Title
A Cognitive Test Battery to Assess General Intelligence in the Pigeon (Columba livia)

Author
Flaim, Mary

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Cognitive Test Battery to Assess General Intelligence in the Pigeon (*Columba livia*)

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Psychology

by

Mary Elizabeth Flaim

2021

ABSTRACT OF THE DISSERTATION

A Cognitive Test Battery to Assess General Intelligence in the Pigeon (*Columba livia*)

by

Mary Elizabeth Flaim
Doctor of Philosophy in Psychology
University of California, Los Angeles, 2021
Professor Aaron P. Blaisdell, Chair

The study of intelligence in humans has been ongoing for over 100 years, including the underlying structure, predictive validity, related cognitive measures, and source of differences. One of the key findings in intelligence research is the uniform positive correlations among cognitive tasks. Factor analysis consistently extracts one factor that can account for approximately half of the variance in performance. This factor is termed *g* and all cognitive tasks positively load onto this factor. This has been replicated with every cognitive test battery in humans. Nevertheless, many other aspects of intelligence research have revealed contradictory lines of evidence. Recently, cognitive test batteries have been developed for animals to examine similarities to humans in cognitive structure. When mice and some avian species are assessed with cognitive test batteries, performance positively correlates and the first component extracted has similar properties to *g*. There are some limitations to the species tested thus far, including comparability in the cognitive domains assessed across species and homogeneous samples. The pigeon is an ideal subject to overcome these issues since pigeons, humans, and other primates are

frequently given similar tasks. We created a test battery for pigeons that assessed different cognitive domains, including associative learning, short term memory, cognitive flexibility, and reaction time. This test battery was administered to 23 subjects that ranged in age from 6 months to 18 years old. The tasks included were sufficiently sensitive to detect individual differences, while still being reliable measures of performance. Despite the strengths of the test battery, we did not consistently extract a *g* like factor. Analyses indicated a two-component structure, where the associative learning and reaction time tasks loaded onto component 1, while short term memory and cognitive flexibility tasks loaded onto component 2. While it is impossible to determine what these components represent from the results of these experiments alone, we speculated that these components could reflect reliance on different underlying cognitive abilities, degree of automaticity, and sensitivity to age related decline. Additional research, including administering test batteries to other species, will be necessary to fully understand why pigeons show two-components instead of *g*.

The dissertation of Mary Elizabeth Flaim is approved.

William Grisham

Alan Castel

James Stigler

Barney Schlinger

Aaron P. Blaisdell, Committee Chair

University of California, Los Angeles

2021

Contents

Contents v

ACKNOWLEDGEMENTS xii

VITA xvi

Chapter 1: Introduction 1

Brief overview of *g* 1

Related Cognitive Factors 8

Short-Term Memory 11

Processing Speed 12

Response Inhibition 14

Associative Learning 16

Related Cognitive Factors - Summary 17

g in Nonhuman Animals 18

Nonhuman Primates 21

Mice 25

Avian Species 29

Discussion 36

Table 1.1. 53

Chapter 2: Transferring Relational Rule Learning: A Potential Problem Between Successive and Simultaneous Choice Procedures when Assessing Pigeons (*Columba Livia*) 58

Abstract 58

Introduction 58

Method 61
Subjects 61
Apparatus 62
Stimuli 63
Matrix Displays 63
Procedure 64
Data Analysis 67
Results 67
First Rule Acquisition and Transfer Testing 67
Second Rule Acquisition and Transfer Testing 68
Order or Rule Effects 70
Discussion 71
Chapter 3: Assessing Associative Learning Using the Symbolic Match to Sample Task 86
Abstract 86
Introduction 86
Method 90
Subjects 90
Apparatus 91
Stimuli 91
Procedure 92

Data Analysis 95
Results 96
Discussion 96
Chapter 4: Serial Reversal Learning 105
Abstract 105
Introduction 105
Methods 110
Subjects 110
Apparatus 110
Stimuli 111
Procedure 111
Data Analysis 112
Results 112
Discussion 115
Chapter 5: The Delayed Match to Sample Task 125
Abstract 125
Introduction 125
Methods 132
Subjects 132
Apparatus 132

Stimuli 133
Procedure 133
Data Analysis 136
Results 137
Simultaneous Match to Sample 137
Delayed Match to Sample 137
Discussion 141
Chapter 6: Choice Reaction Time 154
Abstract 154
Introduction 154
Experiment 1 160
Methods 161
Results 166
Discussion 168
Experiment 2 169
Method 169
Results 170
Discussion 172
Experiment 3 174
Methods 175

Results 175

Discussion 178

General Discussion 179

Chapter 7: A Cognitive Test Battery to Assess General Intelligence in the Pigeon (*Columba livia*) 193

Abstract 193

Introduction 193

Method 203

Subjects 203

Apparatus 204

Procedure 204

Data Analysis 209

Results 210

Individual Cognitive Tasks 210

Cognitive Test Battery 212

Discussion 215

Chapter 8: Conclusion 229

References 234

List of Figures

Figure 1.1. *The relationship between g and cognitive abilities*..... 51

Figure 1.2 *A cladogram of species given cognitive test batteries* 52

Figure 2.1. *Example matrix displays for the relational rule learning task* 76

Figure 2.2. *Acquisition performance for the first rule learned* 77

Figure 2.2. *Transfer session performance for the first rule learned* 78

Figure 2.4 *Acquisition data for the second rule learned.* 79

Figure 2.5. *Transfer session performance for the second rule learned* 80

Figure 3.1 *Stimuli in the symbolic match to sample task* 100

Figure 3.2 *Visualization of a trial in the symbolic match to sample* 101

Figure 3.3 *Number of session to criteria in the symbolic match to sample* 102

Figure 3.4 *Correlation between age and performance in the symbolic match to sample* 103

Figure 4.1 *First session performance on the serial reversal learning task*..... 120

Figure 5.1 *Visualization of a delayed match to sample trial*..... 144

Figure 5.2 *Accuracy on the delayed match to sample at 3 points in training* 145

Figure 5.3 *Log transformed data for the delayed match to sample at 3 points in training*.... 146

Figure 5.4 *Performance on the delayed match based on age* 147

Figure 6.1 *Visualization of a trial in the choice reaction time task* 183

Figure 6.2 *Performance in experiment 1 in the choice reaction time task*..... 184

Figure 6.3 *Performance in experiment 2 in the choice reaction time task*..... 186

Figure 6.4 *Example of stimuli used in experiment 3 for the choice reaction time task*..... 188

Figure 6.5 *Performance in experiment 3 in the choice reaction time task*..... 189

List of Tables

Table 1.1 *Summary of cognitive test battery research in nonhuman animals.* 53

Table 2.1. *Number of sessions excluded for the relational rule learning task* 81

Table 2.2. *Binomial results when the luminosity rule was learned first* 82

Table 2.3. *Binomial results when the size change rule was learned first* 83

Table 2.4. *Binomial results when the size change was learned second* 84

Table 2.5. *Binomial results when the luminosity rule was learned second* 85

Table 3.1. *Number of sessions to reach criterion in the symbolic match to sample* 104

Table 4.1. *Mean trials to reach criterion for a reversal in 4 avian species* 121

Table 4.2. *Number of sessions to reach criterion for each reversal for each subject* 122

Table 4.3. *Correlation matrix for the initial discrimination and each reversal* 123

Table 4.4. *Number of trials needed to meet alternative criteria for reversal learning.* 124

Table 5.1. *Post hoc analyses of performance on the delayed match to sample* 148

Table 5.2. *Analysis of variance of performance on the delayed match to sample* 149

Table 5.3. *Correlation matrix of accuracy performance on the delayed match to sample* ... 151

Table 5.4. *Correlation matrix of log transformed data on the delayed match to sample* 153

Table 6.1 *Number of errors in the choice reaction time task.* 191

Table 6.2 *Results from the linear regression in the choice reaction time task.* 192

Table 7.1. *The number of tasks in the battery completed by each subject* 224

Table 7.2. *Correlation matrix for the serial reversal learning task* 225

Table 7.3. *Correlation matrix for the aggregate measures in the test battery* 226

Table 7.4. *Correlation matrix for the individual measures in the test battery* 227

Table 7.5. *Principal component analyses of the cognitive test battery* 228

ACKNOWLEDGEMENTS

This work is a culmination of many years of support in academic settings that fostered my intellectual curiosity. I am deeply grateful that I was able to pursue my doctoral training at the University of California, Los Angeles since the strong psychology department was instrumental to the creation and execution of the tasks described within this dissertation. Within the department I would first like to thank my advisor, Dr. Aaron Blaisdell, who has been inspiring in terms of his breadth and depth of knowledge. He always provided an interesting perspective and helped me become a more well-rounded researcher. I also appreciate the freedom to pursue such an ambitious and far-reaching project, while providing enough guidance to ensure it was conceptually sound. Having Dr. Blaisdell as my advisor has increased my clarity as a writer and developed my independence as a researcher. I will always appreciate what I have learned from him.

I would also like to thank my other committee members. Dr. Grisham's perspective as a comparative psychologist and neuroanatomist has been helpful in recognizing broader species implications of behavior and the underlying neural circuits. I was also lucky enough to be his teaching assistant one summer, and I was inspired by his commitment to teaching (both the undergraduates and myself), ambition in his research projects within the lab setting, and overall kindness. I always enjoyed the many conversations I've had with Dr. Grisham, no matter if it was about research, his tales from graduate school, or small liberal arts colleges on the east coast. Dr. Stigler and Dr. Castel have also been instrumental in crafting the final project of this dissertation, providing a deeper knowledge of cognitive psychology and educational interventions. Their comments during the final oral exam in particular, were an interesting perspective that I would not have thought of on my own and would be unlikely to hear in animal

research circles. Dr. Schlinger for highlighting the importance of ecological relevance, in addition to his patience with various, potential, collaborations.

Beyond my committee members, I have been fortunate enough to interact with other faculty members, both within UCLA and from my undergraduate university. I appreciate Drs. Alicia Izquierdo and Janet Tomiyama for creating and leading the professional development course, which provided a safe and structure place to become a better mentor and create the various documents necessary to apply to postdoctoral and faculty positions. Dr. Michael Fanselow who's graduate and undergraduate courses have greatly enhanced my understanding of learning theory. Dr. Andrew Wickenheiser who kindly shared his experience as a new faculty member and shows genuine interest in such a wide variety of research. From my undergraduate university, I will be forever indebted to Drs. Vern Bingman and Vincent 'Gino' Coppola, who introduced me to pigeon research. I never thought I would be interested in animal research, but working with pigeons felt like being struck by lightning. I'm grateful they saw my potential and for their continued support.

At the administrative level, I appreciate the work of Lisa Lee, the graduate studies coordinator who has provided so many friendly reminders to keep me on track to complete the doctoral degree requirements. Cheryl Polfus and Diego Garcia, the instructional scheduling coordinators who have always assigned me to classes that were within my area of expertise, but that also expanded my own knowledge. Kevin Nguyen, who has built and repaired many of the devices in our lad. Without these pieces of equipment, many of my experiments would have been impossible to conduct. Numan Interiano for his incredible patience and understanding with animal care protocols and training, in addition to the other animal care staff who maintained the cages and were an extra set of eyes to maintain high standards of care. William Cage who has

done an amazing job of transitioning into a new role of overseeing animal care in Franz. Luis Alavarez and Hamid Zafarani, who ensured the pigeons had clean homes. The registered veterinary technician, Denaya Brown who has been an absolute joy to interact with and key in maintaining the health and wellbeing of the pigeons.

I have also been lucky with my fellow lab members and research assistants. Julia Schroeder, who was the senior graduated student in the Blaisdell lab when I joined, was absolutely vital in my training. Without her it would have been incredibly difficult to understand the rodent operant chambers and UCLA requirements. In particular, I appreciate her coding expertise that made it possible to conduct my first pigeon experiment that served as a cornerstone for the doctoral research presented here. Bianca Landfield was also instrumental in helping me with rodent experiments in my first years of graduate school, both in training and moral support. Ben Seitz, who joined the lab my third year, but was so vital for bringing a fresh perspective and finding open-source programs that have greatly expanded the lab's potential. In addition, I appreciate the comradery within the lab and when we TA'ed together. Jingxuan Guo, who has been an incredible research assistant. Without her support and expertise, it would have been very difficult to conduct the fine-grained analysis in the choice reaction time task. Anna MacFarlane was another amazing research assistant who took on a large role in creating the program for the simultaneous relational rule learning task. Finally, Valeria Gonzalez, now postdoc in the lab, who is one of my favorite people. From research ideas, emotional support, and board games, I cannot express how much our friendship has enhanced my life.

Outside of the lab, I am very lucky to have such supportive friends and family. Nina Lichtenberg and Evan Hart, who demonstrated incredible work life balance (beach in the morning, lab in the afternoon) and offer their hospitality whenever I am in Maryland. Garrett

Blair, who is one of the kindest people I have ever met. Maureen Gray, who's honesty and life perspective deepened our relationship, but has also sharpened my own values. Beth Moroney, who's been equally supportive of runs and nachos. Marisol Lauffer, fellow pigeon enthusiast and goth queen. Allison Burch, who makes me laugh till my cheeks hurt every time we talk on the phone. Hannah Finlayson, a friendship that began a decade ago, who has always encouraged me in my personal and professional life. My parents, Jenny and Mike Flaim, who have encouraged and fostered my independence for my entire life. They have always made me feel like I can accomplish anything I want, but always made having fun a priority too. I can't imagine being where I am today without their love and support. Finally, I would like to thank my partner, Mark Coffin. He has been incredible and words are not enough to express the amount of love and support he continues to show.

This research was supported by the University Fellowship and Graduate Summer Research Mentorship Grant from University of California, Los Angeles graduate division. For the work that has been previously published, I would like to thank my co-authors. Aaron Blaisdell has been instrumental to the initial designs of these experiments and with editing the manuscripts. Jingxuan Guo assisted with data processing and analysis for the choice reaction time task described in chapter 6. Chapter 1 is a version of Flaim, M., & Blaisdell, A. P. (2020). The comparative analysis of intelligence. *Psychological Bulletin*, *146*(12), 1174, reprinted with permission.

VITA

Master of Arts in Psychology 2016
University of California – Los Angeles, Los Angeles, CA

Bachelor of Arts in Psychology with Minor in History 2014
Bowling Green State University, Bowling Green, OH

Publications

Flaim, M., & Blaisdell, A. P. (2020). The comparative analysis of intelligence. *Psychological Bulletin, 146*(12), 1174.

Seitz, B. M., **Flaim, M. E.**, & Blaisdell, A. P. (2020). Evidence That Novel Flavors Unconditionally Suppress Weight Gain in the Absence of Flavor-Calorie Associations. *Learning & Behavior*, 1-13.

Coppola, V. J., Kanyok, N., Schreiber, A. J., **Flaim, M. E.**, & Bingman, V. P. (2016). Changes in hippocampal volume and neuron number co-occur with memory decline in old homing pigeons (Columba livia). *Neurobiology of Learning and Memory, 131*, 117-120.

Coppola, V. J., **Flaim, M. E.**, Carney, S. N., & Bingman, V. P. (2015). An age-related deficit in spatial–feature reference memory in homing pigeons (*Columba livia*). *Behavioural Brain Research, 280*, 1-5.

Conference Presentations

Flaim, M., & Blaisdell, A. P., (2021) *Pigeon (Columba livia) Performance on the Delayed Match to Sample Task (DMTS) as a Function of Age.* (Updated). Talk given at the 28th International Conference on Comparative Cognition, Virtual

Guo, J., **Flaim, M**., & Blaisdell, A. P. (2021*) Hicks' Reaction Time Task Performance by Pigeons.* Talk given at the 28th International Conference on Comparative Cognition, Virtual

Flaim, M., & Blaisdell, A. P., (2018) *The Comparative Analysis of Intelligence.* Talk given at the 18th International Society for Comparative Psychology, Los Angeles CA

Flaim, M., Cai, D. J., Silva, A. J., & Blaisdell, A. P. (2018) *Causal Reasoning in Mice.* Talk given at the 25th International Conference on Comparative Cognition, Melbourne FL

Blaisdell, A.P., Seitz, B.M., Diaz, R., & **Flaim, M.E.** (2018). *The Modified Law of Effect and the Partial Reinforcement Extinction Effect.* Talk given at the 25th International Conference on Comparative Cognition. Melbourne, FL.

Seitz, B.M., **Flaim, M.E.,** & Blaisdell, A.P (2018). *Pavlovian Diet: Investigations on Flavor-Calorie Associations and Their Role in Weight Gain and Food Consumption.* Talk given at the 25th International Conference on Comparative Cognition. Melbourne, FL.

Flaim, M. E., Schroeder, J. E., Bye, J. K., Bedoyan, L., He, R., Tantiwuttipong, P., Cheng, P. W., Blaisdell, A. P. (2016) *Rats' Knowledge that an Outcome is a Continuous Variable Increases Their Use of an Additive Causal-Invariance Function in a Blocking Procedure.* Talk given at the 23rd International Conference on Comparative Cognition, Melbourne FL

Invited Talks

Avian and Mammalian Aging, Learning and Behavior Brown Bag, University of California – Los Angeles, February 7, 2020

Development of a Raven's Progressive Matrices to Examine General Intelligence in the Pigeon, Columba livia, Learning and Behavior Brown Bag, University of California – Los Angeles, February 3, 2017

Intelligence, Causal Reasoning, and Pavlovian Conditioning, Alumnus Talk, Bowling Green State University, Bowling Green Ohio, August 9, 2016

Grants, Scholarships, and Fellowships

University of California, Los Angeles

Graduate Summer Research Mentorship Grant	2016, 2017
University Fellowship	2015-2016

Bowling Green State University, Main Campus

BGSU Success Scholarship	2010-2014

Honors

Graduated Magna Cum Laude	2014
Dean's List College of Arts and Sciences	2010-2014

Chapter 1: Introduction

Brief overview of *g*

Why do some people live longer, healthier lives, or obtain higher levels of education, or gravitate towards more cognitively-demanding careers? What underlies individual differences in reaction time, working memory, or learning? Many researchers have argued that individual differences in intelligence underlie each of these, as it is the best, though not a perfect, predictor of many of these (Brodnick & Ree, 1995; Conway et al., 2002; Deary et al., 2004; Gottfredson, 2002; Gottfredson & Deary, 2004; Jensen, 1998; Ree & Earles, 1992; Schmidt, 2011, 2014; Sheppard & Vernon, 2008; but see Gutman & Schoon, 2013; Heckman et al., 2013 for the importance of 'non-cognitive' factors and Ceci, 1991 on how schooling is a causal factor for performance on intelligence measures). Intelligence is typically measured with a full-scale IQ (FSIQ) test. The FSIQ contains a battery of diverse tasks designed to assess different aspects of cognition, including basic math skills, matrix reasoning, spatial reasoning, verbal comprehension, and memory, though the specific content can vary across tests (Johnson et al., 2004; Reynolds et al., 2013; Schrank & McGrew 2001), and concerns have been raised about how often these tests reflect western education or culture (Ceci, 1991; Nisbett, 2009; Serpell, 2000; Wicherts et al., 2010). Capturing these individual differences across all these tasks with a single metric may appear to overlook important factors. Perhaps a person is terrible at math, for example, but has exceptional verbal comprehension. Nevertheless, for a large majority of people, performance typically correlates across all tasks – despite their diversity (Carroll, 1993; Deary, 2000).

Charles Spearman (1904) was the first to report a uniform positive correlation among diverse cognitive tasks in people and he called this the 'positive manifold'. This finding has

continued to be replicated ever since (Carroll, 1993; Deary, 2000; Jensen, 1998). When these positive correlational matrices are subjected to factor analysis, one factor that explains approximately half of the variance in performance is extracted. It is called the *g* factor (Carroll, 1993; Jensen, 1998; Spearman, 1904). When tasks load onto a factor, it indicates how much variance in that task can be explained by the factor. All cognitive tasks load in the appropriate direction onto *g*, but not all tasks load equally (Nisbett, 2009; Reynolds et al., 2013; Weiss et al., 2013). The tasks that show the highest loading onto this *g* factor involve reasoning, abstraction, task complexity, and task novelty, irrespective of how the information is presented within the tasks themselves (Ackerman & Cianciolo, 2000; Jensen, 1998; but see Ceci, 1996 for difficulties determining how 'abstract' a problem is). Many researchers have described the *g* factor as 'indifferent to the indicator' because loading depends more on the complexity and abstraction of the task rather than on a specific type of measure, or in this usage the 'indicator' (Jensen, 1998; Spearman, 1904). This is why *g* is thought to reflect a general cognitive ability and has predictive validity in a variety of contexts (Deary et al., 2004; Gottfredson, 2002).

While the *g* factor typically accounts for half of the variance in performance, it can depend on the number, reliability, and familiarity of the tasks used, and the variation in the sample tested (Ackerman & Cianciolo, 2000; Colom et al., 2002). All cognitive tasks measure *g*, but they also measure narrower abilities and contain task-specific variance (Carroll, 1993; Gustafsson, 2003; Jensen, 1998). The most accurate measures of *g* will be obtained with a diverse test battery. Even with a diverse test battery, the intercorrelations or extracted factor can be smaller than expected due to measurement (un)reliability and range restriction (Jensen, 1998; Viswesvaran et al., 2014). Range restriction limits the amount of variability in the sample, but the variability across subjects is exactly what factor analysis is attempting to explain (Jensen,

1998)! This lack of variability will lower the value of the extracted factor, but this is primarily due to sample characteristics. Range restriction is a crucial factor when assessing *g* in small samples with a small number of tasks. While a large and representative sample can help with range restriction, the same cannot be said for the effects of measurement reliability. Moderate to low task reliability (.5-.8) attenuates the subsequent correlations, which can impact later factor analysis since more of the variance will be due to random error or transient factors unrelated to *g* (Fan, 2003; Jensen, 1998). Some statistical methods (like structural equation modelling, SEM, and confirmatory factor analysis, CFA) can account for this, but multiple regression does not, increasing the likelihood of false positives (Westfall & Yarkoni, 2016). Nevertheless, many studies have handled these challenges beautifully. The results from various large FSIQ tests conducted with representative samples indicate that, even though the exact test content can vary, the same *g* factor is extracted (Johnson et al., 2004; Reynolds et al., 2013; Schrank & McGrew 2001). The *g* factor is robust against different methods of analysis, populations, cultures, and test batteries (Carroll, 1993, 2003; Chabris, 2007; Deary, 2000; Warne & Burningham, 2019; but see Wicherts et al., 2010) and is relatively stable throughout the lifespan starting at 2 years old (Deary et al., 2013; Gignac, 2014; Spinath et al., 2003).

While *g* can account for a large amount of cross-task variance in individual differences, additional variance can be explained by group factors. Some tasks show stronger correlations with each other, forming a subgroup. For example, in a test battery with three verbal measures and three math measures, there is a stronger correlation within verbal measures and within math measures than between both domains (Carroll, 1993; Jensen, 1998). Group factors are more strongly affected by test battery composition (Cattell, 1987; Carroll, 2003; Johnson et al., 2004). The most commonly found group factors include fluid intelligence (*Gf*), crystallized intelligence

(*Gc*), quantitative reasoning (*Gq*), visual processing (*Gv*), processing speed (*Gs*), and memory, though the exact terminology can vary (Carroll, 1993; Cattell, 1987; Hakstian & Cattell 1978). Most research has focused on *Gf* and *Gc* (Kvist & Gustafsson 2008; Nisbett, 2009; but see Johnson & Bouchard, 2005).

Gf is the ability to solve novel and complex problems, in particular those that require relational reasoning, and frequently use shapes and figures in the tasks as opposed to words. *Gf* loads very highly, and sometimes perfectly, onto *g* (Benson et al., 2010; Bickley et al., 1995; Carroll, 1993; Gustafsson, 1984; Kvist & Gustafsson 2008), though the strength of loading depends on many factors, such as sample homogeneity, number of tests in the assessment, and methods of analysis (Blair, 2006; Carroll, 2003; Kan et al., 2011; Thorsen et al., 2014). It is still debated to what degree measures of *Gf* are dependent on school exposure and culture. Some researchers argue that, because *Gf* tasks typically do not use language or memorized facts, this means it relies less on prior knowledge or schooling, and thus should be viewed as culture free (Cattell & Horn, 1978; Jensen, 1998; Kent, 2017). Other researchers have argued that the increased emphasis on formal schooling and increase in visual stimuli in a given culture has resulted in improvements in these tasks in subsequent generations, indicating that these tasks are dependent on school and culture (Baker et al., 2015; Cahan & Noyman, 2001; Ceci, 1991; Nisbett, 2009; Pietschnig & Voracek, 2015). This debate aside, since many measures of *Gf* do not rely on language, they are an important target assessment for assessing *g* in nonhumans (see below). While the ideal way to measure any construct is with a variety of measures, the Raven's Progressive Matrices (RPM) is a quintessential example of a *Gf* task (Carpenter et al., 1990; Nisbett, 20009). The RPM is a series of partially completed matrices, where the participant is tasked with selecting the choice option that will correctly complete the matrix from a set of

distractors (Raven, 1941). Each item in the matrix is transformed, sometimes in multiple ways, across the rows and columns of the matrix. The participant must infer the underlying rules and correctly apply them to find the correct answer (Raven, 1941, 2008). The test items progressively increase in difficulty, with few people correctly answering the final questions (Carpenter et al., 1990). Despite the strong visuo-spatial component, the RPM is used as a measure of reasoning (Schweizer et al., 2007; but see Gignac, 2015; Stephenson & Halpern, 2013).

Gc reflects the ability to correctly use and apply learned knowledge (Kvist & Gustafsson 2008). Some researchers have emphasized the role of language and verbal storage, and rely strongly on vocabulary measures to assess this ability (Ackerman & Cianciolo, 2000; Reynolds & Turek, 2012; Rolfhus & Ackerman, 1999). When a more diverse battery is used, however, the extent to which verbal comprehension overlaps with *Gc* has varied (Carroll, 2003; Kan et al., 2011; Schipolowski et al., 2014). In humans, knowledge is typically gained and tested through language, which could have led to the debate about what is the nature of *Gc* (Keith & Reynolds 2010; Schipolowski et al., 2014). Nevertheless, despite *Gc* being predominately measured through language, language is not the only way to assess knowledge. As we discuss in more detail later, non-language methods are needed to assess *Gc* in nonhuman animals. Researchers that have utilized a more comprehensive test of knowledge to measure *Gc* have found that it loads highly onto *g* and is a better predictor of academic and job performance compared to *Gf,* particularly for older adults (Postlethwaite, 2011; Schmidt, 2014).

Despite these general issues with group factors, it has been consistently found that *Gf* and *Gc* load highly onto the *g* factor and co-vary with each other (Carroll, 1993; Schipolowski et al., 2014). One potential explanation for this co-variation is provided by investment theory (Kvist & Gustafsson, 2008). Since *Gf* influences a person's ability to understand or learn from novel

problems, it is theorized that *Gf* is used when learning new information. As this information is acquired, it then becomes *Gc*. The initial ability level, determined by *Gf*, will determine the efficacy of this investment or learning when it is specifically directed according to interest, leading to more specific knowledge gains, or during more passive exposure, leading to more general gains (Cattell, 1987; Schmidt, 2011, 2014).

It is possible to discuss *g* at two different levels, as a statistical finding referred to as psychometric *g*, and as a psychological construct. Psychometric *g* is not controversial (Blair, 2006; Carroll, 1993; Jensen, 1998). Positive correlations across diverse cognitive tasks are no longer seen as surprising. Further, many researchers do not argue that a factor analysis will produce one factor that can account for half of the variance (Conway & Kovacs, 2015; van der Maas et al., 2006). The status of *g* as a psychological construct, however, is still heavily debated. Despite the fact that *g* has been consistently reported in over a century's worth of research, and we know which tasks consistently load highly onto *g*, there remains no consensus as to what *g* actually is (Carroll, 1993; Cattell, 1987; Chabris, 2007; Deary, 2000; Gottfredson, 2002; Gustafsson, 1984, 2003; Jensen, 1998; Kovacs & Conway, 2016; van der Maas et al., 2006). *g*'s ontological status remains a mystery. The following is brief overview of many popular theories of *g* today. Our aim is not to review an exhaustive list of all current theories of *g*, nor a nuanced treatment of the theories that are discussed. Additionally, this paper is not an endorsement of any particular theory of *g*. Rather, the goal of this paper is to provide a general background about theories of intelligence for readers outside the expert community. Furthermore, we also do not cover the vast literature on cognitive abilities in infants or the developmental aspects of general intelligence. Again, this is because we ultimately are interested in discussing tests of *g* in adult nonhuman animals. Undoubtedly, once such tests can be reliably developed, they should allow

for the investigation of how developmental processes contribute to *g* in nonhuman animals, but such a discussion would be premature now. For excellent empirical research and theories on the development of *g*, see Blaga et al. (2009); Bornstein et al. (2006); Coyle et al. (2011); Demetriou et al. (2018); Fagan et al. (2007); Rose et al. (2008); and Spinath et al. (2003).

One proposal is that *g* is a single entity that is related to a wide variety of cognitive abilities because it *causes* differences between individuals in those abilities (Brown et al., 2006; Carroll, 1993; Gustafsson, 1984; 2003; Schmidt, 2011; 2014; 2017 right panel of Figure 1). Even though this perspective purports *g* as a single entity, it does not necessarily reflect one physical structure or psychological process (Jensen, 1998). In attempting to identify the physical substrates of *g*, a variety of results have been found including relevant genes (Plomin & von Stumm, 2018), neural networks (Duncan et al., 2000), neural substrates (Schmitt et al., 2020), and developmental processes (Garlick, 2002). It is unlikely that any one of these alone is responsible for *g* and more likely that there is a dynamic interaction between all of these physical substrates and with the environment (Ceci, 1991; Chabris, 2007; Garlick, 2002; Jensen, 1998; Kan et al., 2013; Schmitt et al., 2020; van der Maas et al., 2006). Indeed, some researchers argue that schooling has robust and potentially causal effects on the physical substrates that could underlie *g* (Baker et al., 2015). At the psychological construct level, other researchers theorize that elementary cognitive process underlie *g*, meaning that differences in one or more of these basic abilities is predominately the reason behind differences in *g* (Gignac, 2014; Jensen, 1998, p 260). Working memory (WM), short-term memory (STM), processing speed, associative learning, and response inhibition have all been proposed as components of *g* (Conway et al., 2002; Deary, 2000; Dempster, 1991; Jensen, 1998; Kaufman et al., 2009; Sheppard & Vernon, 2008; left panel of Figure 1).

Other researchers argue that *g* is actually a statistical artifact. These theories state that more complex tasks (which are more *g* loaded) require a broader array of *independent* resources. Even though these processes are independent, the nature of the tasks creates a correlation (Bartholomew et al., 2009; Kovacs & Conway, 2016) or that the developmental trajectory creates mutually beneficial interactions between independent abilities (Rose et al., 2008; van der Maas et al., 2006). In the next section, we review the relationship between *g* and the cognitive mechanisms listed earlier to investigate this issue. These cognitive mechanisms were investigated because of the rich literature that is available to review and because the potential role they play when investigating *g* across species.

Related Cognitive Factors

Working Memory

WM describes the ability to hold a limited amount of information over the short term (seconds to minutes). What differentiates WM from STM is that WM involves manipulating the stored information or engaging in a secondary task while the to-be-recalled information is held in memory (Baddeley 2003; Conway et al., 2002). For example, STM might involve holding in memory a list of items until their recall is requested, while WM would involve performing mathematical operations, counting, or some other transformation while encoding a list of to be recalled items. Some common WM tasks are the complex span task, *n*-back task (Au et al., 2015; Shelton et al., 2010), and reverse span task (Oberauer et al., 2000).

In the complex span task, there is a competing demand that is interspersed between to-be-remembered items (Conway et al., 2002; Engle et al., 1999). For example, in the operation span task, participants must verify if the solution to a given equation is correct or incorrect, before

being presented with the to-be-recalled word, letter, or number (Conway et al., 2002). Different variations of the complex span task include verifying if a sentence is logical or counting the number of squares before being presented with the to-be-remembered items. The degree of interference between the interleaved task and the items for recall can vary by changing the similarity between them. Some experiments found worse recall performance when the interleaved task and the to-be remembered item are highly similar, for example both involve words or visuospatial judgements (Jarrold et al., 2011; Shah & Miyake, 1996), but this is not consistently found across all item types (Bayliss et al., 2003). Performance on these different span tasks are correlated, but the correlation is different from unity (Bayliss et al., 2003; Conway et al., 2002). It is possible that this is partially due to measurement error, but an exploratory factor analysis extracted three factors, which they interpreted to be verbal storage, visuospatial storage, and a general processing factor (Bayliss et al., 2003). This indicates that each span task captures more specific and general properties of WM, which is consistent with theoretical conceptions (Baddeley, 2003).

In the *n*-back task, participants are presented with a continuous stream of items. As each item is presented, the participant must decide whether it matches an item presented *n* trials ago, with the range typically extending from 0-3 (Jaeggi et al., 2008). For example, in the stream: fish, peanut, cup, fish, pipe, dog, dog, car, phone, car; a response would be required to the second presentation of ‘dog’ in a 1-back task, ‘car’ in a 2-back task, and ‘fish’ in a 3-back task. Generally, the larger the *n*, the more difficult the task. Thus, participants must continuously update the items in their WM, while simultaneously comparing each current item to the appropriate item n-back.

In the reverse-span task, participants are presented with a series of letters or numbers, and then they must repeat them in reverse order. This means that participants are simultaneously holding and transposing the information so that it can be presented in reverse order (Jensen, 1998; Oberauer et al., 2000).

Performance on these WM tasks is correlated with measures of *g*, but the reason for this correlation is not well understood. Some researchers have shown that WM training improves performance on WM and highly *g*-loaded tasks, indicating that WM is a subcomponent of *g* (Jaeggi et al., 2008; Schmiedek et al., 2010). Others argue, however, that these improvements are hollow – that is, they stem from non-*g* factors, like test familiarity or strategy adjustments during test battery completion (Colom et al., 2002; Colom et al., 2013; Estrada et al., 2015). Additionally, not all researchers have shown WM-training effects on *g* (Chooi & Thompson, 2012; Harrison et al., 2013; Redick et al., 2013). These mixed effects could mean that WM is used during these *g* loaded tasks in holding necessary information, but the ability to correctly identify which information is necessary is unique to *g*. Being able to hold more information or handle competing demands more effectively does not necessarily indicate an improved ability for abstract reasoning. This dissociation between WM and abstract reasoning could indicate that there is a causal relationship between WM and *g*, but that differences in *g* cause differences in WM, not the other way around. Theoretically, from this perspective an increase in *g* should result in an increase in WM performance as well. Alternatively, there could be another, more general factor that underlies the efficacy of WM and *g*-loaded tasks. Training on WM that fails to improve this underlying factor should result in little impact for performance on *g* loaded measures. Finally, WM and *g* could be independent cognitive abilities and the reason for the correlation is due to task impurity.

The diversity of tasks used to measure WM obstructs determining the relationship between *g* and WM. Each type of WM task places different demands on WM, and thus they may not all be measuring the same construct (Aben et al., 2012). Supporting this is the fact that these tasks do not always strongly correlate with each other (Au et al., 2015; Jaeggi et al., 2008; Kane et al., 2007; but see Schmiedek et al., 2014; Wilhelm et al., 2013). Furthermore, WM tasks sometimes strongly correlate with STM measures (Aben et al., 2012; Colom et al., 2008; Conway et al., 2005; St Clair-Thompson, 2010; Figure 1), further obscuring relationships between tasks and the underlying constructs they purportedly measure. Likewise, WM is not necessarily a unitary construct, but may itself consist of separate processes, such as attention (Baddeley, 2003), processing speed (Unsworth et al., 2009), and STM capacity (Conway et al., 2003), among others (Kovacs & Conway, 2016; Schmiedek et al., 2014; Wilhelm et al., 2013). Support for the relationship between *g* and these other processes have all been reported (Chuderski et al., 2012; Conway et al., 2003; Unsworth et al., 2009; Figure 1). Thus, it is possible that different WM tasks differentially tap into these alternative processes (or subcomponents).

Short-Term Memory

STM is the ability to hold information over a delay period, without an explicit competing task or manipulation requirement (Unsworth & Engle, 2007). The information being held in STM is subject to capacity limits and decay over time (Cowan, 2008). Performance on WM and STM tasks tends to be correlated, likely because both involve the short-term retention of information (Aben et al., 2012; Colom et al., 2008; Conway et al., 2002; Figure 1). STM and WM are not dichotomous constructs; rather, tasks fall on a continuum depending on how demanding is the secondary task (Aben et al., 2012; Engle et al., 1999). For example, requiring

participants to repeat a letter, preventing them from verbally rehearsing the to-be-remembered items, is still considered a STM task since the secondary task is of low difficulty (Conway et al., 2002; Engle et al., 1999). Nevertheless, as discussed above, WM does have some unique properties (Conway et al., 2002; Cowan, 2008). Some researchers have found that STM alone is related to *g* using SEM and CFA (Colom et al., 2008; Martínez et al., 2011; Figure 1). Yet others have failed to find a unique relationship between STM and *g* using SEM, but these utilized different construct measures (Conway et al., 2002). Given these challenges, researchers tend to focus on WM when investigating the relationship between STM processes and *g*.

Processing Speed

The term "processing speed" is used to describe a variety of tasks that can vary in complexity and memory demands (Deary, 2000). These tasks typically assess how quickly a participant can detect a change in the environment, perceive the difference between two stimuli, or transform stimuli. Some tasks that require detecting a change in the environment are based off of Hick's law, that reaction time (RT) will increase linearly with increases in information a task requires (Hick, 1952). The Jensen box is an example of an apparatus that utilizes the principal of Hick's Law (Deary, 2000; Jensen & Munro, 1979). The Jensen box consists of a home key and 1, 2, 4, 6, or 8 stimuli placed equidistant from the home key in a semi-circular arrangement. Participants must keep their finger resting on the home key until one of the stimuli in the array changes (e.g., color or brightness). Participants are instructed to touch the changed stimulus as quickly as possible (Deary et al., 2001; Jensen, 1982; Vickrey & Neuringer, 2000). Even at its most simple, when the array only has one stimulus, there is still a relationship between RT and intelligence, with the correlation ranging between -.18 to -.22 (Deary, 2000; Doebler & Scheffler, 2016; Sheppard & Vernon, 2008). Another processing speed task is the digit-symbol

substitution task (Conway et al., 2002; Hoyer et al., 2004). In this task, participants are given a conversion table of digits and a corresponding symbol. The symbols are usually simple shapes or a series of connected lines that do not resemble letters. Participants must complete a table of numbers with the appropriate corresponding symbol, or they are shown various digit-symbol pairs and must determine if the pairs are valid or invalid according to the conversion table as quickly and accurately as possible (Conway et al., 2002; Hoyer et al., 2004). The conversion table is always present, so this task does not rely on memory processes. This task is commonly included in FSIQ tests (Benson et al., 2010). Even though processing-speed tasks appear simple, they show a consistent, modest relationship to *g*, with correlations typically ranging from -.22- to -.4, such that faster or shorter RTs correlate with higher scores on intelligence tests (Deary, 2000; Doebler & Scheffler, 2016; Sheppard & Vernon, 2008; Vernon, 1983; Figure 1). These two tasks are a very select subset of all the different processing speed tasks that are used (Deary, 2000; Sheppard & Vernon, 2008).

Why processing speed shows a consistent relationship with *g* is not well understood (Deary, 2000). There is some evidence that processing speed influences how quickly a competing task can be performed in WM tasks (Conway et al., 2002; Unsworth et al., 2009; Figure 1). Therefore, processing speed may only be related to *g* because it influences WM. When WM tasks are also included, processing speed is no longer directly related to *g* (Conway et al., 2002). It is also possible that processing speed, WM, and *g* rely on the same underlying mechanism or process. There are a large variety of processing speed tasks, however, so it is unclear if these different tasks measure the same underlying construct (Stankov & Roberts, 1997). Tasks used to show a relationship between processing speed and *g* differ greatly from those used to study how processing speed influences WM (Colom et al., 2008; Conway et al.,

2002; Deary, 2000; Stankov & Roberts, 1997). Thus, it is difficult to determine if processing speed and WM show the same relationship across all of these tasks.

Response Inhibition

Some researchers have suggested that response inhibition is a crucial factor underlying differences in intelligence (Dempster, 1991). Response inhibition is the ability to suppress unwanted motor responses or thoughts and can be measured with anti-saccade, Stroop, Go/No-go (GNG), and stop signal tasks (Friedman et al., 2006; Swick et al., 2011; Verbruggen et al., 2014). A reversal learning task is also used to a lesser degree to measure inhibition (Eagle et al., 2008; Izquierdo & Jenstch, 2012). In the anti-saccade task, participants must avoid moving their eyes towards a target and instead they must move their eyes in the opposite direction (Klein et al., 2010). In the Stroop task, participants must read the ink color of a word out loud, even when it conflicts with the word's meaning (i.e. the word "red" printed in yellow ink; Stroop, 1935). In the GNG and stop signal tasks, participants must make a motor response when they see one type of stimulus and withhold the motor response when they see or hear other types of stimuli. For the GNG task, participants receive successive presentations of two stimuli intermingled within the session. Responses to the positive discriminative stimulus (S+) are rewarded while responses to the negative discriminative stimulus (S-) are not rewarded. Initially participants typically make responses to both stimuli, but with further training learn to inhibit responses to the S-. The stop-signal task is similar to a GNG task. On some trials, only the S+ is presented, and participants are rewarded for responding to the S+. Occasionally a trial will initially present the S+, and after a short delay the S- is also presented. The participant is instructed to withhold responses when the S- is presented. Thus, the stop-signal task measures the ability of the participant to suppress behavior in the midst of preparing or making a response. (Swick et al., 2011; Verbruggen et al.,

2014). Finally, in the reversal-learning procedure, the first phase consists of a GNG procedure in which participants learn that one cue (S+) is associated with a reward, while the other is not (S-; note, the S+ and S- may be presented simultaneously rather than successively). Once discrimination performance stabilizes, the stimulus-outcome assignments are reversed (Eagle et al., 2008; Izquierdo & Jenstch, 2012). For example, after learning to respond to a blue circle (S+) and withhold responding to a yellow circle (S-), the blue circle becomes the S- and the yellow circle becomes the S+. Response inhibition influences how quickly the participant can inhibit the original learned responses, and replace them with new responses.

For reversal learning, no significant relationship has been found between intelligence and the number of trials needed to reverse the initial discrimination in children (Plendleith, 1956) or adults (Stevenson & Zigler, 1957). Using the anti-saccade task, the total number of errors (Friedman et al., 2006) and the errors with a regular latency (Klein et al., 2010) showed a modest correlation with intelligence. A similar result was found with the GNG (Horn et al., 2003) and the stop-signal tasks (Friedman et al., 2006). The Stroop task had low (Friedman et al., 2006) to nonsignificant (Polderman et al., 2009) correlations with measures of intelligence. When intelligence was broken down into *Gf* and *Gc*, only the anti-saccade task had a significant correlation with *Gf*, ranging from .19-.23, while the stop signal and Stroop task were not significantly correlated, with a correlation ranging from .03-.12. For *Gc* the correlations for these three inhibitory tasks were low, .12-.19, though 4 out of 6 were significant (Friedman et al., 2006). These three tasks loaded significantly onto the same 'inhibition' factor, but SEM showed that it did not explain any unique variance for *Gf* or *Gc* when other cognitive abilities were in the model (Friedman et al., 2006; Figure 1). This indicates that the modest correlations were the result of task impurity. Inhibition was not the cause of the correlations, but a correlation was

found because of the other cognitive factors that were also being used, potentially WM. Yet it is not possible to draw a strong conclusion about the relationship between inhibition and *g* due to relatively small amount of research that has been conducted.

