either insight-alone or emotion-alone Conditions.

If emotional processing does not contribute to perceptual insight processing, it can be alternatively hypothesized that there should be no difference in heart rate variability due to insight gained with neutral low-arousing stimuli and insight gained with highly-arousing stimuli whether positive or negative in valence. The strength of insight can also be assessed by self-report. If emotion and insight are intertwined and additive, achievement of insight when

accompanied by emotionally-arousing stimuli should not only result in larger physiological responses, but stronger self-reports of the insight itself.

The experimental task presented here, the Camouflage Affective Response Task (CART), is based on Dallenbach's (1951) stimuli: monochrome images with details removed by Gaussian smoothing so that the images' contents are not readily recognizable without insight. Unlike the Camouflage Perceptual Fluency task presented in Study #2, the purpose of these images is to elicit an affective response, rather than to measure visuoperceptual insight fluency as originally intended by Dallenbach (1951).

Images here ranged the full Valence and Arousal scales of the International Affective Picture System (IAPS; Lang et al., 2008) and included only full-screen images that do not require letter or pillar boxing for display. The full range of IAPS Valence scores indicates that some images in the selected stimuli set will be emotionally negative, some emotionally positive, and some emotionally neutral. The full range of IAPS Arousal scores indicates that some stimuli images will be strongly emotional, with others less-so. Physiological measures of heart rate (to be used in heart rate variability analysis) and respiration were collected to assess the role of emotional state on the reputed affective nature of the insightful "aha!" moment. Galvanic skin response was not collected due to the slow response latency of the measure.

The scientific question of emotional factors in perceptual insight is examined here by a two-day protocol of emotional-camouflage image naming coupled with heart rate monitoring. There are two separate measurements here. It is reasonable to hypothesize that emotional content

of camouflage images will elicit a change in affect, which in turn results in measurable increase in heart rate variability during exposure to those stimuli with stronger emotional content. At the same time, it is hypothesized that participants in this protocol will respond at a higher accuracy rate to these strongly-emotional images. If emotional content does contribute to insight (as measured through heart rate variability), then the effect of heart rate variability due to emotion and insight should be additive. That is, stimuli with higher emotional content should result in higher heart rate variability coupled with higher rate of insight. If they are not contributory, or otherwise unrelated, the addition of emotional arousal should not affect the increased arousal found with successful insight and result in no difference in heart rate variability between neutral and arousing instances of insight.

Study #7 Methods

To assess general emotional state that might affect experimental outcomes, The Profile of Mood States (POMS; Pollock, Cho, Reker, & Volavka, 1979) and Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) were administered. Due to time constraints, emotional state was not assessed with POMS and PANAS after each stimulus presentation.

To test the experimental hypotheses, participants were asked to perceive a set of images that at first glance appear to contain black-and-white blobs scattered about at random (Dallenbach, 1951; Ludmer, Dudai, & Rubin, 2011). After an image was presented, participants

were given the opportunity to respond by giving a short description of what they perceived to be the image's contents.

Camouflage images used in this experiment were based on samples taken from the International Affective Picture System (IAPS; Lang et al., 2008) chosen to represent a cross-section of varying Valence and Arousal scores. In other words, source images ranged positive through negative as well as strongly through weakly emotional. The source IAPS images were then smoothed through a Gaussian filter (sigma = 8 or 16, cutoff = 128, determined for each image through normative testing). Each pixel's RGB channel values were averaged and rounded with equal weight given to each channel. This process resulted in a monochrome black-and-white "camouflage" image (see Figure 25 for an example). Gaussian filter kernel parameters and color-to-monochrome averaging methods used in the Camouflage Perceptual Fluency task are a replication of Ludmer et al. (2011). Image resolution was maintained at 1024 x 768 pixels, as found in the IAPS source, through this entire process. Any images that required letterboxing or pillarboxing for display (i.e., images not originally in the 1024 x 768 aspect ratio) were excluded.

Figure 25: Example IAPS stimulus image 1610 scored positive in Valence, low in Arousal. Less than 100% (but more than 0%) of independent raters recognized the contents of this image during normative testing.

These smoothed monochrome images were then reviewed by three independent raters in a laboratory setting for solvability. Any images where raters were in complete agreement on the content (too easy), were in complete disagreement, or were all incorrect (too difficult) were discarded. Because some combinations of Valence and Arousal values are overrepresented in the IAPS sample, image stimuli were categorized into three sets that balanced on Valence and Arousal values (Table 16 & Table 17). Three sets of 24 images were generated where overall mean Valence score was 5.13 (SD = 2.02); a single factor ANOVA found no significant difference in scores among the three sets ($F(2, 69) = 0.025$, $p = 0.975$). The overall mean Arousal score was 4.83 (SD = 1.13); a single factor ANOVA found no significant difference in scores among the three sets ($F(2, 69) = 1.41$, $p = 0.25$).

	Set 1	Set 2	Set 3
Mean	5.191	5.131	5.059
SD	1.915	2.117	2.095

Table 16: IAPS Valence score means and standard deviations for three stimuli sets. Image sets are generated to represent three different Valence levels.

	Set 1	Set 2	Set 3
Mean	4.545	5.086	4.866
SD	1.237	0.967	1.145

Table 17: IAPS Arousal score means and standard deviations for three stimuli sets used in the experiment.

The selected stimulus images were presented in an experiment via an E-Prime 2.0 computerized task (Psychology Software Tools, Pittsburgh, PA) to participants in a single-session laboratory protocol. This single session was broken into four parts: a Pretest exposure to stimuli, a Feedback exposure to stimuli, a short break, and a posttest exposure to stimuli.

At Pretest, one list of 24 images was presented serially as a randomized-order block with no feedback. Immediately following this was a Feedback presentation of the other list of 24 images presented as a block with performance feedback. This feedback consisted of showing the participant the unaltered original full-color IAPS image for 4.0 seconds which immediately followed the presentation of the black-and-white camouflage stimulus image. The Posttest consisted of 72 images in total; the same image sets from Pretest along with the third set presented as novel images were presented in randomized order. None of the images presented at Posttest were presented with feedback.

After the presentation of a stimulus image, participants were asked whether insight was achieved for the preceding stimulus presentation, and how strong they felt the moment of insight was. Participants were then asked for a short, typed response of what was being depicted in the presented stimulus before continuing to the next image. Participants were told at the beginning of the experiment session to respond quickly and accurately, but the response periods were not deadlined with a time limit. This was intended to avoid penalizing participants who were poor typists. Participants were allowed to make empty responses and continue to the next stimulus by hitting the enter key; this was designed to discourage nonsense or guess responses and to reduce judgement burdens on the independent raters who were later tasked with scoring participant responses.

By dividing the experiment into Pretest and Posttest and including both familiar and novel items at posttest, any physiological changes that accompany insight at posttest can be attributed to either a flash of insight or memory recall. Previous work has demonstrated that when given familiar objects in a difficult-to-recognize format, object perception mechanisms can be trained for better recognition performance (Gauthier, Williams, Tarr, & Tanaka, 1998; Wiesmann & Ishai, 2010). These physiological reactions can also be compared to participants' own self reports of insight. The effects of previous exposure (with or without reported insight) can also be tested by comparing responses against the posttest novel camouflage stimuli.

Whether or not a participant achieved insight on a particular stimulus can be determined by scoring the accuracy of the typed-in descriptions. Accuracy performance can be determined because the presented images are derived from photographs of actual objects and scenes.

Accuracy scores on the image identification task were determined by independent rater who compared the original unprocessed image from the IAPS corpus to what the participant entered during the task. Responses that reasonably describe the main focus of the image were considered "correct". For example, an image of a grove of trees that is called a "nature scene" is correct, but incorrect if called a "wheat field". Similarly, an image of a dog was rated correct if called a "wolf" or "puppy" but rated incorrect if called a "child". Interrater reliability protocols were used to minimize idiosyncratic ratings below a 10% rater disagreement threshold (Lacy & Riffe, 1996).

Participant heart rate was monitored during the experiment session using a Zephyr BioModule BH3 recording device and BioHarness 3 chest straps (Zephyr Technology, Annapolis, MD). Heart rate was monitored and recorded for the duration of all experimental tasks. These signals were processed by Mindware HRV Analysis v3.12 software (Mindware Technologies Ltd., Gahanna, OH). The automated heartbeat detection function was used in conjunction with a human reviewer who checked for and repaired beat detection errors.

Participants were 11 normal healthy adults age 18 or older (M = 21.7 years, SD = 3.5; 8 Female) recruited through the Department of Psychology's online scheduling system (Sona Systems, Ltd., Tallinn, Estonia) and through advertisements posted on the University of Chicago campus. Participants were offered cash or equivalent course credit in compensation for time spent during the study. The experiment protocol was reviewed and approved by the University of Chicago Independent Review Board (IRB Approval #12-1367).

Study #7 Results

The study described in this chapter was designed to assess whether emotion could inform insightful visuoperceptual processes. It is hypothesized that strongly negative or positive camouflage images (i.e., high IAPS Arousal) would elicit increased heart rate variability compared to neutral images during an insight task.

Before testing the hypothesis with heart rate variability, response accuracy rates were assessed. An ANOVA for response accuracy rates was significant across IAPS High, Medium, and Low Valence categories ($F(2, 30) = 5.214$, $p = 0.0114$). Participants were most accurate at responding to items of Medium Valence (
Table 18, below). Due to the stimulus selection method described above, accuracy rates were not compared in terms of IAPS Arousal.

IAPS Valence	Mean Accuracy	SD Accuracy
High	0.397	0.100
Medium	0.552	0.138
Low	0.427	0.117

Table 18: Accurate identification rates by stimuli IAPS Valence category.