Associative Learning

Associative learning is the ability to mentally link or associate specific stimuli together. One measure of associative learning is a simple discrimination task, where the participant must select between two stimuli. One of the stimuli is paired with a reward while the other stimulus is not. Selecting the rewarded stimulus does not seem to be related to intelligence in children or adults, but this is underexplored (Plenderleith, 1956; Stevenson & Zigler, 1957). The paired associates task and the three-term contingency task, however, show a more promising relationship with intelligence. In the paired associates task, participants are told to remember pairs of unrelated one-syllable words. During training, participants see the first word of the pair, then press a key to reveal the second word. A test usually follows immediately after the training phase, where the first word of the pair is given and the participant must type the second word (Alexander & Smales, 1997). A variation on this is the three-term contingency task. During training, one word serves as the stimulus and there are three response keys. When the participant presses the response key, a word is revealed. At test, the participant is shown the stimulus word and must type the correct word for each response key (Williams & Pearlberg, 2006). The paired associates and three-term contingency tasks are significantly correlated with each other, .43-.64, and with *g*, .31-.52 (Alexander & Smales, 1997; Kaufman et al., 2009; Tamez et al., 2008; Williams & Pearlberg, 2006). The paired associates and three-term contingency tasks are not pure measures of associative learning considering how much information needs to be stored and retrieved, which clearly relies on memory processes. As discussed earlier, WM and possibly

STM are related to *g*. Nevertheless, it has been found using SEM that associative learning tasks are uniquely related to *g* independent of the memory and retrieval requirements (Kaufman et al., 2009; Figure 1). This suggests that associative learning is another potential underlying cognitive mechanism of *g*, but the task needs to be difficult or complex to reveal such a relationship.

Related Cognitive Factors - Summary

Four of the cognitive mechanisms discussed, WM, STM, processing speed, and associative learning are all related to *g* to varying degrees (Figure 1). Why these factors are related is still being explored, and the relative importance of each factor is debated. The relationship between response inhibition and *g* is underexplored, but so far response inhibition does not seem to be related to *g* in any significant way. This is difficult to understand in the context of the relationship *g* has with task complexity since some measures of inhibition, like the Stroop task, appear more complicated than measures of processing speed, yet processing speed has consistent correlations with intelligence (Deary, 2000). Recently, how much unique variance processing speed can explain was called into question (Conway et al., 2002), but the consistency of the correlation is undeniable. Nevertheless, merely knowing which cognitive mechanisms are related to *g* does not provide much insight into what, exactly, *g* actually is. We know how to measure *g* and its validity as a predictor of many life outcomes, but over one hundred years of research has yet to elucidate its exact nature. Investigating *g* and its psychological correlates in nonhuman animals (the focus of the next part of our review) would open up new avenues of research into the biological and empirical nature of *g*, and perhaps break through the current impasse in human research on the subject.

g in Nonhuman Animals

The *g* factor has been found consistently in human samples with a variety of measures, but what about other species? Finding a *g* factor in nonhuman animals would enable the study of many important questions about *g*, such as its evolutionary origins, its effects on biological fitness, and questions about mechanism that are challenging or even impossible to study in humans, such as the role of genes, its neural underpinnings, and environmental determinants during development. Useful animal models for studying *g* would allow the latest tools to be applied, such as optogenetics, chemogenetics, other gene-modification techniques (e.g., CRISPR), powerful control of the individual's environment from conception to adulthood, and other forms of neural manipulation. Application of these tools would allow for unprecedented insights into the causal role of genetic, neural, and environmental factors in *g* and intelligence. These insights could, in turn, provide translational significance to understanding *g* in humans. Working with animals, either in a lab or other setting, it is easy to see individual differences in task performance, but it is not clear if these differences would be consistent across a variety of tasks, like what we see in humans with the *g* factor (Macphail, 1987).

Research on nonhuman animals has shown they are capable of extraordinary cognitive feats, like tool use in New Caledonian crows (Auersperg et al., 2011), the range of abilities demonstrated by Alex the African Gray parrot (Pepperberg, 2018), and many more than what can be listed here. While impressive in their own right, extraordinary performance by particular species, be it in a single, specialized task (e.g., the spatial memory of the Clark's nutcracker (Balda & Kamil, 1992), or tool use by the New Caledonian Crow), or by only a few subjects across many tasks (e.g., by the African Gray parrot; the Bottlenose dolphin (Herman, 2010)) does not provide insights into psychometric *g* that would be provided by consistent performances

across many tasks whereby stable individual differences replicate key aspects of psychometric *g* in humans (Macphail, 1987). These exceptional animals cannot be meaningfully discussed in a review about *g* since they have not been given test batteries designed to determine cross-task consistency in performance. The purpose of exploring the potential for psychometric *g* in other species is not to rank species in their intelligence. Rather, developing test batteries that can be applied across species can help illuminate the conditions under which cognitive abilities will show a pattern of positive correlations. Thus, an animal model for measuring *g* would open up new avenues of research into the environmental and neural contributions to psychometric *g*, which can inform on theories of the causes for the correlations that determine *g*.

To meaningfully relate task performance to *g* in nonhuman animals requires reliable measures of performance in standardized behavioral tasks. Recently, researchers have been investigating individual differences in cognition in nonhuman primates, mice, and birds using test batteries (Burkart et al., 2017; Shaw & Schmelz, 2017). Some of these test batteries, however, are inadequate. Some suffer from including too few tasks (Anderson, 1993), or the included tasks lack sufficient variety to derive meaningful individual differences (Locurto & Scanlon, 1998). Other batteries include tasks that are ill-defined, and therefore obscure the underlying constructs (Keagy et al., 2011). Finally, some test batteries do not adequately control for the way in which a particular species interacts with their environment, or non-cognitive differences between species, such as motivation (Bitterman, 1965; Macphail, 1987). For example, it may be difficult for a subject to use a tool with their beak when the tool was designed to be used with a hand (Krasheninnikova et al., 2019). Thus, while many studies of *g* in animals have found positive correlations across tasks, these deficiencies make it difficult to relate these studies to the *general* cognitive ability found in humans.

Nevertheless, some test batteries used in nonhuman animal research have enabled assessment of the underlying cognitive abilities (Table 1). We focus the remainder of this review on these stronger test batteries which provide evidence for a general factor of intelligence. This necessarily restricts our discussion to species for which sufficiently strong data are available. As a reminder, we are also not focusing on species differences in intelligence, but rather individual differences in psychometric *g* for various species. Thus, relatively *smart* species, such as crows, parrots, and dolphins, are not included, while cognitively humble species, such as mice, are. It is beside the point whether parrots are deemed *smarter* than pigeons, or that apes are *smarter* than mice, as we are not concerned with ranking species intelligence against each other, but rather finding in nonhuman populations, similar individual differences as have been consistently found between individual people. Indeed, even demonstrations of differences in cognitive prowess of various species are not sufficient evidence for true species differences in *general* cognitive abilities (Burkart et al., 2017; Macphail, 1987). Until these species are given a diverse battery of tests, it is impossible to comment on the consistency of performance across tasks, which is at the center of *g* in research on humans.

Using appropriate test batteries, evidence for correlations within subject have been found in chimpanzees (Hopkins et al., 2014; Woodley of Menie et al., 2015), cotton-top tamarin monkeys (Banerjee et al., 2009), rhesus macaques (Herndon et al., 1997), orangutans (Damerius et al., 2018), mice (Galsworthy et al., 2002, 2005; Kolata et al., 2005, 2007; Matzel et al., 2003, 2006), robins (Shaw et al., 2015), bowerbirds (Isden et al., 2013), and magpies (Ashton et al., 2018). The general factor found in these studies can explain from 18 to 64% of the variance in individual performance. Performance has been related to WM (Kolata et al., 2005) and is stable over long periods of time (Ashton et al., 2018; Hopkins et al., 2014). Despite this, these test

batteries sometimes fail to extract a general factor, such as studies in chimpanzees (Herrmann et al., 2007, 2010), mice (Locurto et al., 2003, 2006), and song sparrows (Anderson et al., 2017; Boogert et al., 2011). We next explore why evidence for a *g*-like factor in nonhuman species is not as reliable as in the human literature. We also discuss the value of *g* in nonhuman species in predicting fitness related outcomes.

Nonhuman Primates

Herrmann et al. (2007) developed the primate cognitive test battery (PCTB) to assess performance across human children (age 2.5 years) and adult nonhuman primates. The PCTB includes 15 tasks from social and physical cognitive domains, such as the understanding of physical objects, social cues, and causal relationships (Table 1). This test battery was specifically created in order to test different evolutionary theories on why humans seem to show more advanced cognitive abilities compared to nonhuman primates which is why tests of social abilities are included (Herrmann et al., 2007). Approximately the same test battery was given to children, chimpanzees, and orangutans, but the analyses conducted did not allow for the examination of how individuals performed across all tasks. A follow up paper examined the results from the children and chimpanzees in order to determine the structure of these cognitive abilities (Herrmann et al., 2010). Using CFA, for children they found evidence for three factors, physical, social, and spatial, which is surprising considering other research has shown evidence for a general factor for children in this age range (Spinath et al., 2003). For chimpanzees, they found evidence for two factors, spatial and physical-social that account for individual differences in performance. While the initial research indicated there may be a relationship in chimpanzees between boldness and performance on the physical tasks, where bolder chimpanzees had better performance (Herrmann et al., 2007), this relationship was not further elaborated (Herrmann et

al., 2010). The lack of a *g* factor for either species is surprising, though this could be due to a number of factors, particularly when examining the results for the chimpanzees. The authors acknowledge that their test battery contains a much higher proportion of social tasks compared to what is typically found in the literature (Spinath et al., 2003). They also acknowledge that a number of their test items had low variabilities, though they do not specifically state which tasks (p. 108). Finally, the reliabilities of the tasks for the chimpanzee sample ranged from .05-.66, which as discussed earlier, can weaken subsequent correlations. It is not entirely clear how or if this was controlled for in their subsequent analyses.

Another group of researchers used a modified version of the PCTB and found evidence of a *g* factor in chimpanzees using principal component analysis (PCA) that was stable across 2 years and was heritable, consistent to what is seen with humans (Hopkins et al., 2014). However, it is unclear how much variance in performance is explained by this *g* like factor in chimpanzees or if the modified version of the PCTB changed the task reliabilities. Additionally, while age and sex were collected as potentially confounding variables, it is not clear if any personality measures, like boldness, were taken. Using the same data set, a follow up study used different statistical techniques and confirmed both the presence of a single factor and its heritability, but it is still unclear how much variance is explained by this factor (Woodley of Menie et al., 2015). In an attempt to resolve the discrepancy in results, a reanalysis combined both data sets (Kaufman et al., 2019). Using CFA, they found evidence for a *g* factor and group factors for chimpanzees and children, however, the exact structure of these factors was different between the two species. Additionally, it was unclear how much variance in performance was explained by *g*. For chimpanzees, they confirmed that this was relatively stable over time, though performance

tended to improve during the second test. They reported that the stability coefficient for the PCTB was .5 compared to .96 for FSIQ tests given to children.

Evidence of *g* has also been found in cotton-top tamarins and orangutans using different test batteries for each (Banerjee et al., 2009; Damerius et al., 2018). The tamarins were tested on 11 tasks, including social tracking, reaching, and reversal learning (Banerjee et al., 2009). Participation in all tasks was voluntary. Data were collected in the form of ranks and Bayesian latent variable analysis was used. Using this method, they found evidence for a *g* factor, but no evidence for distinct group factors. They acknowledge, however, that the lack of group factors could have been due to low levels of reliability for some of the tasks. The orangutans were tested on five tasks, including response inhibition, causal reasoning, and reversal learning, all showing high levels of variability (Damerius et al., 2018). Using PCA, one factor was extracted that explained 31.28% of the variance in performance and all tasks loaded onto this factor, similar to what is seen in humans. While this research was conducted with orangutans at rehabilitation centers, there was variation in how much of their development occurred in the rehabilitation center versus the wild, which was related to differences in noncognitive factors. For nonwild subjects, there was a positive relationship between curiosity and *g*.

Rhesus macaques have also been given a test battery that included 6 tasks, including delayed nonmatch to sample (DNMS) and reversal learning, but the goal of this study was to determine if there were age related cognitive declines in this species (Herndon et al., 1997). Using PCA, the first component extracted accounted for 48% of the overall variance, but was significantly negatively correlated with age, indicating that older subjects performed worse on all tasks. While *g* is stable across individuals over time in human populations, there is evidence for age related declines in cognitive abilities, and that these declines are independent of *g* (Gow et

al., 2011). The relationship between *g* and age-related cognitive decline is complicated and outside of the scope of this review article, but the research by Herndon et al. (1997) indicates that it is likely that rhesus macaques have a *g* like factor.

One of the key differences from human research in test batteries for nonhuman primates and other species is the inclusion of social tasks (Banerjee et al., 2009; Herrmann et al., 2010; Hopkins et al., 2014; Table 1.1). In human research, intelligence and social ability appear to be separable domains and dissociable. People can show an impairment in social ability while performing normally on IQ tests, and vice versa (Adolphs, 1999). When humans with intact and normal brain functioning were tested on both measures of *g* and social knowledge, the correlation between the two measures was quite low (Derksen et al., 2002). Nevertheless, this low correlation could also result from comparing the subjective self-report measure of social knowledge to the more objectively measured *g* (Derksen et al., 2002). Studies with human children and adolescents indicate that general intelligence and 'Theory of Mind', or the ability to understand the mental state of another, are independent (Cavojová et al., 2013; Rajkumar et al., 2008), but these populations are older than the participants tested by Herrmann et al. (2007, 2010). For nonhuman primates, inclusion of 6 social tasks in the PCTB also failed to find a *g* factor (Hermann et al., 2010). Others suggest, based on reanalysis of these data, that these social tasks could be equivalent to *Gc*, the cultural-knowledge group factor seen in humans (Kaufman et al., 2019). This suggestion is premature, however, given that the operational definition and assessment of *Gc* in humans varies widely across labs (Kan et al., 2011; Keith & Reynolds 2010; Schipolowski et al., 2014). The relationship between social ability, cultural knowledge, and general cognitive abilities should be tested more thoroughly in humans throughout the lifespan in order to better establish their relationship.

Mice

The cognitive abilities for mice have been heavily explored by Locurto, Galsworthy, and Matzel (Table 1.1). Test batteries typically include measures of WM, associative fear learning, olfactory discrimination, and spatial memory, though the content and quantity of tasks varies. Unlike primate test batteries, mouse batteries frequently include measures of anxiety and overall activity levels, likely because these emotional responses are frequently studied in mice, especially in connection to fear learning and drug effects. Across a series of experiments, Locurto et al. (2003; 2006) devised cognitive test batteries for mice consisting of a visual nonmatch to sample (NMTS) task, spatial NMTS, spatial learning (Hebb-Williams Maze), detour problems, WM, place learning, olfactory learning and discrimination, fear conditioning, and operant acquisition. Briefly, the WM tasks have been a 4 (2006) or 8-arm radial maze and a variation of the radial maze task called the 4x4 task (2003). In the radial arm maze, there is a central platform with *n* enclosed arms radiating from it. Each arm contains a food reward and the subject is allowed to freely sample any arm at any time. Subjects entering an arm and failing to obtain the reward or entering an arm again after already obtaining the food reward were counted as WM errors. The 4x4 task also took place in the 8-arm radial maze. In the first phase, four of the arms contained a food reward, while the other four were blocked off. Once the animal had sampled all of the rewards, they were removed from the maze for 30 seconds. In the second phase, all of the arms were open, but only the four previously blocked arms were baited. Entering arms that had been rewarded in the first phase, entering the same arm twice in the second phase, and entering an arm for the first time, but failing to obtain the food reward were all counted as errors. The control procedures measured activity levels on land and in water in an open field chamber, and a light-dark preference test. The number of transitions in the light-dark

chamber and the distance travelled in the open chamber was termed activity. The time spent next to the wall in the open field chamber negatively correlated with the number of center crosses in the open field chamber. Together they were counted as an anxiety measure. The same control procedures were used with all cognitive test batteries (Locurto et al., 2003; 2006).

For the research conducted in 2003, the subjects were trained on the cognitive tasks until their performance reached asymptote. Multiple dependent measures were taken from each task and an aggregate score was used in the analysis, which had a reliability of .88. The average correlation between the cognitive tasks, however, was .12. When the correlational matrix of cognitive tasks and control measures was subjected to PCA, multiple independent factors were extracted (Locurto et al., 2003). In the follow up study of 2006, subjects were given fewer trials on the cognitive tasks, and only one dependent measure was used in subsequent analyses. This reduced reliability to .54 in Experiment 1 and .58 in Experiment 2. The average correlations between the learning tasks for these experiments were -.03 and .15, respectively. The authors state, "The relatively low reliabilities in the present study contributed to the relatively low average correlations observed," yet it does not appear as though these correlations were corrected for measurement unreliability (Locurto et al., 2006 p 382). PCA, including the control measures, revealed a similar result, where multiple independent factors were extracted.

Other researchers have not had similar results, even when using the some of the same tasks. Galsworthy and colleagues have also tested mice on a diverse battery of cognitive tests, but have found evidence for *g*. In 2002, Galsworthy et al., tested mice with two measures of spatial learning (Hebb-Williams and Morris Water Maze), spontaneous alternation in a T-shaped maze, a detour task, contextual memory, and a problem-solving task. Multiple dependent measures were used in the correlational matrix for some of these tasks. The reliabilities of these

tasks ranged from .68-.84. The control procedures measured anxiety with an open field arena, defecation in testing environments, and latency to swim to a visible platform. A correlation matrix with all of the cognitive tasks and the spontaneous alternation task revealed that a majority of the tasks were positively correlated, and some of the positive correlations were nonsignificant, with an average correlation of .2. When a PCA was conducted, the first component explained 31% of the total variance in performance. A separate PCA was conducted on the measures of anxiety. They found that the first component could explain 46% of the variance in anxiety, but this component did not significantly correlate with any of the cognitive measures or their *g*-like factor.

In a follow-up study, Experiment 1 used essentially the same test battery, but for Experiment 2 it was expanded to include a spatial reversal in the Morris water maze, a water plus maze, novel object exploration, and an additional problem-solving task (Galsworthy et al., 2005). Additionally, in Experiment 2, many of the tasks were shortened. For these experiments only one dependent measure per task was used in the correlational matrix and an aggregate performance score was used when appropriate. Reliabilities were only reported for each dependent measure, however, not the aggregate. For Experiment 1, reliabilities ranged from .47-.87, and in Experiment 2 they ranged from .03-.78. The mean correlation was .18 and 0.06 respectively. A principal component factor analysis (PCFA) resulted in one factor that could account for 32% of the variance in Experiment 1, and 19% of the variance in Experiment 2. They acknowledge that the low task reliabilities could have attenuated the subsequent *g* factor, but did not indicate that the correlations had be corrected in order to compensate for this (Galsworthy et al., 2005 p. 688).

Studies conducted in Matzel's lab used a test battery that consisted of egocentric navigation (Lashley III maze), passive avoidance, spatial learning (Morris water maze), odor

discrimination, and fear conditioning (Kolata et al., 2005, 2007, 2008; Matzel et al., 2003; Sauce et al., 2014). These tasks were administered in such a way to ensure variability between subjects and capture differences in learning (Kolata et al., 2008) An open-field arena was used to determine anxiety and activity levels. An analysis similar to Galsworthy et al. (2002, 2005) was conducted. Performance on the cognitive tasks showed a uniformly positive correlational matrix and PCA extracted one component that explained 38% of the variance (Matzel et al., 2003). When the behavior in the open field was analyzed, only the amount of time spent away from the walls was significantly related to the general factor. This type of behavior, spending time in the open part of the open-field arena, is thought to reflect novelty seeking. As with humans, subsequent studies found this factor to correlate with WM, which was assessed with two 8-arm radial mazes (Kolata et al., 2005; Sauce et al., 2014). This factor also correlated with performance on a mouse version of the Stroop task (Kolata et al., 2007). In humans, such a relationship has received only mixed support, however it is underexplored (Friedman et al., 2006; Polderman et al., 2009). Pooling across prior data sets (n=241) produced a sample size with substantially more power. With this sample, the average correlation was .22, a magnitude similar to what they had found in the individual studies, but they did not report the task reliabilities. PCA confirmed a general factor that accounted for 38% of the variance and identified a potential group factor of spatial ability (Kolata et al., 2008). This strengthens the similarity between humans and mice in the structure of cognitive abilities.

To recap, for mice, one lab has had consistent success in capturing a general factor for cognition using their test battery (Kolata et al., 2005, 2007, 2008; Matzel et al., 2003; Sauce et al., 2014), while other labs have had more inconsistent results (Galsworthy et al., 2002, 2005; Locurto et al., 2003, 2006; Table 1.1). One key difference comes from how control measures are

incorporated into the data analysis. When the control measures are entered into the factor analysis, a *g* factor is not extracted (Locurto et al., 2003, 2006). When the control measures are subjected to a separate factor analysis, and a correlational analysis is used to determine if the factors are related, typically a *g* factor that can account for approximately 30% of the variance is found (Galsworthy et al., 2005; Matzel et al., 2003; Kolata et al., 2008). The latter is the method of correlated vectors, and while some human researchers have advocated for its use (Jensen & Weng, 1998) other researchers have identified potential issues with its use (Ashton & Lee, 2005; Wicherts, 2017). It is also not always clear why certain dependent measures are being collected in cognitive tasks with mice (Locurto et al, 2006). Unlike in human intelligence tests where there is one dependent measure for each task, with mouse studies multiple measures are typically collected. If the rodent *g* is as robust as the human *g*, we would expect to see a similar positive correlational matrix in each species, regardless of task battery composition or dependent measures collected. When constructing test batteries for humans, however, tasks are chosen specifically because they are known to load highly onto *g*, and avoided if they don't. This bias could artificially strengthen the correlation between tasks (Locurto et al., 2006). The discrepancy between Galsworthy's, Locurto's, and Matzel's labs in data analysis and success in finding a general factor should be investigated further, possibly by standardizing certain methods to ensure minimum between lab variation.

Avian Species

The structure of cognition has also been explored in a wide variety of avian species, including song sparrows (Anderson et al., 2017; Boogert et al., 2011), robins (Shaw et al., 2015), spotted bower birds (Isden et al, 2013), and Australian magpies (Ashton et al., 2018; Table 1.1). Given the more distant relationship between birds and mammals (~350 mya), investigation of *g*

in birds could provide insight into the phylogenetic depth of general intelligence (Figure 2). Similar results across birds and mammals could also result from convergent evolution, where a general cognitive factor evolves independently across multiple species due to similar environmental conditions or social structures. Likewise, since research with birds, especially pigeons, often uses similar methods and procedures as used in human cognitive research (e.g., behavioral psychophysics experiments using visual touchscreen operant chambers), birds provide a powerful tool, similar to nonhuman primates, with which to tease apart the relationship between *g* and its underlying cognitive components. For non-pigeon avian research, test batteries typically consist of acquisition of novel operant behavior, discrimination learning, reversal learning, spatial/reference memory, and response inhibition (Table 1.1). Response inhibition is assessed with a detour tube task. In this task, subjects are presented with a transparent tube with a visible food reward inside. The tube is positioned such that the subject must inhibit the direct approach to the food, and instead move away from the reward to access it from the side of the tube (Kabadayi et al., 2018; van Horik et al., 2018).

Wild male song sparrows were administered the motor learning, color association, color reversal, and the detour task in a laboratory environment. The number of songs in their repertoire was also collected. Song learning is thought to encompass cognitive abilities due to the process of learning songs from other males, directly or through recordings, during the critical period early in life. Once males reach sexual maturity, they produce crystalized song typical of adults of that species. If song learning was influenced by general cognitive ability, it would be a potential mechanism for mate choice for cognition (Boogert et al., 2011). The correlational matrix for the cognitive abilities was not uniformly positive and the average correlation was .248. PCA extracted 2 components, where the first component accounted for 45% of the variance and the

second component accounted for 33% of the variance. The color association and color reversal learning tasks loaded positively onto the first factor, the motor learning task had a weak negative loading, and detour performance had a strong negative loading. This negative loading indicates that the detour tube task is measuring something different compared to the other tasks. Song repertoire size showed a complicated relationship with these cognitive tasks. Larger song repertoires were associated with faster performance on the detour task, but slower performance on the reversal learning task. Song repertoire size was negatively correlated with detour performance, however, meaning that birds with a larger repertoire were faster at the detour task. The researchers acknowledge that differences in noncognitive factors like personality and experience could have influenced performance on these measures and be a potential factor in why a *g* like factor was not found.

A similar test battery, but with the inclusion of a spatial/reference memory task, was given to hand-reared male and female song sparrows (Anderson et al., 2017). Two measures of song accuracy were assessed in addition to repertoire size. The correlational matrix was not uniformly positive and many correlations were weak. The average correlation for males (n=19) was .101, but this actually decreased to .036 when females were added to create a larger sample size (n=38-41). PCA was conducted with the correlational matrix from the male subjects and two components were extracted from this test battery. Similar to the results with the wild population, the color association, reversal, and spatial learning task loaded positively onto the first component, but the detour task loaded negatively (Anderson et al., 2017; Boogert et al., 2011). All measures of song performance were positively correlated but, in contrast to the wild population, better performance on color reversal was associated with higher song quality while better performance on the detour task was associated with *poorer* song quality. This further

emphasizes that cognitive abilities in sparrows do not show a uniform relationship (Anderson et al., 2017; Boogert et al., 2011). While these studies did not allude to task reliability, a follow up study investigated the consistency of performance across time (Soha et al., 2019). Subjects were tested once a year for two or three years on the test battery used by Anderson et al. (2017). Performance and relative rank were not consistent across years for males or females, with the average correlation across time being .13. The relationship that cognitive performance had with measures of song accuracy also varied across years. This variance over time makes interpreting the initial studies difficult.

Research with other avian species has produced similar correlational matrices to what was found with song sparrows, though with stronger evidence for a general cognitive factor. Bower birds were given a problem-solving task, where they had to remove a novel barrier, novel motor learning, color discrimination and reversal, shape discrimination, and spatial memory. A majority of the correlations were positive and the average correlation was .26. PCA extracted two components, where all tasks loaded positively on the first component and it accounted for 44% of the variance, indicating a general factor. Whether this general factor was related to mating success was also studied. Male bower birds build elaborate nests (bowers), which appears to be a cognitively demanding task, to attract mates. Similar to song sparrows, if this nest building ability is generally related to performance on other cognitive tasks, it could be used as a signal by females for mate selection. Yet no consistent relationship between mating success and cognitive measures has been found (Isden et al., 2013). Wild robins were administered the same test battery as described by Anderson et al. (2017). A majority of the correlations were positive, with an average of .158, and PCA extracted two components. All tasks loaded positively onto the first component and it accounted for 34.46% of the total variance. The loadings onto the first

component were strengthened after removing potential non-cognitive confounds, like innate color preference (Shaw et al., 2015). However, the reliability of these tasks was not given.

A study with Australian magpies administered the same test battery that was given to song sparrows (Anderson et al., 2017) and robins (Shaw et al., 2015), though time to learn a novel motor behavior was not included in the correlational matrix or the PCA. They found uniformly strong positive correlations among cognitive tasks, with an average correlation of .465, and when given similar tasks 2 weeks later, performance was very reliable (.806-.975). PCA extracted one component that explained 64% of the variance in performance. Group size was related to this factor, where subjects living in larger groups performed better on these cognitive tasks. Furthermore, maternal cognitive ability was found to be the best predictor of reproductive success as measured by the number of fledglings produced and the number that survived to adulthood (Ashton et al., 2018). This contrasts with earlier studies that had only looked at the mating performance of males (Anderson et al., 2017; Boogert et al., 2011; Isden et al., 2013).

Thus, as with nonhuman primates and mice, evidence for *g* in avian species has yielded mixed results. In the song sparrow, performance on the detour task has a negative loading on the first factor extracted (Anderson et al., 2017; Boogert et al., 2011). Negative loadings are not seen in human studies of intelligence unless better performance is measured in the opposite direction of the other tasks (Jensen, 1998). In contrast, for the remaining species, robins, spotted bower birds, and Australian magpies, performance on all tasks showed positive loadings on the first factor and the first component accounted for an average of 47% of variance in performance (Ashton et al., 2018; Isden et al., 2013; Shaw et al., 2015). The detour task itself could be the reason for the different pattern of results. Follow-up studied with robins and pheasant chicks

found that better (i.e., healthier) body condition and experience with transparent objects reduced the number of ineffective pecks to the transparent wall (Shaw, 2017; van Horick et al., 2018). Noncognitive factors could be influencing performance on the detour task and obscuring a general factor in sparrows.

It is also possible that these species are under different evolutionary pressures which has created differences in how cognitive abilities are related. The predictive value of *g* in humans has been strongly linked to outcomes that are the products of cultural evolution that themselves can vary substantially across individual, such as occupation and education attainment. Thus, exploring *g* in an ecological/evolutionary context could help illuminate why certain tasks load more highly onto *g* than others. Avian species that show nonsignificant positive and negative correlations on these cognitive tasks might be under different evolutionary pressures than those showing significantly, uniformly positive correlations. Evolutionary theories are elaborated on later, but briefly, Australia is home to spotted bower birds, which has a weak correlational matrix, and magpies, which has uniform, positive correlations. These species differ in terms of how they interact with conspecifics and humans, with bower birds being more isolated, which could be driving differences in how performance on these tasks are related (Ashton et al., 2018; Isden et al., 2013). The consistency of the test batteries given to these different avian species makes it easier to theorize about which factors are causing the differences in performance, a strength of the studies conducted so far. Testing wild subjects allows for a more nuanced understanding of how cognitive ability can impact reproduction and survival, and how this could interact with environment and social structure.

One species that is conspicuously absent from avian studies of intelligence is the pigeon. This is surprising given their long history as research subjects in psychology. Pigeons show

evidence of cognitive processes typically studied in human and nonhuman primates, such as abstract reasoning (Blaisdell & Cook, 2005; Katz & Wright, 2006), rule learning (Garlick et al., 2017), WM (Cook & Blaisdell, 2006; Kangas et al., 2011; Lind et al., 2015), associative learning (Cook et al., 2005), artificial grammar learning (Herbranson & Shimp, 2008), inhibition of return (Cook et al., 2012), memory interference dynamics (Wright et al., 1985), and choice RT (Vickrey & Neuringer 2000). Often these cognitive processes are assessed in similar ways in pigeons as they are in human and nonhuman primates facilitating cross species comparisons. These cognitive processes seem to be supported by similar neuroanatomical structures, indicating that there is some restriction on how certain cognitive abilities evolve (Colombo & Broadbent, 2000; Colombo & Scarf, 2012; Divac et al., 1985; Güntürkün, 2005). These similarities suggest that, if pigeons were given a comprehensive test battery, a *g* factor would emerge. Frequently, pigeon researchers use the same subjects across multiple experiments, so it is likely that many labs already have assessed subjects on a variety of cognitive tasks, making it even more surprising that no lab has yet correlated their performance on different tasks. We are currently assessing pigeons on a test battery to measure *g*, with the addition of a novel reasoning task – a modified version of the RPM (mRPM; Flaim & Blaisdell, 2021)—an assessment of *Gf* in humans (Raven, 2008). This factor loads highly onto *g*, yet is completely absent from any study in nonhumans. Briefly, for the mRPM task, subjects must learn an abstract rule to identify the rewarded stimulus, and transfer learning of that rule to novel stimuli. Preliminary results indicate that the mRPM is sensitive to individual differences in rule acquisition and transfer (Flaim & Blaisdell, 2021).

Discussion

Most animal studies have revealed a similar cognitive structure as found in humans. Nevertheless, weaknesses in animal test batteries make it difficult to determine if they extract the same factor across species. Test batteries for nonhuman animals sometimes assess abilities that are underexplored in humans. In primate studies, for example, social tasks are frequently included (Derksen et al., 2002), while avian batteries always include a measure of inhibition—both of which are underexplored in humans (Dempster, 1991). Even when the same cognitive abilities are tested, the methods are vastly different. The 8-arm radial maze is commonly used in animal studies of WM, but this is not the way WM is assessed in human studies of *g* (Conway et al., 2005). WM is theorized to have domain *general* properties that would result in similar performance across specific task stimuli, as long as the tasks had similar demands (Unsworth et al., 2008). The 8-arm radial maze appears as though it has similar task characteristics to the WM tasks given to humans, since the subject has to maintain and update a list of locations within the trial. This would indicate that it is a valid measure to investigate WM across species, even though the particular task format has been designed to take advantage of the rat's species-specific tendencies. Nevertheless, as discussed earlier, WM is unlikely to be a unitary construct. Different aspects of WM have been emphasized in the different tasks used with humans, and each underlying aspect has shown a relationship with *g* (Kane et al., 2007; Unsworth & Engle, 2007; Unsworth et al., 2008). One group of researchers has investigated how these different aspects of WM are related to the *g* factor in mice, providing further evidence that WM and its relationship to *g* is similar across species (Kolata et al., 2007), but more research needs to be conducted before forming strong conclusions. Human performance on an 8-arm radial maze should be compared with more traditional measures of WM and measures of *g* in nonhumans

(Astur et al., 2004). Research with nonhuman animals should investigate WM with a broader array of tasks to determine if it also shows similar domain general properties and specific underlying processes (Kolata et al., 2007; Shaw & Schmelz, 2017).

Associative learning tasks pose a similar issue. In avian studies, with the exception of the study by Cook et al. (2005), studies of associative learning typically involve the acquisition of two associations (Anderson et al., 2017; Ashton et al., 2018; Boogert et al., 2011; Isden et al, 2013; Shaw et al., 2015). While underexplored, this type of associative learning task does not show a significant relationship with intelligence in humans, in children or adults (Plenderleith, 1956; Stevenson & Zigler, 1957). The associative learning tasks that are sensitive to differences in cognitive ability, the paired associates and three-term contingency tasks, have 10-30 unique stimulus pairs, placing more demands on learning, memory, and retrieval systems (Alexander & Smales, 1997; Kaufman et al., 2009; Tamez et al., 2008; Williams & Pearlberg, 2006). Differences in task design are expected when conducting comparative studies in order to accommodate different physical and sensory capabilities, in addition to other factors like motivation (Macphail, 1987). As mentioned earlier, this means different species will need different parameters in order to ensure that performance is an adequate reflection of cognitive ability, but greater care should be taken to ensure that the underlying construct is the same (see Wright et al., 1985 for a beautiful demonstration of this using the comparative method of systematic variation (Bitterman, 1975)).

Task purity is also a problem, with some tasks included in the batteries unduly influenced by personality, subject experience, experimental conditions, and physical health (Boesch, 2007; Kabadayi et al., 2018; Shaw, 2015; Sorato et al., 2018; van Horick et al., 2018). Finally, tests included in the battery should show high amounts of variability between subjects, but high

reliability within subject. Deficits in either of these elements will hinder detection of a *g* factor (Carroll 1993; Jensen 1998). Some tasks in the PCTB have low levels of between-subject variation, which may contribute to the difficulty in uncovering a general factor (Burkart et al., 2017; Hermann et al., 2010). In the test batteries for sparrows and mice, low levels of reliability may have attenuated correlations and weakened the general factor found (Cauchoix et al., 2018; Fan, 2003; Glasworthy et al., 2005; Soha et al., 2019). Low reliabilities attenuating the subsequent correlations were often mentioned in these experiments, yet these correlations were not corrected to compensate for this issue. For these species, however, the reliability and variability may not be an issue entirely with the tasks, but with the subjects. Task reliability will be higher in populations with higher variance in their true scores, that is, their scores independent from random error. Populations with higher true variance could produce a stronger *g* factor since there is more variance available to be accounted for. While tasks still need to be carefully constructed in order to show between-subject variability on the one hand, and within subject reliability on the other hand, potential differences in true variance across species should be kept in mind.

When a general factor has been found in nonhumans, the correlational matrix across task performance is not as robust as what we see in humans (Banerjee et al., 2009; Carroll, 1993; Galsworthy et al., 2002, 2005; Herndon et al., 1997; Hopkins et al., 2014; Isden et al., 2013; Jensen, 1998; Kolata et al., 2005, 2007, 2008; Matzel et al., 2003, 2006, 2008; Shaw et al., 2015; Woodley of Menie et al., 2015; but see Ashton et al., 2018). This is especially problematic when PCA is used to extract a *g* factor. PCA uses the total variance in the extracted components, even unique and error variance. This can result in overestimating the amount of variance the first extracted component can explain (Jensen & Weng, 1998). Some of the studies yielding poor

correlational matrices used PCA, which may have overestimated general cognitive ability in animal studies (Galsworthy et al., 2002, 2005; Isden et al., 2013; Matzel et al., 2003; Shaw et al., 2015; Table 1.1). This is not to dismiss the *g* factors that have emerged from weaker correlational matrices, but we need to understand why the correlations from nonhuman studies tend to be weaker. This could be due to low task reliability, as mentioned earlier. Sample size, however, is another factor impeding strong correlations, as most animal studies are underpowered (Banerjee et al., 2009; Galsworthy et al., 2002, 2005; Herndon et al., 1997; Hopkins et al., 2014; Isden et al., 2013; Matzel et al., 2001, 2006, 2008; but see Kolata et al., 2008). For PCA and other methods of factor analysis, it is recommended that there should be at least ten subjects for each measure, but few studies have achieved this ideal (Burkart et al., 2017; Costello & Osborne, 2005; Yong & Pearce, 2013). Some researchers have compensated for this by comparing empirical results to the results of a random bootstrapping procedure or randomly simulated data sets (Ashton et al., 2018; Damerius et al., 2018; Shaw et al., 2015), though this is not common practice. Another potential factor is the subject sample. Since the factors are extracted to explain variance in performance, the subject population must be heterogenous (Burkart et al., 2017; Yong & Pearce, 2013). Animal studies often lack heterogeneity, such as when studies of wild animals only test males (Isden et al., 2013), or bold individuals low in neophobia (Shaw & Schmelz, 2017). In lab environments, although outbred strains of mice are used, they are reared in nearly identical conditions, thereby diminishing inter-individual variance (Galsworthy et al., 2002; Kolata et al., 2005; Matzel et al., 2003). Environmental factors can make important contributions to cognitive abilities (Light et al., 2010; Neumann et al., 2007; Nisbett, 2009). Thus, the strength of strong environmental control of laboratory populations is also a weakness. A recent study showed that when mice were exposed to an enriched

environment for two weeks, their performance on a cognitive test battery improved (Sauce et al., 2018). Factors like environmental conditions and population characteristics should be further explored to understand how they could be affecting performance on cognitive tasks.

Given the strong interest in general intelligence in humans, establishing methods for identifying a *g* factor across diverse species should be a top priority of comparative cognition research. The cognitive abilities of many species are starting to be formally recognized and tested, but understanding how those abilities are related to each other remains a mystery. Under what conditions will species show evidence for general cognitive ability versus distinct and nonoverlapping cognitive abilities? What are the costs and benefits of having a generalized versus specialized system? Social structure/group size, diet, and environmental complexity/variability have all been proposed as determinants of cognitive abilities (Ashton et al., 2017; Herrmann et al., 2007; Mettke-Hofmann, 2014; van Horik & Emery, 2011). Group size, for example, has been theorized to increase cognitive abilities because larger groups put more demands on learning about and remembering more individuals, including their status within the group, and inter-individual interactions (van Horik & Emery, 2011). A weak correlational matrix was found in spotted bower birds (Isden et al., 2013), for example, while the correlational matrix found in the Australian magpies was stronger (Ashton et al., 2017). While these species show many behavioral similarities, including vocal imitation, sedentary lifestyle, and diet; they differ in their social interactions, breeding behaviors, and parenting. In bower birds, females select males based on bower attributes and mating display, but males do not assist with parenting (Isden et al., 2013). Additionally, there is evidence to suggest that interaction with conspecifics occurs primarily during breeding and mating in the form of competition, but less research has been published on bower bird behavior outside of bower activities (Madden, 2008).

By contrast, Australian magpie groups involve complex social behaviors, where members help to provision nestlings that are not related to themselves (alloparenting), and work together to defend their territory from predators and out-group members (Farabaugh et al., 1992; Finn & Hughes, 2001). The difference in social complexity might contribute to the different strengths of the correlational matrices in these two species. Group size can also explain differences in *g* within species as well. Within the Australian magpies, Ashton et al. (2017) found that cognitive performance improved as group size increased. This supports the idea that larger, more complex social groups are more cognitively challenging, thereby enhancing cognitive abilities of its members, but its potential explanatory value for magpies does not necessarily mean it will be able to explain differences across species. Enhanced cognition could be general (Ashton et al., 2017), or be restricted to social cognition. Herrmann et al's (2010) findings, and the low correlations between *g* and social ability in humans, suggest little effect of social complexity on *g*. Nevertheless, nuances within group size and social dynamics could help elucidate why these different results are found (Holekamp, 2007; Shultz & Dunbar, 2006).

Other researchers have argued that diet plays a substantial role in shaping cognitive abilities and brain function (DeCasien et al., 2017; Holekamp, 2007; Mettke-Hofmann, 2014; but see Allen & Kay, 2011). Having a varied diet (e.g., omnivorous or frugivorous) is associated with larger brains and/or higher cognitive abilities compared to species with specialized diets (e.g., folivorous). This could be due to increased demands on learning and memory systems posed by an omnivorous diet, improved diet quality, or the combination of the two. Nevertheless, research on the role of diet on cognition usually focuses on a single cognitive ability, such as innovation, or uses brain size as a proxy for cognition, rather than measuring *g* (Chittka & Niven, 2009; Roth & Dicke, 2005; Sol et al., 2016; Snodgrass et al., 2009). There is also evidence that

habitat complexity can influence brain size and rates of learning (Mettke-Hofman, 2014; Sayol et al., 2016; Schuck-Paim et al., 2008). These influences are not necessarily mutually exclusive, and may interact in their contribution to natural behavior (Lefebvre & Sol, 2008; Mettke-Hofman, 2014). It is possible that these influences will consistently co-vary. In cichlid fish, for example, environment complexity positively correlates with number of conspecifics (Pollen et al., 2007). A similar result was seen in African Starlings, where cooperative breeding is observed more frequently in complex environments (Rubenstein & Lovette, 2007). The potential for environment, diet, and social structure to co-vary makes it difficult to determine their independent contributions to brain size or cognitive abilities. Investigating a wider range of species could help answer this question. Noted by Holekamp (2007), spotted hyenas have high quality diets and complex social groups, whereas carnivorous and omnivorous bears also have high quality diets, but are predominately solitary. A better understanding of how diet, environment, and social structure impact specific cognitive abilities and brain size will also facilitate our understanding of how they relate to the underlying cognitive structure.

On the surface, it seems beneficial to have a larger brain and more advanced cognition. Larger brains are more diverse in function and structure (Roth & Dickie, 2005). Yet brains are metabolically costly and so selection for increased brain size usually requires specific environmental conditions and tradeoffs with other metabolically expensive organ systems (Burkart et al., 2017; Byrne & Bates, 2007; Chittka & Niven, 2009; Isler & van Schaik 2006; Iwaniuk & Nelson, 2003; Roth & Dickie, 2005). In humans, brain size increased as our digestive tracts shrank (Aiello & Wheeler, 1995). In birds, there is a negative correlation between brain size and pectoral muscle mass (Isler & van Schaik 2006). Increasing brain size and cognitive ability is not the only solution to meet environmental challenges, however. In birds, there is a

negative correlation between migratory distance and brain size, where the birds that traveled the furthest had the smallest brains (Sayol et al., 2016; Vincze, 2016). Yet, migratory birds show better long-term spatial memory compared to non-migratory birds, indicating that despite having smaller brains specific cognitive abilities can be selected for (Mettke-Hofmann & Gwinner, 2003). Detailed comparative studies can illuminate the conditions that support selection for *general* cognitive abilities versus specific cognitive processes (Chittka & Niven, 2009; Mettke-Hofman, 2014).

In a comparative analysis, better performance on cognitive tasks may not correlate with measures of fitness. In some species, fitness is increased through the selection of traits that attract mates or defeat rivals, yet with a concomitant decrease in brain size (Lefebvre & Sol, 2008). For other species, while potential mates that perform better on cognitive tasks are preferred (Chen et al., 2019; Spritzer et al., 2005a), this does not always result in increased fitness for those males or that females will act on that preference (Spritzer et al., 2005b).

Survival is another potential correlate with better cognitive performance in animals (Sol et al., 2007; Sol et al., 2008), though this correlation is not always found (Kotrschal et al., 2015). There is evidence for a complicated interaction between cognitive abilities and personality that could result in equivalent rates of survival despite differences in cognitive abilities across individuals (Mazza et al., 2018; Mettke-Hofmann, 2014). In great tits (*Parus major*), a species of song bird, individuals who were more competitive in maximizing a particular food resource during winter performed more poorly on a problem-solving task compared to less competitive individuals (Cole & Quinn, 2012). Although intelligence is predictive of health and longevity in humans (Murray et al., 2011), in modern societies this is more dependent on navigating environments that humans have created (especially schooling), not those created by the natural

environment (Flinn et al., 2005). No other animals have created, then subsequently had to resist, high-fat and high-sugar foods in order to prevent disease states. Understanding how cognitive abilities are related to survival in nonhumans will require the integration of multiple factors, including how cognitive abilities are interrelated. Species that show evidence for more interrelated cognitive abilities may have different interactions with noncognitive factors, like personality.

Another strength of animal research is that it could inform on different aspects that influence or are correlated with human intelligence. The benefit of using lab animals is they provide more control over biological factors which can be independently manipulated. The ability to closely monitor or manipulate brain function or genetics in animals can help elucidate which genes, brain regions, neural connections, and neurotransmitters are involved or correlated with cognitive functions, including *g* (Plomin 2001; Matzel et al., 2013). Animal models have already identified some neurobiological correlates related to cognitive ability, such as the importance of dopamine receptor function in mice (Wass et al., 2013, 2018), and cortical thickness and brain size in chimpanzees (Hopkins et al., 2018). As animal test batteries improve, is it likely that more neurobiological correlates will not only be identified, but manipulated to help determine their causal influence on cognitive performance.

While animal and human research investigating the physical substrate of intelligence is important, there are some misconceptions about how deterministic these neurobiological correlates are. This is partially because some researchers have consistently stated that differences in intelligence are due to differences in inherited genes that are not sensitive to environmental

factors (Jensen, 1998; Rushton & Jensen, 2005)[1]. This argument is sometimes supported by heritability estimates that state intelligence is 60-80% heritable (Bailey, 1997; Gillborn, 2016). This theory of intelligence being determined by genes has persisted, yet the theory is continuously criticized due to how heritability estimates are calculated, and because more recent findings on the relationship between genes and intelligence fail to support these heritability estimates. Heritability estimates are used to determine how much variance in a characteristic can be attributed to genetic differences at the population level. Heritability estimates are influenced by how variable the environment is and are unique to populations at the time of estimate, meaning the same characteristic can have different estimates depending on who is sampled and when (Nisbett, 2009; Sauce & Matzel, 2018; Tucker-Drob & Bates, 2016; Turkheimer et al., 2003). These high heritability estimates for intelligence are also difficult to reconcile with current genetic research. For most traits, the number of genes involved in the expression is large, and the effect size of each individual contributing gene is minute on its own (Allen et al., 2010; Beauchamp et al., 2011; Chabris et al., 2012). Furthermore, there is no evidence to suggest that these genes are insensitive to the environment (Bailey, 1997; Chabris et al., 2012; DeYoung & Clark, 2012). Some researchers argue that it is precisely a gene x environment interaction that could explain both high heritability estimates and low identification rates for specific gene variants (DeYoung & Clark, 2012; Sauce & Matzel, 2018). Heritability estimates typically over attribute variance in a trait to genes by including the gene x environment interaction in the estimate of heritability (Jensen, 1998; Sauce & Matzel, 2018). These results indicate that it is unlikely that differences in intelligence are due to immutable genetic factors. When

[1] Many of these theories were created in attempt to explain differences in IQ scores between races. A discussion of race, IQ, and genes is outside the scope of this review, but please see Frank, 2015, Krimsky & Sloan, 2011, and Nisbett, 2009 for discussion on why it is incorrect and harmful to posit race-based differences as innate.

neurobiological factors related to differences in cognitive performance are found they should not be presented as the sole and universal contributor to differences. Reductionist arguments like these could inadvertently perpetuate racist ideas (Gillborn, 2016; Phelan et al., 2013). Instead, these findings should be presented in the context of environmental interactions.

Investigating how neurobiological correlates of intelligence are related to the environment is easier with animal research due to the amount of control a researcher has on the environment. As mentioned earlier, short-term interventions that provide environmental enrichment improve performance on a cognitive test battery in mice (Sauce et al., 2018). Short term, intensive WM training increases dopamine (D1) receptor sensitivity and improves performance on a cognitive test battery in mice, highlighting the importance of even short-term interventions on biological substrates (Wass et al., 2013). Chronic environmental conditions and how that is related to cognitive performance could also be investigated. Animal research has already successfully modeled some of the environmental effects of development in a low socioeconomic status (SES) environment, including its neurological consequences (Hackman et al., 2010). SES correlates with intelligence, thus integrating these two lines of rodent research (environmental and genetic manipulations) could help uncover the causal direction of this correlation (Brooks-Gunn et al., 1996; Hackman et al., 2010; Jensen, 1998; Mani et al., 2013; Schmidt, 2017). Extended environmental manipulations will likely be key to understanding how chronic conditions impact cognitive function and the underlying neurobiological correlates.

Humans are a language-using species, and language enables much greater intelligence in our species than what is found even in other highly intelligent species (Penn et al., 2008). Furthermore, verbal fluency correlates positively with FSIQ (Ardila et al., 2000). Thus, it is difficult to disentangle the contribution of language to *g*. By studying non-language animal

models, we can gain insight into the cognitive processes and capacities that contribute to *g* that do not require, or that are independent of language (Shaw & Schmelz, 2017; Figure 2).

Well-developed test batteries for use in different animals, including humans, can help validate the neuroscience of *g* and its related cognitive mechanisms. Finding a general cognitive factor in animals has so far been only partially successful. Correcting methodological issues discussed in the previous sections will improve the search for a *g* factor in other species. Test batteries across all species, including humans, could be modified to facilitate comparative research. Tasks that have been used with both humans and other species that have not been included in test batteries are ideal targets for development. As discussed earlier, assessments of WM in rodent test batteries typically employ an 8-arm radial maze. Humans have been tested on a virtual radial arm maze, but this has not been incorporated into a larger battery or compared to more traditional measures of WM (Astur et al., 2004; Shaw & Schmelz, 2017). Reversal learning is another example of a cognitive task that is commonly included in animal test batteries, and is commonly used in humans to investigate neuropsychiatric disorders (Izquierdo et al., 2017), yet is underexplored in humans in relation to *g*. Furthermore, nonhuman animals should receive tasks that more closely resemble those used to study *g* in humans. For example, pigeons have shown similar RT effects on a variation of a human task based on Hick's Law (Vickrey & Neuringer, 2000). Including tasks like this in a test battery for animals would allow for increased correspondence between human and nonhuman animal measures of *g*.

Test batteries should also include more tasks where animals have to use previously acquired knowledge to solve novel problems (van Horick & Lea, 2017). Understanding how to apply knowledge beyond the trained situation is thought to explain why *g* is one of the best predictors of job performance (Schmidt, 2014). In the test batteries given to animals, there is a

debate about how ecologically relevant those tasks should be (Burkart et al., 2017; Herrmann et al., 2007). Nevertheless, if the goal is to discover *general* cognitive abilities, then it is not clear how important it is that the tasks in the test battery are ecologically relevant. The more ecologically relevant a task is, the more likely that they will engage highly-conserved behavioral processes (those often labeled as "instinctive"), with little inter-individual variation (Burkart et al., 2017). Using contrived and standardized tasks, such as in an operant chamber, can actually help control for noncognitive factors, like environmental experience, and facilitate comparisons across species (Clarin et al., 2013, but see Shaw, 2017).

Perhaps the most important factor is that test batteries should assess clear and separable domains of cognition as much as possible (Burkart et al., 2017; Shaw & Schmelz, 2017). Many studies, particularly those investigating cognition in the wild, use ill-defined tasks such as 'problem solving' or 'innovation'. This can make it difficult to determine what aspects of cognition are being used to solve the task, whether the same strategy is engaged across subjects, and if the behaviors are related to other cognitive abilities. Ultimately, there should be more communication across labs to determine that test batteries for different species attempt to assess the same underlying constructs, but which constructs should receive the most focus? As reviewed earlier, in humans, WM, processing speed, and associative learning have shown a relationship to *g*, though the causal nature of this relationship *g* is still debated. These basic cognitive processes have been found in just about all vertebrate orders, ranging from birds and mammals to amphibians and fish. Furthermore, these core cognitive processes reflect basic functions of vertebrate brains, often involving collaboration across multiple circuits, such as hippocampus, frontal cortex, and basal ganglia (Papini, 2008). By focusing on these core

cognitive processes, it is reasonable and possible to create a comprehensive cross species test battery that could extract psychometric *g* should it be present.

If psychometric *g* is found in a broad range of taxa, the causal factor may reflect a deep homology of the vertebrate brain despite species-specific brain and cognitive specializations (Güntürkün & Bugnyar, 2016; Osvath et al., 2014). This hypothesis requires testing, but such testing in turn requires the development of a test battery that can reliably assess these core cognitive functions across diverse species of vertebrate, from the human to pigeon to fish. Despite our suggestion that *g* should be assessed with the common set of general core cognitive processes of the vertebrate brain, this does not reject the idea that there are species-specific cognitive specializations found in individual species or groups of species. As an analogy, the 5-digit hand is a deep homology found in all tetrapods, and reflects the ancestral state. As a result, there are some common core functions of the 5-digit hand. There has also been selection for specialization in hand structure and function, such as the opposable thumb of humans that allows for fine motor precision, and even more extreme specializations for specific forms of locomotion, such as the wings of bats, the fins of whales, and the hooves of horses – each reflecting an adaptive specialization to each species' particular locomotor niche. Nevertheless, independent of these specializations, inter-individual variation in hand function within a species should be readily measurable using batteries of functional tasks, such as grip strength, dexterity, and precision, or locomotor functions. Likewise, as we discussed above, some birds that store seeds to be retrieved weeks or months later show specialized adaptation of spatial memory and the supporting brain systems, in particular the hippocampus. There is likely a complex relationship between specialized cognitive abilities and *g* due to differences in ecological constraints across species. Nevertheless, there ought to be inter-individual variation in spatial

memory in a species of food-storing birds, just as there are within a species of non-food storing birds, despite the fact that the food-storing species has an overall greater spatial memory than does the non-storing species. Appropriate tests that assess general cognitive functions are needed to facilitate assessments of *g* across a diverse array of species. Thus, assessments should be focused on the general cognitive processes, such as WM and associative learning, that are found in all vertebrates.

It is inarguable that one factor explaining half of the variance in performance on cognitive tests has been identified in humans (Lubinski, 2004). This factor is a good predictor of mortality, health, level of education, and SES. Furthermore, it is clear that this factor is most strongly related to WM and processing speed (Figure 1). What this factor consists of and what underlies its function is still under intense investigation. Better measures of a general factor in humans and animals could be an important effective tool to shed new light on general intelligence. Only then can we more clearly elucidate the evolutionary and environmental contributors to a general cognitive ability.

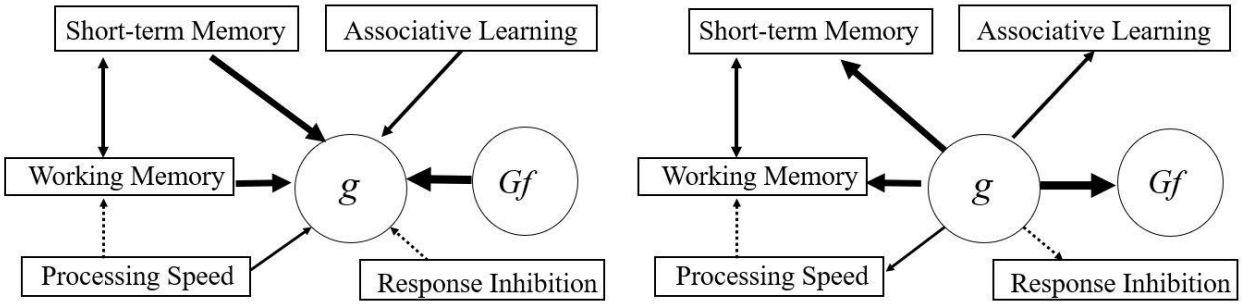


Figure 1.1. A diagram representing the reviewed cognitive abilities and their relationship to *g* and to each other. The thickness of the lines represents the strength of the relationship, while the type of line (solid or dashed) represents the consistency of the relationship. The direction of the arrows indicates the theoretical causal relationship.

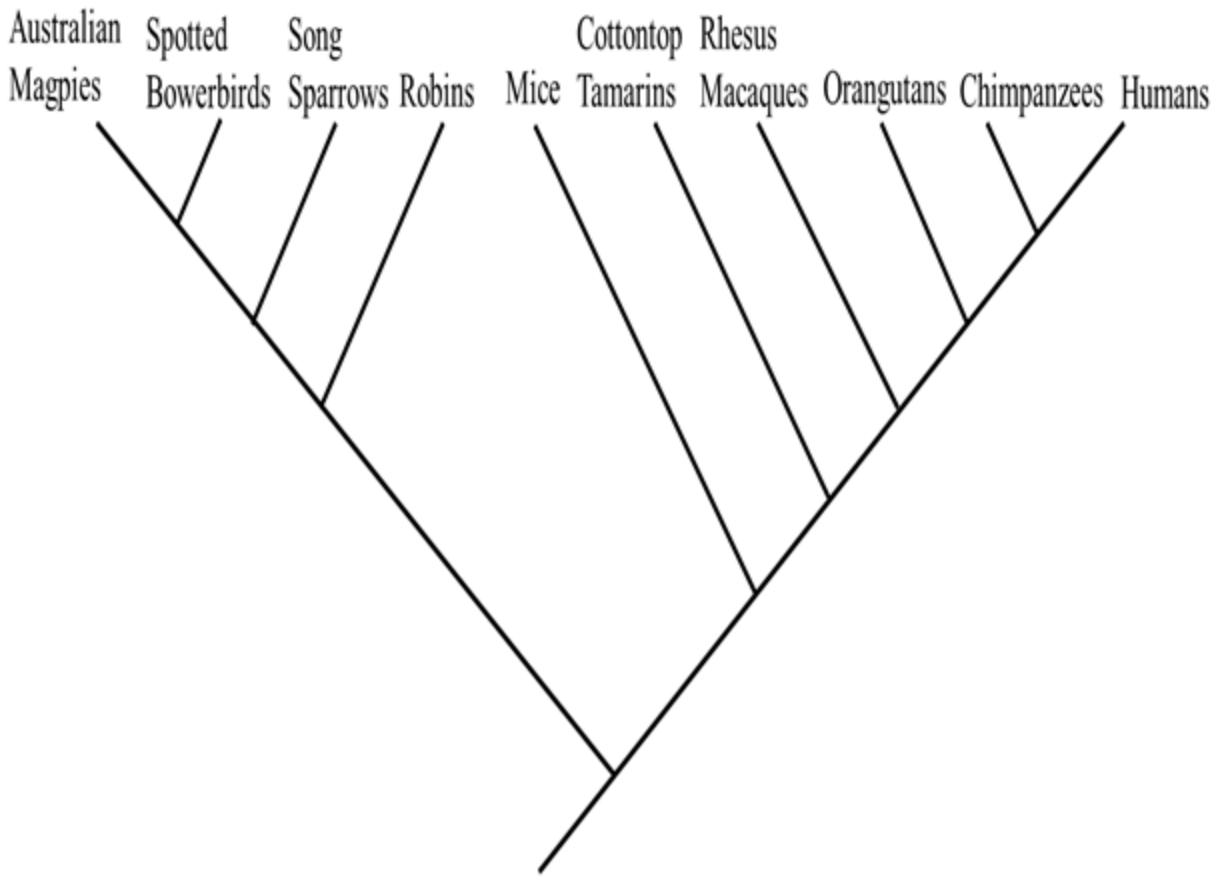


Figure 1.2. A cladogram of the species reviewed that have been given cognitive test batteries.

Table 1.1. Summary of cognitive test battery research in nonhuman primates, rodents, and birds.

Species (*n*)	Tasks	Correlational matrix	Analysis	*g*?	Other notable findings	Reference
Primates						
Rhesus monkeys (30)	Delayed nonmatch to sample (10 and 120s delay), delayed recognition span task – spatial and color, reversal learning – spatial and object	Not shown	PCA	Yes 48%	Performance declined with age	Herndon et al. (1997)
Cotton-top tamarin monkeys (22)	Occluded reach, targeted reach, A-not-B, reversal learning, exploration, numerical discrimination, acoustic discrimination, object tracking social tracking, hidden reward retrieval, food extraction puzzle	Not shown	Bayesian latent variable analysis	Yes, unclear how much the *g* factor accounted for		Banerjee et al. (2009)
Chimpanzees (106)	Spatial memory, object permanence, rotation, transposition, relative numbers, addition numbers, causal noise, causal shape, tool properties, social learning, comprehension, pointing cups, attentional state, gaze following, intentions	Not shown	EFA and CFA	No; 2 factors spatial and physical/social		Herrmann et al. (2010)

Chimpanzees (99)	Spatial memory, object permanence, rotation, transposition, relative numbers, causal noise, causal visual, tool use, tool properties, comprehension, production, attention state, gaze following	Not shown	PCA	Yes 54%	The *g* factor and 2 of the other components were highly heritable; a re-test 2 years later showed consistent performance	Hopkins et al. (2014)
Orangutans (53)	Box task, detour tube task, tube trap task, honey tool task, associative and reversal learning	Majority positive, none reached significance	PCA (confirmed with EFA)	Yes 36%	Curiosity correlated with g in captive only, not wild types	Damerius et al. (2018)
Rodents						
Mice (40)	Open field, spontaneous alteration in a T maze, Hebb-Williams, MWM, burrowing task, contextual memory, plug puzzle	Majority positive, 8/28 significant	PCA	Yes 31%		Galsworthy et al. (2002)
Mice (56)	Lashley maze, passive avoidance, MWM, odor discrimination, fear conditioning, control: open-field exploration, defecation in water/novel environments	Uniformly positive, 2/10 significant	PCA	Yes 38%	Propensity to explore was correlated with 4/5 tasks	Matzel et al. (2003)
Mice (60)	Hebb-Williams, plus maze, radial arm maze, visual nonmatch to sample, detour problems, control: light dark, activity measures in land and water	Majority positive, 3/15 significant (excluding controls)	PCA	No, 4 components that all explained the same amount of variance	Included the control measures in the PCA analysis; counterbalanced the task order	Locurto et al. (2003)

Mice (21)	Lashley maze, passive avoidance, MWM, odor discrimination, fear conditioning, radial arm maze with delay, dual radial arm maze, control: open field exploration	All in the appropriate direction, 9/28 significant	PCFA	Yes 40%	Only Working memory capacity was correlated with performance on the learning battery	Kolata et al. (2005)
Mice (exp. 1 = 84, exp. 2 = 167)	Exp. 1: spontaneous alt. in T-maze, Hebb-Williams, MWM, burrowing puzzle, plug puzzle; exp. 2: all tasks as exp.1 plus MWM reversal, syringe puzzles, water plus maze, object exploration	Exp. 1: uniformly positive, 6/15 significant; Exp. 2: majority positive, 15/55 significant	PCFA (replicated w/ PCA and PFA)	Exp. 1: yes 36%; Exp. 2: yes 22%	Males outperformed females on all tasks	Galsworthy et al. (2005)
Mice (43)	Lashley maze, passive avoidance, MWM, odor discrimination, fear conditioning, plus maze (spatial), control: exploratory behaviors, sensory/motor function, stress, fear, pain reactivity	Not shown	PCFA	Yes 32%	Exploratory behavior loaded onto the general factor	Matzel et al. (2006)

Mice (exp. 1 = 47, exp. 2 = 51)	Exp. 1: detour, win-shift, olfactory discrimination, fear conditioning, operant acquisition, control: light dark test, open field; Exp. 2: detour, Hebb-Williams, radial arm maze, olfactory foraging, fear conditioning, control: light dark test, open field	Exp. 1: half positive, 4/10 significant; Exp. 2: majority positive, 3/10 significant (excluding controls)	PCA	Exp. 1: no, 2 components extracted; Exp. 2: yes 34% (controls excluded)	For exp. 1 3/4 significant correlations were negative. When control measures were included in the PCA, 3 independent components were extracted	Locurto et al. (2006)
Mice (27)	Lashley maze, passive avoidance, MWM, odor discrimination, fear conditioning, mouse Stroop, nonspatial radial arm maze, delayed reinforced alternation, control: open-field exploration	All in the appropriate direction, 5/15 significant	PCFA	Yes 43%	Selective attention had the strongest correlation, short term memory capacity was modest, and duration had the weakest	Kolata et al. (2007)
Mice (241, combined from prior studies)	Lashley maze, passive avoidance, MWM, odor discrimination, fear conditioning, spatial win-stay (n=98), reinforced alternation (n=78), control: open-field exploration, defecation in water/novel environments	Uniformly positive, 9/10 significant	PCFA verified with CFA	Yes 38%	Spatial group factor with a hierarchical design	Kolata et al. (2008)
Mice (26)	Mouse Stroop, T-Maze reversal, latent inhibition, dual radial arm maze, odor discrimination, reinforced alternation, fear conditioning, radial arm maze	Not shown	EFA and CFA	Yes 37%	External attention was significantly related to the general factor	Sauce et al. (2014)

Birds						
Song sparrows (52)	Novel motor task, color association, color reversal (2009, 2010), tube task (2010)	Majority positive, none reached significance	PCA	No, 2 components extracted	Not related to song repertoire size	Boogert et al. (2011)
Spotted bowerbird (14)	Barrier removal, novel motor task, color discrimination, color reversal, shape discrimination, spatial memory	Majority positive, none reached significance	PCA (on 11 subjects)	Yes 44%	Not related to mating success	Isden et al. (2013)
Robins (16)	Motor task, color discrimination, color reversal, spatial memory, tube task, symbol discrimination	Majority positive, none reached significance	PCA	Yes 34%		Shaw et al. (2015)
Song sparrows (41)	Novel motor task, color association, color reversal, tube task, spatial learning	Majority negative, 1/10 significant (positive)	PCA	No, 2 components extracted		Anderson et al. (2017)
Magpies (56)	Color association, color reversal, tube task, spatial learning	Uniformly positive and significant	PCA	Yes 64%	Larger group size was related to better cognitive performance. Better cognitive performance in females resulted in better offspring success	Ashton et al. (2018)

Chapter 2: Transferring Relational Rule Learning: A Potential Problem Between Successive and Simultaneous Choice Procedures when Assessing Pigeons (*Columba Livia*)

Abstract

Raven's Progressive Matrices (RPM) is a nonverbal intelligence test based on relational rule learning. We previously reported a modified RPM (mRPM) task for pigeons by simplifying the relational rules and procedure (Flaim & Blaisdell, 2021). Pigeons were trained to detect a change in size or orientation using a successive or Go/No-Go procedure, where one display was presented at a time. Pecking a display with the rule-based transformation resulted in a food reward, while pecking at other displays were not. This mRPM was successful at revealing individual differences in the rate of acquisition and transfer, but there were potential issues with the procedure. Subjects trained longer when learning the change in orientation, indicating it was more difficult than the size change. Ideally both rules would be equally difficult, so for this procedure a change in luminosity was used instead of a change in orientation. A successive procedure could rely on a mental representation of the reinforced displays, which could make rules more difficult to learn. A simultaneous choice procedure was used instead, where a reinforced display was always presented with a nonreinforced display. These changes did not make the procedure more sensitive to individual differences, but the results have implications for successive and simultaneous choice tasks.

Introduction

Abstract reasoning, the ability to flexibly apply a rule to a novel situation or stimulus, is commonly assessed in human intelligence tests (Carroll, 1993). One procedure that evaluates

abstract reasoning is the Raven's Progressive Matrices (RPM; Raven, 2003) which assess an individual's ability to apply relational rules. In the RPM, participants are presented with a series of partially completed matrices. Each 3x3 matrix is filled with elements except for the lowest right cell. Elements vary from each other according to one or more relational rules, such as transforming the shape, texture, pattern, etc., of each element. A set of options are provided, but only one that when placed into the empty cell completes the matrix according to the relational rule or rules (Carpenter, Just, & Shell, 1990). The RPM increases in complexity as the participant progresses through the problem set, either by increasing the number of rules controlling the transformations or by making the rules more abstract (Carpenter, Just, & Shell, 1990; Raven, 2003).

We recently adapted the RPM to be suitable for nonhuman animals, specifically the pigeon (Flaim & Blaisdell, 2021). To modify the RPM (mRPM) for pigeons we reduced it to a 2x2 matrix, simplified the relational rules, and implemented a gradual discrimination training procedure. Pigeons were trained on two rules, one involving a size transformation and the other an orientation transformation. Subjects were initially trained with just one rule, with stimuli presented in one row or column of the matrix (Figure 2.1). Instead of using a multiple-choice procedure as for with humans, pigeons were trained on a successive discrimination procedure (a.k.a. a Go/No-Go procedure). On each trial, the pigeon was presented with a single matrix containing two elements. If the elements were transformed according to the relational rule in effect during training (e.g., a size change), then pecking at the matrix would be followed by delivery of a food reward. If the elements of the matrix did not involve a relational transformation, then pecking at the matrix was not rewarded. Thus, correctly transformed matrices served as S+ displays, and the remaining matrices were S- displays. Once pigeons had

reached a predetermined criterion of discrimination performance with a specific rule, they were presented with displays containing elements of novel shapes and/or colors to assess the degree of relational transfer. As with humans on a conventional RPM task, individual pigeons differed in transfer performance, ranging from full transfer, to partial transfer, to no transfer. Pigeons then received discrimination training on the other relational rule, followed by transfer tests.

Our results revealed strong individual differences across pigeons in both the ability to acquire discriminative control by relational rules and in the transfer these rules to novel elements. For example, individual pigeons differed in the number of sessions needed to reach criterion. They also differed in transfer performance to displays containing novel elements. We computed a single metric by which to rank-order all pigeons in their relational ability, similar to how the RPM provides a single score by which to rank individual human performance. In addition to uncovering individual differences in relational ability, we also found differences in how difficult each rule was for pigeons, with the orientation transformation being more difficult than the size transformation. Most pigeons required more training sessions to reach criterion on the orientation rule discrimination, with many pigeons never reaching criterion. Thus, the orientation rule was less sensitive to individual differences and therefore not as useful as the size change rule.

The aim of the current experiment was to develop a simultaneous mRPM procedure for the pigeon. The successive discrimination procedure previously used requires the comparison between S+ and S- displays to take place across trials, which may tax memory (Cook, Kelly, & Katz, 2003; Flaim & Blaisdell, 2021). Furthermore, not every trial of a successive discrimination procedure provides an opportunity for reinforcement. Reinforcement is only available on S+ trials, but never on S- trials. This poses the additional challenge of requiring the pigeon to inhibit

pecking on S- trials. These features of a successive discrimination procedure may have made the discrimination more difficult to learn. The simultaneous discrimination procedure used in the current study, however, allows the pigeon to directly compare the S+ to an S- display, which may make the discrimination easier to learn by reducing the burden on memory and avoiding the need to inhibit pecks to the S- display given that the S+ display was simultaneously available and could provide reward if pecked. Another change we adopted for the simultaneous mRPM procedure was to implement a correction procedure for the first 16 trials. In a correction procedure, if the pigeon selects the S-, the trial is repeated until the S+ is selected.

Furthermore, we replaced the orientation change rule with a luminosity change rule. Other research has shown that pigeons detect changes in luminosity in a discrimination procedure (Wills & Mackintosh, 1999). Despite the reasoning that led to the development of a simultaneous mRPM task, our results indicate that these changes were *less* successful at capturing individual differences in performance. We discuss the implications for the differences in successive and simultaneous choice procedures.

Method

Subjects

Eleven homing pigeons (*Columba livia*) from Double T farm served as subjects. Three of the subjects were 16 years old (Hawthorne, Dickinson, and Vonnegut), while the remaining subjects were 2 years old (n = 8; Mario, Luigi, Wario, Waluigi, Peach, Bowser, Yoshi, and Shy Guy) at the start of the experiment. Within the older subjects, one was female (Dickinson) and within the younger subjects two were female (Mario and Waluigi). Groups were counterbalanced as much as possible with respect to age and experience. The older subjects had participated in a wide variety of behavioral experiments, but the current procedures were selected to minimize

transfer from prior experiences. The younger subjects had received instrumental training to peck at a touchscreen. Half of the younger subjects were naïve to any experiment (n = 4), while the other half were briefly trained with the same procedure as described in this manuscript, but with an orientation transformation. All subjects were maintained at 80% of their free-feeding weight, but were allowed free access to water and grit while in their home cages. Testing occurred at approximately the midpoint of the light portion of the 12-hour light-dark cycle. All procedures were approved by the UCLA Institutional Review Board.

Apparatus

Testing was conducted in a flat-black Plexiglas chamber (38 cm wide x 36 cm deep x 38 cm high). All stimuli were presented by computer on a color LCD monitor (NEC MultiSync LCD1550M) visible through a 23.2 x 30.5 cm viewing window in the middle of the front panel of the chamber. The bottom edge of the viewing window was 13 cm above the chamber floor. Pecks to the monitor were detected by an infrared touchscreen (Carroll Touch, Elotouch Systems, Fremont, CA) mounted on the front panel. A food hopper (Pololu, Robotics and Electronics, Las Vegas, NV) was located in the center of the front panel, its access hole flush with the floor. The food hopper contained a mixture of leach grain pigeon pellets and seed (Leach Grain and Milling). All experimental events were controlled and recorded with a Pentium III-class computer (Intel Santa Clara, California). A video card controlled the monitor in the SVGA graphics mode (800 x 600 pixels). Stimuli were presented using the coding language Python (Python Software Foundation, https://www.python.org/) and the extension PsychoPy version 3.0.3 (Peirce, 2007).

Stimuli

A white circle measuring 35 pixels in diameter served as a 'ready' stimulus. The matrix consisted of four black squares separated by a light gray line. Each square was 99 x 99 pixels and the line between the squares was 200 x 1 pixels. The matrices were 1.9 cm away from the left and right screen edge and 7.62 cm away from the top and bottom of the screen edge. The right edge of the left hand matrix was 13.97 cm to the left of the left edge of the right hand matrix. Visual items could be presented in the cells of the matrix. The set of items included four shapes, a rectangle (42 x 81 pixels), equilateral triangle (77 x 68 pixels), right facing arrow (76 x 66 pixels) and heart (91 x 81 pixels). Each shape could appear in red, blue, yellow, or green.

Matrix Displays

The matrix displays could have items presented in two or four of the cells of the matrix. If there were two items in the matrix, they could only be presented along the row or column of the matrix, never along the diagonal. Items could appear in the top or bottom row, or in the left or right column. The transformed stimulus could be in the left or right, top or bottom position of the pair. The partially-filled matrices had three types of displays, S+, S^{diff}-, and S^{id}- and the items inside the S+ matrix could be transformed via a size or luminosity change. The display types, other than the luminosity rule itself, are identical to what is described in Flaim and Blaisdell (2021), but they will be described here briefly.

For the luminosity change rule, the S+ display consisted of presenting one of the shapes in a lighter shade relative to the other darker shape (e.g., Figure 2.1a). The S^{id}- display was similar to the S+ display except that both shapes were presented in the same shade (e.g., both light or both dark, Figure 2.1b). The S^{diff}- display consisted of two different shapes each in a

different color (Figure 2.1c). However, the two different shapes were either both light or dark, thus while the two shapes inside the matrix underwent a change, it was not the relevant change.

For the size change rule, the S+, S^{id}-, and S^{diff}- displays were similar except that instead of a relational rule involving different shades of the same color, the relational rule involved a size change of 50%, with one item being larger than the other for S+ displays, (See Figure 2.1, panels d-f).

Procedure

First rule training.

Each subject was trained on a partially completed matrix that consistently showed the same rule along the same axis. The groups for each rule presentation were denoted by whether the items were in the row or column and what the rule was. This resulted in four groups, luminosity-column (n = 3; Hawthorne, Yoshi, and Shy Guy), luminosity-row (n = 3; Mario, Luigi, and Dickinson), size-column (n = 2; Wario and Waluigi), and size-row (n = 3; Peach, Bowser, and Vonnegut). All of the subjects in the luminosity-row group and one subject in the luminosity-column group, Yoshi, were briefly trained with an orientation change rule. The orientation rule is described in Flaim and Blaisdell (2021) and was presented along the same axis as their group assignment for this experiment. The procedure for learning the orientation change, however, was as described in this manuscript.

Pigeons received one session per day, five days a week, with each session consisting of 128 trials. Each trial was initiated by a single peck to the ready stimulus. Once the trial was started, two matrices were presented on the left and right side of the screen. One of the matrices was always an S+ while the other could be an S^{diff}- or S^{id}-, resulting in two trial types, 'Different' and 'Identical'. The S+ was presented alongside each type of S- matrix an equal number of times

per session. The side of the screen on which matrices were presented was counterbalanced, so each type of matrix appeared equally often on each side of the screen. During training, the items inside the matrix could be the equilateral triangle or rectangle in red or blue. When the pigeon completed four consecutive pecks (Fixed-Ratio 4, or FR4) to one of the matrices, the trial was terminated. Pecks to the other matrix display would reset the peck requirement. There was no time limit to complete this peck requirement outside of the session length. If the pigeon completed the peck requirement to the S+ display, the food hopper was illuminated and the pigeons could access the food hopper for 3-s. If the pigeon completed the peck requirement to the S-, the trial was terminated, and the 1-s intertrial interval (ITI) began.

During the first 16 trials, a correction procedure was implemented, where the subject would repeat the trial until they completed the peck requirement to the S+, including the peck to the start stimulus and the ITI. In the first and last 16 trials of a session, completing the peck requirement to the S+ always resulted in a food reward. Within the remaining 112 trials, however, 16 of the trials were not reinforced, even if the pigeon completed the peck requirement to the S+. This was to prevent disruption for future non-rewarded probe trials. Trials were organized into 18-20 blocks and each block had two or four randomly placed nonreinforced trials. The maximum number of consecutively nonreinforced trials was eight, but the randomization made such an event unlikely. Nonreinforced trials were equally distributed across trial type (S^{id}- or S^{diff}-), item type (rectangle or triangle, blue or red) and location (left or right of the midline).

Selection of the S+ was calculated separately for each trial type, when the S+ was presented with the S^{id}- or S^{diff}-. Training continued until subjects reached criterion, 80% accuracy

for both trial types on two consecutive sessions, or until subjects had received 100 sessions of training. Then subjects were given transfer tests with novel displays.

Transfer testing.

When subjects reached criterion or reached 100 sessions of training, they received transfer test sessions. Each transfer session had 112 reinforced baseline trials with the same displays and reinforcement contingency as in training, eight nonreinforced baseline trials, and eight nonreinforced novel probe trials for a total of 128 trials in each session, equally distributed between the display types. Nonreinforced trials could not appear in the first or last 16 trial block. Subjects received a total of five transfer sessions. If transfer sessions were separated by 72 hours, subjects received another training session without probe trials before continuing with the remaining transfer sessions.

Probe displays contained a novel shape (right facing arrow or heart), a novel color (yellow or green), or both. Probe trials only compared a probe S+ to a probe S^{diff}- or S^{id}-. Subjects never had to choose between a baseline S+ and a probe S+. Similarly, pigeons never had to make a choice between a probe S+ and baseline S- or a baseline S+ and a probe S-.

Second rule training and transfer testing.

After transfer testing for the first rule learned, subjects were trained with the other rule presented along the other axis they had not been exposed to. For example, if a subject was in the size-row group during the first rule training, they would now be in the luminosity-column group. Training and testing on the second rule proceeded as described above.

Data Analysis

Sessions were only included in the analysis if the subject advanced through at least 30 trials. A session with less than 30 trials indicated either a computer error or low motivation. Correction trials were not analyzed or counted towards session inclusion criteria. The number of sessions excluded for each subject is detailed in Table 2.1. Training data were analyzed in 5-session blocks. Accuracy was calculated for each trial type. Number of sessions to criterion for each rule was used to determine if there were order effects. Effects were collapsed across presentation axis (i.e., horizontal and vertical) to maintain sufficient power for analyses. Additionally, previous research indicated that presentation axis did not impact performance (Flaim & Blaisdell, 2021). Testing data were analyzed across all transfer sessions. During transfer sessions, only the accuracy during the nonreinforced baseline trials was compared to probe trials. Data were analyzed using JASP, version 0.14.1 (JASP Team, 2020).

Results

First Rule Acquisition and Transfer Testing

All subjects in the luminosity groups (n = 6) received the maximum duration of training, 20 5-session blocks, without reaching criterion (Figure 2.2a). For most subjects this was due to worse performance when discriminating the S+ from the S^{id}-. All pigeons in the size group reached criterion (Figure 2.2b). For the two subjects in the size-column group, Wario reached criterion in eight blocks, while Waluigi reached criterion in 14 blocks. For the three subjects in the size-row group, Peach reached criterion in five blocks, Bowser in 10, and Vonnegut in 11. To compare the number of sessions during training between the luminosity and size groups, variance was artificially created for the luminosity group by adding or subtracting 0, 10, 15, or 20. These

numbers were selected to create equal variance across the luminosity and size groups, while maintaining a median value of 100 for the luminosity group. Lavene's test showed that variance was not significantly different across groups (F (1) = 0.31, p = .59). A nonparametric Mann-Whitney U test indicated that subjects in the luminosity group experienced significantly more sessions (*Mdn* = 100, *SD* = 18) compared to subjects in the size group (*Mdn* = 49, *SD* = 18.38), U = 30, p = .008.

In the luminosity group, even though no subjects reached criterion during training, performance was variable across subjects during transfer sessions (Figure 2.3a). A one-tailed binomial test was used to determine the probability of correct choices was greater than chance, or .5, for the nonreinforced baseline trials and probe trials for each trial type. None of the subjects performed greater than chance on the probe trials, but three of the subjects (Hawthorne, Dickinson, and Mario) performed significantly better than chance on the nonreinforced baseline trials despite not reaching criterion during training (Table 2.2).

In the size group, all subjects reached criterion during training, but performance on the novel probes during transfer sessions was uniformly low (Figure 2.3b). Again, using a one-tailed binomial test, almost all subjects performed significantly better than chance on the nonreinforced baseline trials, but none of the subjects performed better than chance on the probe trials (Table 2.3).

Second Rule Acquisition and Transfer Testing

Almost all of the subjects that had first trained on the luminosity rule reached criterion on the size change rule (Figure 2.4a). The subject that did not reach criterion, Hawthorne, was in the size-row group. Hawthorne had a 4-month break in training after completing 61 sessions due to COVID-19 safety related restrictions. For the other two subjects in the size-row group, Yoshi

reached criterion after 16 blocks of training, while Shy Guy reached criterion in 6 blocks. For the subjects in the size-column group, Dickinson reached criterion in 8 blocks, Mario in 4, and Luigi in 9 (Figure 2.4a).

For the subjects that had first been trained on the size change rule, approximately half reached criterion on the luminosity change rule (Figure 2.4b). The two subjects in the luminosity-row group, Wario and Waluigi, received the maximum duration of training without reaching criterion. Similar to Hawthorne, Waluigi had a 4-month break in training after completing 86 sessions. The other subject in the luminosity-row group, Wario, did not have a break in their training, but still did not reach criterion with the maximum duration of training. In the luminosity-column group, Peach and Bowser reached criterion in 12 blocks of training, while Vonnegut reached criterion in 14 (Figure 2.4b). Since subjects showed more variability in the amount of training received and Levene's test did not indicate a significant difference in the variance across groups (F (1) = 0.53, p = .485), the data were not transformed. A nonparametric Mann-Whitney U test did not show a statistically significant difference in the number of sessions experienced in the size (Mdn = 39, SD = 29.7) and luminosity (Mdn = 69, SD = 20.18) groups, U = 6, p = .116.

For subjects who learned the size rule second, even though most subjects reached criterion, performance on the probe trials was consistently poor (Figure 2.5a). A One-tailed binomial test indicated that most subjects performed significantly better than chance on baseline trials, but none of the subjects performed better than chance on the probe trials (Table 2.4). A similar result was found for the subjects that learned the luminosity rule second (Figure 2.5b). A one-way binomial test indicated that most subjects performed significantly better than chance on

baseline trials, including the subjects that did not reach criterion. Yet, similar to the previous results, no subject performed significantly better than chance on the probe trials (Table 2.5).

Order or Rule Effects

To determine if the order in which each rule was learned had an impact on the number of sessions needed to reach criterion, performance on one rule was compared to the group that learned it first to the group that learned it second. To compare the luminosity groups, the data from the group that learned it first has variance added to it, similar to what was described above, and Levene's test showed there was no significant difference in the variance, F (1) = 0.033, p = .9. A Mann-Whitney U test indicated there was no significant difference in the number of sessions experienced between the group that learned the luminosity rule first (*Mdn* = 100, *SD* = 20) and the group that learned it second (*Mdn* = 69, *SD* = 20.18), U = 24, p = 0.118.

A similar result was found for the size rule. Levene's test showed there was no significant difference in the variance, F (1) = 1.346, p = .276, and a Mann-Whitney U test indicated that there was no significant difference in the number of sessions experienced for the group that learned the size change first (*Mdn* = 49, *SD* = 18.38) and the group that learned it second (*Mdn* = 39, *SD* = 29.7), U = 18, p = .647.