An additional diagnostic ANOVA was run to determine whether heart rate variability differed between correct and incorrect recognitions. There was no significant difference in heart

rate variability between correct and incorrect responses ($F(1, 1267) = 1.021$, $p = 0.310$). A Spearman's rank-order correlation was also not significant ($r = -0.006$, $p = 0.822$). No significant difference is found when the test is limited to only novel Pretest items $F(1, 477) = 1.162$, $p = 0.282$).

According to the hypothesis, emotion and insight stimuli should result in the highest heart rate variability. Here, heart rate variability (HRV) in response to these somatic emotion inputs is calculated with the SDNN measure, the standard deviation of the normal-to-normal RR (R-peak to R-peak) heartbeat interval (Park, Oh, Noh, Kim, & Kim, 2018). High and Low Valence images were kept separate in this analysis. ANOVA was significant for image Valence (High, Medium, or Low) and heart rate variability (SDNN, $F(2, 1266) = 3.522$, $p = 0.0298$) across both Pretest and Posttest (Table 19). While the Low Valence (i.e., negative) stimuli resulted in the highest mean SDNN (variability mean by stimulus) as predicted, High Valence (i.e., positive) stimuli resulted in lower mean variability than Medium (i.e., neutral) stimuli.

Valence	Mean	SD
High	70.1	33.5
Medium	74.1	35.2
Low	77.2	47.5

Table 19: SDNN for camouflage images categorized by IAPS Valence

No interaction was identified for image Valence and Session ($F(2, 1257) = 1.887$, $p = 0.152$), and no interaction was identified for image Valence and response accuracy ($F(2, 1257) = 0.147$, $p = 0.863$). Similarly, a mixed-design ANOVA also failed to find significant differences in SDNN between Pretest and Posttest ($F(1, 1259) = 0.082$, $p = 0.700$) or an interaction between

experiment Session and response accuracy ($F(1, 1259) = 0.166$, $p = 0.684$).This suggests that the hypothesis regarding emotion-plus-insight resulting in higher variability than insight-alone or emotion-alone might not be correct.

At Posttest, there was no interaction between Valence category, image novelty, or response accuracy ($F(2, 766) = 2.487$, $p = 0.084$). There was also no interaction between image novelty and response accuracy ($F(1, 766) = 0.021$, $p = 0.884$), Valence category and response accuracy ($F(2, 766) = 0.419$, $p = 0.658$), or Valence category and novelty ($F(2, 766) = 0.029$, $p = 0.9711$). Although a main effect was found for image novelty at Posttest ($F(1,766) = 0.142$, $p = 0.033$; Table 20), this result further argues against accepting the hypothesis.

Novelty	Mean	SD
New Presentation	70.7	33.3
Old Presentation	71.7	39.0

Table 20: Heart rate variability (SDNN) at Posttest, categorized by whether participants had been shown the source IAPS image at Pretest or not.

Could these results be influenced by individual variability? Furthermore, could these results have been influenced by demographic factors such as Age and Gender? To identify the role of demographic factors, a series of tests was conducted starting with an ANOVA on the SDNN measure using Participant as a factor. The results suggest that some participants had much more variable heart rates than others ($F(1, 1267) = 27.27$, $p < 0.001$). When participant age was median-split as a factor, the interaction suggests that the variability is due to differences between older and younger individuals (interaction $F(1, 1263) = 9.306$, $p = 0.002$).

While the age range of participants was compressed (the youngest participant was 19, while the eldest was 29 years old) across very few participants, statistical analysis suggests Age might be an important factor in the perception of these stimuli. A significant interaction between Session and Age (Figure 26) suggests that the older participants in the dataset maintain a higher variability in heart rate during Session #1 and Session #2 (SDNN; M = 86.922, SD = 57.229 and M = 87.442, SD = 34.880, respectively) than younger participants, whose variability drops from Session #1 to Session #2 (SDNN; M = 71.823, SD = 27.139 and M = 57.874, SD = 33.430, respectively).

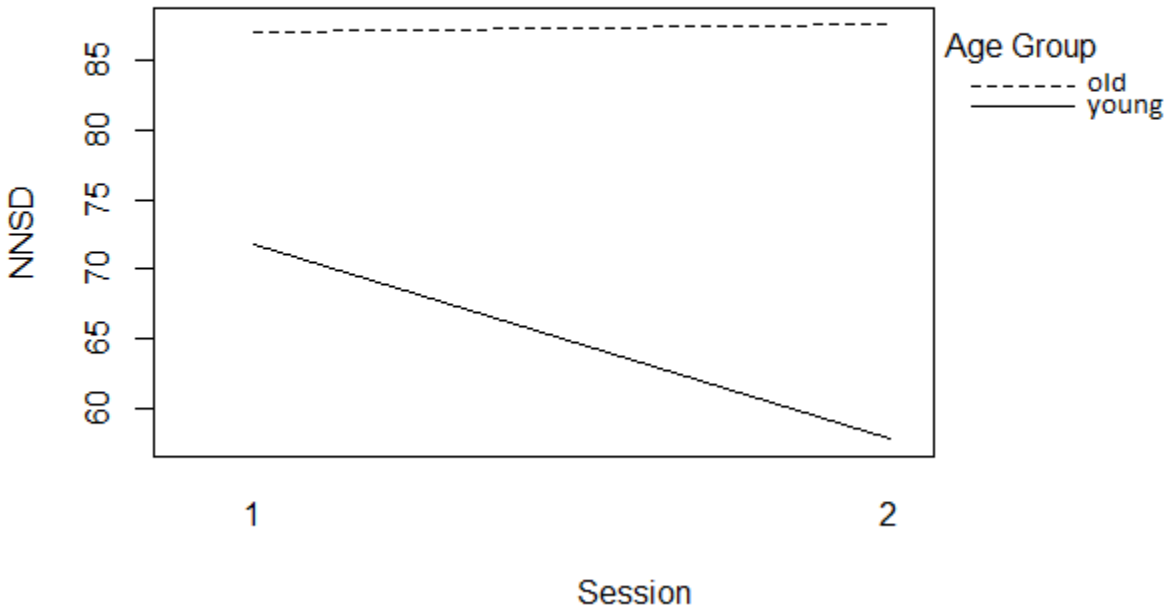

Figure 26: Age Group comparison of heart rate variability (SDNN) at Sessions #1 and #2.

No significant interaction was found between Age and Session for response accuracy ($F(1, 1314) = 0.119$, $p = 0.3702$) in a mixed-design ANOVA, suggesting the age effect found earlier has to do with heart rate variability and not accuracy. No main effect in the test was found

for Session (F(1, 1314) = 3.145, p = 0.0764), but a significant main effect was found for Age (median split; (1, 1314) = 5.131, p = 0.024).

Surprisingly, Gender also appears to be a significant factor in the insightful perception of emotional camouflage images (Figure 27). A significant interaction between Gender and Session (SDNN; F(1, 1261) = 30.437, p < 0.001) suggests that while females have a heart rate that is more variable than males in Session #1 (M = 85.281. SD = 44.890 for females and M = 60.367, SD = 29.376 for males), the variability appears to converge toward similar values at Session #2 (M = 72.764, SD = 38.542 for females and M = 67.585, SD = 32.856 for males), which may be due to an exposure effect.

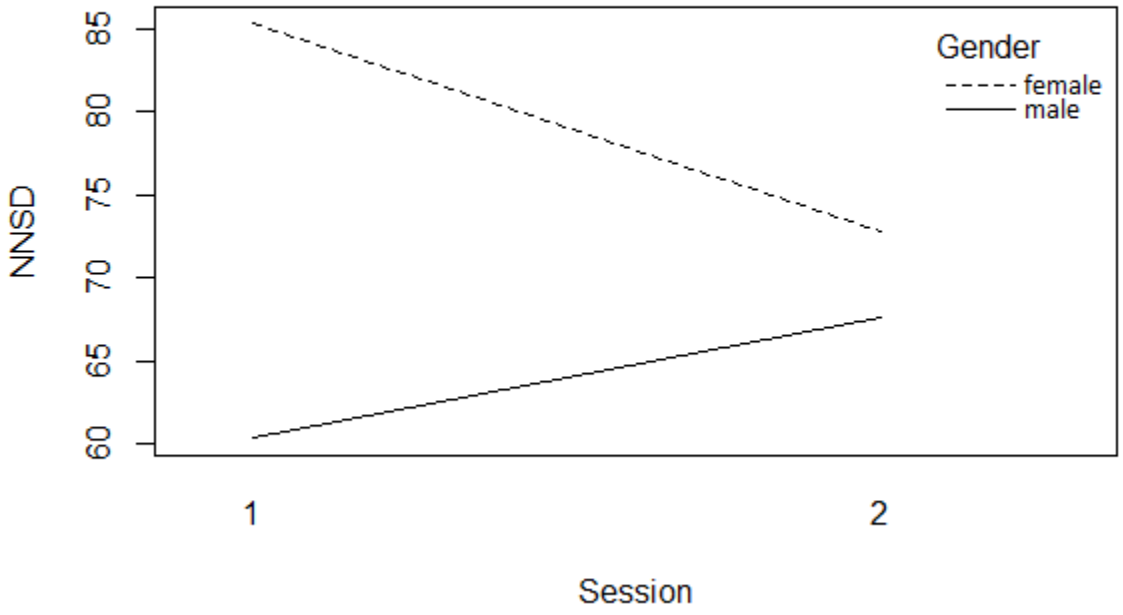

Figure 27: Gender comparison of heart rate variability (SDNN) at Sessions #1 and #2

No significant interaction was found between Gender and Session for response accuracy ($F(1, 1314) = 0.012$, $p = 0.911$), and no main effect was found of Gender on response accuracy ($F(1, 1314) = 3.147$, $p = 0.076$). However, a main effect was found for Gender and response accuracy ($F(1, 1314) = 6.362$, $p = 0.012$), which was not predicted as part of the hypothesis. Males scored a mean of 1.6 points (SD = 1.2), and females scored a mean of 1.5 points (SD = 1.3) across all images.