Because there were no significant differences in the number of sessions experienced based on the order in which each rule was learned, data were collapsed across training order to determine if there was a significant difference in how many sessions were experienced for the luminosity versus the size change rule. A Wilcoxon signed rank test indicated that subjects received significantly more sessions of training on the luminosity rule (*Mdn* = 100, *SD* = 17) compared to the size rule (*Mdn* = 40, *SD* = 24), U = 55, p = .006.

Discussion

The goal of this experiment was to increase the sensitivity of the mRPM described in Flaim and Blaisdell (2021). One of the changes in this experiment was to replace the orientation relational rule with a luminosity relational rule. Unfortunately, pigeons failed to acquire discriminative control by the luminosity rule. Subjects experienced significantly more training sessions for the change in luminosity compared to the change in size, with none of the subjects trained on the luminosity rule first reaching criterion, and only 3 of 5 pigeons trained on the luminosity rule second reaching criterion. This indicates that the luminosity change was more difficult to discriminate. The uniformly poor performance for the luminosity change rule suggests not including it in assessments of individual differences in relational control of behavior. The variability in how many sessions were experienced with the size change rule, along with the fact that 10 out of 11 pigeons reached criterion during training, however, indicates that this is still a useful discrimination for assessing individual differences in relational control of behavior.

The primary difference from the procedure of Flaim and Blaisdell (2021) was the use of a simultaneous choice task instead of a successive discrimination procedure. The successive procedure was successful in revealing individual differences in the number of sessions to reach criterion and how performance transferred to novel probes (Flaim & Blaisdell, 2021). While the choice procedure used here revealed individual differences in the number of sessions required to reach criterion for the size change rule, performance was uniformly poor on novel probes. None of the subjects showed evidence that they had learned a relational rule that could be flexibly applied to novel stimuli. This contrasts markedly with the results from Flaim and Blaisdell using

the successive discrimination procedure, where 3 out of 5 of the pigeons that reached criterion on the size change rule showed above chance transfer to novel stimuli.

The mRPM is not the first task in which different patterns of transfer were reported for successive versus simultaneous procedures. The properties of stimulus class formation, specifically symmetry, have shown similar results (Frank & Wasserman, 2005; Urcuoli, 2008). In these experiments, pigeons were first shown a sample stimulus. After completing an observing response to the sample, pigeons are then presented with one or more comparison stimuli. Subjects are trained to choose the comparison that matches the sample. For example, if the sample is a red circle, the subject should select the red circle comparison, and not the green circle comparison stimulus. Subjects are also trained to 'arbitrarily' match, so during this phase if the sample was a red circle, subjects should select the comparison that is a horizontal line and not a vertical line (a.k.a, a conditional discrimination or symbolic match to sample). Using the conditional discrimination procedure, pigeons demonstrate symmetry if they can spontaneously reverse the associated pair, such that if the sample is a horizontal line, then the pigeon should select the red circle comparison. If pigeons receive training using the simultaneous presentation of the sample and comparisons, they do not demonstrate symmetry. If, however, pigeons receive training on a successive presentation of the sample and comparison(s), they *do* demonstrate symmetry (Frank & Wasserman, 2005; Urcuioli, 2008). While a theory has been developed to explain why a successive procedure will result in symmetrical matching, it is not clear how such a theory could explain the difference in transfer performance on the mRPM (Urcuioli, 2008).

Our study involves training pigeons on a relational rule, where the relation, such as a size change or a change in illumination, between stimulus elements defines the rule. There is a long history of studying relational control of behavior in animals, such as pigeons and rats. Most of

these prior studies, however, use the transposition task, where the subject is reinforced for selection the larger or brighter (or smaller or dimmer) of two stimuli. In these transposition experiments, it has generally been found that simultaneous presentation of the S+ and S- leads to more rapid acquisition of the discrimination, as well as greater transfer to stimuli extrapolated along the same dimension (e.g., by choosing a test stimulus that is larger (smaller) or brighter (dimmer) than the S+) than does a successive presentation of S+ and S- on separate trials (see Wills & Mackintosh, 1999, for a demonstration in pigeons and review of the literature in rats). While both involve stimulus control by an intradimensional relational change, there are a number of differences, however, between transposition experiments and our mRPM procedure. Our procedure involves selecting a stimulus display (the matrix) that contains two elements that exemplify the relational rule (e.g., choose the matrix where one of the pair of stimuli is larger than the other), whereas transposition experiments require the subject to select one of the elements over the other (e.g., choose the larger stimulus). In transposition experiments, the S- is an element on the same dimension as the S+, whereas in the mRPM, the S- is a pair of identical stimuli, or a pair of stimuli that differ on a number of factors (e.g., color and shape). Finally, transposition experiments assess transfer to novel stimuli that are the same as the S+ and S-, only above or below the S+ along the dimension of relational control. Our procedure, on the other hand, involves transfer of the same relational rule to pairs of stimuli that differ in non-relevant dimensions, such as novel colors and shapes. Thus, it is difficult to pinpoint what the cause is of the simultaneous procedure leading to better stimulus control in a transposition task, whereas the successive procedure leads to better stimulus control in the mRPM task.

While there has not been a systematic investigation on how the use of a simultaneous versus successive procedure could be impacting performance, there are interesting implications if

this pattern is replicated with additional relational learning procedures. There would need to be a more general theory as to why successive procedures improve transfer performance in certain tasks, whereas simultaneous procedures improve transfer in others. One possibility is that during a successive procedure, the subject has to compare the stimulus presented on the screen to a memory of previously reinforced S+ and nonreinforced S- stimuli. Memories are often imprecise, due to the failure to fully encode the stimulus, the inability to maintain a precise representation, and/or retrieving incorrect memories resulting in confusion errors (Jasnow, Cullen, & Riccio, 2012; Wright, Kelly, & Katz, 2018). An incomplete S+ memory could still result in accurate pecking behavior, during training and transfer tests, if the memory contains the relevant information (the transformation rule) instead of the irrelevant information (shape, color, and location of stimuli within the matrix). In fact, this incomplete representation that only contains the relevant information could result in better transfer, compared to having a more detailed mental representation containing both relevant and irrelevant information. Computational models that use a few, 'best' predictors transfer to novel situations better than models that use all cues present (Hertwig & Todd, 2003). Successive procedures could result in an incomplete, but selective, mental representation that transfers well to novel stimuli, while simultaneous procedures could result in a more detailed representation, containing both relevant and irrelevant details, that does not transfer well.

While this experiment was ultimately unsuccessful at increasing the sensitivity of the mRPM procedure, there are two important results. First, learning a relational size rule was still a viable method to assess individual differences in relational learning. Second, successive procedures may result in more robust transfer to novel stimuli compared to simultaneous procedures, at least for versions of the mRPM task. This procedural difference could have major

implications when assessing the cognitive abilities of pigeons and the underlying neurobiological correlates.

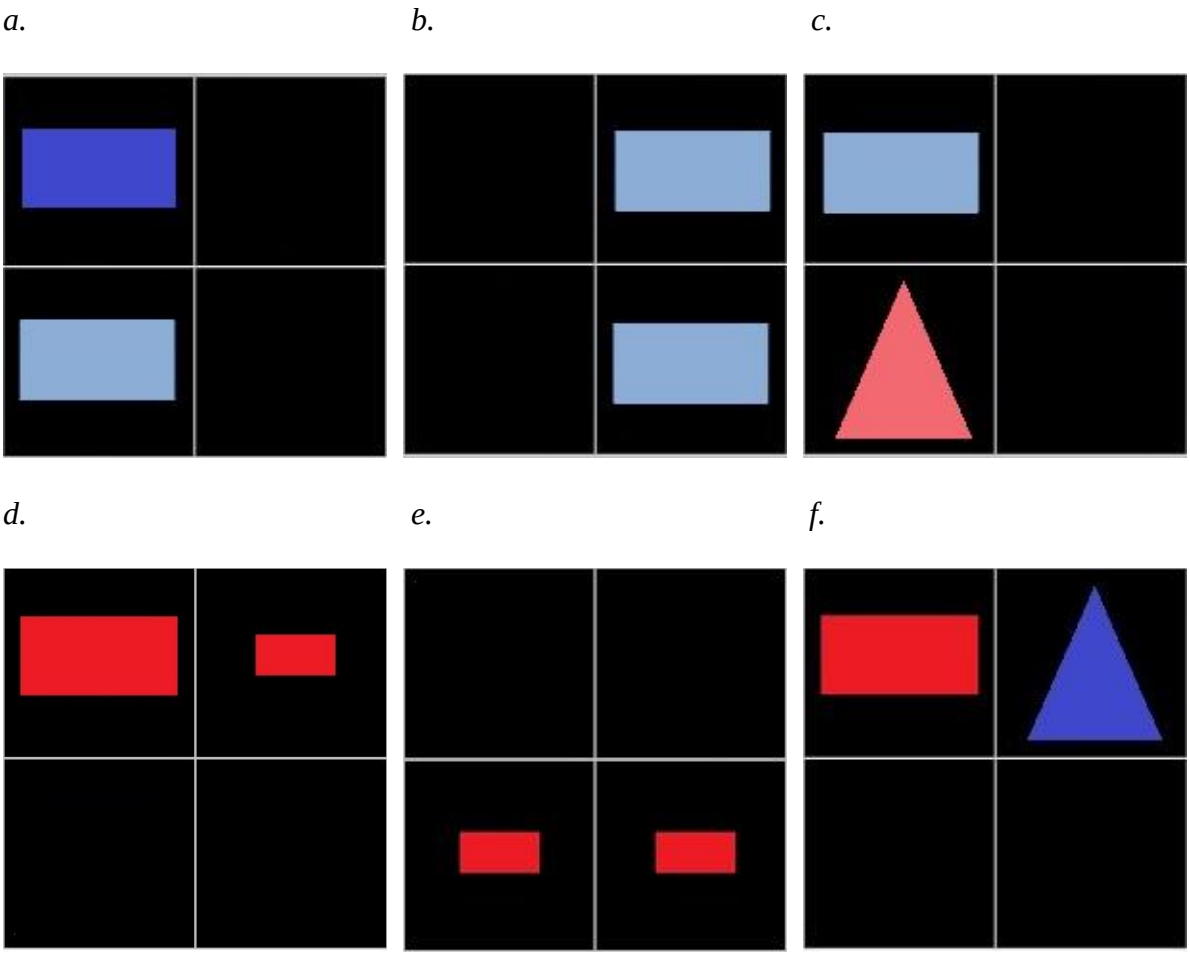


Figure 2.1. The top row is an example of an S+, S^{id}-, and S^{diff}- display for the luminosity-column group. The second row is an example of an S+, S^{id}-, and S^{diff}- display for the size-row group.

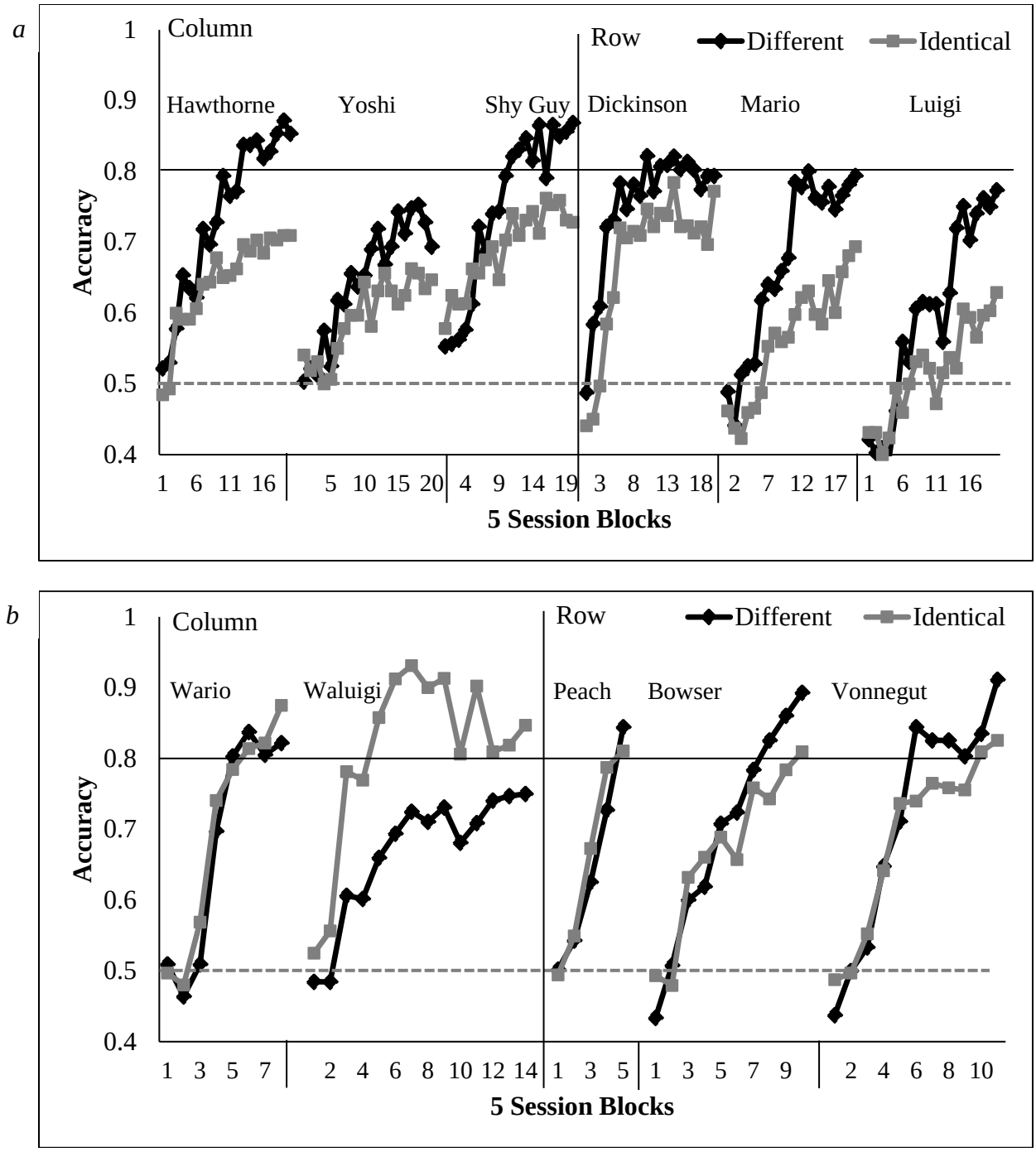


Figure 2.2. Acquisition data for the first rule, with panel a showing the luminosity rule and panel b showing the size rule. The dotted line indicates chance performance and the solid line indicates criterion level of performance, accuracy of 0.8. The vertical line separates the groups based on the presentation axis. Data were blocked by 5 sessions, which may obscure criterion level performance.

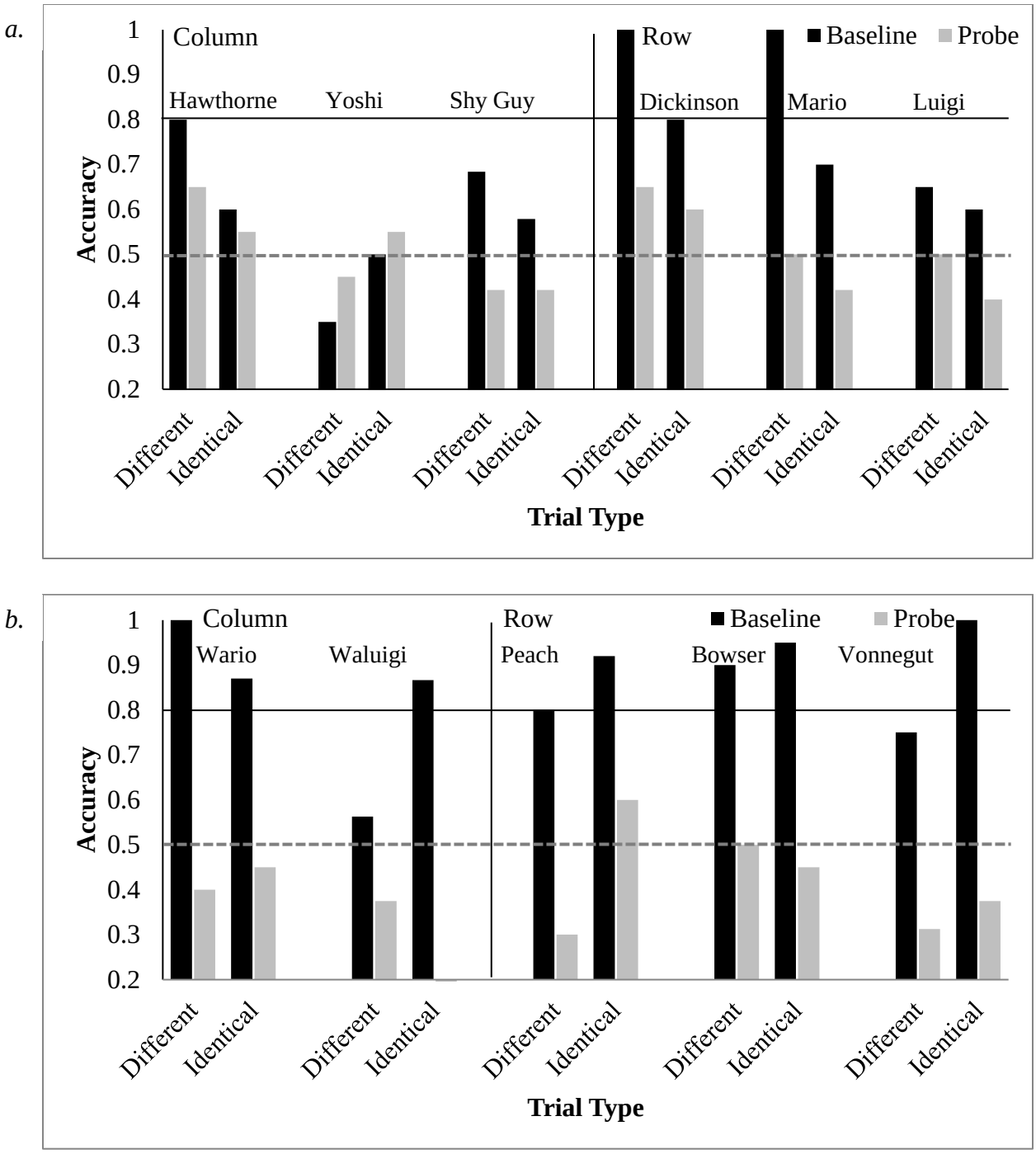


Figure 2.3. Performance on transfer sessions for the first rule learned, with panel a showing the luminosity rule and panel b showing the size rule. The dotted line indicates chance level of performance and the solid horizontal line indicates criterion level of performance, accuracy of 0.8. The solid vertical line separates the groups based on the presentation axis. Data were blocked by 5 sessions, which may obscure criterion level performance.

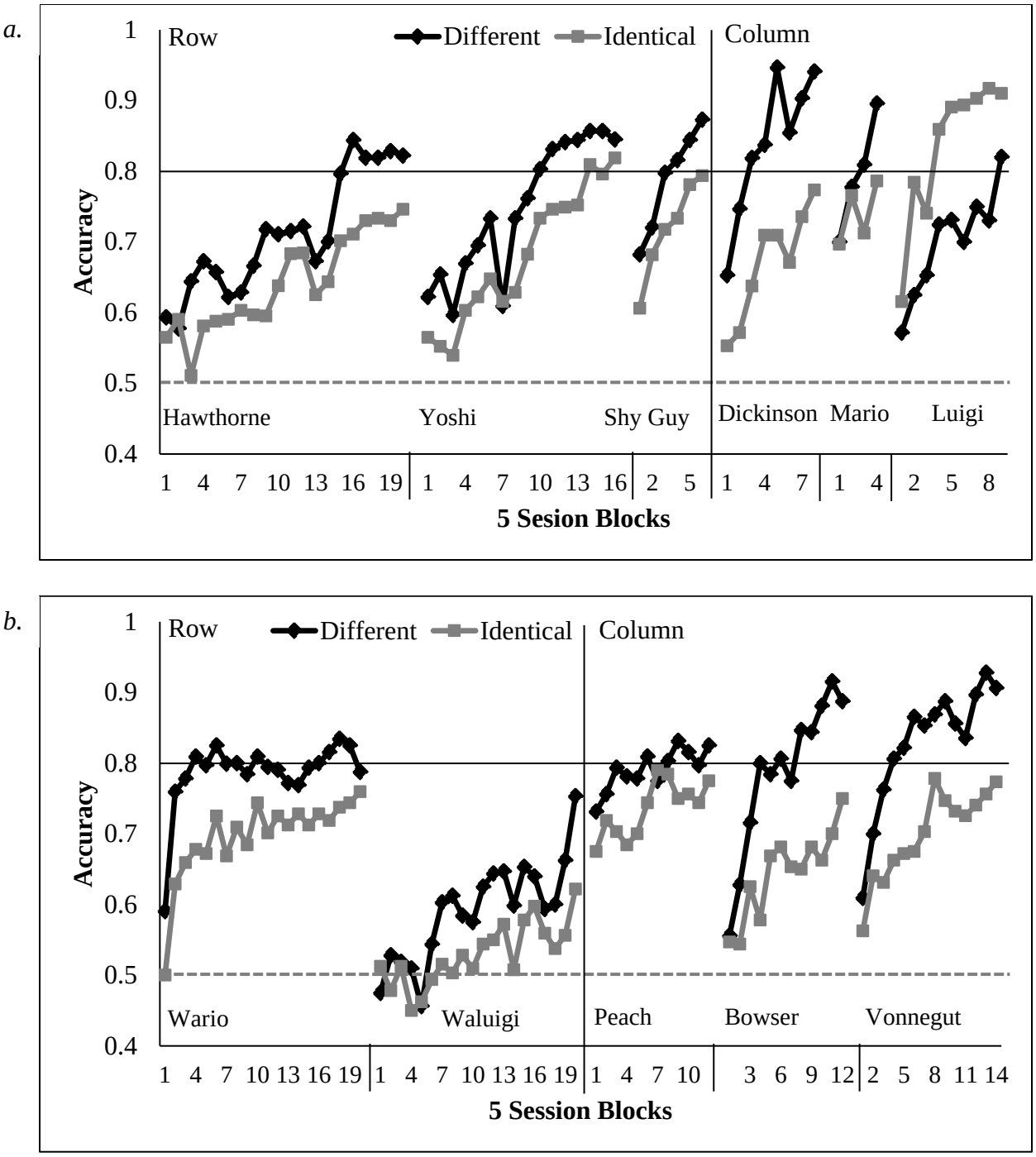
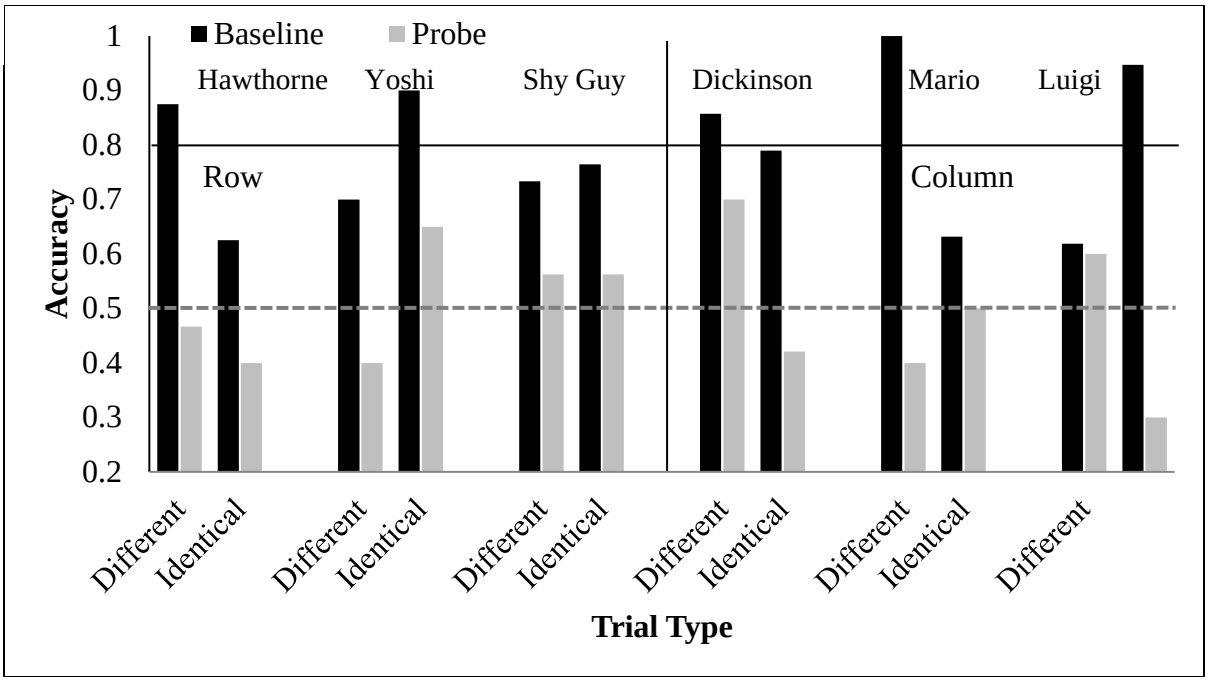


Figure 2.4. Acquisition data for the second rule, with panel a showing the size rule and the panel b showing the luminosity rule. The dotted line indicates chance performance and the solid line indicates criterion level of performance, accuracy of 0.8. The vertical line separates the groups based on the presentation axis. Data were blocked by 5 sessions, which may obscure criterion level performance

a.



b.

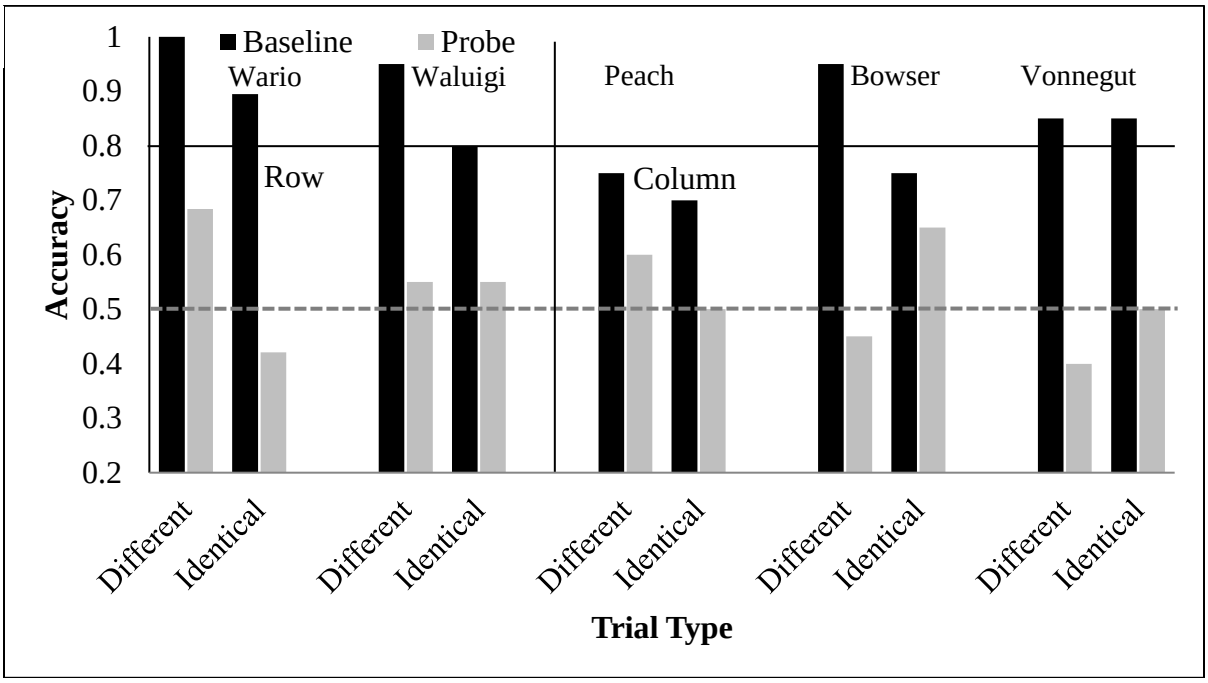


Figure 2.5. Performance on transfer sessions for the second rule learned, with panel a showing the size rule and the panel b showing the luminosity rule. The dotted line indicates chance level of performance and the solid horizontal line indicates criterion level of performance, accuracy of 0.8. The solid vertical line separates the groups based on the presentation axis.

Sessions Excluded		
	First Rule	Second Rule
Subjects	Luminosity	Size
Hawthorne	3	2
Yoshi	1	0
Shy Guy	5	5
Mario	11	1
Luigi	11	2
Dickinson	9	0
	Size	Luminosity
Wario	9	0
Waluigi	25	7
Peach	2	1
Bowser	2	1
Vonnegut	3	9

Table 2.1. Number of sessions excluded for each subject for each rule learned. Sessions were excluded if the subject completed less than 30 trials, not including correction trials.

Luminosity						
Axis	Subject	Trial Type	Display Type	*n*	*k*	*p*
Column	Hawthorne	Different	Probe	20	13	.13
		Identical	Probe	20	11	.41
		Different	Baseline	20	16	>.01
		Identical	Baseline	20	12	.25
Column	Yoshi	Different	Probe	20	9	.75
		Identical	Probe	20	11	.41
		Different	Baseline	20	7	.94
		Identical	Baseline	20	10	.5
Column	Shy Guy	Different	Probe	19	8	.82
		Identical	Probe	19	8	.82
		Different	Baseline	19	13	.08
		Identical	Baseline	19	11	.32
Row	Dickinson	Different	Probe	20	13	.13
		Identical	Probe	20	12	.25
		Different	Baseline	20	20	>.01
		Identical	Baseline	20	16	>.01
Row	Mario	Different	Probe	20	10	.5
		Identical	Probe	20	8	.87
		Different	Baseline	20	19	>.01
		Identical	Baseline	20	14	.06
Row	Luigi	Different	Probe	20	10	.5
		Identical	Probe	20	8	.87
		Different	Baseline	20	13	.13
		Identical	Baseline	20	12	.25

Table 2.2. The binomial test results during transfer session when a change in luminosity was the first rule learned, where n is the number of trial and, k is the number of correct choices. The probability of the number of correct choices being greater or equal than reported out of the number of total trials was tested against a probability of 0.5.

Size						
Axis	Subject	Trial Type	Display Type	*n*	*k*	*p*
Column	Wario	Different	Probe	20	8	.87
		Identical	Probe	20	9	.75
		Different	Baseline	25	25	>.01
		Identical	Baseline	15	13	.01
Column	Waluigi	Different	Probe	16	6	.89
		Identical	Probe	16	3	.99
		Different	Baseline	16	9	.4
		Identical	Baseline	16	13	.01
Row	Peach	Different	Probe	20	6	.98
		Identical	Probe	20	12	.25
		Different	Baseline	15	12	.02
		Identical	Baseline	25	23	>.01
Row	Bowser	Different	Probe	20	10	.58
		Identical	Probe	20	9	.75
		Different	Baseline	20	18	>.01
		Identical	Baseline	20	19	>.01
Row	Vonnegut	Different	Probe	16	5	.96
		Identical	Probe	16	6	.23
		Different	Baseline	16	12	.04
		Identical	Baseline	16	16	>.01

Table 2.3. The binomial test results during transfer session when a change in size was the first rule learned, where n is the number of trial and, k is the number of correct choices. The probability of the number of correct choices being greater or equal than reported out of the number of total trials was tested against a probability of 0.5.

Size						
Axis	Subject	Trial Type	Display Type	*n*	*k*	*P*
Row	Hawthorne	Different	Probe	15	7	.7
		Identical	Probe	15	6	.85
		Different	Baseline	16	10	.23
		Identical	Baseline	16	14	>.01
Row	Yoshi	Different	Probe	20	8	.87
		Identical	Probe	20	13	.13
		Different	Baseline	20	14	.06
		Identical	Baseline	20	18	>.01
Row	Shy Guy	Different	Probe	15	9	.3
		Identical	Probe	15	9	.3
		Different	Baseline	14	11	.03
		Identical	Baseline	16	13	.01
Column	Dickinson	Different	Probe	20	14	.06
		Identical	Probe	20	8	.87
		Different	Baseline	21	18	>.01
		Identical	Baseline	19	15	.01
Column	Mario	Different	Probe	20	10	.5
		Identical	Probe	20	8	.87
		Different	Baseline	20	19	>.01
		Identical	Baseline	20	14	.06
Column	Luigi	Different	Probe	20	12	.25
		Identical	Probe	20	6	.94
		Different	Baseline	21	13	.19
		Identical	Baseline	19	18	>.01

Table 2.4. The binomial test results during transfer session when a change in size was the second rule learned, where n is the number of trial and, k is the number of correct choices. The probability of the number of correct choices being greater or equal than reported out of the number of total trials was tested against a probability of 0.5.

Luminosity						
Axis	Subject	Trial Type	Display Type	*n*	*k*	*p*
Row	Wario	Different	Probe	19	13	.08
		Identical	Probe	19	8	.87
		Different	Baseline	19	19	>.01
		Identical	Baseline	19	17	>.01
Row	Waluigi	Different	Probe	20	11	.41
		Identical	Probe	20	11	.41
		Different	Baseline	20	19	>.01
		Identical	Baseline	20	16	>.01
Column	Peach	Different	Probe	20	12	.25
		Identical	Probe	20	10	.5
		Different	Baseline	20	15	.02
		Identical	Baseline	20	14	.06
Column	Bowser	Different	Probe	20	9	.75
		Identical	Probe	20	13	.13
		Different	Baseline	20	19	>.01
		Identical	Baseline	20	15	.02
Column	Vonnegut	Different	Probe	20	8	.87
		Identical	Probe	20	10	.5
		Different	Baseline	20	17	>.01
		Identical	Baseline	20	17	>.01

Table 2.5. The binomial test results during transfer session when a change in luminosity was the second rule learned, where n is the number of trial and, k is the number of correct choices. The probability of the number of correct choices being greater or equal than reported out of the number of total trials was tested against a probability of 0.5.

Chapter 3: Assessing Associative Learning Using the Symbolic Match to Sample Task

Abstract

Intelligence research in humans has consistently shown that performance positively correlates across almost all cognitive tasks. Despite the long history of intelligence research, how associative learning relates to intelligence has only been investigated recently. More complex measures of associative learning are related to intelligence, while simpler measures are not. In animals, associative learning has also been shown to be related to general cognitive abilities or intelligence, but only simple measures have been used. This experiment uses the symbolic or arbitrary match to sample task with four unique pairs as a complex associative learning task more similar to what has been used in human research. This task is sufficiently sensitive to individual differences in performance to be incorporated into a cognitive test battery for pigeons. Our results indicate that this task could also be used to investigate reliability and age-related declines of cognitive performance.

Introduction

When people are given a diverse battery of cognitive tests, performance positively correlates across the measures, meaning that if a person performs well in one task, they are likely to perform well in another (Carroll, 1993; Deary, 2000). This consistent performance across tasks is within-subject reliability, but there are differences in performance across subjects or between-subject variability. If a dimension reducing technique, like factor analysis or principal component analysis, is applied to the positive correlational matrix, it will extract one factor that can account for approximately half of the variance in performance between people. In addition, all tasks load onto this factor, meaning that performance on the task can be accounted for by the

factor. This factor is called *g* due to the way it is generally related to almost all cognitive tasks (Carroll, 1993; Deary, 2000; Spearman, 1904). *g* is highly related to the concept of intelligence since they are both related to individual differences in cognitive performance. While intelligence does not have a universally accepted definition, definitions typically include the ability to achieve goals and behave optimally in a wide variety of environments or situations (Legg & Hutter, 2007). One definition of optimal behavior includes maximizing the number of rewards received with the least amount of effort (Legg & Hutter, 2007; Mettke-Hofmann, 2014; Zentall, 2015). Associative learning, learning the contingency between stimuli or stimuli and responses, is a key factor in guiding optimal behavior (Heyes, 2012; Mettke-Hofmann, 2014; Veksler et al., 2104; Wasserman & Miller, 1997). Even though individual differences in associative learning are thought to be involved in, or serve as a marker of intelligence, support for this relationship is relatively new (Alexander & Smales, 1997; Kaufman et al., 2009; Tamez et al., 2008; Williams & Pearlberg, 2006; but see Harootunian, 1966). Initial investigations of associative learning and intelligence in humans found no differences across groups known to differ in cognitive abilities (Stevenson & Zigler, 1957), and broader investigations of learning ability showed weak to negative correlations with performance on intelligence tests (Woodrow, 1946). One reason why these early investigations between learning and intelligence failed to show a relationship is due to the simple tasks used to assess learning, like responding to one size of a block (Stevenson & Zigler, 1957). More recent experiments have identified a relationship between *g* and task complexity; the more complex the task, the more it will load onto *g* (Marshalek et al., 1983; Sheppard & Vernon, 2008; Stankov & Crawford, 1993). The more complex associative learning procedures that have shown a relationship to *g* are word-pairs (Alexander & Smales, 1997) and the three-term contingency task (Kaufman et al., 2009; Tamez et al., 2008; Williams &

Pearlberg, 2006). Both of these tasks involve simple words, typically 1 syllable or three letters, that are semantically unrelated to each other. In word-pairs, participants are first shown a word that serves as the cue, then the second word is shown, pairing these words together. At test, they are shown the cue word and the participant must recall the second word to correctly complete the pair. The list can vary from 12 to 30 pairs of words (Alexander & Smales, 1998; Kaufman et al., 2009). In the thee-term contingency task, participants are first presented with a word that serves as a cue and three response keys. Pressing each response key reveals a different word. At test, participants are shown the cue word and must type the word associated with each response key (Kaufman et al., 2009 Tamez et al., 2008; Williams & Pearlberg, 2006). Performance on these tasks is positively correlated with performance on intelligence tests, showing that associative learning ability is positively related to *g* (Alexander & Smales, 1997; Kaufman et al., 2009; Tamez et al., 2008; Williams & Pearlberg, 2006).

While intelligence research in humans has been ongoing for over a century, general cognitive abilities have only been recently investigated in nonhuman animals (hereafter animals; Flaim & Blaisdell, 2020). A variety of animals, including mice (Kolata et al., 2008), robins (Shaw et al., 2015), spotted bowerbirds (Isden et al., 2013), and magpies (Ashton et al., 2018), have been given test batteries that assess a wide range of cognitive abilities. Typically, these test batteries include a simple measure of associative learning, where one stimulus is followed by a food reward and another stimulus is not, in addition to other measures of memory, inhibition, and problem solving (Ashton et al., 2018; Kolata et al., 2008; Isden et al., 2013). Similar to what is seen in humans, performance typically correlates across tasks and one factor extracted can account for 22-64% of the variance in performance across subjects (Flaim & Blaisdell, 2020). Associative learning loads onto this factor, similar to what has recently been seen in people.

These results indicate that intelligence has similar properties across species, but it is difficult to say conclusively because of the differences in the cognitive test batteries. The associative learning tasks administered to animals are very simple compared to what is given to humans, yet they still load onto the general factor extracted. The positive loading seen in animals may be due to their inexperience with experimental procedures since novelty can also increase a task's *g* loading (Carroll, 1993; Sternberg & Gastel, 1989). Only experimentally-naïve animals have been given this simple associative learning task, so the unnatural apparatus and stimulus-outcome contingencies may be sufficiently novel for the task to load onto the *g* factor (Ashton et al., 2018; Isden et al., 2013; Matzel et al., 2003; Shaw et al., 2015). Therefore, these results could indicate that novelty is related to *g* for human and animals, strengthening the idea that the same factor is being extracted across species. Yet, if the general intelligence found in animals is the same as what is found in humans, we should also see a relationship between complexity and intelligence.

It would be impossible to administer the exact same complex associative learning task to humans and animals since animals do not have the same capacity for language. It is possible, however, to use other rich, but distinctive stimuli paired together in a similar, arbitrary manner as in the word-pairs task. Using complex visual stimuli, such a task has already been established for the pigeon, the symbolic or arbitrary match to sample (SMTS; Rodewald, 1974, Velasco et al., 2010). The SMTS task has been primarily used to explore stimulus class formation (Urcuioli, 2015), but the goal for our experiment was to determine if the SMTS has sufficient variability in performance to be incorporated into a cognitive test battery for the pigeon. In the SMTS procedure, the subject is first shown a sample stimulus. When the subject completes the peck requirement to the sample, two comparison stimuli appear. One of the comparison stimuli is followed by a food reward when selected (e.g., pecked) while the other is not. Pictures of foods

and animals from the food-pics database (Blechert et al., 2015) were used as the stimulus set where one category consistently served as the sample while the other category served as the comparisons. Subjects were trained with 4 stimulus pairs, similar to the word pairs task given to humans. Subjects were trained on the stimulus pairs until they reached 80% accuracy on all 4 pairs in a single session or until they had been trained for 35 sessions. There was variability in how many sessions subjects needed to reach criterion, but the age of the subjects may be an important factor for inclusion in a cognitive test battery.

Method

Subjects

Seventeen pigeons served as subjects. All subjects had been previously trained to eat from the food hopper. All subjects, except for Wenchang, had prior experience with cognitive tasks administered via an operant touchscreen. One subject, Estelle, had prior training with a different version of the SMTS, which did use two of the same stimuli as this experiment. The two stimuli served a different function and had different pairings compared to the previous experiment to minimize transfer. Additionally, approximately one year had elapsed between the two experiments. There were eight females ranging in age from 0.5-18 years old and nine males ranging in age from 3-18 years old. Pigeons were individually housed in steel home cages with metal wire mesh floors in a vivarium. They were maintained at 80% of their free-feeding weight, but were allowed free access to water and grit while in their home cages. Testing occurred at approximately the midpoint of the light portion of the 12-hour light-dark cycle.

Apparatus

Testing was conducted in a flat-black Plexiglas chamber (38 cm wide x 36 cm deep x 38 cm high). All stimuli were presented by computer on a color LCD monitor (NEC MultiSync LCD1550M) visible through a 23.2 x 30.5 cm viewing window in the middle of the front panel of the chamber. The bottom edge of the viewing window is 13 cm above the chamber floor. Pecks to the monitor were detected by an infrared touchscreen (Carroll Touch, Elotouch Systems, Fremont, CA) mounted on the front panel. A custom-built food hopper (Pololu, Robotics and Electronics, Las Vegas, NV) was located in the center of the front panel, its access hole flush with the floor. The food hopper contained a mixture of leach grain pigeon pellets and seed (Leach Grain and Milling). All experimental events were controlled and recorded with a Pentium III-class computer (Intel, Santa Clara, California). A video card controlled the monitor in the SVGA graphics mode (800 x 600 pixels). Stimuli were presented using the 3.6 version of Python with the psychopy toolbox, version 3.0.3 (Peirce, 2007).

Stimuli

The stimulus set consisted of eight food and eight animal images from the food-pics database for a total of 16 images (Blechert et al., 2014; Figure 3.1). The food items consisted of a slice of cupcake, three overlapping strawberries, a sandwich, a salad in a white bowl, a pile of Brussel sprouts with a basil leaf and carrot stick, a top-down view into a bowl of tortellini noodles, mixed vegetables consisting of peas, corn kernels, Brussel sprouts, carrots sliced into discs, a cauliflower floret, and peeled potatoes, and a pile of candies with different colored exteriors. The animals were a frog, butterfly, bird, fish, penguin, turtle, kitten and elephant. The image was presented on a white background. The specific values for the images were measured

and provided by Blechert et al. (2014) and the color composition, intensity, contrast, spatial layout, and complexity were approximately equal across the animal and food images. Each picture from one set was assigned to a picture from the other set, for example the kitten was always paired with the mixed vegetables. The difference in color, intensity, contrast, spatial layout, and complexity was controlled for within each pair with the intent that no other feature could be used to perform the task. The stimuli were all square, measuring 120 x 120 pixels. The background was dark gray during all phases of the trial including delivery of the food reward, and completely black during the ITI.

Procedure

Autoshaping and instrumental training.

Each subject received one session per day, five days per week. Each session terminated after the completion of 96 trials or 120 minutes had elapsed, whichever came first. This was consistent throughout the entire experiment. The images were divided into two categories, foods and animals. Four images from each category were used to create two training sets of stimuli, set A and B. Training set A consisted of the elephant, butterfly, bird, fish, candy, Brussel sprouts, cupcake, and sandwich. Training set B consisted of the penguin, turtle, kitten, frog, salad, tortellini, mixed vegetables, and strawberries. Ten subjects trained with set A and seven subjects trained with set B (Table 3.1). The stimuli were consistently presented in three locations, arranged in a triangular formation (Figure 3.2). The sample location was in the middle of the screen. The comparisons were shown offset to the left and right of the midline. If a stimulus was not presented during a trial, the location was marked by a white square outline.

Pigeons were initially trained with a mixed autoshaping and instrumental procedure. All stimuli from set A or B appeared in the sample, left comparison, or right comparison position an

equal number of times. Only one stimulus was presented at a time and the other locations were marked with a white outline (Figure 3.2a). During the first 48 trials, the stimulus was presented for 10 s. If the pigeon pecked on the stimulus (FR1) the trial would end, then the food port was illuminated and the hopper was raised for 3 seconds. The food reward delivery was consistent throughout the duration of the experiment. Pecks within 25 pixels of the outer border of the stimulus were considered on-target. If the pigeon did not peck within the target region, the food reward would automatically be delivered after 10 seconds. Pecks to the background or where the locations were marked by a white outline were neither reinforced nor punished. After food delivery terminated, there was a 13-s ITI with a black screen. During the last 48 trials, the stimulus would stay on the screen until the pigeon completed the FR1 peck requirement to the stimulus; that is, only the instrumental schedule was in place for the last 48 trials of each session. Once a pigeon was consistently pecking at the stimulus (pecking on the stimulus on 80% of the trials for 2 consecutive sessions), the autoshaping procedure was discontinued and an instrumental contingency was enforced for the entirety of each session. During the instrumental procedure, the stimulus would stay on the screen until the pigeon completed the peck requirement. The pigeon was trained with an FR1 until it completed the session within 120 minutes. Then the peck requirement was gradually increased from an FR1 using a series of variable ratio (VR) schedules, starting with VR3 +/- 2 (actual values 1, 2, 3, 4, 5), VR6 +/- 2 (4, 5, 6, 7, 8), then VR9 +/- 2 (7, 8, 9, 10, 11). Subjects had to finish the session within 120 minutes on each VR schedule before advancing to the next schedule. When subjects had completed all of the VR schedules, the number of trials that could be followed with reinforcement was reduced to 72 (75% of trials). Each stimulus in each location was presented without reinforcement once per session, but never in the first or last block of 24 trials. When subjects completed 2 consecutive

sessions within 120 minutes on this reduced reinforcement schedule, subjects began the SMTS task.

Symbolic match to sample.

Each trial had two phases, a sample phase and a choice phase. Which category of images, food or animal, was used as the sample and which was used as the comparison was counterbalanced across subjects (Table 3.1). The sample and comparisons were consistently drawn from the same food or animal category. For example, a subject would only see animal images as the sample and only food images as the comparisons. The stimulus pairs were kept constant across subjects. For training Set A, the stimulus pairs were elephant – candy, butterfly – Brussel sprouts, bird – cupcake, and fish - sandwich. For training Set B, the stimulus pairs were penguin – salad, turtle – tortellini, kitten – mixed vegetables, and frog – strawberries. Subjects only saw images from one set during training. Each sample stimulus was presented along with each of the three incorrect comparisons and equal number of times. The correct comparison stimulus was presented equally often as the left or right comparison. This resulted in a total of 24 unique stimulus configurations. Subjects experienced each stimulus configuration four times per session for a total of 96 trials.

At the onset of the trial, all three stimulus locations were presented, with the center location showing one of the four sample images from the food or animal category while the comparisons were marked with a white outline. Similar to during instrumental training, the sample stimulus was presented until the subject completed an FR10 peck requirement. Once the peck requirement was completed, the choice phase began and two stimuli from the remaining category were presented as comparison stimuli (Figure 3.2b). If the subject completed the FR1 peck requirement to the correct comparison stimulus, the choice phase would end immediately

after the peck, the subject would receive a food reward, and then the ITI would begin. If the subject pecked the incorrect comparison, the choice phase would end, the ITI would begin, and the trial would be repeated starting at the sample phase (i.e., a correction procedure). Correction trials were not included in the data analysis. During the choice phase, pecks to the sample and the background were neither reinforced nor punished. The subject had an unlimited amount of time to complete this peck requirement. Training was continued until subjects were 80% accurate on all stimulus pairs in a single session or until they had trained for 35 sessions.

Data Analysis

Sessions were only included in the analysis if the subject completed all 96 trials. One session was excluded for Wenchang, Mario, Estelle, Waluigi, Jubilee, and Luigi (n = 6), two sessions were excluded for Cousteau, and three sessions were excluded for Dickinson for failing to complete the session. The number of sessions to reach criterion was the primary measure of interest. The number of sessions needed was compared across the training sets to ensure that the different images used did not lead to differences in performance. A similar analysis was conducted based on which category of images served as the sample or comparison.

While the primary goal of this experiment was to determine if there was sufficient variability in performance to detect individual differences in associative learning ability, the effect of age on performance was also investigated. To investigate potential effects of age, subjects were divided into two groups, young and old. The subjects in the young group were between 0.5-4 years old at the start of the experiment (n = 9) and the subjects in the old group were between 11-18 years old (n = 8). Age was also investigated as a continuous variable. Data were analyzed using JASP, version 0.14.1 (JASP Team, 2020).

Results

Subjects needed a variable number of sessions to reach criterion, 80% accuracy on all four pairs in a single session, with some subjects reaching criterion in as few as eight sessions while others received the maximum amount of training, 35 sessions, without reaching criterion (Figure 3.3). To ensure that there were no differences based on training set, which category served as the comparison or sample, sex, or age, an independent samples t-test was used (De Winter, 2013). There was no significant difference for subjects that trained with set A (n = 10, M = 20.8, SD =9.95) compared to set B (n = 7, M = 21, SD = 8.96), t (15) = -0.042, p = .967. Similarly, there was no significant difference for subjects that trained with animal pictures as the sample stimuli (n = 8, M = 17.5, SD = 8.82) compared to food pictures as the sample stimuli (n = 9, M = 23.89, SD = 9.06), t (15) = -1.47, p = .162. There was no significant difference in performance for male (n = 9, M = 18.89, SD = 8.94) compared to female subjects (n = 8, M = 23.13, SD = 9.7), t (15) = .94, p = .363. There was also no significant difference in performance in young subjects, ranging from 0.5-4 years old (n = 9, M = 17.44, SD = 7.13) compared to old subjects, ranging in age from 11-18 years old (n = 8, M = 24.75, SD = 10.29), t (15) = 1.72, p = .106. How age impacted performance was also investigated as a continuous variable instead of separating subjects into two group using a two-tailed Pearson's correlation. There was a significant positive correlation between the age of the subject and how many sessions were needed to reach criterion, r (15) = .605, p = .01 (Figure 3.4). Thus, it appears that older pigeons took longer to learn the associations compared to younger pigeons.

Discussion

The SMTS task is a more complex associative learning task, where subjects need to learn four different pairs of pictures, compared to the type of associative learning procedure that is

typically administered to animals in which one stimulus is associated with an appetitive outcome and another is not (Flaim & Blaisdell, 2020). This more complex associative learning task is more similar to the word-pairs task given to humans. The variability in performance on the SMTS task indicates it is sensitive to individual differences in associative learning ability. By incorporating the SMTS task into a cognitive test battery, we can determine if associative learning ability and task complexity is related to a general cognitive factor, similar to what we see in people.

The different training sets and different categories serving as the sample and comparison did not seem to impact performance. This indicates that the SMTS provides a general assessment of associative learning. In addition, these results suggest that subjects could be trained with the stimulus set they to which they had not been exposed to test the reliability of the SMTS task. If this task is a reliable measure of general associative learning ability, then performance should be similar across the different training sets, whereas if the SMTS task is unreliable then individual performance should not be consistent across training sets. Unreliable measures are more heavily impacted by random error or transient factors, unrelated to the construct of interest (John & Benet-Martinez, 2000). Thus, reliable measures provide a more consistent result. Additionally, the strength of the correlation between any measure is restricted by their reliability (Jensen, 1998; John & Benet-Martinez, 2000). As measures become more unreliable, the correlations will be closer to zero, even if the 'true' relationship between the measures is strong (Trafimow, 2015). How unreliable measure attenuate correlations is especially important when investigating *g* in animals. A uniformly positive correlational matrix is the first indication of *g* (Deary, 2000; Jensen, 1998). If the correlation matrix found in animals does not have strong positive correlations, it could be because they do not have a *g* factor, because the measures used are not

reliable, making the correlations appear weaker than they really are, or a combination of these two factors (Soha et al., 2018; Trafimow, 2015). For example, song sparrows and Australian magpies have been given very similar cognitive test batteries that include a simple color discrimination and reversal to measure associative learning (Anderson et al. 2017; Ashton et al., 2018; Boogert et al. 2011; Soha et al., 2019). In Australian magpies, performance on the tasks in the battery was reliable and the subsequent correlation matrix had strong, positive correlations (Ashton et al., 2018). In contrast, performance on these tasks was not reliable in song sparrows, and the correlation matrix was weak and not uniformly positive (Soha et al., 2019). The results from song sparrows suggests that the tasks used are not appropriate measures of cognitive ability or that cognitive ability is not a stable trait for this species specifically (Soha et al., 2019). Either reason makes it less likely that song sparrows have a *general* cognitive ability as seen in Australian magpies, mice, and humans (Anderson et al. 2017; Ashton et al., 2018; Boogert et al. 2011; Flaim & Blaisdell, 2020; Kolata et al., 2008; Soha et al., 2019). Still, it is unclear what could cause a species difference in the task reliability and potential cognitive structure. Future research investigating differences in reliability or differences in the factors that impact reliability would be very informative to understanding how cognitive abilities vary across species (Cauchoix et al., 2018; Colombo & Scarf, 2020). This is why the SMTS task presented here, with no statistically significant differences in initial acquisition based on training set, would be ideal for exploring reliability in performance in the pigeon.

Another result worth closer investigation is how age of the subject impacts the number of sessions needed to reach criterion. When subjects were split into groups based on age, there was no significant difference in performance. Age, however, is a continuous variable, and when investigated from that perspective, there was a significant positive correlation between the

number of sessions needed to reach criterion and the age of the subject. Older subjects generally needed more training sessions to reach criterion compared to younger subjects (Figure 4). This indicates a decline in associative learning abilities with age, similar to what is seen in humans using the word-pairs task (Old & Naveh-Benjamin, 2008). Even with declines in performance due to age, intelligence still impacts performance (Ratcliff et al., 2011). Therefore, this task could be vital for exploring general cognitive abilities or intelligence, how associative learning changes with age, and how these interact in the pigeon (and potentially other nonhuman animals) compared to what is reported for humans.

This experiment indicates that the SMTS is sensitive to individual differences in performance and should be included in cognitive test batteries. Including this task in test batteries could assess potential similarities between the general factor extracted in animals and *g* in humans, like a higher loading for more complex tasks. Additionally, this task could be used to explore reliability and age-related declines in cognitive performance.

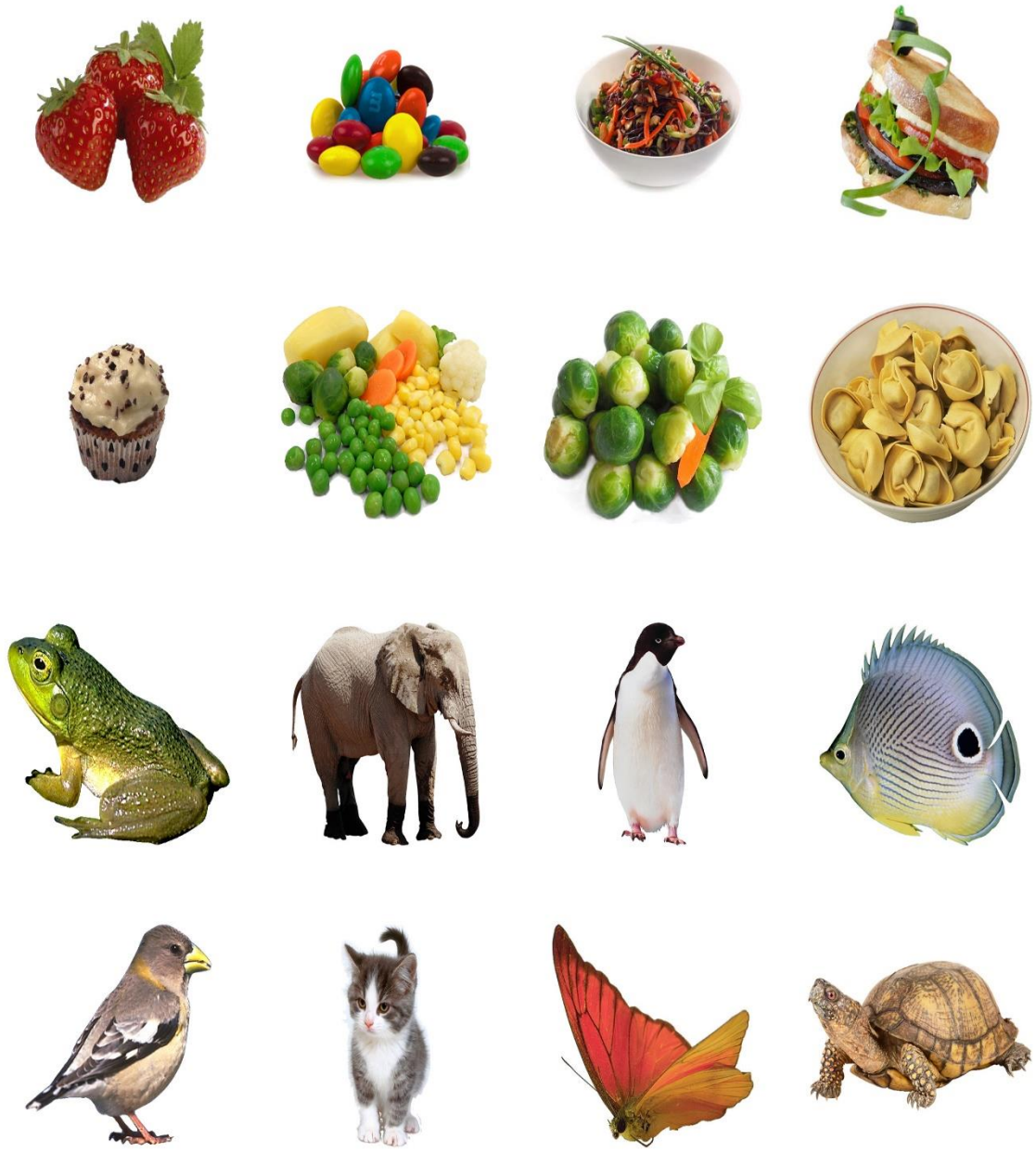
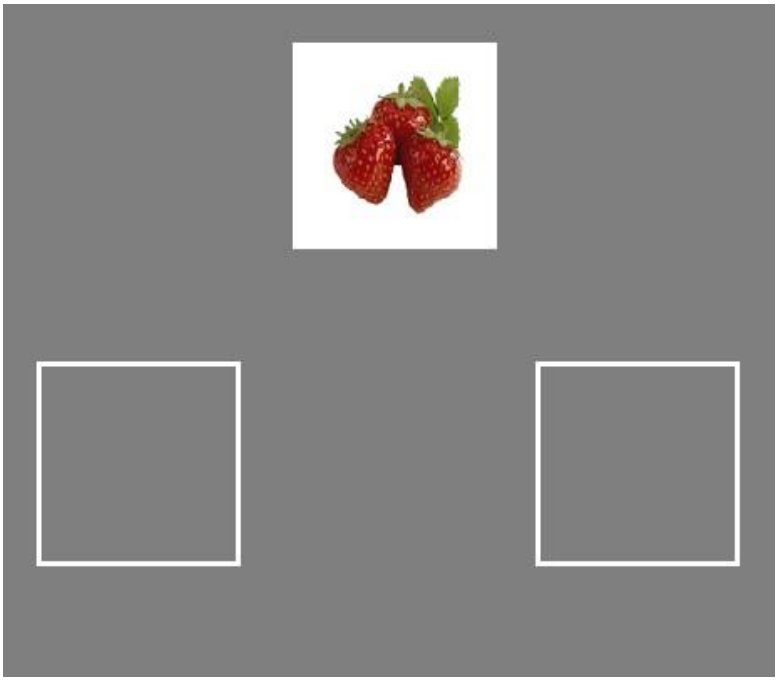


Figure 3.1. Pictures of the foods and animals that were used to create training sets A and B. The pictures are from the food-pic database (Blechert et al., 2014).

a.



b.

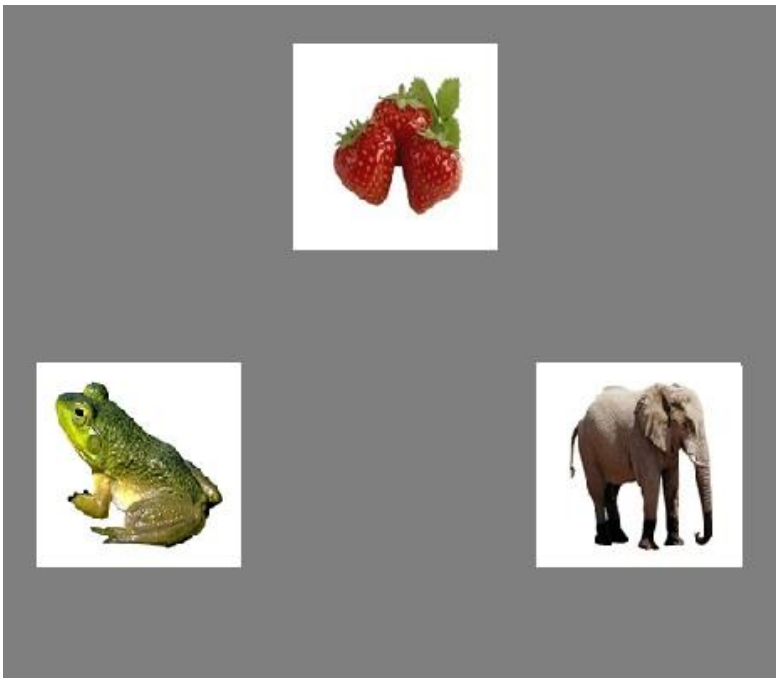


Figure 2. Panel a depicting either a stimulus in the sample location only, while the remaining locations are marked with a white outline, as seen in the sample phase during the symbolic match to sample task. Panel b depicts the choice phase during the symbolic match to sample task.

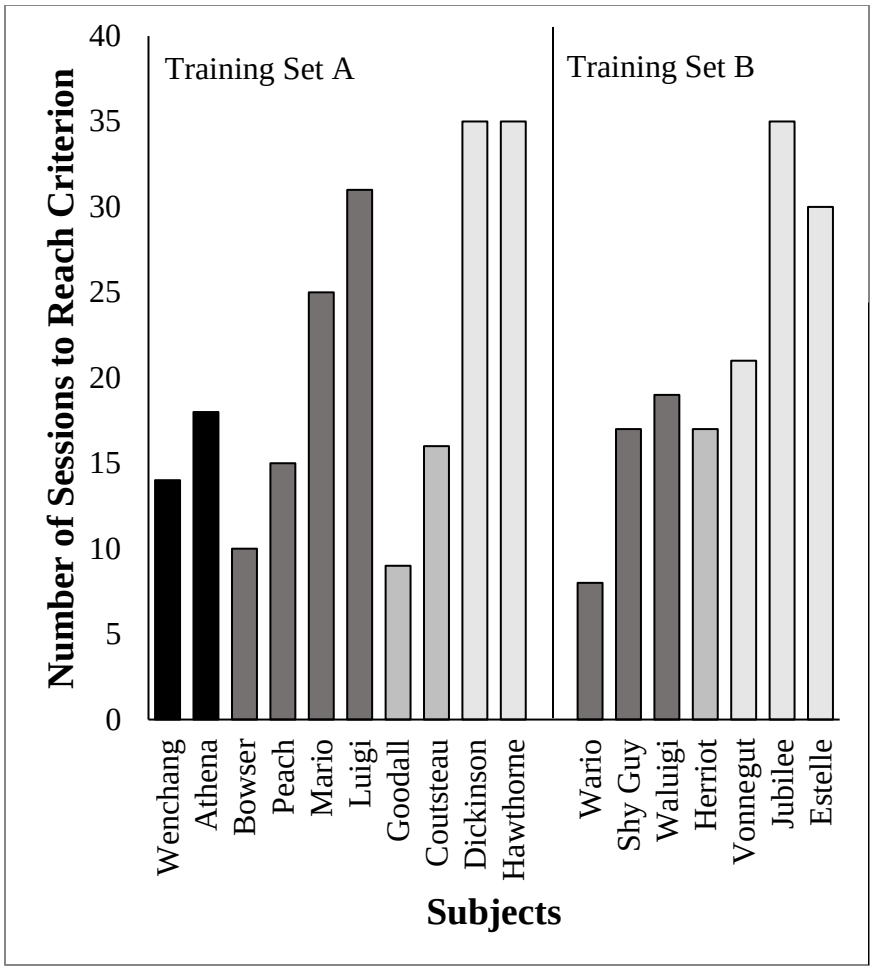


Figure 3. The number of sessions needed to reach criterion based on which training set subjects received. Subjects are organized by age, from youngest to oldest, where younger subjects are depicted in a darker shade and older subjects are depicted in a lighter shade.

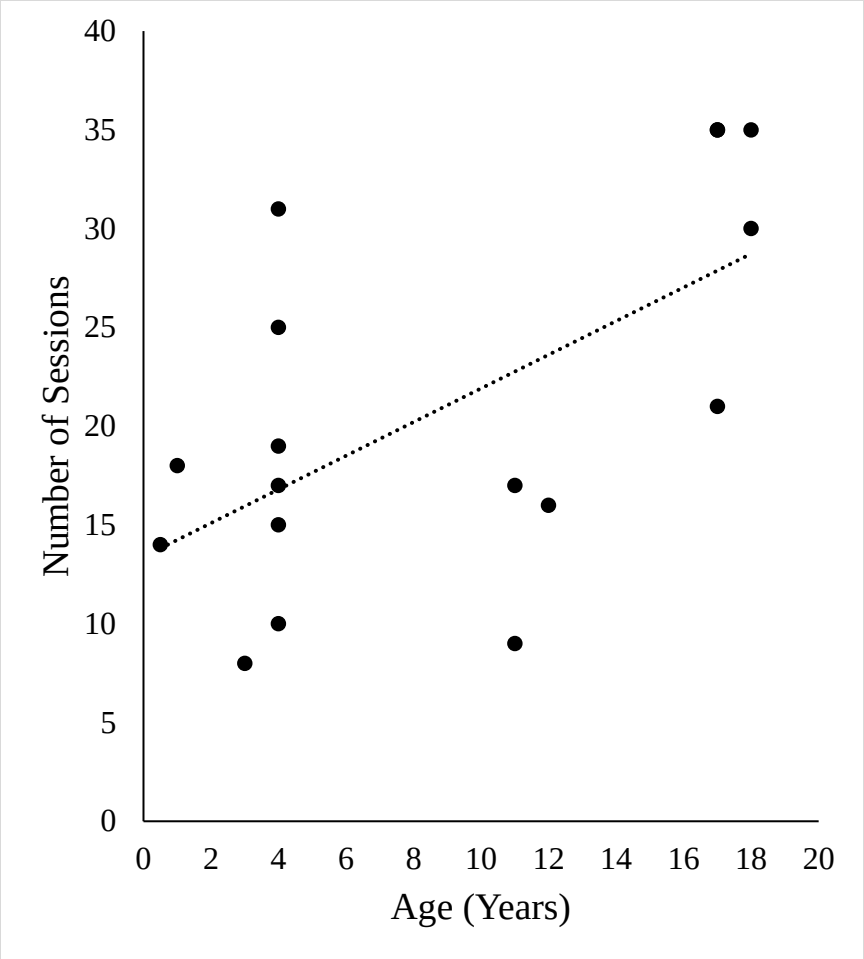


Figure 4. The correlation between the number of sessions experienced during training and the age of the subject in years. Each data point is an individual subject, while the line indicates the strength and direction of the correlation.

Acquisition During the Symbolic Match to Sample Task						
Name	Age	Sex	Set	Sample	Comparison	Sessions to Criterion
Wenchang	0.5	F	A	Animal	Food	14
Mario	4	F	A	Animal	Food	25
Peach	4	M	A	Animal	Food	15
Goodall	11	F	A	Animal	Food	9
Hawthorne	18	M	A	Animal	Food	35
Athena	1	F	A	Food	Animal	18
Luigi	4	M	A	Food	Animal	31
Bowser	4	M	A	Food	Animal	10
Cousteau	12	M	A	Food	Animal	16
Dickinson	17	F	A	Food	Animal	35
Wario	3	M	B	Animal	Food	8
Shy Guy	4	M	B	Animal	Food	17
Herriot	11	M	B	Animal	Food	17
Waluigi	4	F	B	Food	Animal	19
Vonnegut	17	M	B	Food	Animal	21
Jubilee	17	F	B	Food	Animal	35
Estelle	18	F	B	Food	Animal	30

Table 3.1. Number of sessions to reach criterion for each subject based on age in years, sex, training set, and which category of images served as the sample or comparison

Chapter 4: Serial Reversal Learning

Abstract

Reversal learning assesses cognitive flexibility since it requires subjects to change their response after sudden and unsignaled changes in previously learned contingencies. This task has been frequently used to examine the cognitive abilities in a wide variety of human populations, but it is notably absent in assessments of general intelligence. In contrast, reversal learning is commonly included when investigating general intelligence in avian species. A wide variety of avian species have been tested, but the pigeon has yet to assessed in this systematic way. This experiment outlines a serial, or multiple reversal, learning task to determine if it is appropriate to reliably detect individual differences in the pigeon. The results indicate that, although almost all subjects improve their performance over time, there is sufficient variability across subjects for inclusion in a cognitive test battery. While other avian species assessed thus far showed positive correlations between the initial discrimination and first reversal, our subjects only showed a positive relationship from the third reversal onwards.

Introduction

Quickly learning which stimulus or response will result in a reward is important for effectively navigating the world. Since the world is not static, however, responding flexibly to your environment and updating your learning if the stimulus or response contingencies change is equally important (Mettke-Hofmann, 2014). Therefore, the learning processes underlying behavior should be sensitive to changes in the environment, such as shifts in previously established reward contingencies, and modify behavior accordingly (Izquierdo et al., 2017; Racey et al., 2011). There are a wide variety of procedures to assess how behavior changes with fluctuating reward contingencies, but one of the most commonly used procedures is

discrimination reversal learning (Izquierdo & Jentsch, 2012). Discrimination reversal learning is where subjects first learn that one stimulus is followed by a reward while the other is not, then, after reaching a pre-determined criterion, the contingencies switch. When contingencies switch multiple times, this is referred to as serial reversal learning. While reversal learning procedures assess cognitive flexibility, this term lacks precision (Audet & Lefebvre, 2017). More specifically these tasks assess the ability to learn which stimuli are or are not followed by a reward, estimate the probability of a reversal, and, in the case of serial reversal learning, understand the overall task structure (Izquierdo et al., 2017).

Despite the clear cognitive components underlying performance on serial reversal learning tasks, how it is related to intelligence is rarely investigated in humans (Flaim & Blaisdell, 2020). Human intelligence test batteries consistently show that performance positively correlates across diverse cognitive tasks. When this positive correlational matrix is subjected to factor analysis, one factor that can account for half of the variance in performance is consistently extracted. This factor is termed *g* (Carroll, 1993; Deary, 2000). While this *g* factor is well replicated, it is not well understood (Conway & Kovacs 2015; Deary, 2000). There are a number of difficulties in explaining why performance is correlated across cognitive tasks, but two are of particular interest. The first is the multifaceted nature of identified cognitive domains and the second is task impurity. More specific cognitive domains have been theorized to be the primary cause of *g*, but even 'specific' cognitive domains have subcomponents (Kovacs & Conway, 2016). For example, working memory, the ability to store information while processing other information, includes multiple attention and storage systems (Baddeley, 2002). Even if the specific subcomponents are identified, it is impossible to create a task that assesses only one cognitive ability, which is referred to as task impurity. Tasks always assess multiple cognitive

abilities and the identity of the specific cognitive abilities engaged by any specific task is not always clear (Burgoyne et al., 2019; Conway et al., 2003; Redick & Lindsey, 2013). One theory of *g* argues that task impurity is a primary cause for the positive correlations between different tasks (Kovacs & Conway, 2016). Why a *g* factor is found is still debated, but including reversal learning procedures in human intelligence assessments could provide interesting test of the theories. As described earlier, reversal learning relies on multiple cognitive abilities, but they have already been identified. This would make it easier to understand why reversal learning relates to other cognitive abilities, and help clarify when the strongest correlations between different tasks will be found (Conway et al., 2003). In the unlikely event that reversal learning does not correlate with other cognitive abilities, this would highlight an important limitation of *g* and require an explanation.

While reversal learning has been neglected in human intelligence research, it is frequently included in avian test batteries. For example, Australian magpies were given a color discrimination task where one shade of blue was consistently followed by a food reward, while the other was not. Once subjects reached criterion, 10 correct choices out of 12 total choices, the contingency was reversed. The mean number of trials to reach criterion on the initial discrimination was 22.77 and the mean number of trials for the reversal was 30.12, an increase of approximately 32% (Table 4.1). Performance on the initial association and subsequent reversal were positively correlated with each other and with an inhibitory control and spatial memory task. Principal component analysis (PCA), a dimension reduction technique similar to factor analysis, extracted a component similar to the *g* factor in humans. Reversal learning contributed to this component, indicating that procedure can measure general cognitive ability in Australian magpies (Ashton et al., 2018). Using similar procedures, similar results have been found with

robins (Shaw et al., 2015) and spotted bowerbirds (Isden et al., 2013). While a similar procedure was ultimately not successful in uncovering a *g* like factor in song sparrows, performance on the initial discrimination and reversal was positively correlated (Boogert et al., 2011). How different cognitive tasks load onto a general factor or positively correlate could indicate differences in what underlies general intelligence across species, but comparisons across species are difficult due to differences in the test batteries. Part of the reason why there are differences in cognitive test batteries across species is because the testing apparatus must be appropriate for the physical and sensory abilities for the species in question. The pigeon could facilitate comparisons across species because they are a commonly used species for behavioral studies of cognition using touchscreen operant tasks designed to be similar to tasks used in research with human and nonhuman primates (Güntürkün et al., 2017). Yet pigeons have never been given a comprehensive test battery. The goal of this experiment was to determine whether a serial reversal learning procedure would be appropriate to include in such a battery. Other research has confirmed that pigeons are able to perform multiple reversals (Lissek et al., 2002; Durlach & Mackintosh, 1986). Durlach and Mackintosh (1986) initially trained pigeons to discriminate between color or line orientation stimuli. Subjects were trained on this contingency until they reached criterion, 9 out of 10 consecutive trials correct. Once they reached criterion, the contingency was reversed on the next session, so reversals only occurred between sessions. Contingencies were reversed every time subjects met criterion on three consecutive sessions. Their results indicate that subjects who were trained on the color stimuli were more accurate on the subsequent reversals compared to subjects trained on the line orientation stimuli. While these results indicate that subjects can improve their performance over multiple reversals, and that color stimuli are easier for subjects to learn, it is unclear if there were individual differences

across subjects. Serial reversal learning using color stimuli has been used more recently to investigate the role of NMDA receptors in the nidopallium caudolaterale, the avian equivalent of the prefrontal cortex in mammals (Lissek et al., 2002). In this experiment, subjects were trained with red and green color stimuli and when subjects reached criterion, 15 correct consecutive choices, the contingencies were reversed in the next sessions for a total of six between-session reversals. Subjects were more accurate on later reversals, replicating the results from Durlach and Mackintosh (1986), but there was also evidence for between subject variability, particularly for the first three reversals. These results indicate that a serial reversal learning task using color stimuli should be sensitive to individual differences.

The current experiment used blue and yellow circles, and the contingency was reversed after subjects reached criterion, 90% accuracy on two consecutive sessions for a total of five between-session reversals. To compare to other avian experiments, performance was also measured as the number of trials needed to make 10 consecutive correct choices on the initial discrimination and first reversal. The results indicate that all subjects improved their performance on the task over time, and, most importantly for our goals, there was sufficient variability in this task to detect individual differences in cognitive abilities. In contrast to previous experiments, however, performance on the initial discrimination and first reversal was not correlated. Performance was positively correlated across the third, fourth, and fifth reversals. These results could indicate an important difference between pigeons and other avian species, though whether this is because of experience or underlying neurological differences is not clear. These results also highlight the disassociation between early and late reversal performance.

Methods

Subjects

Twenty-three pigeons served as subjects. These subjects ranged in age from 0.5-17 years old at the start of the experiment and there were 10 females (Table 4.2). All subjects had variable exposure to other cognitive tasks. For all subjects at least one of the previously experienced tasks included a 'correction' procedure where trials would repeat until the subject pecked the correct stimulus. Pigeons were individually housed in steel home cages with metal wire mesh floors in a vivarium. They were maintained at 80% of their free-feeding weight, but were allowed free access to water and grit while in their home cages. Testing occurred at approximately the midpoint of the light portion of the 12-hour light-dark cycle.

Apparatus

Testing was conducted in a flat-black Plexiglas chamber (38 cm wide x 36 cm deep x 38 cm high). All stimuli were presented by computer on a color LCD monitor (NEC MultiSync LCD1550M) visible through a 23.2 x 30.5 cm viewing window in the middle of the front panel of the chamber. Stimuli were presented using the coding language Python (Python Software Foundation, https://www.python.org/) and the extension PsychoPy (Peirce, 2007). The bottom edge of the viewing window was 13 cm above the chamber floor. Pecks to the monitor were detected by an infrared touchscreen (Carroll Touch, Elotouch Systems, Fremont, CA) mounted on the front panel. A food hopper (Pololu, Robotics and Electronics, Las Vegas, NV) was located in the center of the front panel, its access hole flush with the floor. All experimental events were controlled and recorded with a Pentium III-class computer (Dell, Austin, TX). A video card controlled the monitor in the SVGA graphics mode (800 x 600 pixels).

Stimuli

The stimulus set consisted of two circular stimuli, each with a diameter of 6 cm. The circle could be blue or yellow. The background screen color was dark gray.

Procedure

Subjects had been exposed to other tasks that required a peck response, so they were not shaped to peck at the stimuli set or trained to eat from the magazine during this experiment. Each subject was given one session per day, five days per week. Each session terminated after completion of 50 trials or after 60 minutes had elapsed, whichever came first. Each trial consisted of the presentation of a blue stimulus on one side of the screen and a yellow stimulus on the other side, with the left-right position counterbalanced within session. The stimuli were located in the horizontal center of the screen and positioned to the left and right of the midline with 15 cm between the stimuli. If the subject made a total of three pecks (FR3) to the rewarded stimulus (S+), the trial would end, they would receive access to the hopper for three seconds, and the 15 s inter-trial interval (ITI) would begin. If they completed the peck requirement to the other, nonrewarded stimulus (S-), the trial would simply end and the ITI would begin. Pecks within 1.5 cm outside of the edge of the stimulus were considered on-stimulus. Pecks to the background had no consequence. There was no time limit to meet this peck requirement.

Each subject was randomly assigned which color would initially be rewarded, counterbalanced according to age and sex as much as possible. Once subjects were pecking the S+ on 90% of the trials on two consecutive sessions, the contingency reversed, now the previous S- was the S+ and vice versa. Every time this criterion was met, the contingencies were reversed. Reversals only occurred between sessions. Subjects went through a total of five reversals.

Similar procedures have been successful at revealing individual differences in avian species (Lissek et al., 2002).

Data Analysis

Accuracy on the first session of each reversal, number of sessions to reach criterion performance, and the total number of sessions in each phase of training were analyzed. Sessions were only included in the accuracy performance and the number of sessions to reach criterion if the subject completed all 50 trials. One session was excluded for Wenchang, Itzamná, Luigi, Mario and Bowser, and two sessions were excluded for Waluigi for failing to reach 50 trials. A Spearman correlation was used to determine if performance was consistent across the initial discrimination and reversals for accuracy on the first session. To compare performance with other avian test batteries, performance was also investigated as the number of trials needed to make 10 consecutive correct choices for the initial discrimination and first reversal. For the number of trials to criterion, sessions were included even if subjects did not complete 50 trials To investigate potential effects of age, subjects were divided into two groups, young and old. The subjects in the young group were between 0.5-4 years old at the start of the experiment (n = 12) and the subjects in the old group were between 11-17 years old (n = 11). Data were analyzed using JASP, version 0.14.1 (JASP Team, 2020), but the goal of this experiment was to determine if this task had sufficient variability across subjects to detect individual differences.

Results

Most subjects received initial discrimination training and 5 reversals, but there were a few computer errors that impacted data collection. Mario advanced from the second to the third reversal before reaching criterion, while Durrell did not receive training on a fifth reversal (Table 4.2). The first session performance for the first reversal was not recorded correctly for Herriot

and Cousteau. For Durrell and Luigi, the number of trials needed to make 10 consecutive correct choices during the first reversal could not be calculated since some of the data files were missing trial number information. Thus, only 19 subjects were included when analyzing the number of trials criterion.

The number of sessions to reach criterion for the initial discrimination and each reversal were analyzed. Due to how the criteria was set, 90% accuracy on two consecutive sessions, there was less variability across subjects, but generally subjects needed fewer sessions to reach criterion for each reversal (Table 4.2). The data violated Mauchly's test of sphericity (X^2 (14) = 35.637, p = .001) so a three-way 2 x 2 x 6 mixed ANOVA, with age and sex as the between subject factors and the reversal number as the within-subjects factor, with a Greenhouse-Geisser correction was used. The results indicated there was no significant main effect of age (F (1, 18) < 1), or of sex (F (1, 18) < 1), nor a significant interaction between age and sex (F (1, 18) < 1).

There was a significant main effect of reversal number (F (2.635, 47.422) = 10.298, p < .001, partial eta squared = .364), but no significant interaction between reversal number and age (F (2.635, 47.422) < 1), or sex (F (2.635, 47.422) < 1), or age and sex (F (2.635, 47.422) < 1). Post-hoc tests using a Bonferroni correction on reversal number indicated that this main effect was due to the difference between the number of sessions to reach criterion on the initial discrimination (M = 2.826, SD = 0.778) and the number of sessions to reach criterion on the first (M = 4.391, SD = 0.891; p < .001), second (M = 4, SD = 0.953; p < .001), third (M = 4.435, SD = 1.472; p < .001), fourth (M = 4.174, SD = 1.193; p < .001) and fifth reversal (M = 4.136, SD = 1.356; p < .001). There were no significant differences in the number of sessions between the other reversals.

The performance on the first sessions for each reversal showed more variability across subjects, but there was a general trend for improving over time (Figure 4.1). A three-way 2 x 2 x 5 mixed ANOVA, with age and sex as between subject factors and reversal number as the within subject factor, indicated there was no significant main effect of age (F (1, 16) < 1), or of sex (F (1, 16) < 1), or a significant interaction between age and sex (F (1, 16) < 1). There was a significant main effect of reversal number, F (4, 76) = 21.535, p < .001, partial eta squared = .531), but no significant interaction between reversal number and age (F (4, 64) < 1), or sex (F (4, 64) < 1), or age and sex (F (4, 64) = 1.494, p = .215). Post-hoc tests using a Bonferroni correction for the accuracy on the first session of each reversal showed a significant difference in performance on the first (M = .167, SD = .159) and third (M = .422, SD = .22; p < .001), first and fourth (M = .456, SD = .206; p < .001), and first and fifth reversal (M = .53, SD = .207; p < .001). The post-hoc tests also showed a significant difference between the second (M = .257, SD = .144) and third (t = -3.905, p = .002), second and fourth (t = -4.494, p < .001), and second and fifth reversal (t = -5.732, p < .001).

A Spearman correlation was used to determine if accuracy on the first session was similar across all phases of the task. Performance on the initial discrimination and the first and second reversals had weak and nonsignificant correlations with all measures. Performance on the third, fourth, and fifth reversals were significantly, positively correlated with each other (Table 4.3).

The data were also analyzed by how many trials were needed for subjects to make 10 consecutive correct choices on the initial discrimination and first reversal. The number of trials received after reaching this criterion were also recorded (Table 4.4). A paired samples t-test showed that subjects needed significantly fewer trials to reach criterion on the initial discrimination compared to the first reversal, t (18) = -8.85, p < .001 (Table 4.1). A Spearman

correlation indicated there was no relationship between performance on the initial discrimination and first reversal, r_s (17) = .06, p = .81. A Spearman correlation was also used to investigate if the number of trials experienced after reaching criterion facilitated subsequent reversal performance (Williams, 1967), but, again, there was no relationship r_s (17) = .1, p = .68.

Discussion

This experiment replicates previous findings that performance improves across multiple contingency reversals, particularly when using performance on the first session as the dependent variable (Durlach & Mackintosh, 1986; Lissek et al., 2002). This experiment also demonstrates the sensitivity of the serial reversal learning procedure to individual differences by showing individual subject data for the number of sessions to reach criterion and performance on the first session for each reversal (Table 4.2, Figure 4.1). The number of sessions to reach criterion was a less sensitive measure, since there was less variability across subjects and post-hoc analyses were not able to distinguish improvements in performance over sessions. The low variability was partially due to how criterion was set, 90% accuracy on 2 consecutive sessions. Another reason for the low variability for this measure is that subjects reached criterion relatively quickly. The initial discrimination and five reversals were completed in 17-36 sessions (Table 4.2). It is difficult to determine if this is similar to other pigeon experiments because of the differences in the criterion. In addition, only the number of errors committed by each subject is reported (Durlach & Mackintosh, 1986; Lissek et al., 2002). It is still surprising from a species perspective because pigeons typically need more training compared to mammals to reach criterion on cognitive tasks (Güntürkün et al., 2017; Mackintosh & Cauty, 1971). The relatively low number of sessions to reach criterion could be due to previous experience with correction procedures, where a trial repeats until the subjects chooses the correct stimulus. All subjects had

experience with correction procedures and this experience could have resulted in a general strategy that transferred to the current experiment.