Female heart rate variability remained high compared to males (Figure 28) and was highest for Low (i.e., negative) Valence images when compared with High (i.e., positive) Valence images (High Valence M = 75.08, SD = 36.469; Medium Valence M = 79.186, SD = 48.492; Low Valence M = 77.893, SD = 38.327). Males, however, demonstrated highest heart rate variability to Low Valence images (M = 57.482, SD = 19.557) and less variability to Medium and High Valence images (M = 72.320, SD = 44.668 and M = 64.391, SD = 23.370, respectively).

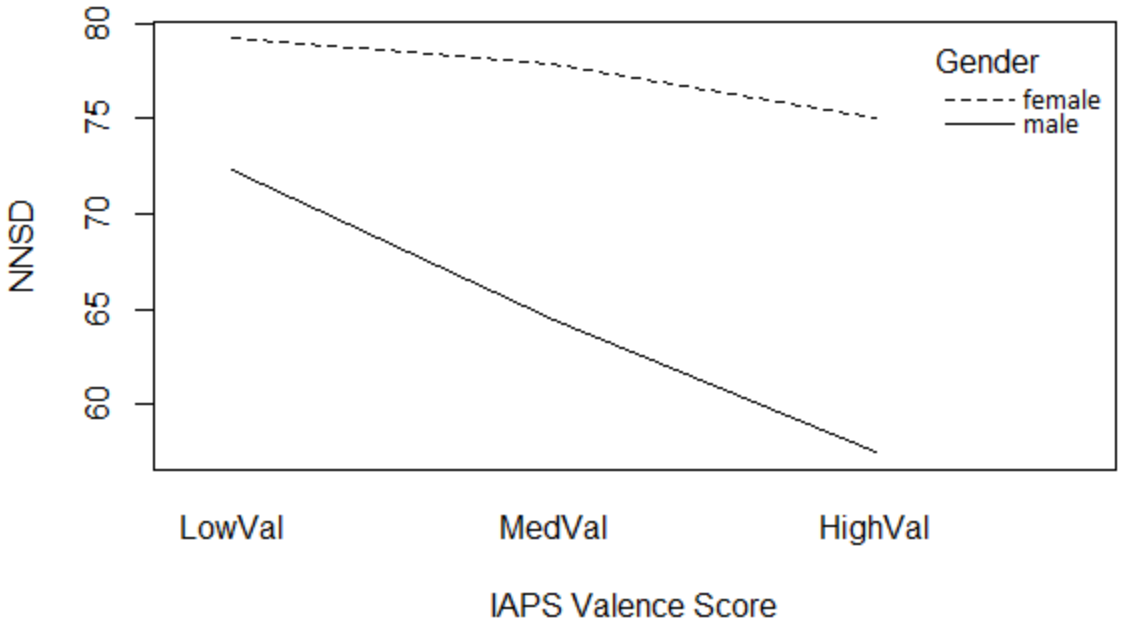

Figure 28: Gender comparison of heart rate variability (SDNN) by camouflage image Valence score

In particular, comparatively higher heart variability seem to be limited to older female participants (M = 105.790, SD = 44.495) relative to younger females (M = 62.566, SD = 30.646), younger males (M = 97.967, SD = 36.756) or older males (M = 63.075, SD = 28.729; Figure 29).

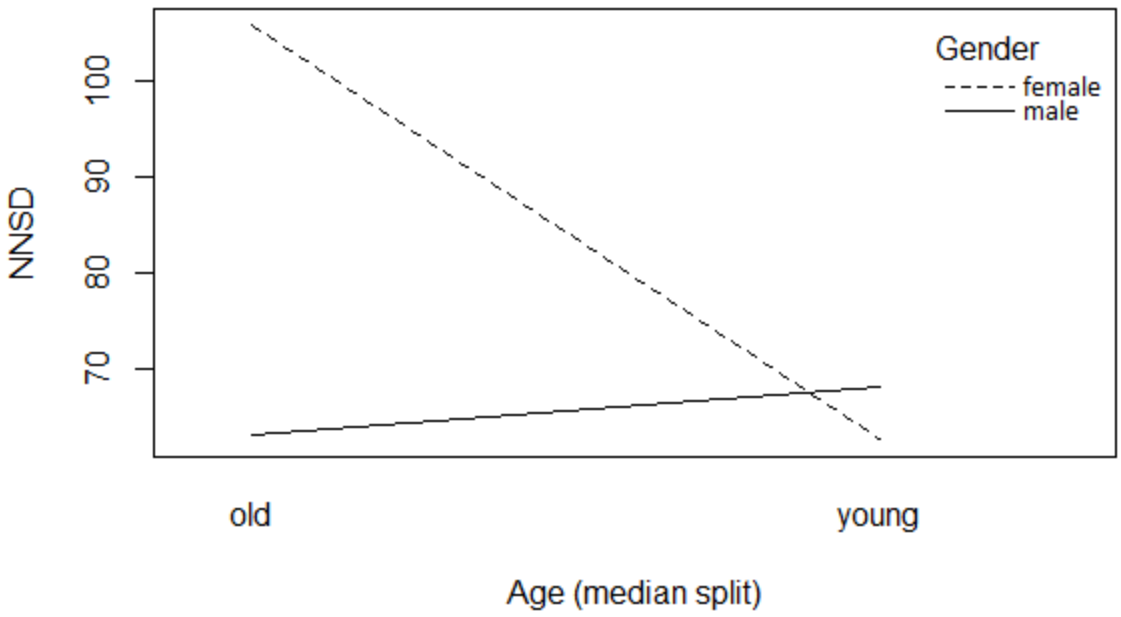

Figure 29: Gender Group comparison of heart rate variability (SDNN) by Age Group

Overall, heart rate was not significantly faster for correct vs incorrect responses ($F(1, 1265) = 0.737$, $p = 0.391$). Heart rate variability was also not significantly different by response accuracy ($F(1, 1265) = 2.377$, $p = 0.123$).

One part of the hypothesis can be tested with Posttest recognition: whether insight alone or emotion alone elicits higher heart rate variability. When limited to Posttest heart rate variability, no interaction was found for response accuracy and whether images were new or old ($F(1, 782) = 0.096$, $p = 0.7568$). No significant main effect was found for response accuracy ($F(1,782) = 0.689$), but was significant for image novelty ($F(1, 782) = 4.683$, $p = 0.031$). Heart rate variability was greater for images that were old at Posttest ($M = 71.7$, $SD = 39.0$) than for images that were new ($M = 70.7$, $SD = 33.3$). In these results, participants would already know

the contents of the image due to the previous exposure and would require no insight problem solving.

Discussion

The current study had two primary aims: this experiment was designed to determine whether insight be found in perceptual behavior (Specific Aim #2) and determine whether insight is affected by autonomic activity (Specific Aim #3). The study attempted to assess these two aims by determining whether emotion, as measured through physiological changes by exposure to emotionally arousing images, can benefit perceptual insight. Previous work has demonstrated that emotional valence can benefit visual perception, but the question of emotion and perceptual insight has remained open. A positive result would accomplish two aims: it would provide further evidence that insight can occur at the level of visual perception (Specific Aim #2), and would suggest that perceptual insight is affected by, and therefore connected to, mechanisms of emotional assessment (Specific Aim #3).

Differences in heart rate variability suggest that emotional information in the imagery was being processed. In younger participants, variability decreased at Posttest, possibly due to learning (e.g., the images may not be as arousing the second time around). However, studies that employed general heart rate monitoring for longer periods than the current study suggest a negative correlation between Age and heart rate variability: as age increases, heart rate variability decreases (Umetani, Singer, McCraty, & Atkinson, 1998).

Male participants displayed greater heart rate variability to Low Valence images (negative emotional content). While these results are in relation to emotion, males have also been previously identified in general heart rate monitoring to display higher variability in heart rate than females (Jensen-Urstad et al., 1997; Silvetti, Drago, & Ragonese, 2001).

Even so, no interaction was found for Valence and response accuracy on heart rate variability. Furthermore, nor were any measures found to significantly correlate with performance. Despite statistically significant evidence that participants are reacting to the emotional content of the camouflage images, there is no significant evidence that participants are utilizing this information to their advantage when forming responses.

Although the study was inconclusive in terms of the Specific Aims, the methods applied here may be useful in further research in insightful problem solving (especially with higher participant recruitment). For example, the current study employed stimuli where the emotional manipulation was endogenous. This may be problematic for several reasons. First, it is possible the emotional stimuli were not salient enough to be detected in the HRV measures. The stimuli were chosen from the IAPS stimulus set; it is possible that the emotional content is not strong enough. Second, the stimuli are visual. Would HRV measures be sensitive enough to detect responses to auditory stimuli, such as the Sinewave Speech and Noise Vocoded Speech found in Chapter Four?

The low participant count is another obvious issue. While no significant differences were identified in key tests, the gender differences suggest some sensitivity in the protocol. Future work would require both a revision of the test protocol (including new stimuli, perhaps with a different sensory modality; a mood manipulation exogenous to the stimuli may also be in order) in addition to a larger participant population based on power analysis and a comparison to participant recruitments found elsewhere in the emotion-insight literature.