The other dependent measure, performance on the first session of each reversal, showed a greater amount of variability across subjects, and post-hoc analyses detected more significant differences between reversals. This indicates that the serial reversal learning task is sensitive to individual differences and would be appropriate to include in a cognitive test battery for pigeons. While between-subject variability is important when assessing individual differences, the task should also be reliable. Only performance on the third, fourth, and fifth reversals were positively correlated with each other, indicating that the initial discrimination and first and second reversals were assessing something different compared to the later reversals. A single reversal relies on extinguishing the previously reinforced response and learning a new response. Multiple reversals, however, could involve acquisition of a more general 'win-stay, lose-shift' rule, which would indicate subjects were learning the structure of the task itself (Izquierdo et al., 2017). Therefore, the initial discrimination and early reversal could measure the ability to form excitatory and inhibitory connections while later reversals measure more abstract rule learning. Alternatively, the increasing rapidity of subsequent reversals could be due to the buildup of excitatory and inhibitory properties to each stimulus, allowing for smaller differences in reward to trigger the reversal of choice. The possibility of using a higher-order, rule-based strategy is an additional measure of cognitive ability not assessed by previous avian test batteries. It is still possible, however, to compare the different species based on the initial discrimination and first reversal.

In robins, song sparrows, spotted bower birds, and magpies, performance on the initial discrimination and reversal was positively correlated (Ashton et al., 2018; Boogert et al., 2011;

Isden et al., 2013; Shaw et al., 2015), whereas for pigeons there was no relationship when using first session accuracy (Table 4.2) or number of trials to make 10 consecutive correct choices. The number of trials to criterion highlighted another difference between pigeons and the other species assessed, specifically how much additional training was required to reach criterion on the reversal. While all species needed more training for the reversal, this was more extreme in pigeons. In robins, song sparrows, and magpies, subjects needed, on average, an additional 18.1, 7.5, and 7.35 trials respectively to reach criterion on the reversal compared to the initial discrimination. This was an increase of approximately 24-32% compared to the number of trials needed for the initial discrimination. Pigeons needed, on average, an additional 75.64 trials to reach criterion on the reversal, or 70% more trials to reverse compared to trials to acquire the initial discrimination (Table 4.1). The absence of a relationship between the initial discrimination and first reversal and the amount of training needed for the first reversal indicates there may be important differences in the cognitive abilities of pigeons compared to robins, song sparrows, and magpies. It may be more difficult for pigeons to form inhibitory connections or extinguish previously learned contingencies. In addition, the ability to form excitatory connections may not be related to the ability to form inhibitory connections, which weakens the hypothesis that pigeons have a *g* like factor. These cognitive differences may be related to life and species differences between pigeons and the other avian species. There is evidence that pigeon domestication began with Neanderthals (Blasco et al., 2014) and most of the subjects (n = 21) had been in captivity for over a year, participating in a variety of experiments. In contrast, the other avian species were wild, assessed in the field, and were relatively naïve to experimental procedures. These species differences could be related to cognitive abilities, but it is also possible that these differences impact what the task assesses. Personality, particularly boldness,

may have had a stronger impact on performance for wild species because the procedures were so novel (Shaw & Schmelz, 2017). Administering a full cognitive and personality assessment to pigeons and other species, wild and domesticated, will help clarify if these differences are related to the presence or absence of a *g* like factor.

While not the main goal of this experiment, age was included as a factor during the analysis. Subjects were divided into a young group, ranging in age from 0.5-4 years old, and an old group, ranging in age from 11-17 years old. Pigeons have shown age related declines in performance on other cognitive tasks starting at 10 years old (Coppola et al., 2014, 2015), but we did not find any effect of age in this experiment in either the number of sessions to reach criterion (Table 4.2) or the performance on the first session of each reversal (Figure 4.1). This could be due to the differences in the tasks, the previous experience of the subjects, or a combination of both of these factors. This is a surprising result that should be investigated more thoroughly to understand how cognition changes with age in the pigeon. For the purposes of the investigating general intelligence in the pigeon, however, this result indicates that age should not be a reason to exclude subjects. Research with humans has shown that intelligence is a stable trait throughout the lifespan (Deary et al. 2013; Deary & Brett, 2015), meaning that the older subjects in this experiment should still provide results relevant to the investigation of general intelligence.

In conclusion, the results from this experiment indicate that serial reversal learning is sensitive to individual differences, irrespective of age. This sensitivity means it is appropriate to include in a cognitive test battery for pigeons. In addition, the task can reliably measure a subject's performance over time, but only after the third reversal. This contrasts with the other avian species assessed thus far, which showed positive correlations in the initial discrimination

and subsequent reversal (Ashton et al., 2018; Boogert et al., 2011; Isden et al., 2013; Shaw et al., 2015). Only by administering a diverse test battery to pigeons will it be possible to determine if this difference extends to the underlying cognitive structure.

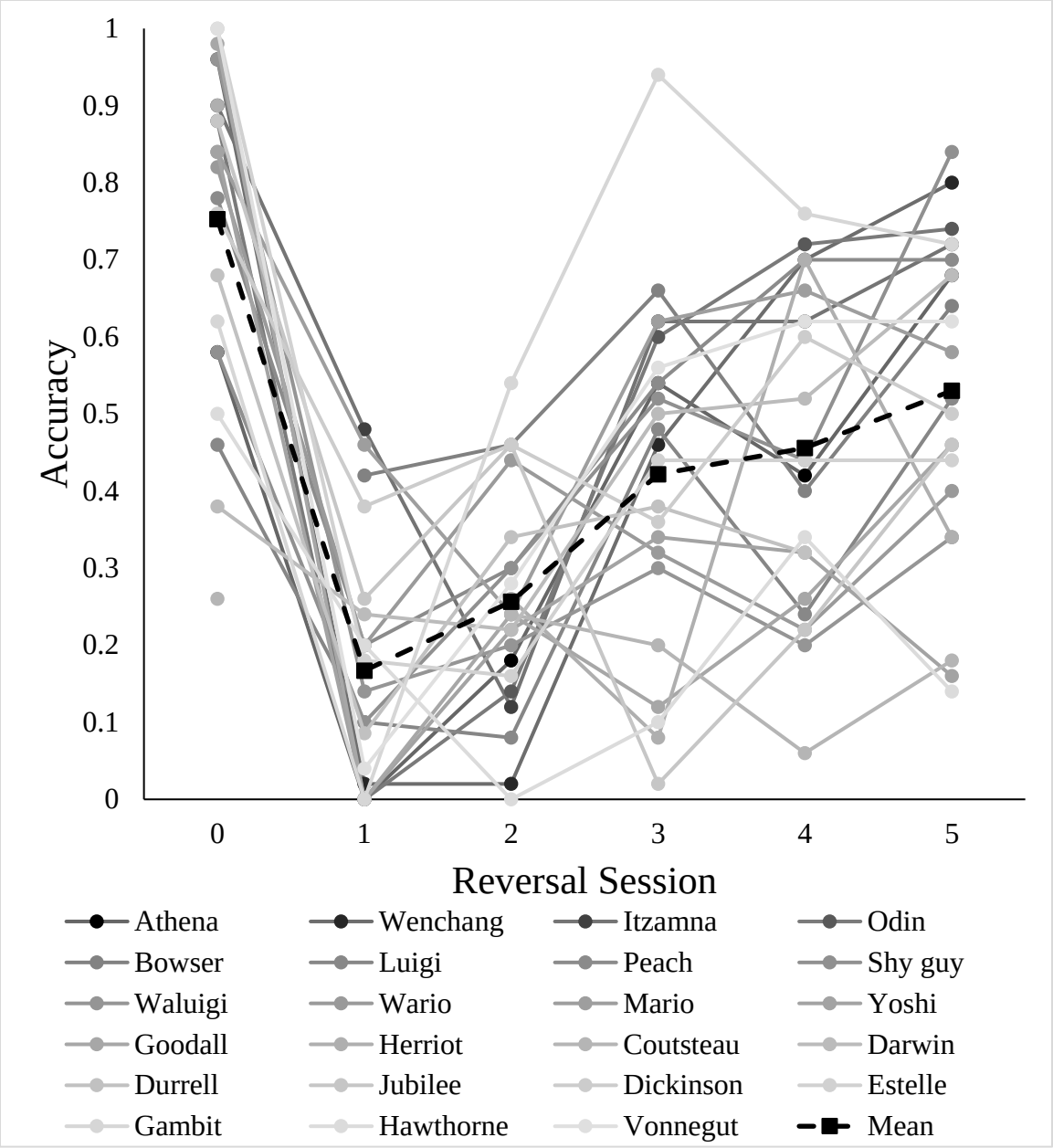


Figure 4.1. Performance on the first session for each reversal for each subject, organized by age from youngest to oldest. Younger subjects are in darker shades, while older subjects are in a lighter shade. Performance on the initial discrimination is not shown for one subject, Bowser, and performance on the first reversal is not shown for two subjects, Herriot and Cousteau, due to a computer error.

Species				
Robins (Shaw et al., 2015)				
	Initial Discrimination	First Reversal	Percent Increase	Criterion
Mean	40.5	58.6	30.89	10 out of 12
Standard Deviation	19.33	15.54		
Range	12, 80	33, 89		
Song Sparrows (Boogert et al., 2011)				
Mean	16.2	23.7	31.65	6 out of 7
Standard Deviation	6.2	7.2		
Range	8, 36	13,40		
Australian Magpies (Ashton et al., 2018)				
Mean	22.77	30.12	24.4	10 out of 12
Standard Deviation	2.08	3.07		
Range	10,65	10, 94		
Pigeons (Flaim & Blaisdell, 2021)				
Mean	31.89	107.53	70.34	10
Standard Deviation	19.41	34.31		
Range	10, 72	46, 186		

Table 4.1. The mean number of trials needed to reach criterion for the initial discrimination and first reversal for robins, song sparrows, Australian magpies, and pigeons. Percent increase refers to the relative increase in trials needed to reach criterion on the reversal compared to the initial discrimination. The criterion was always consecutive correct choices.

Number of Sessions									
			Reversal Number						
Sex	Age (Years)	Subject	0	1	2	3	4	5	Total
F	0.5	Athena	3	5	4	4	3	3	22
F	0.5	Wenchang	2	5	4	3	3	3	20
M	2	Itzamná	2	5	4	3	3	3	20
M	2	Odin	3	5	4	3	4	3	22
M	3	Bowser	2	3	3	4	3	3	19
M	3	Luigi	5	4	5	4	5	5	28
F	3	Peach	3	4	3	4	3	3	20
M	3	Shy guy	3	4	4	4	4	3	22
M	3	Waluigi	2	4	4	7	6	5	28
M	3	Wario	3	4	5	6	6	5	29
F	4	Mario	3	3	2	3	3	3	17
M	4	Yoshi	3	7	4	8	6	8	36
F	11	Goodall	2	4	4	5	5	5	25
M	11	Herriot	2	4	4	4	6	4	24
M	12	Cousteau	3	4	5	4	4	4	24
F	12	Darwin	3	4	4	4	4	4	23
F	12	Durrell	3	5	4	4	4		20
F	16	Jubilee	3	6	6	8	6	7	36
M	17	Dickinson	4	4	4	4	4	4	24
F	17	Estelle	2	4	3	4	3	4	20
M	17	Gambit	3	4	3	3	3	3	19
M	17	Hawthorne	4	5	6	5	5	5	30
F	17	Vonnegut	2	4	3	4	3	4	20
		Mean Young	2.83	4.42	3.83	4.42	4.08	3.92	23.58
		Mean Old	2.82	4.36	4.18	4.45	4.27	4.40	24.09
		Mean All	2.83	4.39	4.00	4.43	4.17	4.14	23.83

Table 4.2. Number of sessions for each reversal and total number of sessions for each subject. One subject, Durrell, only experienced 4 reversals and another subject, Mario, advanced to the third reversal before reaching criterion on the second.

Reversal Learning						
		Initial	First	Second	Third	Fourth
First	Spearman's rho	-0.091 (20)	—			
	p	0.703	—			
Second	Spearman's rho	-0.077 (22)	0.164 (21)	—		
	p	0.734	0.476	—		
Third	Spearman's rho	-0.047 (22)	0.045 (21)	0.055 (23)	—	
	p	0.837	0.848	0.804	—	
Fourth	Spearman's rho	0.198 (22)	-0.053 (21)	-0.007 (23)	**0.597 (23)**	—
	p	0.376	0.818	0.975	**0.003**	—
Fifth	Spearman's rho	0.004 (22)	-0.113 (20)	0.047 (22)	**0.744 (22)**	**0.639 (22)**
	p	0.985	0.635	0.834	< **0.001**	**0.001**

Table 4.3. Correlation matrix between the measures of the serial reversal learning task. The number inside the parenthesis is the sample size. Bolded values indicate the result was significant.

Number of Trials Criterion					
Subject	Age	Initial Discrimination	Trials Past Criterion	First Reversal	Trials Past Criterion
Athena	0.5	56	94	137	113
Wenchang	0.5	10	100	125	125
Itzamná	2	24	76	73	216
Odin	2	30	120	160	110
Bowser	3	20	90	68	82
Peach	3	30	120	71	129
Shy guy	3	58	92	95	105
Waluigi	3	15	149	89	111
Wario	3	10	150	113	137
Mario	4	31	119	46	104
Yoshi	4	43	107	186	164
Goodall	11	11	89	118	82
Darwin	12	52	98	72	128
Jubilee	16	21	129	119	181
Dickinson	17	48	152	93	107
Estelle	17	10	90	94	106
Gambit	17	55	95	112	88
Hawthorne	17	72	128	149	101
Vonnegut	17	10	90	123	77

Table 4.4. The number of trials needed for each subject to make 10 consecutive correct choices on the initial discrimination and the first reversal. The number of trials the subject experienced after reaching this criterion are also included. Age is provided in years.

Chapter 5: The Delayed Match to Sample Task

Abstract

Performance on intelligence tests and working memory tasks are consistently related, but the reason why is poorly understood. This is partially because working memory is supported by multiple cognitive processes, including attention. There is evidence that attention, the ability to prevent irrelevant information from impacting performance, is uniquely related to differences in intelligence in humans and mice. How intelligence, memory, and attention are related in other species is less understood. This experiment investigates if the delayed match to sample (DMTS) task could be used to investigate such a relationship in pigeons. Subjects were trained for 30 sessions on 0, 2, 4, and 8 second delays using red and green stimuli. This procedure was successful at finding individual differences, which means it could be used to investigate intelligence in the pigeon. Attention and its relationship to working memory, as understood in the human and rodent literature, did not seem to impact performance on this task. Implications for the DMTS and memory processes in the pigeon are discussed.

Introduction

Intelligence research has a long history of investigating individual differences in performance (Carroll, 1993; Deary, 2000). A consistent pattern has been found, where people will perform differently from each other, but individuals will perform consistently across diverse cognitive tasks. This between subject variability and within subject reliability results in a positive correlational matrix. Dimension reduction techniques, like factor analysis or principal component analysis, will extract one factor or component that can account for approximately half of the between subject variance in performance. This factor is termed *g* since it is related to almost all cognitive abilities, though not all cognitive abilities are related to the same degree (Carroll, 1993;

Deary, 2000; Spearman, 1904). The cognitive tasks that are most strongly related to *g* involve reasoning, abstraction, and generalization (Ackerman & Cianciolo, 2000). Despite the decades of research on intelligence and *g*, there is no consensus on what *causes* this pattern of results (Conway & Kovacs, 2015; Deary, 2000; Flaim & Blaisdell, 2020). One heavily investigated possibility is that differences in working memory (WM) is the primary cognitive ability underlying this pattern (Conway et al., 2003). WM is the ability to store a limited number of items for later recall while simultaneously processing a competing task or manipulating the stored items (Adams et al., 2018; Baddeley, 2002). One of the primary ways of measuring WM in humans is with complex span tasks (Conway et al., 2005; Flaim & Blaisdell, 2020). In a complex span task, the participant is given to-be recalled items interspersed with a competing task. Frequently, the to-be recalled items are words, letters, numbers, or the number of specific stimuli presented, while the competing task could be verifying if a sentence is logical, a mathematical operation is solved correctly, or verbally counting all stimuli presented on the screen (Conway et al., 2005). Another method of measuring WM is with the visual array task. In this task, participants are briefly shown an array of stimuli, then, after an inter-stimulus interval (ISI) where no stimuli are present on the screen, the array reappears and participants have to state if the array is the same as before or if one of the stimuli within the array has changed (Shipstead et al., 2015). Complex span and visual array tasks are positively correlated with each other, indicating domain general properties of WM, but they are also uniquely related to intelligence (Shipstead et al., 2015). This leads to a theoretical issue with WM as an explanatory factor for causing differences in intelligence.

The primary issue is that WM is not a unitary cognitive ability, but rather contains subcomponents (Conway et al., 2003). While the nature of the specific subcomponents has not

been agreed upon (Adams et al., 2018; Nairne, 2002; Oberauer et al., 2012), different theoretical perspectives have provided competing evidence for which subcomponent of WM has the strongest relationship with *g* or intelligence. Retrieval (Mogle et al., 2008) and temporary storage of information (Colom et al., 2006; Cowan, 2001) have experimental support, but for this experiment we will focus on the role of attention or inhibition of competing information (Bunting, 2006; Unsworth & Engle, 2006; Unsworth et al., 2009). Part of maintaining an accurate memory of the to-be recalled item means either being able to focus attention solely on the most recently presented items or by actively inhibiting memories of previously presented, but no longer relevant items and inhibiting irrelevant external and internal stimuli (Bunting, 2006; Conway et al., 2001; Unsworth & Engle, 2006). It is not clear if this subcomponent is attention or inhibition, but manipulating the amount of proactive interference have been key in demonstrating why it is related to intelligence (Bunting, 2006; van Moorselaar & Slagter, 2020). Proactive interference is previously learned information interfering with learning new information (Teague et al., 2011). Proactive interference is affected by how similar the new and old information is, the similarity of the learning contexts, and the amount of time between learning and retrieval (Bunting, 2006). The more similar the information and context are and the more time between learning and retrieval will result in more proactive interference and less accurate performance during retrieval. If the information and the context are very similar there will be more proactive interference and less accurate performance during retrieval. For example, switching the to-be recalled items from words to numbers (or vice-versa) during a complex span task improved accuracy because there was less proactive interference by making the items less similar (Bunting, 2006). Reducing the amount of proactive interference also reduced the correlation between performance on the complex span and intelligence (Bunting, 2006). The

relationship between time and proactive interference is more complex because it depends on the time between the stimuli and the time between the trials. As shown with the visual array task, the greatest effects of proactive interference are seen when the inter-trial interval (ITI) is short and the ISI is long (Shipstead & Engle, 2013). With respect to differences in intelligence, however, more intelligent individuals benefitted more from a longer ITI, presumably because they were better able to inhibit or remove irrelevant information thus there was less interference during retrieval (Shipstead & Engle, 2013). These results highlight how individual differences in combatting proactive interference are related to differences in intelligence.

This relationship between the attentional process of WM and intelligence has been demonstrated in multiple ways in people, but there is evidence for a similar relationship in mice. When mice are given a wide variety of learning tasks, there is a positive correlation in performance across tasks and one factor is extracted from this positive correlational matrix that can account for 22-43% of the variance in performance (Flaim & Blaisdell, 2020; Galsworthy et al., 2005; Kolata et al., 2007). How attention and WM are related to the general factor in mice has been investigated using the radial arm maze. In the radial arm maze, there is a central hub and *n* arms radiating out from the center. Some or all of these arms contain a food reward and errors can be categorized by subjects entering an arm and failing to obtain the food reward or re-entering an arm where the food reward has already been obtained. The radial arm maze assesses WM because animals need to update where they have already been and where they still need to go (Dudchenko, 2004). Performance on a 4 or 8 radial arm maze correlates with other cognitive measures in mice (Kolata et al., 2007; Locurto et al., 2006), but this task has been modified to more closely resemble the dual storage and processing demands seen in the complex span task. This modification is called the dual radial arm maze task, where two different radial mazes are

placed in the same room so there is overlap between the spatial cues used to navigate the maze. Subjects start navigating one of the mazes, but after they make 3 correct choices, they are taken out and put in the other maze. Once again, after making 3 correct choices in the second maze, they are placed back in the first maze. Subjects alternate between the mazes until all of the food rewards are obtained (Kolata et al., 2005). By alternating which maze mice are navigating and by having overlapping cues in each maze, mice need to maintain two similar lists of locations of where they have been and where they still need to go. Performance on this dual radial arm maze is positively correlated with performance on other cognitive tasks (Kolata et al., 2005). Follow up investigations have indicated that this relationship is primarily due to the ability to deal with interference and not the amount of information that needs to be stored or retrieved (Kolata et al., 2005, 2007; Matzel & Kolata, 2010).

So far, the relationship between intelligence and WM has been most heavily investigated in humans and mice, but the pigeon could be an additional model species to investigate this relationship. Memory processes have been heavily investigated in the pigeon with the delayed match to sample (DMTS) task (Anderson & Colombo, 2019; Kangas et al., 2011; Lind et al., 2015; Roberts, 1972; Zentall & Smith, 2016). In these experiments, pigeons are first shown a sample stimulus. After completing an observing response to the sample, there is a delay period where no stimuli are presented. Pigeons are then presented with one or more comparison stimuli, but the sample stimulus is not present (Figure 5.1). Subjects are reinforced for choosing the comparison that matches the sample. For example, if the sample is a red circle, the subject should select the red circle comparison, and not the green circle comparison stimulus. While multiple cognitive processes contribute to performance accuracy at choice (Zentall & Smith, 2016), manipulations of the delay length and ITI indicates that the ability to resist proactive

interference is important. Performance is less accurate with longer delay lengths and shorter ITIs (Hogan et al., 1981; Roberts & Kraemer, 1982), similar to the results of the visual array task given to humans (Shipstead & Engle, 2013). Another potential source of proactive interference is the stimulus set size. A small set size (for example two colors) would cause the most interference over a session while using unique stimuli every trial should cause the least amount of interference (Anderson & Colombo, 2019). This idea has been investigated with rhesus monkeys, where smaller set sizes require more cognitive effort, compared to larger set sizes (Basile & Hampton, 2013; Brown & Hampton, 2020), but stimulus set size and how interference might build over a session has rarely been investigated with this procedure in pigeons. One experiment in pigeons that used two colors as the stimulus set did not find evidence for proactive interference increasing over the session, rather they found that accuracy *improved* over the session (Edhouse & White, 1988). An important caveat to this result not addressed in the article is that the subjects had years of experience with the DMTS and may have learned how to resist proactive interference. Despite the wealth of information that has been obtained via the DMTS thus far, some additional properties should be assessed before it could be used to investigate intelligence and memory in the pigeon. The first is investigating the potential buildup of proactive interference within a session. Presumably, if proactive interference is accumulating, accuracy will actually be worse at the end of a session compared to the beginning of a session. If these results are found, the effect of practice should be investigated to determine if improvements in performance across sessions are partially due to learning how to resist proactive interference. The second is determining what amount of training and what measure of performance will be the most sensitive to individual differences in the DMTS task. While

previous research has reported individual subject data, it is not clear when the largest difference between subjects was observed (Kangas et al., 2011).

While the primary focus of this experiment was to determine if this task is appropriate to include in a test battery to assess general cognitive abilities, it was also possible that this task would be sensitive to age-related declines in performance. In human and nonhuman primates, older subjects show worse performance on the DMTS compared to younger subjects (Lamar & Resnick, 2004; Rodriguez & Paule, 2009). Further, performance on the DMTS relies on the prefrontal cortex (PFC), an area of the brain particularly sensitive to age in mammals (Bizon et al., 2012; Lamar & Resnick, 2004). In pigeons, performance on the DMTS relies on the avian equivalent of the PFC, the nidopallium caudolateral (NCL; Karakuyu et al., 2007), but it is not known if the NCL is affected by age in a similar way as mammals. How subjects perform on this task could be an indicator of underlying neurobiological changes as a function of age. Therefore, the age of the subject will also be investigated as a potential factor impacting performance. In the current task, pigeons were initially trained on a simultaneous MTS until they reached a predetermined criterion, then they were trained on the DMTS task with 0, 2, 4, and 8 second delay lengths for 30 sessions. The number of sessions to reach criterion during the simultaneous MTS was investigated. Performance, using percent correct and a log transformation (Kangas et al., 2011), was examined at the beginning (sessions 1, 2, and 3), middle (sessions 14, 15, and 16) and end (sessions 28, 29, and 30) of training. The effect of proactive interference was also examined at those time points. Similar to previous research, performance was least accurate at the longest delay length, but improved across sessions. Subjects varied in how much they improved with training and the most variability in performance across individuals was seen at the end of training with the log transformed data. This indicates that this would be appropriate to

include in a cognitive test battery for pigeons. Proactive interference, however, did not seem to impact performance over a session at any point in training. Additionally, the subject's age did not to impact performance. Implications for interpreting DMTS performance in relation to a general cognitive ability in pigeons is discussed.

Methods

Subjects

Eighteen pigeons served as subjects. Subjects ranged in age from 0.5-18 years old at the start of the experiment and there were 10 females. All subjects were trained to peck on the touchscreen and eat from the food hopper. All subjects had previous experience with other cognitive tasks, except for Athena. Subjects were individually housed in steel home cages with metal wire mesh floors in a vivarium. They were maintained at 80% of their free-feeding weight, but were allowed free access to water and grit while in their home cages. Testing occurred at approximately the midpoint of the light portion of the 12-hour light-dark cycle.

Apparatus

Testing was conducted in a flat-black Plexiglas chamber (38 cm wide x 36 cm deep x 38 cm high). All stimuli were presented by computer on a color LCD monitor (NEC MultiSync LCD1550M) visible through a 23.2 x 30.5 cm viewing window in the middle of the front panel of the chamber. The bottom edge of the viewing window is 13 cm above the chamber floor. Pecks to the monitor were detected by an infrared touchscreen (Carroll Touch, Elotouch Systems, Fremont, CA) mounted on the front panel. A custom-built food hopper (Pololu, Robotics and Electronics, Las Vegas, NV) was located in the center of the front panel, its access hole flush with the floor. The food hopper contained a mixture of leach grain pigeon pellets and

seed (Leach Grain and Milling). All experimental events were controlled and recorded with a Pentium III-class computer (Intel, Santa Clara, California). A video card controlled the monitor in the SVGA graphics mode (800 x 600 pixels). Stimuli were presented using the 3.6 version of Python with the psychopy toolbox, version 3.0.3 (Peirce, 2007).

Stimuli

The stimulus set consisted of two circular stimuli, 60 pixels in diameter. The stimuli could be a 1-pixel white outline filled with a red or green color. The background was gray during all phases of the trial and food reward and black during the ITI.

Procedure

Autoshaping and instrumental training.

Each subject received one session per day, five days per week. Each session terminated after the completion of 96 trials or 90 minutes had elapsed, whichever came first. The number of trials and time to complete the session were consistent throughout all phases of the experiment. The stimuli were consistently presented in three locations, arranged in a triangular formation (Figure 5.1). The sample was shown in the center location and the comparison stimuli were offset to the left and right of the midline below the sample, serving as the left and right comparisons respectively. If a stimulus was not presented during a trial, the location was marked by a white circular outline.

Pigeons were initially trained with a mixed autoshaping and instrumental procedure. The red and green stimuli appeared in the sample, left comparison, or right comparison position an equal number of times. Only one stimulus was presented at a time and the other locations were marked with a white outline (Figure 5.1). During the first 48 trials, the stimulus was presented

for 10 s. If the pigeon pecked on the stimulus (FR1) the trial would end, then the food port was illuminated and the hopper was raised for 3 s (food delivery was 3 s throughout the entire experiment). Pecks within 25 pixels of the stimulus were considered on-target. If the pigeon did not peck on the stimulus, the food reward would still be delivered after 10 s. Pecks to the background or where the locations were marked by a white outline were neither reinforced nor punished. After the food delivery was terminated, there was a 13-s ITI with a black screen. During the last 48 trials, the stimulus would stay on the screen until the pigeon completed the FR1 peck requirement to the stimulus. When pigeons were consistently pecking at the stimulus (pecking on the stimulus on 80% of the trials for 2 consecutive sessions), the autoshaping procedure was discontinued and an instrumental contingency was enforced. During the instrumental procedure the stimulus would stay on the screen until the pigeon completed the peck requirement. The pigeon was trained with the FR1 until they reached criterion, finishing the session within 120 minutes on two consecutive sessions. Then the peck requirement was gradually increased from an FR1 to an FR10 using a series of VR schedules, starting with VR3 +/- 2 (actual values 1, 2, 3, 4, 5), VR6 +/- 2 (4, 5, 6, 7, 8), then VR9 +/- 2 (7, 8, 9, 10, 11). Subjects had to reach criterion on each VR schedule before advancing to the next. When subjects had reach criterion on the VR9 schedule, the number of trials that could be followed with reinforcement was reduced to 72 (75% of trials). Each stimulus in each location was presented without reinforcement once per session, but never in the first or last block of 24 trials. When subjects reached criterion on this reduced reinforcement schedule, subjects began the simultaneous MTS task.

Simultaneous match to sample

During the simultaneous MTS, each trial had two phases, a sample phase then a choice phase. During the sample phase, a stimulus was only presented in the sample location, while the comparisons were marked with a white outline (Figure 5.1a). Once subjects completed the observing response to the sample stimulus (FR10), the choice phase began. During the choice phase, the sample stimulus remained on the screen and the comparison stimuli were presented. If the subject pecked (FR1) the comparison that matched the sample, they received a food reward and then the ITI would begin. If they pecked the comparison that did not match the sample the trial would end, the ITI would begin, and the trial would be repeated starting at the sample phase (correction procedure). Correction trials were not used in the data analysis. During the choice phase pecks to the sample or background were neither reinforced nor punished. Subjects had an unlimited amount of time to complete the peck requirement during the sample and choice phases. The correct comparison stimulus was presented equally often as the left or right comparison. Red and green were presented as the correct comparison an equal number of times. This resulted in four unique stimulus configurations. Subjects experienced each stimulus configuration 24 times per session for a total of 96 trials. Subjects trained on the simultaneous MTS until they were 80% accurate on two consecutive sessions. Subjects were then trained on the DMTS.

Delayed match to sample

During the DMTS, each trial had three phases, a sample, delay, and choice phase. Similar to the simultaneous MTS, during the sample phase a stimulus was only presented in the sample location. Once subjects completed the observing response to the sample, the delay phase began. During the delay phase, no stimuli were presented on the screen, but the locations were marked (Figure 5.1b). The delay could be 0, 2, 4, or 8 s long. When the delay had elapsed, only the

comparison stimuli were presented, the sample stimulus was no longer presented with the comparisons (Figure 5.1c). If subjects pecked the comparison that matched the sample, they received a food reward before the ITI began. If subjects pecked the comparison that did not match the sample, the trial ended and the ITI began. The correction procedure was discontinued. Each stimulus configuration was presented with each delay length an equal number of times. Subjects experienced each stimulus configuration with each delay six times per session for a total of 96 trials. Subjects trained on the DMTS for 30 sessions then they received four transfer sessions.

Data Analysis

Sessions were only included in the analysis if the subject completed all 96 trials. During the simultaneous MTS, one session was excluded for Waluigi, Wario, and Dickinson, and 11 sessions were excluded for Darwin. During the DMTS, one session was excluded for Athena, Shy guy, Estelle, Durrell, Jubilee, Wenchang, and Herriot (n = 7). Two sessions were excluded for Waluigi and four sessions were excluded for Darwin. Correction trials during the simultaneous MTS were not analyzed. The number of sessions to reach criterion during the simultaneous MTS was used a potential measure of interest.

Performance was analyzed using percent correct and a log transformation of the data, log $d_t = ½ \log ([c_r/e_r]*[c_g/e_g])$. The log transformation was performed for each retention interval (t) and is the geometric mean of the ratios of correct (c) and incorrect (e) responses to the red (r) and green (g) stimuli respectively. A log transformation was used because it is free from the response bias and the range of values can extend beyond 1, unlike percent correct (Kangas et al., 2011). This result is a wider range of values that is better at differentiating between subjects, even when performance is very high. Performance was examined at the beginning (sessions 1, 2,

and 3) middle (sessions 14, 15, and 16), and end of training (sessions 28, 29, and 30) for each delay length. This means that performance was averaged over 72 trials for each delay. Performance was compared across the different amount of training at each delay length to determine if performance improved over time. For example, performance with a 0 s delay was compared at the beginning, middle, and end of training. The role of proactive interference was analyzed by examining performance within the session at each point in training for each delay length. Specifically, at each point in training, trials were categorized as being in the beginning, middle, or end of each session for each duration. Data were analyzed using SPSS version 27.

Results

Simultaneous Match to Sample

Subjects needed 2-7 sessions of training before reaching criterion on the simultaneous MTS task. The average number of sessions needed was 3.5 and the standard deviation was 1.04. Due to the low amount of variability across subjects it was not possible to analyze this further.

Delayed Match to Sample

Performance over training.

Using accuracy as the dependent measure, performance was consistently the best at the 0 s delay, but performance improved with training on all delay lengths (Figure 5.2). A two-way repeated measures ANOVA was used to investigate if there were differences in performance at each delay length in the beginning, middle, and end of training. There was a main effect of delay (F (3, 51) = 163.23, $p < .001$, partial eta squared = .906), a main effect of the amount of training (F (2, 34) = 47.3, $p < .001$, partial eta squared = .736), and a significant interaction (F (6, 102) = 2.97, $p = .01$, partial eta squared = .149). Post hoc tests with a Bonferroni correction indicated

that there were significant differences in performance at all delays between all amounts of training, except when comparing performance at the beginning and middle of training in the 0 s delay condition and when comparing performance at the middle and end of training in the 8 s delay condition (Table 5.1).

Performance was also investigated on log-transformed data. Similar to the accuracy data, performance was always the best at the 0-s delay, but performance across all delays improved with training (Figure 5.3). The same analyses as described above were used to determine if there were differences in performance across the different delay conditions and amount of training. The delay performance failed Mauchly's test of sphericity (x^2 (5) = 26.71, $p < .001$), as did the interaction between delay and amount of training (x^2 (20) = 50.09, $p < .001$), so a Greenhouse-Geisser correction was used. Similar to the previous analysis, there was a main effect of delay (F (1.64, 27.93) = 112.78, $p < .001$, partial eta squared = .869) and amount of training (F (2, 34) = 35, $p < .001$, partial eta squared = .672), and a significant interaction (F (3.52, 59.79) = 2.72, p = .044, partial eta squared = .138). Post hoc tests with a Bonferroni correction indicated that there were significant differences in performance at all delays between all amounts of training, except when comparing performance at the middle and end of training in the 0 s and 8 s delay condition (Table 5.1).

While there was a main effect of subjects improving over training, individual differences became *more* pronounced by the end of training compared to the beginning (Figure 5.2, 5.3). Since the primary focus of this experiment was to determine if there was sufficient individual variability in the task to include in a cognitive test battery, this increase in variability is directly relevant to our goals. Unfortunately, the most appropriate statistical test available to investigate variability for repeated measures data is not robust against deviations from normality (Derrick et

al., 2018). Analyses for independent observations, however, are more robust, so the data at the beginning and end of training were investigated as independent groups using Levene's test for equality of variances (Table 5.2). For the accuracy data, performance on the 4 and 8 s delay condition had significantly different variances, while for the log transformed data there were significant differences in the 2, 4, and 8 s delay conditions, indicating that variance was significantly larger at the end of training at these delay lengths.

Consistency of performance.

While performance at all delay lengths improved with training for all subjects, it was not clear if subjects were consistent over these conditions. For example, if a subject performed well on during the 8 s delay at the beginning of training, do they also perform well during the 2 s delay at the end of training? To investigate this possibility, a Spearman's correlation was conducted on performance during all delay conditions at all points in training for the accuracy and log transformed data. The correlation matrix was almost identical across the data types (Table 5.3, 5.4). The correlations were almost uniformly positive and over half were significant. The mean correlation for the accuracy data was .55 and, for the log transformed data, the mean correlation was .54. While there were significant positive correlations for performance at the beginning of training, the significant positive correlations were primarily found between the middle and end of training. These results indicate that performance was consistent across the delay conditions and training.

Proactive interference.

The role of proactive interference was investigated by comparing accuracy at each delay length at the beginning, middle, and end of each session in each point in training. While no changes in accuracy within a session were detected by a visual inspection of the data, this was

confirmed with a three-way repeated measures ANOVA. To avoid redundancy with the previous sections, the main effects and interaction between the delay condition and amount of training are not reported here. There was no main effect of when in the session performance was measured (F (2, 34) < 1). There was no interaction with amount of training (F (4, 68) = 1.25, p = .299), with delay (F (6, 102) = 2.05, p = .066), or with amount of training and delay (F (12, 204) = 1.07, p = .385).

Age effects.

The individual subject data in Figures 5.2 and 5.3 were organized by age using a gradient, with younger subjects in darker shades and older subjects in lighter shade. There were no obvious age-related effects, but to further investigate this possibility, a Spearman correlation between age of the subject in years and performance at each delay over different amounts of training was conducted for the accuracy (Table 5.3) and log-transformed data (Table 5.4). Correlations were close to zero, except for accuracy data during the 0 s delay condition in the middle of training, which had a significant *positive* correlation (r_s (16) = .496, p = .036). This indicates that, generally, age did not impact performance and when it did, older subjects had better performance than younger subjects.

To better visualize performance and age, subjects were divided into two groups. Subjects younger than 4 years old were in the 'young' group (n = 9) and subjects older than 11 years old were in the 'old' group (n = 9). Performance across the groups almost completely overlapped, though the old group tended to outperform the young group (Figure 5.4). A 2x4x3 mixed ANOVA, with age group as the between subject factor and delay and amount of training as the within subject factors, was used to further investigate age on the accuracy and log transformed data. To avoid redundancy, only the age group results are reported. For the accuracy data, there

was no main effect (F (1, 16) < 1), or interaction between age and delay (F (3, 48) = 1.27, p = .294), age and amount of training (F (2, 32) = 1.82, p = .179), or between all three factors (F (6, 96) < 1). The results for the log transformed data were virtually identical.

Discussion

The goal of this experiment was to investigate individual differences in performance on the DMTS to determine if it would be appropriate to include in a cognitive test battery for pigeons. Before subjects trained on the DMTS, they were initially trained on the simultaneous MTS. The simultaneous MTS could track variation in learning stimulus configurations, but there was not enough variability across subjects in the number of sessions to reach criterion to detect individual differences in cognition. This may have been due to the correction procedure, where the trial would repeat until subjects pecked the correct comparison. This correction procedure was used to reduce side biases (consistently pecking the left or right key; Kangas et al., 2011), but may have been at the expense of reducing individual differences.

To investigate the DMTS, performance was analyzed using accuracy and a log transformation at each delay length after different amounts of training. The results replicated previous research where performance was always progressively worse with longer delays between the sample and comparison, but performance significantly improved with training (Figure 5.2, 5.3; Kangas et al., 2011). This indicates that this procedure is effective at improving performance, but not all subjects improved equally. When accuracy data were used, variance was significantly higher in the 4 and 8-sec delay conditions at the end of training compared to the beginning. When the log transformed data were used, variance was significantly higher in the 2, 4, and 8-sec delay conditions (Table 5.2). This variability indicates that this procedure and training length is sensitive enough to detect individual differences in performance at all delay

lengths and could be used in a cognitive test battery for pigeons. Despite the increase in variability, performance was positively correlated across the different training points, indicating that subjects were consistent across training. This means the task captures both between subject variability and within subject reliability.

While the DMTS is sensitive to individual differences, it is crucial to understand why there are differences across subjects. The ability to combat proactive interference was investigated as a potential mechanism causing individual differences in subject performance. As described in the introduction, proactive interference is related to WM and intelligence in humans and mice, but there is no compelling evidence that it impacts performance on the DMTS for pigeons. If proactive interference was impacting performance, then accuracy would have been lower later in the session after it had time to build up (Bunting, 2006), especially with a small set size (Anderson & Colombo, 2019). Yet, performance on the delays did not show any significant differences in performance across the session.

Even though proactive interference does not seem to impact performance, the DMTS would still be a valuable addition to a cognitive test battery for pigeons because it shows within subject reliability and between subject variability and has long experimental history. While any speculations on how a general cognitive factor differs in pigeons compared to other species are premature, these results suggest interesting follow up experiments to better understand memory in the pigeon. Memory is not a unitary cognitive ability, and even though the DMTS is used extensively, there are other tasks that also assess memory performance (Wright et al., 2010; Spetch &Edwards, 1986). Administering a more specific memory test battery could be helpful in understanding what aspects of memory are shared across tasks and which are unique (Shaw & Schmelz, 2017). This would enhance our understanding of the processes underlying memory,

similar to the memory research conducted in humans (Conway et al., 2003; Shipstead & Engle, 2013). Age was also investigated as a potential factor that could impact performance since previous research with mammals has shown age related impairments (Lamar & Resnick, 2004; Rodriguez & Paule, 2009). Surprisingly there were no differences in performance based on subject age for any delay length at any point in training (Figure 5.4). While it is outside of the scope of this paper, these results highlight an important difference in avian and mammalian aging that should be investigated further. For the cognitive test battery, these results indicate that this task could be used for subjects of any age.

Overall, a DMTS procedure that administers 30 sessions of training for delays of 0, 2, 4, and 8 s with red and green stimuli, is appropriate to include in a cognitive test battery for pigeons. Performance was distinct at each delay length, but there was sufficient variability across subjects to detect individual differences, particularly with a log transformation. Even though there was variability across subjects, this was not due to proactive interference, which could indicate an important difference in procedures used in humans and mice compared to pigeons.

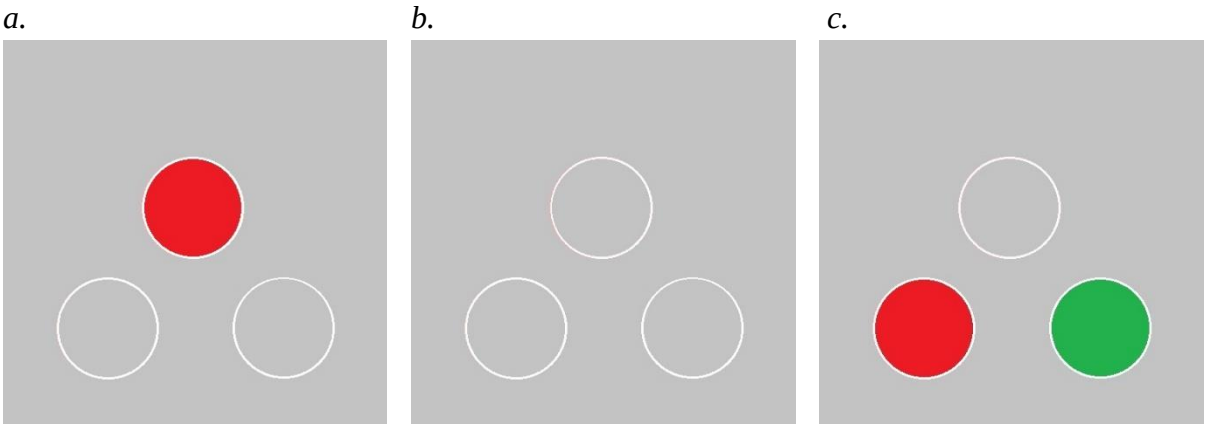


Figure 5.1. An example of a trial during the delay match to sample procedure. Panel a depicts the sample phase, when only one stimulus is presented, panel b depicts the delay phase when no stimuli are presented, and panel c depicts the choice phase when the two comparison stimuli are presented. In this example, subjects should choose the red comparison since it matches the sample.