CHAPTER SIX: PRIMING INSIGHT THROUGH SLEEP-DEPENDENT MEMORY CONSOLIDATION

In Study #1, different insight tasks such as the RAT and Duncker-Maier FF were shown to be related to the use of working memory as measured by R-SPAN. By inference, the similarity of the Camouflage task to the RAT and Duncker-Maier FF raises the possibility that this task as well depends on working memory, whereas the NRT seems to depend much less on working memory, based on Study #1. Given the possible importance of working memory in this one class of insight problems, the present study (Study #8) addressed this question more directly more generally in terms of the role of memory in insight.

All of these tasks differ substantially from the NRT in one particular way related to working memory. Each trial requires use of long term memory and previous experience and of the stimuli from the past and relationships among the stimuli whether the words and associates (RAT) or the objects and their functions (Duncker-Maier FF) or the images and their appearances (Camouflage). In each case, possible solutions must be held in working memory while tested out in the context of the task. This suggests the importance of memory in these tasks in a way that is not true for the NRT—the patterns are new, are present on screen and so need not be in working memory. To understand the role of memory as a potentially important

part of insight problem solving, Study #8 investigates how priming of memory and sleep consolidation of memory play a role in insight.

Previous research has shown that sleep potentiates insight (Wagner, Gais, Haider, Verleger, & Born, 2004) but this was shown in the NRT and other studies have not provided clear evidence for a role of sleep in other tasks like the RAT task (Cai, Mednick, Harrison, Kanady, & Mednick, 2009). Study #8 was designed to investigate the interaction of working memory and long-term memory by attempting to prime solutions to the Duncker-Maier FF problems and testing for sleep consolidation of the effects of priming.

Chrysikou (2006) hypothesized that experience thinking about alternative uses for objects might aid subsequent insight development for solving Duncker-Maier Functional Fixedness problems. In one group, participants thought of alternatives uses for objects that would later appear in the problems. In a second group, participants thought of alternative uses for objects other than those used in the problems to be solved. If the first but not the second group showed improved performance, it would suggest the importance of activating memories of unusual uses for the problem objects. But for the second group, benefits of this prior experience would not be attributable to activating object memories but instead exercising a general strategy for thinking divergently.

Participants were asked to complete a Pretest that consisted of either an alternative uses task in which participants had to describe different uses for common household objects, or a word associations task in which participants for a list of words had to write a related word for

each. This was immediately followed by seven Functional Fixedness problems (Duncker, 1945; Maier, 1970). If insight can be primed, participants who were asked to consider different uses for a brick (for example), might be more likely to overcome functional fixedness and solve one of Duncker and Maier's problems than those who were asked to complete a (non-priming) word-associates task at Pretest.

While Chrysikou (2006) demonstrates that priming can increase the odds of achieving insight, a question remains of whether this type of learning is processed and stored in memory like other kinds of learning. On a longer time span, does priming decay? Can sleep help contribute to memory consolidation and lead to increased odds of achieving insight?

While early work suggested that sleep could help memory retention by reducing interfernce (cf. Jenkins & Dallenbach, 1924), a body of recent research suggests that sleep instead plays an active role in memory consolidation where connections between memories can be created and strengthened (cf. Marshall & Born, 2007). A twelve-hour delay has been shown to decrease performance in the sensorimotor domain (Brawn, Fenn, Nusbaum, & Margoliash, 2008), and in declarative tasks like verbal learning (Gais, Lucas, & Born, 2006) particularly when influenced by an interfering task (Ellenbogen, Hulbert, Jiang, & Stickgold, 2009). Using an animal model, Brawn, Nusbaum, and Margoliash (2013) demonstrated that interfering information received during the day can reduce performance in the evening, yet performance can recover and improve after a night of sleep. Additionally, individuals have been found to be faster and more accurate when typing (Walker, Brakefield, Morgan, Hobson, & Stickgold, 2002), make new inferences (Ellenbogen, Hu, Payne, Titone, & Walker, 2007), and even improve emotional

declarative memory (Hu, Stylos-Allan, & Walker, 2006; Payne, Stickgold, Swanberg, & Kensinger, 2008).

Given the findings from Chapter Two that suggests a cluster of insight tasks that show similarity to working memory span tasks, it is reasonable to explore whether an insight mechanism or process might be subject to sleep-dependent memory consolidation processes. This relates to a secondary question regarding working memory which may also be tested experimentally: does working memory capacity affect the course of sleep-dependent memory consolidation? Using a word-associates task, Fenn and Hambrick (2012) found that working memory capacity predicted performance after a sleep delay, but not after a wakeful delay between testing. Chrysikou's task, in turn, can be easily modified to test whether or not working memory capacity predicts Posttest performance in a test of insight priming.

Combined, it is reasonable to ask whether the priming of insightfulness follows a similar path as these declarative and nondeclarative memory tasks. This evidence, in turn, may help in the understanding of insight as a unique cognitive process or mechanism that is not prone to consolidation, or if it is a process or mechanism that is at least in part dependent on processing during sleep (Specific Aim #1). Using a Pretest-Posttest protocol with twelve-hour Wake or Sleep delays, this experiment aims to determine whether sleep consolidates priming and promotes insight.

Study #8 Methods

The experimental task consisted of two Sessions: Pretest and Postttest. At the Pretest session, participants were asked to complete informed consent, a demographic questionnaire, an in-lab sleep log, the Affect Grid, and the Stanford Sleepiness Scale. Participants were then asked to complete the R-SPAN task, presented using E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA) on a standard Windows desktop computer.

Participants were then given the Pretest task. Depending on the Condition, this was either a word association task (WA; Chrysikou, 2005), twelve items of Guilford's Alternative Uses Task (AUT; Guilford, Christensen, Merrifield, & Wilson, 1978), or twelve items of Guilford's Alternative Uses Task where seven of the standard objects were replaced with the key items from Duncker (1945) and Meier's (1970) Functional Fixedness Problems (AUT-FF; Guilford et al., 1978; Chrysikou, 2005). A complete listing of measures associated with each Condition can be found in Table 21.

Condition	Pretest Task	Pretest Time	Posttest Task	Posttest Time
1 (Wake AUT)	Alternative Uses Task	9:00 AM	Functional Fixedness	9:00 PM
2 (Sleep AUT)	Alternative Uses Task	9:00 PM	Functional Fixedness	9:00 AM
3 (Wake AUT-FF)	Alternative Uses with FF Items	9:00 AM	Functional Fixedness	9:00 PM
4 (Sleep AUT-FF)	Alternative Uses with FF Items	9:00 PM	Functional Fixedness	9:00 AM
5 (One Session)	Alternative Uses Task	n/a	Functional Fixedness	n/a
6 (One Session)	Alternative Uses with FF Items	n/a	Functional Fixedness	n/a
7 (One Session)	Word Association	n/a	Functional Fixedness	n/a

Table 21: Condition protocols. AUT-FF is Chrysikou's (2005) modification of the Alternative Uses Task (AUT) with items from the Functional Fixedness problems substituted in.

At Posttest, all participants were asked to complete the Affect Grid, Stanford Sleepiness Scale, and the seven Duncker-Maier Functional Fixedness problems.

For participants in Wake Conditions, the Pretest session started at 9:00am and the Posttest session started at 9:00pm, twelve hours later. By virtue of the delay occurring over daytime, participants were expected to complete their normal daily routine.

The Alternative Uses Task used in this experiment was a modification of Guilford et al. (1978) by Chrysikou (2005, 2006) administered using pencil and paper, consisting of twelve commonly recognizable items. These twelve items were given across three pages, with four items per page. Participants were given 5 minutes to complete each page. For each item, participants were given a common category for the item (e.g., "A NEWSPAPER is something used for reading"), then told to list as many as six other categories to which the item would belong. Participants were instructed to notice that all of the categories listed were different from each other and different from the primary use of a newspaper: the alternative uses they write down must be novel from each other and from the prompt example.

Furthermore, participants were also directed not to spend too much time on each item. The instructions noted that participants may move back and forth across items on the same page and were permitted return to the incomplete items in a given page if time for that page if time remained.

Responses for the AUT and AUT-FF were scored according to Guilford et al. (1978): Flexibility, number of categories, and quality-of-elaboration scores were judged by three independent raters. Responses for Functional Fixedness were scored by three independent raters following Lacy and Riffe's (1996) interrater reliability guidelines.

Participants were 164 adults aged 18-37 (M = 20.9, SD = 3.4), 90 female, who drank 1.1 servings of caffeine on an average day. The study was reviewed and approved by the University of Chicago Independent Review Board under approval #H11217. Participants were compensated with cash or course credit (0.5 course credit-hours or $5 per half-hour completed plus an additional 0.5 credits or $5 compensation due to the added burden of returning for a second experiment session) in return for their time spent during the study.

Study #8 Results

There were no significant Pretest-Posttest differences for Affect Grid measures for Pleasantness ($F(1, 212) = 0.619$, $p = 0.432$) or Energy ($F(1, 212) = 1.547$, $p = 0.215$). There were also no significant differences between Stanford Sleepiness Scale scores between Sessions ($F(1, 235) = 0.940$, $p = 0.333$). Elaboration Scores at Pretest were significantly different across Conditions ($F(5,133) = 3.874$, $p = 0.003$), due to AUT-FF Conditions scoring higher than AUT Conditions (Table 22).

Pretest Elaboration Scores

Condition	Pretest	Mean	SD
One-Session	AUT	17.467	9.627
Wake	AUT	17.375	9.426
Sleep	AUT	17.792	9.041
One-Session	AUT-FF	25.714	15.333
Wake	AUT-FF	22.154	6.619
Sleep	AUT-FF	27.817	7.054

Table 22: Pretest Elaboration Scores by Experiment Condition

Overall, there was no effect of condition on scores in the Posttest Functional Fixedness problems ($F(6, 148) = 0.417$, $p = 0.867$). When Posttest scores were tested for an interaction between Delay and Pretest type, the interaction was not significant at the 0.05 level ($F(2, 148) = 0.312$, $p = 0.733$). The main effect of Pretest type ($F(2, 148) = 0.165$, $p = 0.848$) and the main effect of Delay ($F(2, 148) = 0.776$, p 0 0.462) were also not significant. This indicates that performance was not significantly different for participants who received a Word Associates task or the two versions of the Alternative Uses Task. Similarly, there was no significant difference when comparing only participants who received the Alternative Uses Task (AUT) as Pretest in the Wake and Sleep Conditions ($F(1, 54) = 2.153$, $p = 0.148$) or those who received the Alternative Uses Task with Functional Fixedness primes (AUT-FF; $F(1, 21) = 0.001$, $p = 0.979$).

Although participants affirmed to the experimenter that they were native speakers of American English, questionnaire responses suggested this may not have been the case (a concern raised by the independent scorers). Of the recruited participants, 120 were native American English speakers with no reported psychoactive drug use (medically prescribed or otherwise) who slept at least 6 hours of sleep the night before the 2nd experiment session. This reduced participant subset were 68 males who were on average 20.0 years old (SD = 3.3). The subset also

drank on average 1.0 caffeinated beverages a day (SD = 1.1), and slept an average of 8.0 hours per night (SD = 1.4).

Even with this reduced subset, however, there was no significant of Condition on Posttest Functional Fixedness scores ($F(6, 129) = 0.565$, $p = 0.758$). The interaction for Delay Condition and Pretest type was not significant ($F(2, 129) = 0.690$, $p = 0.503$), nor were the main effects for Pretest type ($F(2, 129) = 0.042$, $p = 0.959$) or Delay Condition ($F(2, 129) = 0.690$, $p = 0.385$) significant at the 0.05 level.

Interestingly, the Elaboration scores for the Alternative Uses Pretest correlated with the Functional Fixedness scores across all participants ($r(129) = 0.180$, $p = 0.038$; Table 23). When comparing within Conditions, this effect appears to be due to participants in the Sleep and One-Session Conditions who received the standard Alternative Uses Task. Pretest scores for any Condition were not significantly correlated if the Pretest was the AUT-FF, which contained items critical to solving the Functional Fixedness problems presented at Posttest. This did not appear to be a result due to different performance at Pretest across Conditions ($F(2, 91) = 0.015$, $p = 0.985$). Posttest differences across Conditions were also not significantly different ($F(2, 83) = 1.091$, $p = 0.341$; 8 participants excluded due to previous exposure to Functional Fixedness problems).

Alternative Uses - Elaboration Scores			
Condition	**# participants**	**p-value**	**Pearson's r**
All Participants	131	0.039	0.181
Wake (9:00am - 9:00pm), AUT	35	0.786	0.048
Wake (9:00am - 9:00pm), AUT-FF	12	0.878	0.050
Sleep (9:00pm - 9:00am), AUT	21	0.037	0.458
Sleep (9:00pm - 9:00am), AUT-FF	11	0.621	-0.168
One Session, AUT	30	0.035	0.387
One Session, AUT-FF	21	0.293	0.241

Table 23: Pearson's Rank-Order Correlations between Alternative Uses Elaboration Scores at Pretest and Functional Fixedness Scores at Posttest.

The Category Score, which reflects the number of unique object-use categories a given participant generates, was not found to correlate with Posttest Functional Fixedness performance (Table 24).

Alternative Uses - Unique Categories			
Condition	**# participants**	**p-value**	**Pearson's r**
All Participants	131	0.221	0.108
Wake (9:00am - 9:00pm), AUT	35	0.968	-0.007
Wake (9:00am - 9:00pm), AUT-FF	12	0.232	0.373
Sleep (9:00pm - 9:00am), AUT	21	0.693	0.092
Sleep (9:00pm - 9:00am), AUT-FF	11	0.332	0.323
One Session, AUT	30	0.154	0.267
One Session, AUT-FF	21	0.788	0.063

Table 24: Pearson's Rank-Order Correlations between Alternative Uses Category Scores at Pretest and Functional Fixedness Scores at Posttest.

Discussion

The hypothesized differential benefit of sleep versus wake for achieving insight was not found with this experiment (Specific Aim #1). While this may be due to various factors, such as

low power requiring more participant data, the most likely explanation may be rooted in aspects of the Functional Fixedness (FF) problems themselves.

The FF problem set consists of seven problems which must be completed in an expository format. Because it is not a keypress response, reaction times cannot be reliably recorded. In this protocol, participants were asked to read the question prompt and hit the spacebar as soon as they had thought of a response and were about to begin writing. There were two problems with this method. First, there was variable time in reading that is conflated with time to insight. Not only do different problems take differing amounts of time to read, each individual participant can be expected to read at a different pace. Because the spacebar press occurred after reading and thinking (time to insight), it is not possible to separate the two. The second problem was noted by experimenter observation during debriefing: many participants responded that they did not press the spacebar when they thought of their solution; they simply began writing and only touched the keyboard when it was time to move onto the next problem. Because of these issues, reaction times for the seven Functional Fixedness problems were not analyzed in this experiment.

Working memory capacity correlated only with Posttest performance for Sleep and One Session Conditions where participants are Pretested with the AUT, but not with the AUT-FF. One possible explanation is that the AUT-FF primes solutions at Posttest in a different way than the AUT: priming participants with the key objects found at Posttest reduces the need for working memory, thereby negating the need for extra capacity. These results support Stickgold, Scott, Rittenhouse, & Hobson (1999), which found that performance on weakly-priming—but

not strongly priming—word pairs was dependent on whether forced-awake from REM or NREM sleep. These particular findings additionally identify a small but significant role of working memory capacity in insightful problem solving.

The data and analyses demonstrate a correlation between Pretest priming performance (on the AUT) and Posttest performance in the Functional Fixedness problems for the One-Session and Sleep Conditions; the correlation between scores was not significant for Wake participants. This is not evidently due to differences in Pretest scores (though scores were significantly different between AUT and AUT-FF Pretests), nor is it evidently due to differences in Posttest scores across Conditions. Working memory appears to be a factor in Posttest performance when the prime immediately occurs beforehand because of the short timespan between tasks.

For participants in the Sleep Condition, working memory capacity appears to have an effect on behavior relative to participants in the Wake Condition. This may be a result of sleep-dependent mechanisms that are responsible for consolidating the Pretest task material. The results may also be due to working memory capacity allowing for better performance on the later Posttest. Participants in the Wake group, in contrast, experienced a twelve-hour wakeful delay between Pretest and Posttest that may have reduced the usefulness of a large working memory capacity when solving Functional Fixedness problems.

CHAPTER SEVEN: GENERAL DISCUSSION

Is there one kind of insight or several? If there is one kind, there may be a single common process. As a starting point there are clearly two different types of insight tasks. In the NRT, an insight is developed over successive responses and then over subsequent trials. Implicit insight is identified when responses on trials get faster. Explicit insight is identified when participants simply provide the final answer skipping all the intermediate steps and having learned the rule. Thus, there is one insight—the operation of the hidden rule—and this develops over successive trials each of which requires a pattern of responses. By comparison, in the Duncker-Maier Functional Fixedness problems, each problem has a single complex solution. The solutions require thinking of unusual or atypical uses for objects (e.g., hammer as pendulum weight). A second problem would be entirely different with new objects. Thus, while solving a second problem would also require understanding the novel use of a common object that can provide the key to the solution, there is nothing else in common across problems. Similarly, for the RAT, each triplet of words has a unique and unrelated solution compared to the other trials.

The results of Study #1 suggest that tasks such as RAT and Duncker-Maier Functional Fixedness show similar performance patterns and thus may be mediated by more similar processes than the NRT. This is further supported by the grouping of R-SPAN working memory performance with the RAT and Duncker-Maier Functional Fixedness tasks rather than the NRT.

While other interpretations are possible, this suggests the possibility that working memory may play a critical role in these insight tasks more than others such as the NRT. If we consider for a moment the task differences, this leads to one interesting account of insight. For some insight tasks, problem solution depends on a novel approach to using known materials such as known words (RAT) or familiar objects (Dunker-Maier Functional Fixedness). People have to come up with memories for these materials, associations for them, and search for low probability related information. This is a memory retrieval task and could easily depend on using working memory throughout the process, holding possible solutions in mind during problem solving.

Furthermore, this description fits with the result of Study #2 showing that the camouflage task also shares common variance with the RAT and the Duncker-Maier FF task. As with those tasks, the camouflage images are not unusual just visually masked. Observers need to see through the camouflage to recognize a familiar object or scene. This is clearly a memory task for familiar objects with the insight coming when the masking is defeated perceptually; observers hold hypotheses in mind while looking at features of the masked image.

Both of these are quite different from the NRT in which the hidden rule is novel and the patterns are novel and the responses are novel. While the constituents of the patterns are known (e.g., numbers) the patterns are not. Observers might hold the patterns in working memory but all information is available on screen so there is no need to do so. As a result, while insights in the NRT have been shown to be potentiated by sleep (Wagner, Gais, Haider, Verleger, & Born, 2004), this has not been directly linked to working memory. In Study #8, there was no evidence for sleep consolidation of insights in the Duncker-Maier Functional Fixedness task. Although

this is a null result, and further research is needed, if taken at face value it suggests another difference between the types of insight.