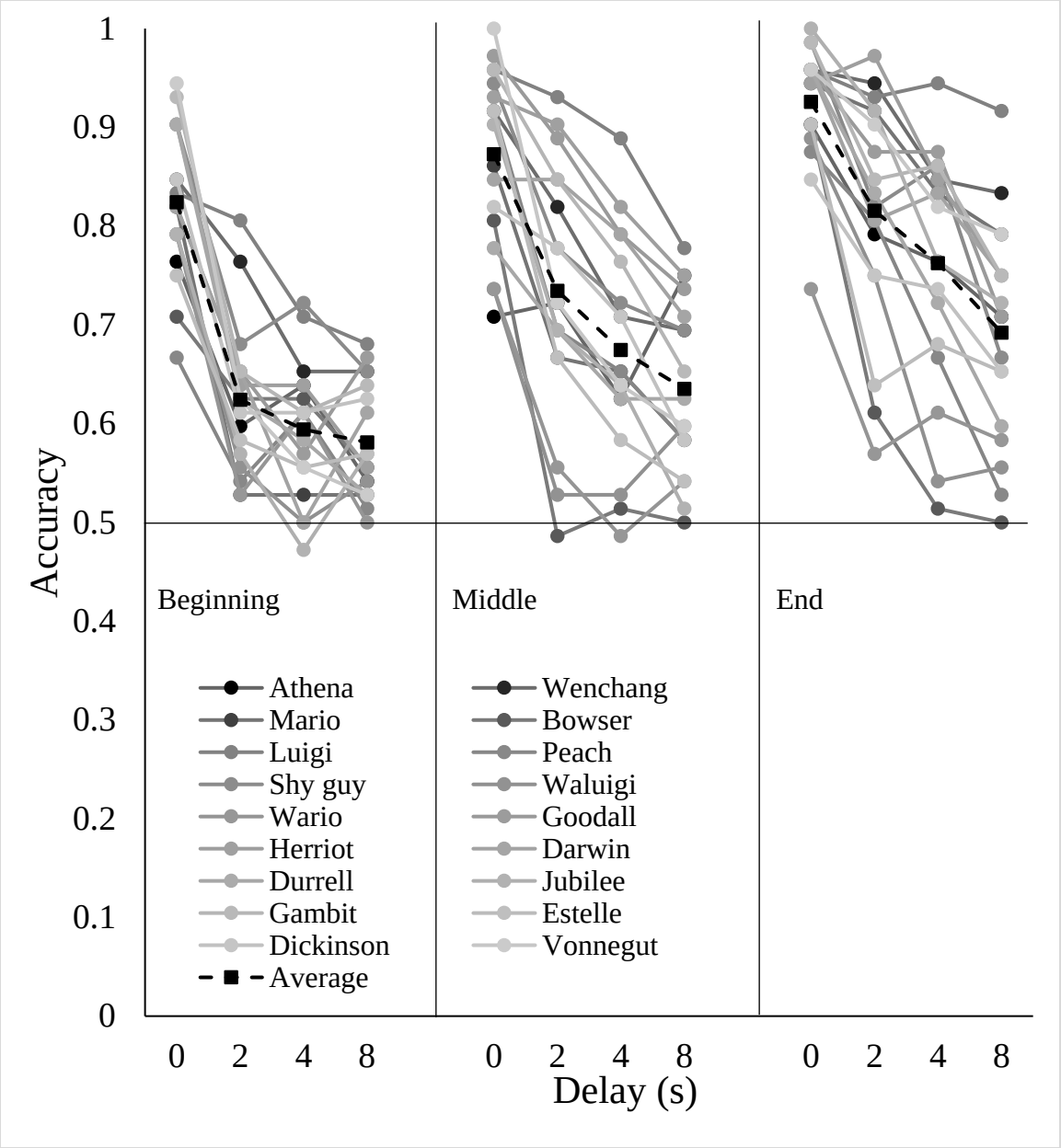


Figure 5.2. Accuracy of performance at each delay length at the beginning, middle, and end of training. The horizontal line indicates chance performance. Data were organized by age, where younger subjects are represented by darker shades and older subjects are represented by lighter shades.

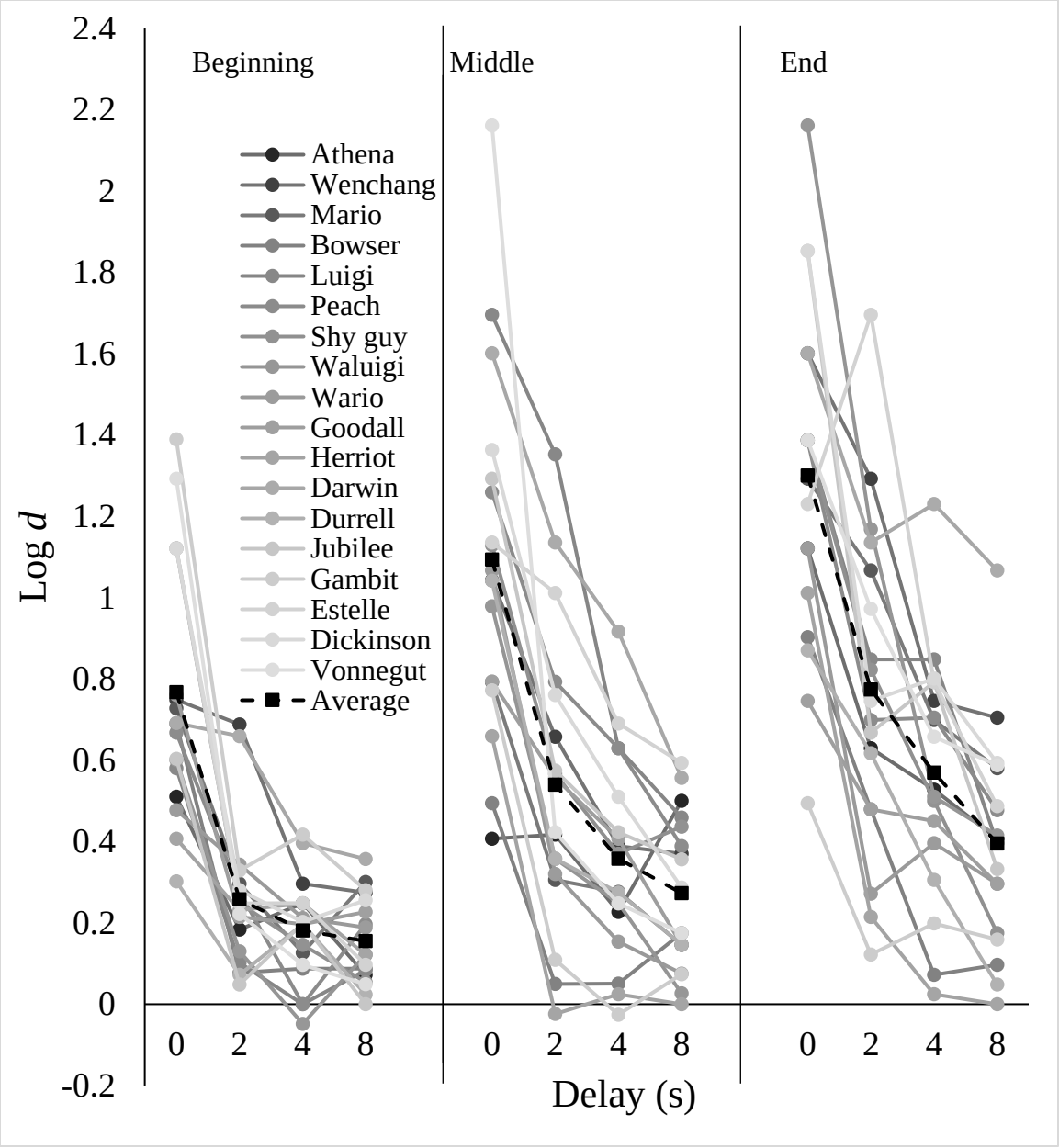
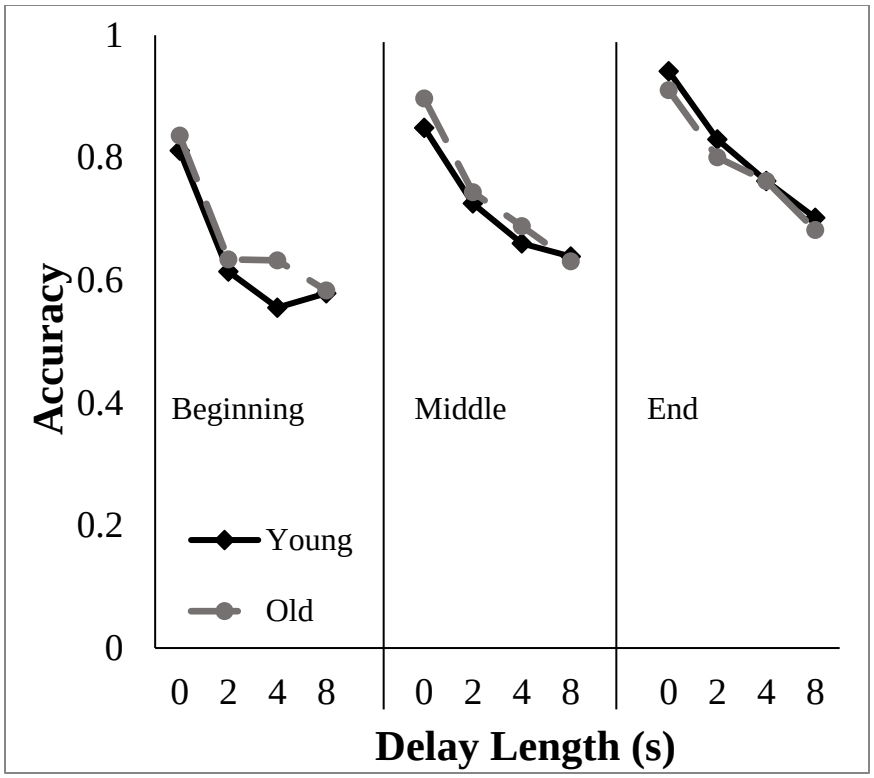


Figure 5.3. Log transformed data of performance at each delay length at the beginning, middle, and end of training. Data were organized by age, where younger subjects are represented by darker shades and older subjects are represented by lighter shades.

a.



b.

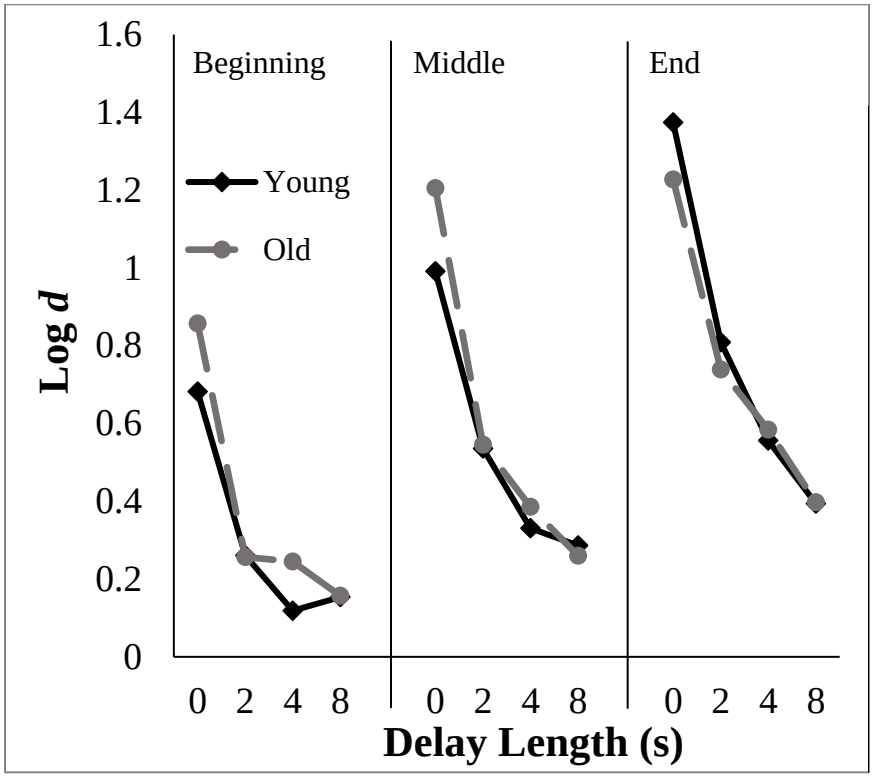


Figure 5.4. Accuracy and log data for subjects based on age at each point in training for each delay length

Post hoc Analyses with a Bonferroni Correction							
		Point in Training			Comparisons		
	Delay (s)	Beginning	Middle	End	Beginning and Middle	Beginning and End	Middle and End
Accuracy	0	(*M* = 0.82, *SD* = 0.08)	(*M* = 0.87, *SD* = 0.09)	(*M* = 0.93, *SD* = 0.06)	*p* = .123	**p < .001**	**p = .019**
	2	(*M* = 0.62, *SD* = 0.07)	(*M* = 0.73, *SD* = 0.13)	(*M* = 0.82, *SD* = 0.12)	**p < .001**	**p < .001**	**p = .006**
	4	(*M* = 0.59, *SD* = 0.07)	(*M* = 0.67, *SD* = 0.11)	(*M* = 0.76, *SD* = 0.12)	**p = .019**	**p < .001**	**p < .001**
	8	(*M* = 0.58, *SD* = 0.06)	(*M* = 0.64, *SD* = 0.09)	(*M* = 0.69, *SD* = 0.11)	**p = .019**	**p < .001**	*p* = .052
Log Transformed	0	(*M* = 0.77, *SD* = 0.31)	(*M* = 1.1, *SD* = 0.44)	(*M* = 1.3, *SD* = 0.42)	**p = .004**	**p > .001**	*p* = .21
	2	(*M* = 0.26, *SD* = 0.18)	(*M* = 0.54, *SD* = 0.37)	(*M* = 0.77, *SD* = 0.4)	**p = .004**	**p > .001**	**p = .037**
	4	(*M* = 0.18, *SD* = 0.13)	(*M* = 0.36, *SD* = 0.25)	(*M* = 0.57, *SD* = 0.3)	**p = .021**	**p > .001**	**p > .001**
	8	(*M* = 0.16, *SD* = 0.11)	(*M* = 0.27, *SD* = 0.19)	(*M* = 0.4, *SD* = 0.26)	**p = .036**	**p = .001**	*p* = .112

Table 5.1. Post hoc analyses comparing differences in performance at each delay length during each point in training. Bolded p values indicate significant differences

Levene's Test for Equality of Variances							
		Accuracy			Log Transformed		
Delay (s)	Based On	*F*	DF	*p*	*F*	DF	*p*
0	Mean	0.51	(1, 34)	0.48	1.092	(1, 34)	0.303
	Median	0.708	(1, 34)	0.406	1.338	(1, 34)	0.256
	Median with adjusted df	0.708	(1, 33.7)	0.406	1.338	(1, 32.86)	0.256
	Trimmed Mean	0.659	(1, 34)	0.423	1.177	(1, 34)	0.286
2	Mean	3.025	(1, 34)	0.091	9.018	(1, 34)	**0.005**
	Median	2.887	(1, 34)	0.098	8.301	(1, 34)	**0.007**
	Median with adjusted df	2.887	(1, 30.15)	0.1	8.301	(1, 25.35)	**0.008**
	Trimmed Mean	2.84	(1, 34)	0.101	8.835	(1, 34)	**0.005**
4	Mean	5.055	(1, 34)	**0.031**	9.83	(1, 34)	**0.004**
	Median	4.376	(1, 34)	**0.044**	9.899	(1, 34)	**0.003**
	Median with adjusted df	4.376	(1, 28.08)	**0.046**	9.899	(1, 24.24)	**0.004**
	Trimmed Mean	4.899	(1, 34)	**0.034**	9.813	(1, 34)	**0.004**
8	Mean	5.675	(1, 34)	**0.023**	6.196	(1, 34)	**0.018**
	Median	4.633	(1, 34)	**0.039**	6.244	(1, 34)	**0.017**
	Median with adjusted df	4.633	(1, 25.87)	**0.041**	6.244	(1, 21.88)	**0.02**
	Trimmed Mean	5.816	(1, 34)	**0.021**	6.524	(1, 34)	**0.015**

Table 5.2. Comparing the variance of performance at the beginning and end of training. Bolded values indicate a significant difference

Accuracy		Beginning				Middle				End			
Delay (s)		0	2	4	8	0	2	4	8	0	2	4	8
0	r_s	--											
	p	.											
2	r_s	**0.567**	--										
	p	0.014	.										
4	r_s	0.199	**0.516**	--									
	p	0.428	0.028	.									
8	r_s	0.295	**0.801**	0.323	--								
	p	0.235	>.001	0.19	.								
0	r_s	**0.502**	**0.56**	0.162	**0.499**	--							
	p	0.034	0.016	0.521	0.035	.							
2	r_s	**0.579**	**0.779**	0.382	**0.681**	**0.62**	--						
	p	0.012	>.001	0.118	0.002	0.006	.						
4	r_s	**0.516**	**0.697**	0.286	**0.669**	**0.657**	**0.931**	--					
	p	0.029	0.001	0.25	0.002	0.003	>.001	.					
8	r_s	0.445	**0.678**	0.437	0.467	0.324	**0.817**	**0.711**	--				
	p	0.064	0.002	0.07	0.051	0.189	>.001	0.001	.				
0	r_s	**0.549**	**0.621**	-0.017	**0.531**	**0.568**	**0.494**	**0.494**	0.337	--			
	p	0.018	0.006	0.947	0.023	0.014	0.037	0.037	0.172	.			
2	r_s	**0.555**	0.461	0.129	0.334	**0.602**	**0.629**	**0.671**	**0.505**	**0.64**	--		
	p	0.017	0.054	0.61	0.175	0.008	0.005	0.002	0.033	0.004	.		
4	r_s	**0.613**	**0.737**	0.325	**0.702**	**0.715**	**0.873**	**0.875**	**0.736**	**0.69**	**0.741**	--	
	p	0.007	>.001	0.188	0.001	0.001	>.001	>.001	>.001	0.002	>.001	.	
8	r_s	**0.56**	**0.508**	0.094	0.431	**0.546**	**0.663**	**0.646**	**0.544**	**0.581**	**0.807**	**0.796**	--
	p	0.016	0.031	0.711	0.074	0.019	0.003	0.004	0.019	0.012	>.001	>.001	.

Age	r_s	0.371	0.084	0.236	-0.062	**0.496**	0.153	0.169	-0.118	0.069	0.034	0.111	0.007
	p	0.13	0.742	0.345	0.807	0.036	0.543	0.503	0.642	0.786	0.893	0.662	0.977

Table 5.3. Correlation matrix of performance over each point in training at each delay length for the accuracy data.

Log		Beginning				Middle				End			
Delay (s)		0	2	4	8	0	2	4	8	0	2	4	8
0	r_s	--											
	p	.											
2	r_s	0.423	--										
	p	0.08	.										
4	r_s	0.178	**0.529**	--									
	p	0.479	0.024	.									
8	r_s	0.337	**0.787**	0.355	--								
	p	0.172	>.001	0.148	.								
0	r_s	**0.672**	**0.635**	0.12	**0.476**	--							
	p	0.002	0.005	0.636	0.046	.							
2	r_s	**0.618**	**0.663**	0.26	**0.629**	**0.779**	--						
	p	0.006	0.003	0.298	0.005	>.001	.						
4	r_s	**0.581**	**0.567**	0.214	**0.667**	**0.694**	**0.927**	--					
	p	0.011	0.014	0.393	0.002	0.001	>.001	.					
8	r_s	0.433	**0.472**	0.308	0.3	**0.473**	**0.799**	**0.702**	--				
	p	0.073	0.048	0.213	0.227	0.047	>.001	0.001	.				
0	r_s	**0.557**	**0.591**	0.077	**0.574**	**0.625**	**0.561**	**0.529**	0.326	--			
	p	0.016	0.01	0.76	0.013	0.006	0.015	0.024	0.187	.			
2	r_s	**0.517**	0.33	0.001	0.31	**0.501**	**0.606**	**0.623**	**0.519**	**0.692**	--		
	p	0.028	0.181	0.997	0.211	0.034	0.008	0.006	0.027	0.001	.		
4	r_s	**0.642**	**0.618**	0.282	**0.662**	**0.742**	**0.885**	**0.863**	**0.711**	**0.713**	**0.728**	--	
	p	0.004	0.006	0.257	0.003	>.001	>.001	>.001	0.001	0.001	0.001	.	
8	r_s	**0.542**	0.467	0.147	0.457	**0.549**	**0.619**	**0.597**	**0.474**	**0.677**	**0.787**	**0.818**	--
	p	0.02	0.051	0.56	0.056	0.018	0.006	0.009	0.047	0.002	>.001	>.001	.

Age	r_s	0.363	-0.052	0.252	-0.069	0.411	0.084	0.115	-0.143	0.035	-0.048	0.113	0.069
	p	0.139	0.837	0.312	0.785	0.09	0.741	0.649	0.57	0.889	0.851	0.655	0.785

Table 5.4. Correlation matrix of performance over each point in training at each delay length for the log transformed data.

Chapter 6: Choice Reaction Time

Abstract

How quickly a person can detect a change in their environment or make a decision about simple stimuli is consistently related to performance on intelligence tests, where more intelligent people are consistently faster. This relationship between reaction time (RT) and intelligence has been replicated with numerous tasks. Recently, a variety of species have shown evidence for a general cognitive factor. A key similarity between the general cognitive factor in animals and intelligence in humans is consistent performance across a variety of cognitive tasks. This indicates that intelligence has similar properties across species. Replicating the speed and intelligence relationship in animals would provide additional evidence that intelligence has similar features across species. Yet, measures of speed are rarely included when investigating general cognitive abilities in animals. The goal of these experiments was to create a procedure to assess RT in pigeons using a touchscreen in a way that was similar to previous research with humans. The task was based on Hick's law, for which RT increases as the number of binary choices increases. While other research has shown that pigeons conform to Hick's law, our procedures failed to replicate this effect. In Experiment 1, subjects showed an 'anti-Hick's' effect that was an artifact of stimulus location on the monitor. After controlling for location, RT did not increase with the number of choices. Potential reasons why these procedures did not produce a Hick's effect and how the results are still relevant when investigating general cognitive abilities in animals are discussed.

Introduction

When people are given a diverse battery of cognitive tasks, there will be differences in performance across people, but there will be consistent performance across tasks, such that if a

person performs well in one task, they are likely to perform well in another (Deary, 2000; Spearman, 1904). This pattern of between-subject variability and within-subject reliability results in a positive correlation matrix. Using dimension reduction techniques, like factor analysis or principal component analysis, on this positive correlation matrix will extract one factor that can account for approximately half of the variance in performance across people (Carroll, 1993; Deary, 2000). This factor is referred to as *g* since it is related to almost all cognitive abilities (Deary, 2000; Spearman, 1904). A *g* factor has been extracted with different intelligence test batteries in different cultures and is one of the most well-replicated findings (Johnson et al., 2004; 2008). Despite how consistently *g* is replicated, we still do not understand why performance positively correlates across tasks (Conway & Kovacs, 2015). This is partially because of restrictions on the types of causal manipulations that can be administered to people. Animal models could be used to investigate potential causes of *g* that would not be ethical or feasible to investigate in people. First, however, it needs to be determined that animals have a similar general cognitive ability. So far, when many species are given a diverse cognitive test battery, performance positively correlates across tasks and one factor is extracted (Ashton et al., 2018; Flaim & Blaisdell, 2020; Isden et al., 2013; Kolata et al., 2008; Shaw et al., 2015). The initial results are promising, but a similar pattern of results does not necessarily indicate the same causal factor. One of the issues is the difference in cognitive test batteries applied to humans compared to those applied to nonhuman animals (hereafter just "animals").

While there are similarities in cognitive test batteries across species, for example they almost always include a learning and memory task, only human test batteries assess *processing speed* (Carroll, 1993; Deary, 2000; Flaim & Blaisdell, 2020). Processing speed is a broad term used to generally describe how quickly participants can react to a stimulus or make a decision.

While many tasks could assess processing speed as described here, the tasks generally fall into five categories, reaction time (RT), general speed of information-processing, speed of short-term memory processing, speed of long-term memory retrieval, and inspection time (Sheppard & Vernon, 2008). For example, the Posner letter matching task assesses speed of long-term memory retrieval by asking participants to quickly decide whether two sets of letters match physically (AA is a match while Aa is not) or semantically (AA and Aa are a match; Posner & Mitchell, 1967). While each speed task has unique aspects, there is evidence for a domain general speed factor that is related to *g* and intelligence (Carroll, 1993; Neubauer & Bucik, 1996; Schubert et al., 2015). People who perform well on intelligence tests are consistently faster on many different types of speeded tasks. The average correlation between performance on each speed task and intelligence tests is modest (r = -.24) but consistent (Deary, 2000; Sheppard & Vernon, 2008). The reason for this correlation is unclear. Other cognitive mechanisms, like attention (Longstreth, 1984; Stankov & Roberts, 1997), and biological mechanisms, like neural processing speed (Schubert et al., 2019) have been theorized to explain the correlation between speed and intelligence, but so far no one explanation has received unequivocal support. Even though the underlying mechanism responsible for the speed and intelligence relationship has yet to be identified, the consistency of the correlation indicates a key property that needs to be replicated in cognitive test batteries given to animals. Choice RT tasks based on Hick's Law, that RT will increase linearly as the number of binary choices increases, utilize relatively simple stimuli and instructions (Hick, 1952), that would be ideal for adapting for use with animals.

In one variation of a Hick's RT task, there is a center 'home' button and an array of lights or potential targets (PTs) that varies in number from 1, 2, 4, and 8. Each array represents 0, 1, 2, and 3 bits of information (Jensen, 1982; Sheppard & Vernon, 2008; Widman & Carlson, 1989).

Participants must rest their finger on the center home button until they hear a brief auditory cue. That indicates that in 1-4 seconds, one of the stimuli in the array will light up or change, becoming the target. Then participants remove their finger from the home key and touch the target or a button below the target as quickly as possible (Jensen, 1982; Widman & Carlson, 1989). Despite the simplicity of the procedure, a variety of dependent measures from the task have been investigated with respect to intelligence. The most obvious is the bit condition and RT. Even in the 0-bit condition, when there is only one PT present, there is negative correlation between RT and intelligence, meaning that faster participants perform better on intelligence tests, and this correlation tends to get stronger across the bit conditions, that is, with increases in information (Sheppard & Vernon, 2008). The variability in RT, often reported as the standard deviation, also correlates with intelligence, where more consistent participants perform better on intelligence tests (Doebler & Scheffler, 2016). The slope of the RT as the bit condition increases has also been investigated. For many participants, RTs increase as predicted by Hick's law (Neubauer, 1991). Some researchers have found that more intelligent participants have a shallower slope, meaning there is a smaller increase in the RT as the bits of information increase (Jensen, 1982), but this has not always been replicated (Longstreth, 1984; Widaman & Carlson, 1989). As already mentioned, the simplicity of the task and variety of dependent measures that can be obtained, make this an ideal task to modify for animals. One modification for pigeons has already been successful, but so far it is the only one.

Vickrey and Neuringer (2000) developed a touchscreen version of the Hick's RT task for pigeons. They found that the RT of pigeons increased with number of choices as predicted by Hick's Law, but the slope was much shallower compared to humans given a similar version of the task. While their version of the task was effective at revealing a Hick's effect, there are some

procedural differences that could impact performance. First, in the pigeon procedure, as soon as the subject pecked the home key, one of the PTs would become the target. This means there was no uncertainty about *when* the target would appear, unlike in the procedures usually used with humans. Previous research in humans has shown that RTs are slower when there is a variable amount of time between the 'warning' stimulus and the onset of the target (Broadbent & Gregory, 1965). When the interval between the warning and target is short and constant, RT could reflect participants learning the timing instead of detecting the target (Crabtree & Antrim, 1988). Additionally, some research indicates that differences in the ability to sustain attention during the interval between the warning stimulus and target could influence the correlation between RT and intelligence (Carlson et al., 1983). Second, the PTs could appear above or below the home key, resulting in a full circle of potential locations. This is a wider spatial range for the target and more potential locations than typically seen in human studies. Previous research in humans indicated a positive trend between RT and distance between stimuli (Widman & Carlson, 1989). While these effects were not significant, the potential distance between stimuli is larger in the Vickrey and Neuringer (2000) paradigm, which could have a larger impact on performance. These two procedural changes could have opposing effects on RT in humans, where the predictable arrival of the target could make participants faster, while the spatial arrangement could make them slower. It is unclear if these differences would have a similar effect on pigeon RTs since the properties of pigeon RTs are not well established. Administering variations of the Hick's task would be helpful for understanding which parameters will impact RT and if the Hick's effect is robust in pigeons.

The remaining procedural differences would make it difficult to incorporate into a larger cognitive test battery due to the length of training. In the Vickrey and Neuringer (2000)

procedure, subjects had to peck a reinforcing star after pecking the target. Training a sequence of pecks (home key, target, then reinforcing star) could extend the initial instrumental training needed for subjects to perform the task. Similarly, pigeons received extensive training with their modified Hick's procedure, a marked contrast from the human intelligence literature. Human participants receive 5-10 practice trials and 15-30 experimental trials at each bit-condition during a single session for a total of 64-160 trials (Carlson & Jensen, 1982; Widman & Carlson, 1989). Pigeons received at least 64 practice trials, though the exact number is difficult to determine, and 16-32 experimental trials at each bit condition for 25 sessions for a total of 2,496 experimental trials. While animal subjects will always need more training since they cannot be given verbal instructions like humans (Zentall, 1997), it is not clear if such extensive training is necessary to compare the results from pigeons to humans. If the initial training and experimental procedure can be completed in a shorter amount of time, it would be easier to include a modified Hick's task in a larger test battery.

The goal of these experiments was to create a touchscreen version of the task that was more comparable to that given to humans, and one that is more streamlined so it could be incorporated into a larger cognitive test battery. Even though the focus of this task has been on general cognitive ability, it could also be instrumental in investigating age related changes in cognitive ability as well (Salthouse, 1996). It has been consistently demonstrated that older adults are slower at a variety of speed related tasks, including those based off Hick's law (Sleimen-Malkoun et al., 2013). The experimental sample ranges in age from 0.5-18 years old and age-related changes in cognitive performance have been found at 10 years old (Coppola et al., 2014, 2015; Table 6.1). Therefore, age was a potential factor, besides general cognitive ability, that could also result in a slower RT.

We present the results from three experiments in pursuit of these goals to create a more efficient RT task for pigeons. While many of the procedural details were chosen to be more similar to what is typically used in human experiments, we did maintain a key difference. In many human experiments, the bit conditions are presented in blocks in ascending order. Under these conditions, the bit condition correlates with the amount of practice on the task, confounding the results (Longstreth, 1984; Neubauer, 1991; Widman & Carlson, 1989). To avoid this confound, the bit conditions were presented in a pseudorandomized order for all experiments. In Experiment 1, bit condition and the screen locations of the PTs were fixed. For example, in the 1-bit condition, the two PTs would always appear to the left and right of the midline (Figure 6.1). In Experiment 2, the screen location of the PTs was pseudorandomized such that the PTs could appear in any of the eight locations, irrespective of bit condition. In Experiment 3, the screen locations of the PTs were pseudorandomized as in Experiment 2, but the difference between the PT and the target was less obvious by reducing target salience. While these experiments were successful in creating a touchscreen version of the Hick's RT task that was more similar to what had been given to humans, and more streamlined in terms of training, we did not replicate the Hick's effect in pigeons with these procedures. Potential reasons for these results and implications for assessing RT in animals are discussed.

Experiment 1

For this experiment, the home key was centered near the bottom of the screen, while the PTs could appear in a semi-circle arrangement above. The bit condition and location of the PTs was fixed, meaning that the screen locations of the PTs were fixed for each bit condition. Additionally, we instituted a variable peck requirement to the home key to activate target presentation. This created uncertainty around when a PT would become a target. These

parameters were chosen to match what has typically been given to human participants. Having the screen locations fixed for each bit condition results in subjects receiving more exposure to and reinforcement of some screen locations relative to others, which could bias RTs (Longstreth, 1984; Neubauer, 1991; Widman & Carlson, 1989). While we recognize this confound, the goal of this experiment was to create a procedure more similar to what has been given to human participants (Jensen, 1982).

Methods

Subjects.

Six pigeons served as subjects. Subjects ranged in age from 0.5 – 12 years and three were male (Table 6.1). One subject, Odin, had only received training to peck the touchscreen. Two subjects, Goodall and Darwin, were initially trained with a slightly different, pilot version of the task where the PTs were in a higher position, which were difficult for some subjects, including Darwin, to reach. This led to differences in RT that were not due to differences in detecting a change in stimuli. The remaining subjects, Wenchang, Luigi, and Wario, had experienced tasks that emphasized associative learning. Additionally, these tasks did not have time sensitive trials, so they were naïve to tasks that had a timed component. Subjects were individually housed in steel home cages with metal wire mesh floors in a vivarium. They were maintained at 80% of their free-feeding weight, but were allowed free access to water and grit while in their home cages. Testing occurred at approximately the midpoint of the light portion of the 12-hour light-dark cycle. All procedures were approved by the UCLA Institutional Review Board.

Apparatus.

Testing was conducted in a flat-black Plexiglas chamber (38 cm wide x 36 cm deep x 38 cm high). All stimuli were presented by computer on a color LCD monitor (NEC MultiSync

LCD1550M) visible through a 23.2 x 30.5 cm viewing window in the middle of the front panel of the chamber. The bottom edge of the viewing window was 13 cm above the chamber floor. Pecks to the monitor were detected by an infrared touchscreen (Carroll Touch, Elotouch Systems, Fremont, CA) mounted on the front panel. A custom-built food hopper (Pololu, Robotics and Electronics, Las Vegas, NV) was located in the center of the front panel, its access hole flush with the floor. The food hopper contained a mixture of leach grain pigeon pellets and seed (Leach Grain and Milling). All experimental events were controlled and recorded with a Pentium III-class computer (Intel, Santa Clara, California). Stimuli were presented using the 2.7.11 version of Python with the Psychopy toolbox, version 3.0.3 (Peirce, 2007).

Stimuli.

There could be 1-9 circular stimuli, measuring 2 cm in diameter present on the screen during a trial. Each stimulus could either be a white outline or filled with white. The white outline was 1 mm thick. The background of the screen and of the white outline stimulus was dark gray.

Procedure.

Preliminary instrumental training.

Each subject received one session per day, 5-7 days a week. Each session consisted of 80 trials. The number of sessions and trial number was consistent throughout the duration of the experiment. During each trial, one stimulus was always present, centered 2.5 cm above the bottom of the viewing window and served as the home key. In the first phase of preliminary training, the home key was filled with white and subjects had to peck the home key to receive a food reward. Pecks within 0.7 cm of the edge of the home key were accepted and the subject had to make a mean response of three pecks (VR3, 3 +/- 2 pecks). The pigeon had 15 s to complete

this peck requirement. After meeting the peck requirement, the trial was terminated, the hopper area was illuminated, and the hopper was raised for 3 s. After the hopper was lowered, there was a 3 s inter-trial interval (ITI). If the pigeon failed to meet this requirement, the trial was terminated and the ITI began. These consequences were consistent throughout the entire experiment. When a pigeon reached criterion, completing the peck requirement for 90% of the trials on two consecutive sessions, it was moved to the second phase of pretraining. This criterion was used throughout the pretraining.

In the second phase of pretraining, each trial had two phases, the home key phase (HKP) and the choice phase. During the HKP, there was a white outline of a circle 5 cm above the home key. This was a PT. Subjects had to complete the same peck requirement to the home key as described above (VR3). If subjects failed to meet this requirement, the choice phase was not started. Instead, the subject went immediately into the ITI and afterwards the trial was repeated (HKP correction). If subjects met the peck requirement for the home key, the home key was replaced with a white outline, and the PT was filled with white, thereby becoming the target. This marks the beginning of the choice phase. The pigeon had to peck the target once (FR1) to receive a food reward. Pecks within 0.7 cm of the target were accepted. The pigeon had 15 seconds to complete the choice peck requirement. If it failed to complete the choice peck requirement, it did not receive a food reward. Pecks to the home key and screen background during the choice phase were neither punished nor reinforced. When a pigeon reached criterion, the available time to peck the target during the choice phase decreased in 5 second increments, until the pigeon only had 5 seconds to complete the peck requirement. After reaching criterion under these conditions, the pigeon moved on to the modified Hick's paradigm (MHP) described next.

Modified Hick's paradigm.

During the MHP, the home key was always present at the start of the trial, and 1, 2, 4, or 8 PTs were present (Figure 6.1). These represented 0, 1, 2, and 3 bits of information (binary choices), respectively. All stimuli were 7 cm distance from the home key as measured from edge to edge. In the 0-bit condition, the PT was placed directly above the home key (Figure 6.1, panels A and B). When more than one PT was present, there were always 2 cm between each, measuring from edge to closest edge. During the 1-bit condition, the two PTs were offset to the left and right of the center (Figure 6.1, panels C and D). As the bits of information increased, PTs were added to the left and right, forming a semi-circular arrangement (Figure 6.1, panels E and F, and G and H). Each location is referred to by which side of the screen it is located and its ordinal position in the direction away from the midline of the screen. This resulted in the following labeled PTs: bottom left (-4), bottom middle left (-3), top middle left (-2), top left (-1), top right (+1), top middle right (+2), bottom middle right (+3) and bottom right (+4). The PTs were only presented in these arrangements.

Each trial in the MHP consisted of two stages, the HKP (Figure 1, left panels) and the choice phase (Figure 1, right panels), similar to what was described for preliminary training. If the subject failed to meet the peck requirement to the home key, the choice phase was not started. Instead, the trial went immediately into the ITI and then was repeated (HKP correction). Correction trials were not included in calculations of total trial number. During the choice phase, when multiple PTs were present, there was only one target. If the subject pecked the target, it was followed with the food reward and ITI. If the subject pecked a PT, the trial was terminated and was counted as an error. Pecks to the background screen and home key were neither reinforced nor punished. Subjects had 5 s to complete the choice phase peck requirement. If they

did not peck any of the targets within 5 s, the trial was terminated and counted a ‘Miss’. The RT for ‘Miss’ trials was the ceiling value of 5 s.

The bit conditions were presented in a pseudorandomized order. The trials were organized into ten 8-trial blocks. Each bit condition could be presented twice without replacement in each block. This means that each bit condition had the potential to be presented for four consecutive trials, but the randomization made it highly unlikely. The bit conditions were presented in a pseudorandom order to ensure that the amount of experience with the task was equal across each bit condition (Widman & Carlson, 1989). There were 80 trials per session, 20 trials for each bit condition. Subjects received 10 sessions of training for a total of 800 trials. Trials where subjects did not meet the home key peck requirement were not included in calculations of total trial number.

Data analysis.

The median and standard deviation RT in seconds was collected for each PT number and each target screen location. Only trials where subjects pecked the target were included in analysis, trials with errors of commission or omission were excluded. Practice effects were examined by comparing the median RT collapsed over the first three to RT collapsed over the last three sessions. Data were analyzed using Python3, version 3.8.3, with the Jupyter Notebook interface, version 6.0.3, R version 3.6.2 with RStudio interface version 1.2.5033, and SPSS, version 27.

Results

Bit condition.

The type of errors made during each bit condition throughout training was collected for each subject (Table 6.1). All subjects made at least one error, but generally there were more errors of omission, particularly for the 0-bit condition. A visual inspection of the data indicated that the error rate was relatively constant over training and, collapsed across all sessions, errors occurred on less than 10% of trials. Because subjects had relatively few errors throughout training and it would be difficult to determine why the errors had been committed, these trials are excluded from the subsequent analyses. The total number of trials excluded for each subject are reported in Table 6.1.

Median RTs were calculated for each subject for each bit condition for the first and last three sessions and over all 10 sessions (Figure 6.2a). In general, subjects were slowest during the 0-bit condition and fastest in the 3-bit condition. Training did not have a consistent effect across subjects because some subjects, like Wenchang, were faster in the last three sessions, while other subjects, like Luigi, were slower. A two-way repeated measures ANOVA was used to determine if there was a significant difference in median RT across bit conditions and amount of training. The amount of training included performance collapsed over all 10 sessions and the first and last three sessions. The bit condition failed Mauchly's test of sphericity (x^2 (5) = 17.6, p = .005) so a Greenhouse-Geisser correction was used. There was no main effect of bit condition (F (1.14, 5.68) = 10.28, p = .018) or of how much training subjects had (F (2, 10) < 1). There was no interaction between bit condition and amount of training, F (6, 30) < 1. Because there was no significant effect of training, performance collapsed all 10 sessions of training were used in the subsequent analyses.

A simple linear regression was used to investigate the relationship between bit condition and RT for each subject in more detail, where bit condition was the predictor and RT was the dependent variable (Table 6.2). For all subjects, the slope of the line was negative and the average slope was -0.06, indicating that as bit condition increased, RT decreased by .06 seconds. Although all subjects showed a significant negative slope in RT, the adjusted R^2 values were small, with a mean value of .041.

Location.

Median RTs as a function of screen location of the target was also investigated. Across bit conditions, subjects were generally slowest at the most central locations and faster the further the target location was from the center (Figure 6.2b). A two-way repeated measures ANOVA was used to investigate whether there were any differences in RT based on screen location of the target and bit-condition. The screen locations in the analyses were restricted to the locations shared across the bit conditions. For the most central locations (-1, +1), the 1, 2, and 3-bit conditions were included in the analysis. There was no main effect of bit-condition (F (2, 10) < 1) or of screen location (F (1, 5) < 1), nor was there an interaction (F (2, 10) < 1). For the -2, -1, +1, and +2 locations, the 2 and 3-bit conditions were included in the analysis. There was no main effect of bit condition (F (1, 5) < 1) or of screen location (F (3, 15) = 3, p = .064), nor an interaction between bit condition and screen location (F (3, 15) < 1).

A repeated measures ANOVA was used to investigate whether there were any differences in RT in the 3-bit condition in which the target could appear in any of the eight screen locations. There was not a main effect of screen location on RT in the 3-bit condition, F (7, 35) = 1.69, p = .144.