Although two Aims of the thesis were to determine a role for perception and emotional arousal in insight, the studies addressing these aims produced no clear significant results. Therefore we cannot draw any conclusions about these aspects of insight. While there is some suggestion in the results of perceptual insight into talker source from the speech Study #6, more research will be needed to establish an effect and then determine that such an effect is really a kind of insight as opposed to perceptual recognition and learning.

The primary findings of the present research then focus on Specific Aim #1 and the evidence that tasks that are called insight can show different patterns of performance grouping into two different categories. At least one insight task, the NRT, develops over time, shows evidence of an implicit insight prior to conscious and explicit insight into the problem solution. Moreover the pattern of performance for the NRT suggests this is not similar to the pattern of performance on a working memory task. Whether a different use of working memory (e.g., sequence memory rather than verbal memory) might be more similar will depend on future research.

For those tasks that cluster together in patterns of performance, the RAT, the Duncker-Maier Functional Fixedness, and the Camouflage tasks, similarities amongst the do suggest that there may be a common process underlying these. Given that they are related to working memory performance it is possible to hypothesize that all three tasks depend on stimulus-cued

retrieval of the items in the problems regardless of task specifics as well as low probability features or associates or possible views. In this retrieval process, a hypothesis-test use of working memory may operate in order to systematically determine solutions to specific problems.

Thus the results reported in Chapter Two (in Studies #1 and #2) suggest that there may be multiple ways of achieving insight rather than a single kind as the uniformity of description of the experience might suggest. Those features that have been ascribed to achieving insight may well just represent the experiential properties of suddenly solving any difficult and relatively opaque problem. However the way in which such problems are solved may depend on the particulars of the stimulus, past experience, and the nature of the problem to be solved.

REFERENCES

Abele-Brehm, A. (1992). Positive and negative mood influences on creativity: Evidence for asymmetrical effects. *Polish Psychological Bulletin*, *23*(3), 203-211.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261-295.

Andreou, C., Bozikas, V. P., Luedtke, T., & Moritz, S. (2015). Associations between visual perception accuracy and confidence in a dopaminergic manipulation study. *Frontiers in Psychology*, 6(414), 1-7. doi:10.3389/fpsyg.2015.00414.

Appelhans, B. M., & Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. *Review of General Psychology*, *10*(3), 229-240.

Auble, P. M., Franks, J. J., & Soraci, S. A. (1979). Effort toward comprehension: Elaboration or "aha"? *Memory & Cognition*, *7*(6), 426-434.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600-609.

Barber, T. X., & Calverley, D. S. (1964). An experimental study of "hypnotic" (auditory and visual) hallucinations. *The Journal of Abnormal and Social Psychology*, *68*(1), 13-20.

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and economic behavior*, *52*(2), 336-372.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. Science, 275(5304), 1293-1295.

Bentall, R. P., & Slade, P. D. (1985). Reality testing and auditory hallucinations: A signal detection analysis. *British Journal of Clinical Psychology, 24*(3), 159–169.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351-368.

Berntson, G. G., Bigger, J. T., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., & der Molen, M. W. (1997). Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology, 34*(6), 623-648.

Boucher, J., & Osgood, C. E. (1969). The Pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior, 8*(1), 1-8.

Boucsein, W., Fowles, D.C., Grimnes, S., Ben-Shakhar, G., Roth, W.T., Dawson, M.E., & Filion, D.L (2012). Publication recommendations for electrodermal measurements. *Psychophysiology, 49*, 1017-1034.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers, 35*(4), 634-639.

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences, 9*(7), 322-328.

Bowers, P. (1979). Hypnosis and creativity: The search for the missing link. *Journal of Abnormal Psychology*, *88*(5), 564-572.

Brainard, D. H. (1997). The Psychophysics Toolbox, *Spatial Vision, 10*(4), 433-436.

Brawn, T. P., Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2008). Consolidation of sensorimotor learning during sleep. *Learning & Memory, 15*(11), 815-819.

Brawn, T. P., Nusbaum, H. C., & Margoliash, D. (2013). Sleep consolidation of interfering auditory memories in starlings. *Psychological Science, 24*(4), 439-447.

Brimijoin, W. O., Akeroyd, M. A., Tilbury, E., & Porr, B. (2013). The internal representation of vowel spectra investigated using behavioral response-triggered averaging. *The Journal of the Acoustical Society of America, 133*(2), EL118-EL122.

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*(2), 563–570.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin, 109*(2), 204-223.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 325-337.

Bruner, J. S., & Potter, M. C. (1964). Interference in visual recognition. *Science*, *144*(3617), 424-425.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences*, *106*(25), 10130-10134.

Chambers, A. M., & Payne, J. D. (2014). Laugh yourself to sleep: Memory consolidation for humorous information. *Experimental Brain Research, 232*(5), 1415-1427.

Chrysikou, E. G. (2005). *When a shoe becomes a hammer: Problem solving as goal-derived, ad-hoc categorization* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3202996)

Chrysikou, E. G. (2006). When shoes become hammers: Goal-derived categorization training enhances problem-solving performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 935-942.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407-428.

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*(5), 769-786.

Creswell, J. D., Dutcher, J. M., Klein, W. M., Harris, P. R., & Levine, J. M. (2013). Self-affirmation improves problem-solving under stress. *PLoS One*, *8*(5), 1-7.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004, May). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (Vol. 1, No. 1-22, pp. 1-2).

Cunningham, J. B., MacGregor, J. N., Gibb, J. L., & Haar, J. M. (2009). Categories of insight and their correlates: An exploration of relationships among classic-type insight problems, rebus puzzles, remote associates and esoteric analogies. *Journal of Creative Behavior, 43*(4), 262-280.

Dallenbach, K. M. (1951). A Puzzle-picture with a new principle of concealment. *The American Journal of Psychology, 64*(3), 431-433.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Phil. Trans. R. Soc. Lond. B, 351*(1346), 1413-1420.

Damasio, A.R., Tranel, D., & Damasio, H.C. (1991). "Ch. 11: Somatic markers and the guidance of behaviour: theory and preliminary testing". In Levin, Harvey S.; Eisenberg, Howard M.; Benton, Arthur Lester. Frontal Lobe Function and Dysfunction. Oxford University Press. pp. 217–229.

Daneman, M., & Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450–466.

Darsaud, A., Wagner, U., Balteau, E., Desseilles, M., Sterpenich, V., Vandewalle, G., & Maquet, P. (2011). Neural precursors of delayed insight. *Journal of Cognitive Neuroscience, 23*(8), 1900-1910.

Darwin, C. J. (2005). Turning speech into music in a two-dimensional space by varying the bandwidth and rate of tone pulses placed along formant tracks. *The Journal of the Acoustical Society of America, 117*(4), 2570-2570.

Duncker, K. (1945). On problem-solving (L. S. Lees, Trans.). *Psychological Monographs, 58*(5).

E-Prime 2 [computer software]. Sharpsburg, PA: Psychology Software Tools, Inc.

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope, 58*(9), 955-991.

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

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences, 104*(18), 7723-7728.

Ellenbogen, J. M., Hulbert, J. C., Jiang, Y., & Stickgold, R. (2009). The sleeping brain's influence on verbal memory: Boosting resistance to interference. *PLoS One, 4*(1), e4117.

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition, 2*(2-3), 101-118.

Estrada, C. A., Isen, A. M., & Young, M. J. (1994). Positive affect improves creative problem solving and influences reported source of practice satisfaction in physicians. *Motivation and Emotion, 18*(4), 285-299.

Fenn, K. M., & Hambrick, D. Z. (2012). Individual differences in working memory capacity predict sleep-dependent memory consolidation. *Journal of Experimental Psychology: General*, *141*(3), 1-7.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*, 614-616.

Fodor, J. (1986). Input systems as modules. In *The modularity of mind* (pp. 47-100). Cambridge, MA: MIT Press.

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion, 19*(3), 313-332.

Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America*, *110*(2), 1150-1163.

Gais, S., Lucas, B., & Born, J. (2006). Sleep after learning aids memory recall. *Learning & Memory*, *13*(3), 259-262.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training 'greeble'experts: A framework for studying expert object recognition processes. *Vision Research, 38*(15), 2401-2428.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155-170.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*(3), 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1), 1-38.

Gick, M. L., & Lockhart, R. S. (1995). Cognitive and affective components of insight. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 197-228). Cambridge, MA, US: MIT Press.

Glicksohn, J., & Barrett, T. R. (2003). Absorption and hallucinatory experience. *Applied Cognitive Psychology, 17*(7), 833-849.

Glucksberg, S., & Weisberg, R. W. (1966). Verbal behavior and problem solving: Some effects of labeling in a functional fixedness problem. *Journal of Experimental Psychology, 71*(5), 659-664.

Gosselin, F., Bacon, B. A., & Mamassian, P. (2004). Internal surface representations approximated by reverse correlation. *Vision Research, 44*(21), 2515-2520.

Gosselin, F. & Schyns, P. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science, 14*(5), 505-509.

Gross, O., Toivonen, H., Toivanen, J. M., & Valitutti, A. (2012, November). Lexical creativity from word associations. In *2012 seventh international conference on knowledge, information and creativity support systems (KICSS)*, (pp. 35-42). IEEE.

Grossberg, S. (2000). How hallucinations may arise from brain mechanisms of learning, attention, and volition. *Journal of the International Neuropsychological Society, 6*(5), 583-592.

Guilford, J. P., Christensen, P. R., Merrifield, P. R., & Wilson, R. C. (1978). *Alternate uses: Manual of instructions and interpretation.* Orange, CA: Sheridan Psychological Services.

Hänsel, A., & von Känel, R. (2008). The ventro-medial prefrontal cortex: a major link between the autonomic nervous system, regulation of emotion, and stress reactivity? *BioPsychoSocial Medicine, 2*(1), 21.

Hecht, S., Shlaer, S., & Pirenne, M. H. (1942). Energy, quanta, and vision. *The Journal of General Physiology*, *25*(6), 819-840.

Henrich J., Heine S.J., Norenzayan A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 62–83.

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 283-295.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. C. (1973). Quantification of sleepiness: A new approach. *Psychophysiology*, *10*(4), 431-436.

Holyoak, K. J., & Thagard, P. R. (1989). A computational model of analogical problem solving. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 242-266). New York, NY: Cambridge University Press.

Hu, P., Heald, S., Malonis, P., & Nusbaum, H. (2017, July). *Gaining Insight for Innovation*. Paper presented at the International Open and User Innovation Conference, Innsbruck, Austria.

Hu, P., Stylos-Allan, M., & Walker, M. P. (2006). Sleep facilitates consolidation of emotional declarative memory. *Psychological Science*, *17*(10), 891-898.

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk (New York University No. CEDER-10-01). New York, NY: New York University.

Isen, A. M., Daubman, K. A., & Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology, 52*(6), 1122-1131.

James, W. (1893). *The Principles of Psychology* (Vol. 1). New York, NY: Holt.

Jausovec, N. (1989). Affect in analogical transfer. *Creativity Research Journal, 2*(4), 255-266.

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology, 35*(4), 605-612.

Jensen-Urstad, K., Storck, N., Bouvier, F., Ericson, M., Lindbland, L. E., & Jensen-Urstad, M. (1997). Heart rate variability in healthy subjects is related to age and gender. *Acta Physiologica, 160*(3), 235-241.

Jones, G. V. (1989). Back to Woodworth: Role of interlopers in the tip of-the-tongue phenomenon. *Memory & Cognition, 17*(1), 69-76.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*(11), 4302-4311.

Kosslyn, S. M. (1981). The medium and the message in mental imagery: A theory. *Psychological Review*, *88*(1), 46-66.

Kounios, J., & Beeman, M. (2009). The Aha! Moment the cognitive neuroscience of insight. *Current Directions in Psychological Science*, *18*(4), 210-216.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*(3), 1065-1081.

Lacy, S., & Riffe, D. (1996). Sampling error and selecting intercoder reliability samples for nominal content categories. *Journalism & Mass Communication Quarterly*, *73*(4), 963-973.

Lang, A., Dhillon, K., & Dong, Q. (1995). The effects of emotional arousal and valence on television viewers' cognitive capacity and memory. *Journal of Broadcast and Electronic Media*, *39*(3), 313-327.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *The IAPS: Technical manual and affective ratings*. Gainesville, FL: NIMH Center for the Study of Emotion and Attention.

Lang, S., Kanngieser, N., Jaśkowski, P., Haider, H., Rose, M., & Verleger, R. (2006). Precursors of insight in event-related brain potentials. *Journal of Cognitive Neuroscience, 18*(12), 2152-2166.

Launay, G., & Slade, P. (1981). The measurement of hallucinatory predisposition in male and female prisoners. *Personality and Individual Differences, 2*(3), 221-234.

Leppänen, J. M., & Hietanen, J. K. (2004). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological Research, 69*(1-2), 22-29.

Levenson, R. W. (1992). Autonomic nervous system differences among emotions. *Psychological Science, 3*(1), 23-27.

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 1-10.

Liu, J., Li, J., Feng, L., Li, L., Tian, J., & Lee, K. (2014). Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. *Cortex, 53*, 60-77.

Ludmer, R., Dudai, Y., & Rubin, N. (2011). Uncovering camouflage: Amygdala activation predicts long-term memory of induced perceptual insight. *Neuron, 69*(5), 1002-1014.

Lynn, S. J., & Rhue, J. W. (1986). The fantasy-prone person: Hypnosis, imagination, and creativity. *Journal of Personality and Social Psychology, 51*(2), 404-408.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PloS One, 4*(3), e4638.

MacGregor, J. N., & Cunningham, J. B. (2008). Rebus puzzles as insight problems. *Behavioral Research Methods, 40*(1), 263-268.

Maier, N. R. F. (1970). *Problem solving and creativity in individuals and groups*. Belmont, CA: Brooks/Cole.

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in Cognitive Sciences, 11*(10), 442-450.

Mason, O., & Claridge, G. (2006). The Oxford-Liverpool Inventory of Feelings and Experiences (O-LIFE): Further description and extended norms. *Schizophrenia Research, 82*(2), 203-211.

Mason, O., Claridge, G., & Jackson, M. (1995). New scales for the assessment of schizotypy. *Personality and Individual Differences, 18*(1), 7-13.

MATLAB Release 2011 [computer software]. Natick, MA: The MathWorks, Inc.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS), 8*(1), 3-30.

McCraty, R., Atkinson, M., Tiller, W. A., Rein, G., & Watkins, A. D. (1995). The effects of emotions on short-term power spectrum analysis of heart rate variability. *The American Journal of Cardiology*, *76*(14), 1089-1093.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220-232.

Merckelbach, H., & van de Ven, V. (2001). Another White Christmas: fantasy proneness and reports of 'hallucinatory experiences' in undergraduate students. *Journal of Behavior Therapy and Experimental Psychiatry*, *32*(3), 137-144.

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition, 15*(3), 238-246.

Meyer, A. S., & Bock, K. (1992). The tip-of-the-tongue phenomenon: Blocking or partial activation? *Memory & Cognition*, *20*(6), 715-726.

Moritz, S., Ramdani, N., Klass, H., Andreou, C., Jungclaussen, D., Eifler, S., Englisch, S., Schirmbeck, F., & Zink, M. (2014). Overconfidence in incorrect perceptual judgments in patients with schizophrenia. *Schizophrenia Research: Cognition, 1*(4), 165-170.

Nestor, A., Vettel, J. M., & Tarr, M. J. (2013). Internal representations for face detection: An application of noise-based image classification to BOLD responses. *Human Brain Mapping*, *34*(11), 3101-3115.

Nusbaum, H. C., & Schwab, E. C. The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Volume 1, Speech Perception*. New York: Academic Press, 1986.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355–376.

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*(3), 466-478.

Olton, R. M., & Johnson, D. M. (1976). Mechanisms of incubation in creative problem solving. *The American Journal of Psychology*, *89*(4), 617-630.

Palomba, D., Angrilli, A., & Mini, A. (1997). Visual evoked potentials, heart rate responses and memory to emotional pictorial stimuli. *International Journal of Psychophysiology, 27*(1), 55-67.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411-419.

Park, H., Oh, S., Noh, Y., Kim, J. Y., & Kim, J. H. (2018). Heart rate variability as a marker of distress and recovery: The effect of brief supportive expressive group therapy with mindfulness in cancer patients. *Integrative Cancer Therapies.* Advance Online Publication. doi:10.1177/1534735418756192

Parra, A. (2006). Seeing and feeling ghosts: Absorption, fantasy proneness, and healthy schizotypy as predictors of crisis apparition experiences. *Journal of Parapsychology, 70*(2), 357-372.

Payne, J. D., Stickgold, R., Swanberg, K., & Kensinger, E. A. (2008). Sleep preferentially enhances memory for emotional components of scenes. *Psychological Science, 19*(8), 781-788.

Pecchinenda, A., & Smith, C. A. (1996). The affective significance of skin conductance activity during a difficult problem-solving task. *Cognition & Emotion, 10*(5), 481-504.

Perky, C. W. (1910). An experimental study of imagination. *American Journal of Psychology. 21*(3), 422-452.

Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological Science, 17*(4), 292-299.

Pollock, V., Cho, D. W., Reker, D., & Volavka, J. (1979). Profile of Mood States: The factors and their physiological correlates. *The Journal of Nervous and Mental Disease, 167*(10), 612-614.

Porges, S. W. (1995). Orienting in a defensive world: Mammalian modifications of our evolutionary heritage. A polyvagal theory. *Psychophysiology, 32*(4), 301-318.

Porges, S. W. (2001). The polyvagal theory: Phylogenetic substrates of a social nervous system. *International Journal of Psychophysiology, 42*(2), 123-146.

Posey, T. B., & Losch, M. E. (1983). Auditory hallucinations of hearing voices in 375 normal subjects. *Imagination, Cognition, and Personality, 3*(2), 99-113.

Pumprla, J., Howorka, K., Groves, D., Chester, M., & Nolan, J. (2002). Functional assessment of heart rate variability: Physiological basis and practical applications. *International Journal of Cardiology, 84*(1), 1-14.

Razumnikova, O. M. (2007). Creativity related cortex activity in the remote associates task. *Brain Research Bulletin, 73*(1), 96-102.

Remez, R. E., Rubin, P. E., Pisoni, D. B., Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*(4497), 947–950.

Rieth, C. A., Lee, K., Lui, J., Tian, J., & Huber, D. E. (2011). Faces in the mist: Illusory face and letter detection. *i-Perception*, *2*(5), 458-476.

Rowe, G., Hirsh, J. B., & Anderson, A. K. (2007). Positive affect increases the breadth of attentional selection. *Proceedings of the National Academy of Sciences*, *104*(1), 383-388.

Rubin, N., Nakayama, K., & Shapley, R. (2002). The role of insight in perceptual learning: Evidence from illusory contour perception. In M. Fahle & T. Poggio (Eds.), *Perceptual insight* (pp. 235-251). Cambridge, MA: MIT Press.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493-502

Schooler, J. W., & Melcher, J. (1995). The ineffability of insight. In Smith, S. M., Ward, T. B., & Finke, R. A. (Eds.), *The Creative Cognition Approach* (pp. 97-133). Cambridge, MA: MIT Press.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, *27*(4), 395-408.

Schwartz, B. L. (1999). Sparkling at the end of the tongue: The etiology of tip-of-the-tongue phenomenology. *Psychonomic Bulletin & Review, 6*(3), 379-393.

Schwartz, B. L., & Smith, S. M. (1997). The retrieval of related information influences tip-of-the-tongue states. *Journal of Memory and Language, 36*(1), 68-86.

Schwartz, G. E., Weinberger, D. A., & Singer, J. A. (1981). Cardiovascular differentiation of happiness, sadness, anger, and fear following imagery and exercise. *Psychosomatic Medicine, 43*(4), 343-364.

Segal, S. J. (1972). Assimilation of a stimulus in the construction of an image: The Perky effect revisited. In P. W. Sheehan & J. S. Antrobus (Eds.), *The function and nature of imagery* (pp. 203-230). New York, NY: Academic Press.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423 & 623–656.

Shannon R.V., Zeng, F-G., Kamath, V., Wygonski, J., Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(6), 1447-1496.

Silvetti, M. S., Drago, F., & Ragonese, P. (2001). Heart rate variability in healthy children and adolescents is partially related to age and gender. *International Journal of Cardiology*, *81*(2), 169-174.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2014). Top-Down Influences of Written Text on Perceived Clarity of Degraded Speech. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 186-199.

Sona Systems [computer software]. Tallinn, Estonia: Sona Systems, Ltd.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-292.

Stea, J. N., & Hodgins, D. C. (2012). The relationship between lack of control and illusory pattern perception among at-risk gamblers and at-risk cannabis users. *The Social Science Journal*, *49*(4), 528-536.

Stevens, K. N. (2000). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Stickgold, R., Scott, L., Rittenhouse, C., & Hobson, J. A. (1999). Sleep-induced changes in associative memory. *Journal of Cognitive Neuroscience, 11*(2), 182-193.

Subramaniam, K., Kounios, J., Parrish, T. B., & Jung-Beeman, M. (2009). A brain mechanism for facilitation of insight by positive affect. *Journal of Cognitive Neuroscience, 21*(3), 415-432.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*(6), 401-409.

Teasdale, J. D., & Fogarty, S. J. (1979). Differential effects of induced mood on retrieval of pleasant and unpleasant events from episodic memory. *Journal of Abnormal Psychology, 88*(3), 248-257.

Tellegen, A., & Atkinson, G. (1974). Openness to absorbing and self-altering experiences ("absorption"), a trait related to hypnotic susceptibility. *Journal of Abnormal Psychology, 83*(3), 268-277.

Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders, 61*(3), 201-216.

Thorndike E.L. (1898). Animal intelligence: an experimental study of the association processes in animals. *Psychological Monographs, 8*.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs.*

Topolinski, S., & Reber, R. (2010). Gaining insight into the "Aha" experience. *Current Directions in Psychological Science, 19*(6), 402-405.

Tracy, J. L., & Robins, R. W. (2008). The automaticity of emotion recognition. *Emotion, 8*(1), 81-95.

Umetani, K., Singer, D. H., McCraty, R., & Atkinson, M. (1998). Twenty-four hour time domain heart rate variability and heart rate: Relations to age and gender over nine decades. *Journal of the American College of Cardiology*, *31*(3), 593-601.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498-505.

Vaschillo, E. G., Bates, M. E., Vaschillo, B., Lehrer, P., Udo, T., Mun, E. Y., & Ray, S. (2008). Heart rate variability response to alcohol, placebo, and emotional picture cue challenges: Effects of 0.1-Hz stimulation. *Psychophysiology*, *45*(5), 847-858.

Vosburg, S. K. (1998). The effects of positive and negative mood on divergent-thinking performance. *Creativity Research Journal, 11*(2), 165-172.

Vrana, S. R., Cuthbert, B. N., & Lang, P. J. (1989). Processing fearful and neutral sentences: Memory and heart rate change. *Cognition and Emotion, 3*(3), 179-195.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature, 427*(6972), 352-355.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: Sleep-dependent motor skill learning. *Neuron, 35*(1), 205-211.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063-1070.

Whitson, J. A., & Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. *Science, 322*(5898), 115-117.

Wiesmann, M., & Ishai, A. (2010). Training facilitates object recognition in cubist paintings. *Frontiers in Human Neuroscience, 4*(11), 1-7.

Witkin, H. A., Oltman, P. K., Raskin, E., & Karp, S. A. (1971). *A manual for the embedded figures test.* Palo Alto, CA: Consulting Psychologists Press.

Wulfert, E., Roland, B. D., Hartley, J., Wang, N., & Franco, C. (2005). Heart rate arousal and excitement in gambling: Winners versus losers. *Psychology of Addictive Behaviors, 19*(3), 311–316.

Yaniv, I., & Meyer, D. E. (1987). Activation and metacognition of inaccessible stored information: Potential bases for incubation effects in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(2), 187-205.

Yokoi, K., Nishio, Y., Uchiyama, M., Shimomura, T., Iizuka, O., & Mori, E. (2014). Hallucinators find meaning in noises: Pareidolic illusions in dementia with Lewy bodies. *Neuropsychologia*, *56*, 245-254.

Yordanova, J., Verleger, R., Wagner, U., & Kolev, V. (2010). Patterns of implicit learning below the level of conscious knowledge. *Journal of Psychophysiology*, *24*(2), 91-101.

Zenasni, F., & Lubart, T. (2002). Effects of mood states on creativity. *Current Psychology Letters: Behaviour, Brain & Cognition, 8,* 33-50.

Zhang, H., Liu, J., Huber, D. E., Rieth, C. A., Tian, J., & Lee, K. (2008). Detecting faces in pure noise images: a functional MRI study on top-down perception. *Neuroreport, 19*(2), 229-233.

APPENDICES

This Appendix contains questionnaires used in experiments not available in the literature; these were used to assess demographics, sleep, and alertness at experiment sessions. These questionnaires were reviewed and approved by the University of Chicago Independent Review Board (IRB) under the respective approvals for each study where they were used.

Appendix A: Demographics Questionnaire

The following Demographics Questionnaire was completed by participants at the beginning of the first (or only) experiment session, after informed consent and before experimental tasks.

CENTER FOR COGNITIVE AND SOCIAL NEUROSCIENCE

Demographic Information

How old are you? _____ years

Are you: □ Male □ Female

Are you: □ Left-handed □ Right-handed

Are you a student? □ Yes □ No

If yes, what year?

What is your major / what department are you in?

Appendix B: Stanford Sleepiness Scale

The Stanford Sleepiness Scale (SSS; Hoddes et al., 1973) was completed by participants in Studies where sleepiness or wakefulness could be a factor in performance. In the version here, the Scale was a laminated Letter-size card with which the experimenter could ask the participant which sentence best described how they felt at that moment. When paired with the Sleep Log (Appendix C), participants were asked to respond to the SSS by writing the numeral from the card into the Sleep Log entry.

STANFORD SLEEPINESS SCALE

Code: *1 – feeling active, vital, alert, or wide awake*

2 – functioning at high levels, but not at peak; able to concentrate

3 – relaxed; awake; not at full alertness; responsive

4 – a little foggy, not at peak; let down

5- fogginess; beginning to lose interest in remaining awake; slowed down

6 – sleepiness; prefer to be lying down; fighting sleep, woozy

7 – almost in reverie; sleep onset soon; lost struggle to remain awake

X – asleep

Appendix C: Sleep Log

The Sleep Log is kept in-laboratory by the experimenter and completed at the beginning of a session for experiments with sleep/wake delays. Participants were instructed to fill in the top half at the first session and to save the second half rod the second session.

Sleep Log

Participant ID: *[Filled-in by experimenter]*

Bring to experiment session on _____

Fill out for Night of [ ]

1. What time did you go to sleep last night? _____

2. What time did you wake up this morning? _____

3. What was the quality of your sleep last night? *(circle one)*

poor average good

4. Did you take any naps during the day? *(circle one)*

yes no

If yes, how long was it and when was it? _____

Fill out for Night of [ ]

1. What time did you go to sleep last night? _____

2. What time did you wake up this morning? _____

3. What was the quality of your sleep last night? *(circle one)*

poor average good

4. Did you take any naps during the day, or after your session yesterday? *(circle one)*

yes no (circle one)

If yes, how long was it and when was it? _____

Appendix D: Participant Questionnaire

The following Participant Questionnaire was collected in-laboratory at beginning of the experiment session after informed consent but prior to any experimental tasks.

PARTICIPANT QUESTIONNAIRE

1.) What time do you usually go to sleep?_____

2.) Do you fall asleep easily?_____

3.) How deeply do you sleep?_____
(Light, Medium, Deep)

4.) What time do you usually wake up?_____

5.) Do you usually feel well rested?_____

6.) Have you ever sought medical
attention for a sleep disorder?_____

7.) Do you have any disabilities
that disrupt your sleep? _____

8.) Are you currently taking any
medications to help you sleep?_____

9.) Are you taking any other medications?_____
(excluding oral contraceptives)

10.) Do you have a history of substance abuse
or diagnosed major mental illness? _____