Discussion

Training did not have a clear or significant effect on median RT since some subjects were slower in the first three sessions compared to the last three, while other subjects showed the opposite effect. It is not as surprising that some subjects became faster with experience since that result has been found in the previous literature (Vickrey & Neuringer, 2000). It is more difficult to understand why some subjects were slower with experience. Median RT may have increased due to fatigue, boredom, or lower motivation due to satiation, but it is not possible to determine which, if any, impacted performance. It is unlikely that a slower RT is due to a speed-accuracy trade off considering the overall error rate was small and constant during training. To minimize potential effects of practice, the median RT was collapsed across all 10 sessions. Subjects showed the opposite of the predicted results, subjects were slowest in the 0-bit condition and fastest in the 3-bit condition (Figure 6.2a), though the difference between bit-conditions was not significant in this analysis. All subjects had a negative slope, but bit condition accounted for a relatively small amount of the variance in performance, indicating that other factors had better explanatory value, in particular the target screen location. Subjects could have been slowest to peck and had the most errors of omission in the 0-bit condition because the target was at the highest physical location on the screen, possibly as a result of having to stretch their necks up to reach those locations (personal observation). While we confirmed that subjects could physically reach the target in the 0-bit condition, it may have been more difficult for subjects to coordinate and execute the peck response to that location. Almost all subjects had the longest latencies for targets that were higher on the screen, irrespective of the bit condition (Figure 6.2b). Even though most subject showed a similar pattern of RT based on screen location, the variability across subjects may have prevented finding significant differences. Since the available number

of PTs was confounded with screen location, it was unclear whether we uncovered a real 'anti-Hick's' effect or if it was an artefact of the display setup. To distinguish between these accounts, the next experiment separated screen location of the target and the number of PTs.

Experiment 2

In Experiment 1, we found an 'anti-Hicks' effect that may have been due to the screen locations of the PTs. In this experiment, PTs were still restricted to the semi-circular arrangement described in Experiment 1, but there was no restriction on the screen location of PTs across bit conditions. Thus, a PT could appear in any of the 8 locations. To equalize the effect of location across bit conditions, the target would appear equally often in each location for each bit condition. Seven new subjects were trained on this 'random' version of the task. Each session included more trials so the target could be presented at each location for each bit condition, but subjects received fewer sessions to keep the overall amount of training similar to Experiment 1.

Method

Subjects.

Seven new pigeons served as subjects. Subjects ranged in age from 0.5 – 18 years old and four were male (Table 6.1). The amount of experience subjects had with other cognitive tasks varied, but none of the subjects had experience with a similar RT task. One subject, Thoth, had only received training to peck the touchscreen. Subjects were housed and maintained as described in Experiment 1.

Apparatus and stimuli.

The testing apparatus and stimuli were the same as described in Experiment 1.

Procedure and data analysis.

The preliminary training for the random version of the task was the same as described in Experiment 1. The random MHP was very similar to what was used in Experiment 1, in terms of the HKP, choice phase, and food reward. The key difference in the random MHP, was that PTs could be presented in any of the eight screen locations, irrespective of bit condition, and the target was presented equally often at each location for each bit condition. Subjects trained for nine sessions and there were 96 trials per session for a total of 864 trials. Within each session there were 24 trials for each bit condition, and the target was presented in each location three times for each bit condition. The trials were organized into three 32-trial blocks. Each bit condition could be presented with the target at each location once without replacement in each block. The data were analyzed as described in Experiment 1.

Results

Errors

The total number of omission trials was higher in the random variation of the task (M =19.14, SD = 23.57) compared to the first experiment (M = 16.33, SD = 18.79, Table 6.1). Additionally, in this variation, omissions were more evenly distributed among the bit-conditions, whereas in the first experiment omissions mostly occurred in the 0-bit condition. The number of commission errors was lower in this experiment (M = 4.71, SD = 6.78) compared to the first experiment (M = 9, SD = 12.31). The commission errors were distributed in a similar way across experiments, occurring more frequently during the 2 and 3-bit conditions (Table 6.1). A 2x4 mixed ANOVA was used to compare the number of omission and commission errors across the bit conditions in each experiment. The bit-condition for the omission data failed Mauchly's test of sphericity (x^2 (5) = 11.71, p = .04), so a Greenhouse-Geisser correction was used. There was

no main effect of bit-condition (F (1.74, 19.17) = 3.27, p = .066), or of experiment (F (1, 11) < 1), nor an interaction (F (3, 33) = 2.06, p = .159). Similar results were found with the commission errors. There was no main effect of bit condition (F (2, 22) = 1.98, p = .161) or of experiment (F (1, 11) < 1), nor an interaction (F (2, 22) < 1).

A visual inspection of the data confirmed that the error rate was constant across the training. Due to the difficulty in interpreting errors, as explained earlier, only the median RT from hit trials was used in the subsequent analyses. The number of trials excluded for each subject can be found in Table 6.1.

Bit condition.

In general, the median RT did not change across the bit conditions across training (Figure 6.3a). The amount of training did not have a similar effect across subjects and a two-way repeated measures ANOVA, with bit-condition and amount training as the within-subject factors, was used to investigate the potential effects of practice. The amount of training failed Mauchly's test of sphericity (x^2 (2) = 9.59, p = .008), so a Greenhouse-Geisser correction was used. There was no main effect of bit condition (F (3, 18) < 1) or of amount of training (F (1.08, 6.48) < 1), nor an interaction (F (6, 36) < 1). Therefore, the median RT collapsed across all 9 sessions of training was used in the subsequent analyses.

A simple linear regression was used to further investigate the relationship between bit condition and RT for each subject (Table 6.2). For all subjects, the slope was near 0 and bit condition could not account for any of the variance in RT.

Location.

The median RT based on target location showed a 'W' pattern, where RT was generally longest at the most central locations, was faster as the target was presented along the periphery of the semi-circle, before increasing again at the most distal locations for all bit conditions (Figure 6.3b). A two-way repeated measures ANOVA was used to investigate the median RT with bit condition (1, 2, and 3-bits) and screen location as the within-subject factors. There was no main effect of bit-condition (F (2, 12) < 1) and no interaction between bit condition and screen location (F (14, 84) = 1.58, p = .101). There was a main effect of screen location (F (7, 42) = 2.99, p = .012, partial eta squared = .33). Post hoc tests with no correction indicated that RT was significantly slower to the top left (-1) target compared to the BML (-3, p = .014), TML (-2, p = .029), and TMR (+2, p = .043) targets. Similarly, RT was significantly slower to the top right (+1) target compared to the TMR (+3, p = .04) location. These differences did not survive a Sidak or Bonferroni correction

Discussion

In Experiment 1, when the PTs were fixed to certain screen locations in each bit condition, an 'anti-Hick's' effect was found. The goal of Experiment 2 was to determine whether RT differences were a function of screen location or bit condition. This was accomplished by varying the screen location of PTs across all eight locations, irrespective of bit condition. With this manipulation, we did not replicate the 'anti-Hick's' of Experiment 1, *nor* the Hick's effect as normally reported in human studies. In general, subjects were still slowest when the location of the target was at the most central point, but this was consistent across all bit conditions. The median RT was consistent across all bit conditions. While this contradicts the previous pigeon and human intelligence literature, these results may be related to the stimulus intensity of the

target and the contrast between the target and PTs. In Experiments 1 and 2, the target was a completely-filled white circle, while the PTs were white outlines, resulting in a high contrast between the target and PTs. The target being presented likely resulted in a change in the luminance in that location, though this wasn't directly measured. Previous research with humans has shown RT will decrease as stimulus luminance increases (Pins & Bonnet, 1996).

The contrast between the target and the PT may also influence visual search processes. While visual search processes are typically investigated with a slightly different paradigm, participants must find the target in an array of nontarget stimuli versus detecting a simple change in the environment, there are relevant similarities between the visual search tasks and the choice RT task presented here (Blough, 1979; Teichner & Krebs, 1974). Visual search can be guided by either of two processes that operate in tandem; a parallel search process, where the participant is viewing all of the stimuli simultaneously, and a serial search where the participant is viewing each stimulus individually (Moran et al., 2015). How the number of stimuli impacts RT can be used as an indicator of which process is controlling behavior. A flat slope, where RT does not increase with the number of stimuli, is evidence for the parallel process while an increasing slope is evidence for the serial search process. Typically, flatter slopes are found when the target 'pops out' or is easily discriminable from the nontarget stimuli. Differences in luminance between the target and nontarget stimuli seem to be particularly susceptible to a pop out effect (Theeuwes, 1995). This may also be a factor in choice RT since procedures using a light tend to have shallower slopes compared to procedures that use other stimuli, like numbers (Teichner & Krebs, 1974). The similarity in results across different types of experiments indicates that visual search processes are impacting performance in Hick's RT tasks in humans. There is also evidence that pigeons have similar parallel and serial search processes that are impacted by target salience and

stimulus set size in a similar way as for humans (Blough, 1979). Therefore, it is likely that the flat slope found in Experiment 2 is due to a stimulus driven, parallel search process.

To investigate this idea, in Experiment 3, we varied the procedure again by dramatically reducing the salience of the target. This should make the target much more difficult to distinguish from the remaining PTs. By making it more difficult to discern the PTs from the target, it should encourage the subject to more actively monitor the PTs to detect the change when a target is presented, rather than relying on reflexive responding to any sudden onset in the periphery which can be done through parallel processing (Blough, 1979). The more PTs that require monitoring, presumably the longer it should take the pigeon to detect target onset.

Experiment 3

In Experiment 2, there was no effect of bit condition on median RT, failing to replicate the 'anti-Hick's' effect from Experiment 1, and failing to replicate the Hick's effect previously reported in pigeons (Vickrey & Neuringer, 2000) and people (Sheppard & Vernon, 2008). It was possible that subjects were relying on salient changes in their peripheral vision to guide choice behavior, which could utilize parallel processing and, thus, would not be impacted by the number of PTs available. To attenuate this strategy, in this experiment the difference between the PT and the target was more subtle. Instead of the PTs being a white outline filled with gray and the target being a completely filled white circle (Figure 6.1), the PTs were filled with a semi-transparent white and the target was an opaque white (Figure 6.4). This subtle distinction should make it more difficult to rely on a change in peripheral vision, and instead encourage active monitoring of the available PTs to detect target onset – a serial processing strategy. Six new subjects were used to test this manipulation. If this manipulation prevented subjects from using

parallel processing to guide choice behavior, we would expect subjects to more actively monitor the number of PTs available, increasing RT as an increasing function of bits of information.

Methods

Subjects.

Six new pigeons served as subjects. Subjects ranged in age from 2 – 17 years old and three were male (Table 6.1). The amount of experience subjects had with other cognitive tasks varied, but none of the subjects had experience with a similar RT task. Subjects were housed and maintained as described in Experiment 1.

Apparatus and stimuli.

The testing apparatus and stimuli were the same as described in Experiment 1 except that the stimuli were always filled with white. The stimuli could either be semi-transparent or completely opaque (Figure 6.4).

Procedure and data anlysis.

The preliminary training and MHP was the same as described in Experiment 2, except the new stimulus set was used. The data were analyzed as described in Experiment 2.

Results

Errors.

In the previous experiments, subjects were fairly accurate but, for this variation, one subject, Yoshi, had an unusually high number of omission (70) and commission (72) errors (Table 6.1). Errors were related to the screen locations of the target, primarily occurring on the left half of the stimulus configuration, which makes it unlikely that the high error rate is due to the manipulation. Commission errors tended to increase with bit-condition, while omission errors

were more evenly distributed. The total amount of commission errors for Yoshi was more than twice the standard deviation (M = 20.83, SD = 25.76), was outside of the interquartile range calculation to detect outliers (Rousseeuw & Croux, 1993), and more than twice the second highest commission error total, which was 33. Because Yoshi was committing errors in a systematic way and met the criteria to be classified as an outlier, his data were excluded from analysis. Without Yoshi's data, the mean number of omission (M = 13.2, SD = 15.59) and commission (M =10.6, SD = 6.66) errors was similar to the previous experiments.

A 3 x 4 mixed ANOVA was used to compare the number of omission and commission errors across the bit conditions in each experiment. The bit-condition for the omission data failed Mauchly's test of sphericity (x^2 (5) = 12.75, p = .026), so a Greenhouse-Geisser correction was used. There was no main effect of bit condition (F (1.86, 27.9) = 2.27, p = .125), or of experiment (F (2, 15) < 1), nor an interaction (F (6, 45) = 2.08, p = .074). For commission errors, there was no main effect of experiment (F (2, 15) < 1), nor an interaction (F (4, 30) < 1). There was a main effect of bit-condition (F (2, 30) = 6.86, p = .004, partial eta squared = .314) and post-hoc tests with a Bonferroni correction indicated that there were significantly more errors in the 3-bit condition (M = 4.36, SD = 4.84) compared to the 1 (M = 1.45, SD = 1.9, p = .032) and 2 (M = 2.29, SD = 4.84, p =.011).

As in the previous experiments, the subsequent analysis investigated the potential effects of practice on median RT for hit trials only due to overall low error rate and difficulty in interpreting errors.

Bit condition.

Median RT tended to increase with bit-condition, though this was not consistent across all subjects or across amount of training. For example, Durrell initially showed an increase in

median RT with bit-condition, but at the end of training median RT decreased with bit condition (Figure 6.5a). A two-way repeated measures ANVOA, with bit condition and amount of training as the within-subject factors, was used to investigate potential training effects. There was no main effect of the amount of training (F (2, 8) = 2.62, p = .133) nor an interaction (F 6, 24) < 1). There was a main effect of bit condition (F (3, 12) = 4.7, p = .022, partial eta squared = .54). Post hoc tests with no correction indicated that RT was significantly faster in the 1-bit condition compared to the 2-bit (p = .022) and 3-bit (p = .015) conditions, but these differences did not survive a Sidak or Bonferroni correction. Since there were no main effects of amount of training, the median RT collapsed across all nine sessions of training was used in the subsequent analyses.

The effect of bit condition on RT was investigated further using simple linear regression for each subject. Similar to Experiment 2, the slope for all subjects was close to 0 and bit condition could not account for any variance in performance (Table 6.2).

Location.

Similar to Experiment 2, the median RT based on screen location showed a 'W' pattern (Figure 6.5b). A two-way repeated measures ANOVA was used to investigate the median RT with bit condition (1, 2, and 3-bits) and screen location as the within-subject factors. There was no main effect of bit condition (F (2, 8) = 4.35, p = .053), or of location (F (7, 28) = 1.45, p = .224), nor an interaction (F (14, 56) = 1.19, p = .311).

Age.

While each experiment included birds that ranged from young to old, there wasn't sufficient power to investigate age related differences in each individual experiment. To increase power, age effects were investigated across experiments. Only performance in the 3-bit condition was analyzed because it had a similar range across experiments (M = 1, SD = 0.26). To verify

there were no significant differences due to experiment, a 3 x 2 ANOVA with experiment and age as between subject factors was performed. To investigate age, two groups were created with a 'young' group ranging in age from 0.5-4 years old (n =10) and an 'old' group ranging in age from 11-18 years old (n = 9). There was no main effect of experiment (F (2, 13) < 1), nor of age (F (1, 13) < 1). There was no significant interaction, F (2, 13) = 3.06, p = .08. Age was also investigated as a continuous variable with a Pearson correlation. There was a positive correlation between age and median RT, indicating that as age increased, RT also increased, but it was not significant (r (17) = .36, p = .14). The correlation was also performed after excluding 18-year-old Dickinson because it's RT, 1.76 s, was 2 standard deviations away from the mean. While the subsequent correlation was still positive, it was even weaker (r (16) = .16, p = .54).

Discussion

In Experiment 3, the difference between the target and the PT was made more subtle to encourage serial processing. If serial processing was controlling behavior, median RT should increase with bit-condition, which would also conform to Hick's Law. The median RT from almost all subjects increased as predicted, particularly early in training, though the amount of training did not have a significant effect (Figure 6.5a). While there was a significant main effect of bit condition, post hoc analyses did not survive correction. In addition, when collapsed across all training sessions, the slope for all subjects was practically 0. This is similar to the results from Experiment 2, where the median RT was consistent across bit conditions (Figure 6.3a). To further emphasize the similarities across experiments, the median RT was within the same range (0.6 – 1.8 s) and there were no significant differences in the error rate, which suggest that speed-accuracy trade-off is similar across experiments. This indicates subjects were either using the same process across experiments, that a serial process does not cause large differences in RT

with this procedure, or that this procedure was not successful at biasing pigeons to utilize a serial process (Figure 6.3a). While the modification did not accomplish the intended goal, it did allow us to investigate the effect of age on RT using cross experiment analysis. There was a weak, positive correlation between age and median RT, meaning that RT increased with age, but this relationship was not significant.

General Discussion

The goal of these experiments was to create a streamlined RT task for pigeon's based on Hick's Law that was similar to what had been used in human research (Vickrey & Neuringer, 2000). While the experiments were successful in these respects, only Experiment 3 showed some weak evidence for replicating previous research where RT increased as bits of information increased (Hick, 1952; Jensen, 1982; Vickrey & Neuringer, 2000). The results from Experiment 1 indicated that the physical screen location of the target was an important factor in determining RT. Across all experiments, subjects were slowest when the target was in the most central location, which happened to be the highest on the screen, though there was some evidence to suggest that subjects were also slower when the target was in the screen locations furthest from the center (Figures 6.2b, 6.3b, 6.5b). The difference in RT based on location could not be due to differences in experience with the particular locations for Experiments 2 and 3. It is possible that it was more difficult to execute a motor response to these locations, though it was confirmed that all subjects could reach all locations and all locations were equidistant from the home key. Ideally, RT would be similar across the target locations, but there is some evidence to suggest that, for humans, differences in RT based on absolute screen location do not affect the RT based on bit-condition (Wright et al., 2007), which is supported by the results from Experiments 2 and 3.

When the target screen location was presented at equal frequencies across bit conditions, as in Experiments 2 and 3, median RT was consistent across bit conditions in Experiment 2 and the slope for each subject was 0 (Figure 6.3a, Table 6.2). In Experiment 3, while there was a significant main effect of bit condition, it did not survive correction and the slope for each subject was 0 (Figure 6.5a, Table 6.2). The results from Experiments 2 and 3 indicated that either the manipulation in Experiment 3 was not sufficient for subjects to use serial processing instead of parallel processing, that serial processing does not dramatically increase RT in this procedure, or that subjects were using a different search process during this task. How visual search processes impact performance in humans is not frequently discussed in the context of Hick's RT and intelligence, but comparisons of the slope across different stimuli indicate that these processes also impact human performance on these types of tasks (Teichner & Krebs, 1974). While there was a main effect of bit condition on median RT in Experiment 3, it would be beneficial for future investigations if the difference between bit conditions was larger. A potential factor not explored here is stimulus-response compatibility.

The congruency of stimulus and the required response influences RT, with high congruence resulting in faster RTs (Neubauer, 1991). For example, RTs are faster if a stimulus appearing on the left side of the screen requires pushing a button on the left compared to a button on the right (Lien & Proctor, 2002). There is evidence to suggest that high levels of stimulus-response compatibility can attenuate or eliminate increases in RT based on Hick's law (Proctor & Schneider, 2018; Wright et al., 2007). In our version of the task, subjects had to directly peck the target, which likely has very high stimulus-response compatibility, considering the propensity of pigeons to peck visual stimuli associated with reward (Brown & Jenkins, 1968). This is not a wholly satisfactory explanation, however, since Vickrey and Neuringer (2000) also

required pigeons to directly peck the target and the RT of their subjects conformed to Hick's law. Another difference in the Vickrey and Neuringer (2000) procedure compared to what is typically given to humans, was that pigeons had to peck a 'reinforcing star' after pecking the target to receive a food reward or conditioned reinforcer. It is possible that the additional peck to the 'reinforcing star' reduced the stimulus-response compatibility of pecking the target, but it is not immediately apparent why. Alternatively, placing the extra step between pecking a target and delivery of food reward resulted in a sufficiently long target response-reward interval to reduce stimulus control by the target. Future investigations should manipulate the stimulus-response compatibility to better understand how this impacts RT for pigeons specifically. It is possible that requiring the peck to be made slightly off target would sufficiently reduce compatibility for subjects to conform to Hick's law (Proctor & Schneider, 2018).

Decreasing the compatibility between the stimulus and the response might strengthen the relationship between age and RT. With this set of experiments, a positive, but weak and nonsignificant, correlation was found between age and RT in the 3-bit condition. In humans, reducing the compatibility of the response had a stronger negative impact on the RT of older adults compared to younger adults (Sleimen-Malkoun et al., 2013). A similar increase in RT may be found for older pigeons if stimulus-response compatibility is reduced. It is also possible that a relationship between RT and age is only consistently seen when subjects are very old. The four oldest pigeons were among the slowest RTs (Figure 3a, 5a). The maximum observed lifespan of a pigeon is 35 years; thus a stronger relationship may be found when even older subjects are included in the procedure (Carey & Judge, 2000).

Even though RT did not change as predicted in Experiment 2, it is possible that this task still captures individual differences in speed that are relevant to general cognitive abilities. As

mentioned in the introduction, performance on a wide variety of speed tasks seems to rely, at least in part, on a domain general speed ability. A similar domain general factor has also been implicated when investigating the differences in RT based on stimulus-response compatibility. A relationship between RT and intelligence is typically found in people, even when using tasks with a high level of stimulus-response compatibility (Neubauer, 1991). This indicates that this task could still be useful for determining if speed is related to other cognitive abilities in pigeons, similar to what is seen in humans. This straightforward procedure would be relatively easy to administer to other visually-guided species. This could advance comparative investigations of the role of processing speed in a general cognitive factor.

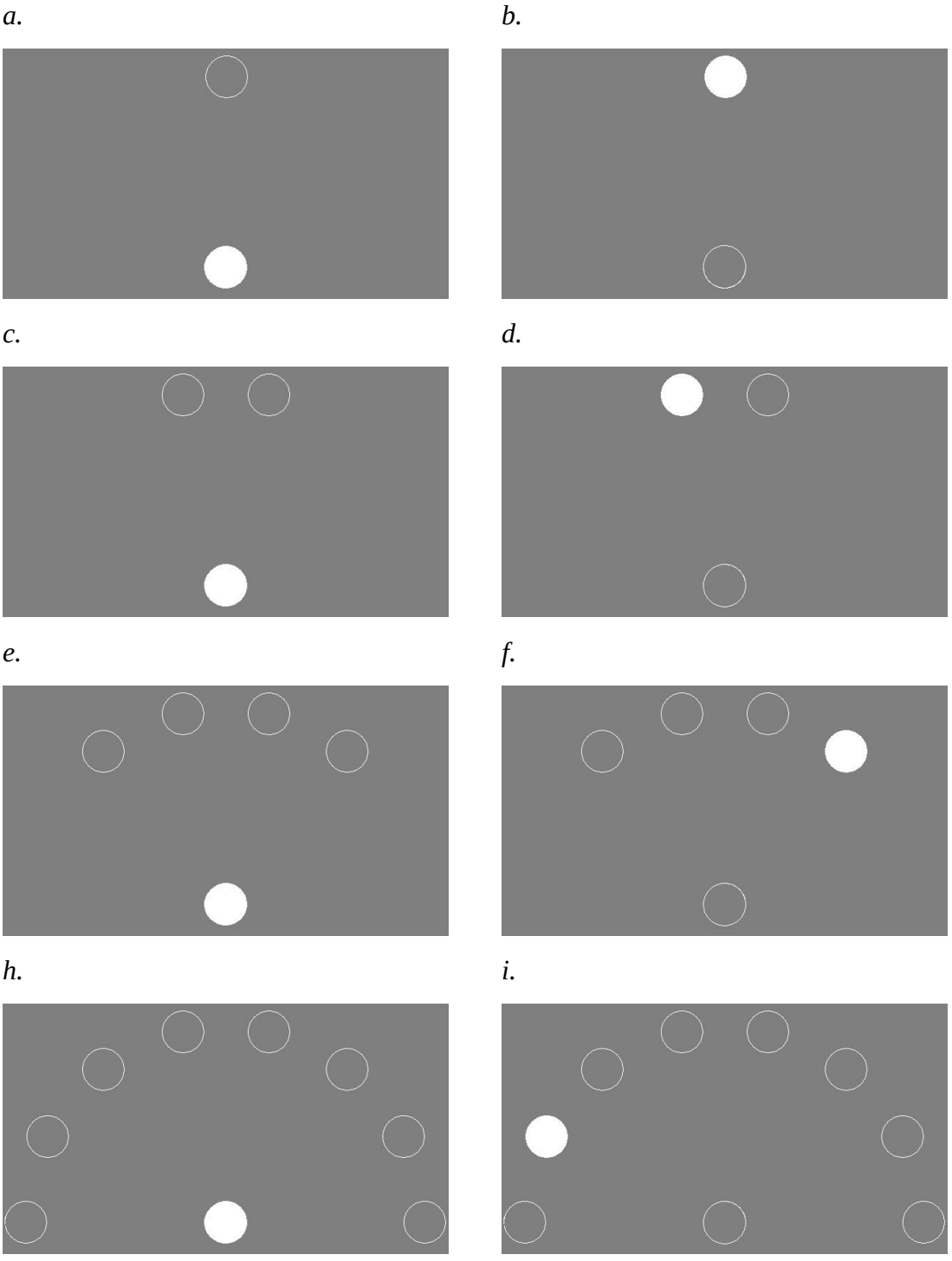
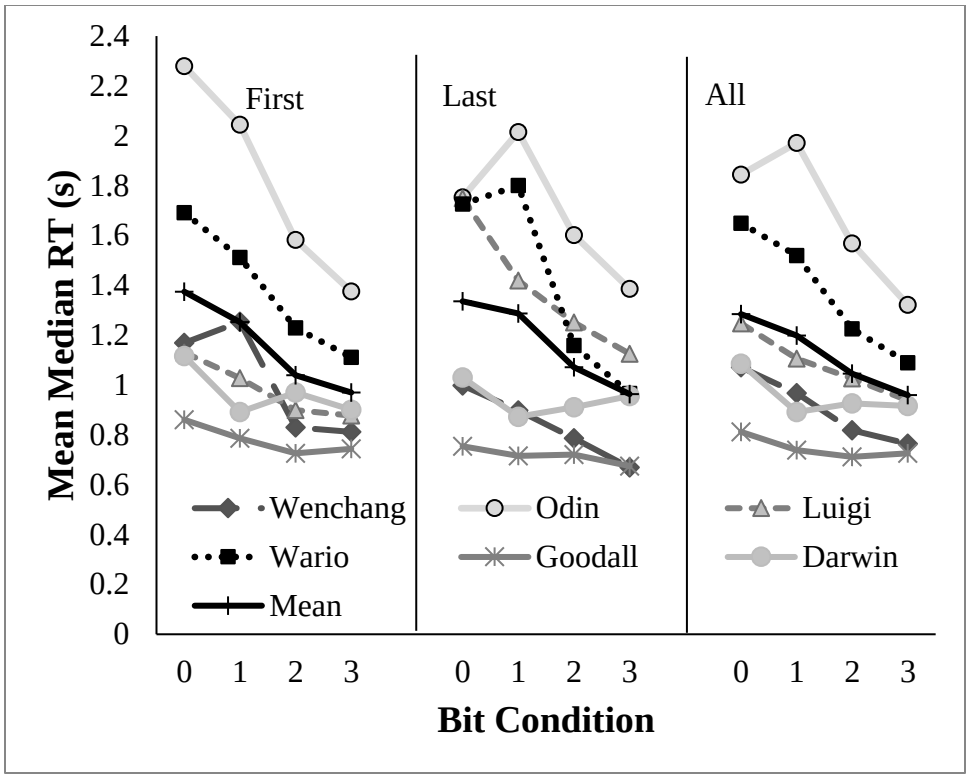


Figure 6.1. Examples of trials during the Modified Hick's Procedure from Experiment 1. The rows represent different bit conditions, or the number of binary choices. From the top, the bit conditions are 0, 1, 2, and 3. The left column shows the home key phase and the right column shows the choice phase.

a.



b.

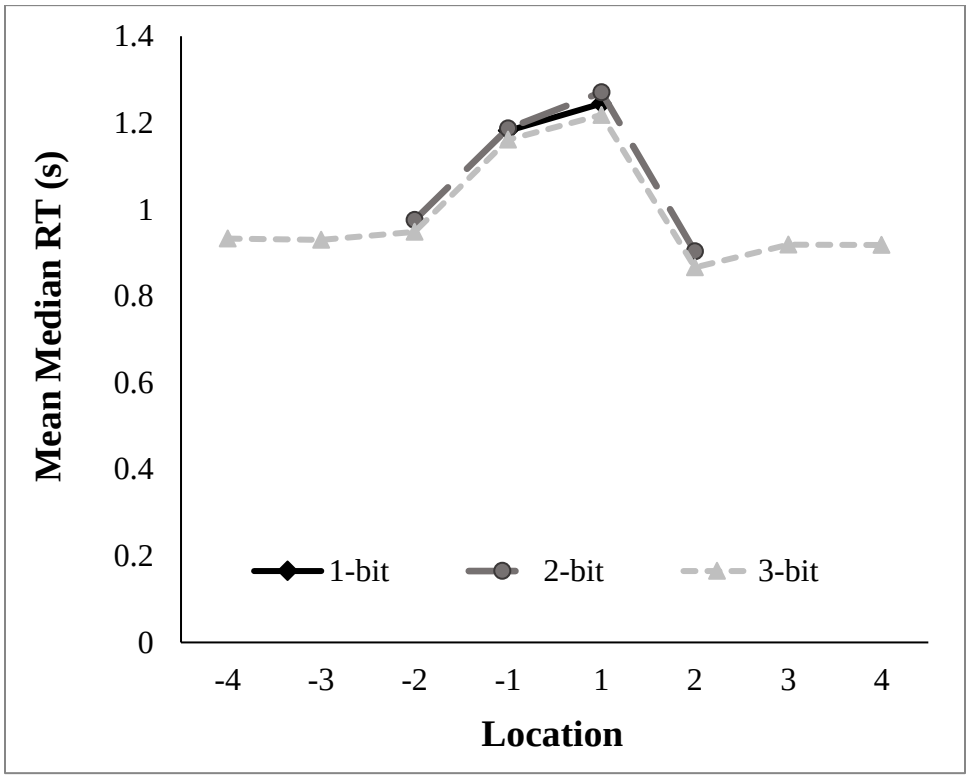
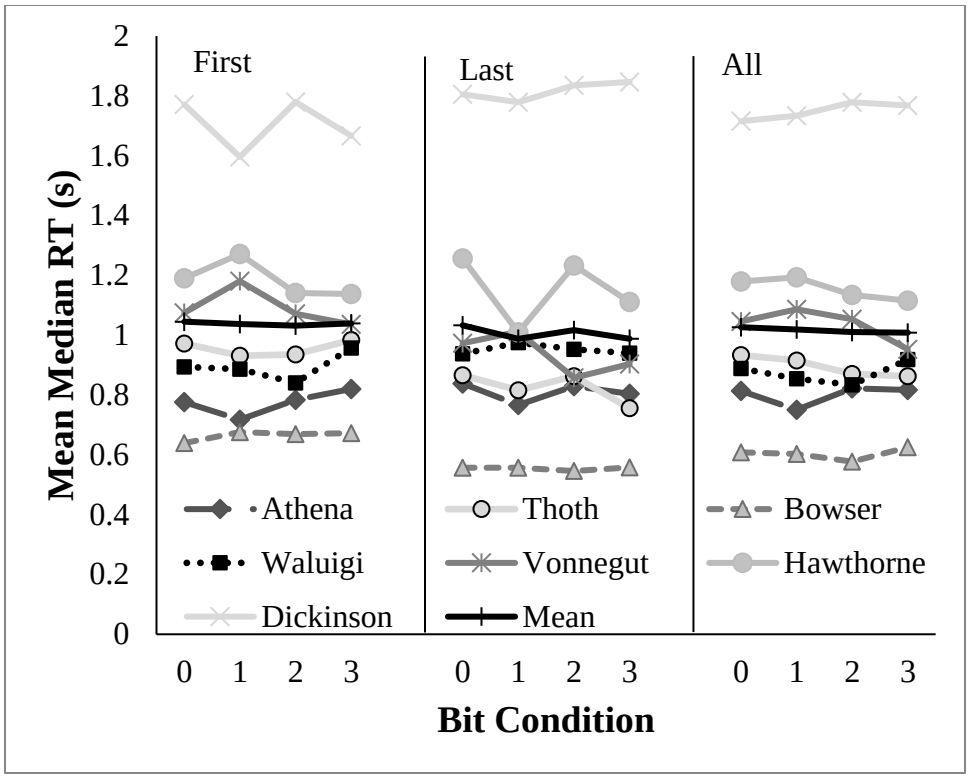


Figure 6.2. Panel a depicts the mean median reaction time (RT) to each bit condition of Experiment 1, where first and last depict performance collapsed over the first and last three sessions respectively. All is performance collapsed over all training sessions. Panel b depicts the

mean median RT to each target location by bit condition collapsed over all training sessions. For location, negative numbers represent the left half of the stimulus display and numbers further from 0 represent locations further from center.

a.



b.

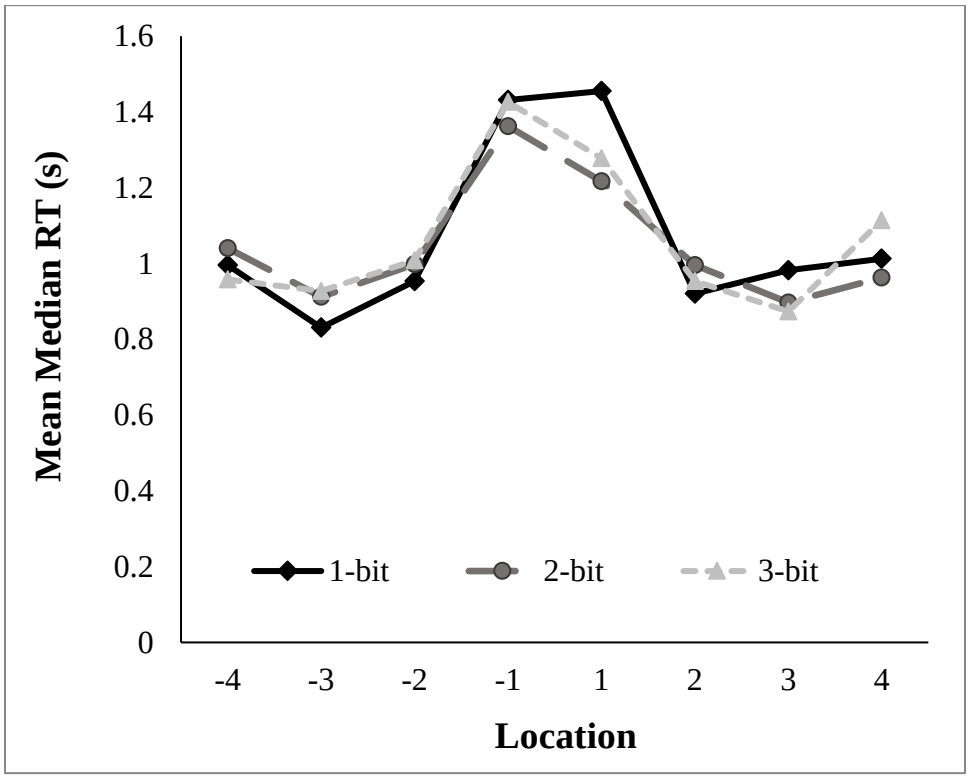


Figure 6.3. Panel a depicts the mean median reaction time (RT) to each bit condition from Experiment 2, where first and last depict performance collapsed over the first and last three sessions respectively. All is performance collapsed over all training sessions. Panel b depicts the

mean median RT to each target location by bit condition collapsed over all training sessions. For location, negative numbers represent the left half of the stimulus display and numbers further from 0 represent locations further from center.

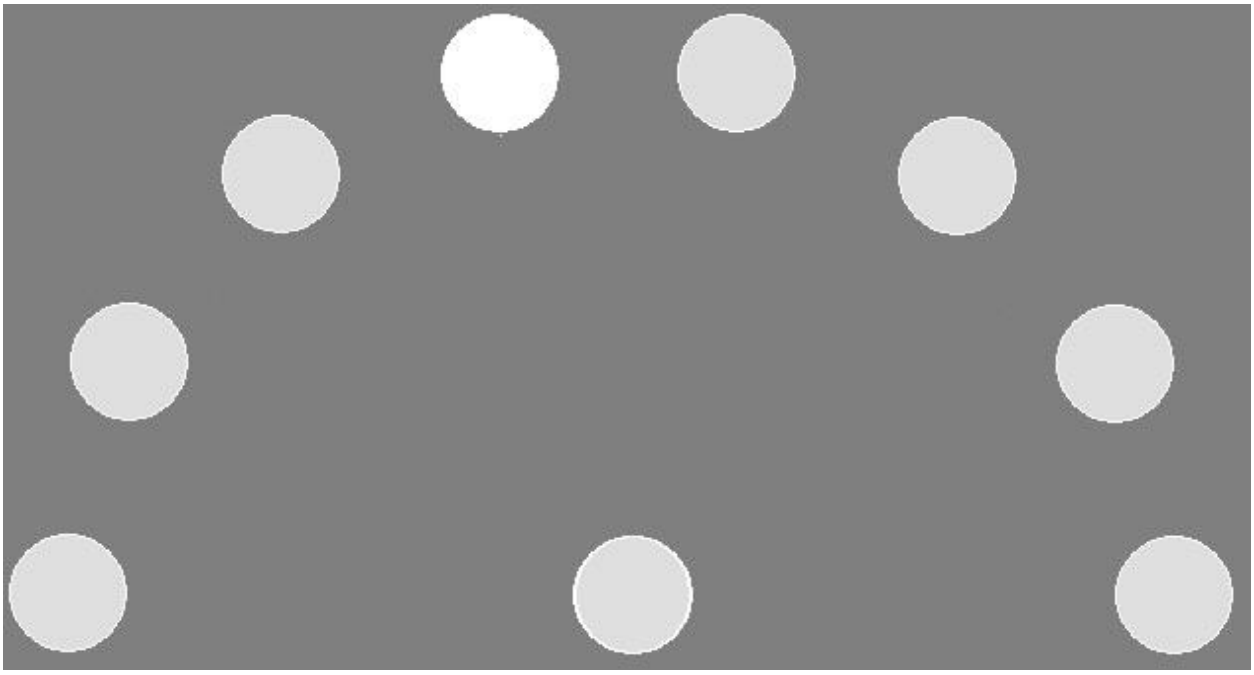
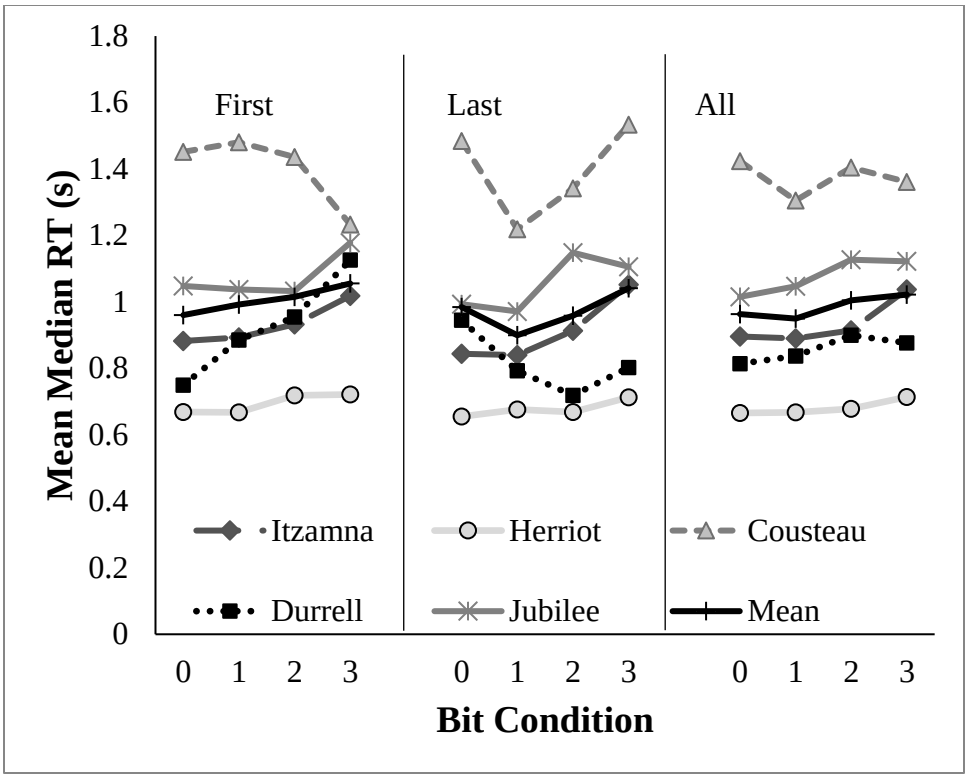


Figure 6.4. An example of a trial during the choice phase from Experiment 3, where the top left location (-1) is the target.

a.



b.

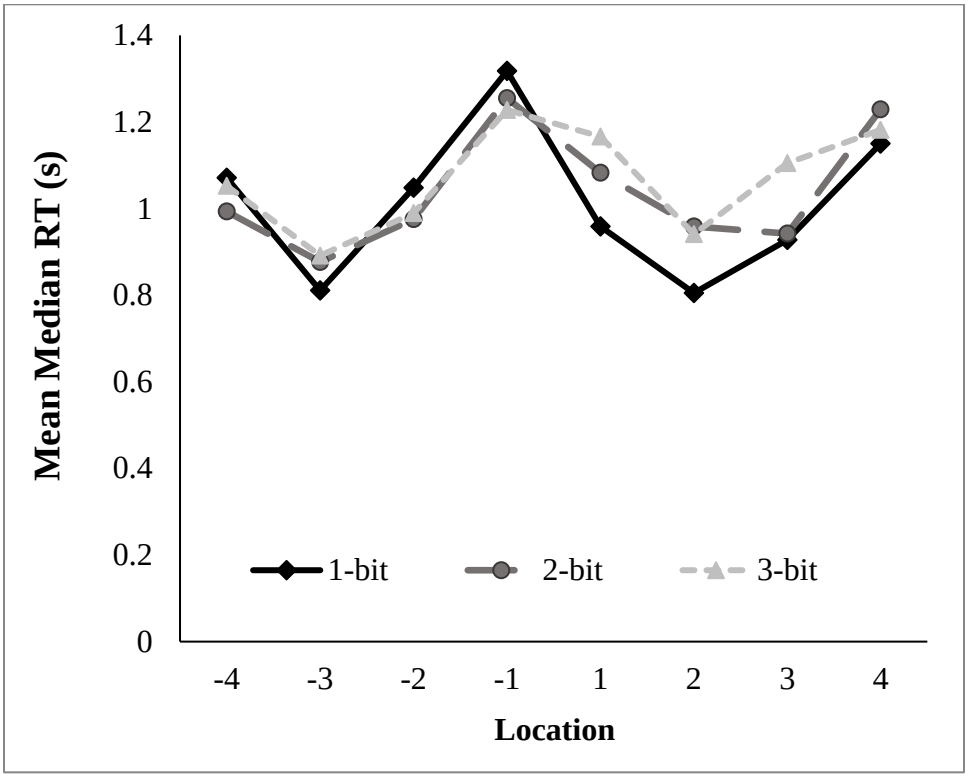


Figure 6.5. Panel a depicts the mean median reaction time (RT) to each bit condition from Experiment 3, where first and last depict performance collapsed over the first and last three sessions respectively. All is performance collapsed over all training sessions. Panel b depicts the mean median RT to each target location by bit condition collapsed over all training sessions.

For location, negative numbers represent the left half of the stimulus display and numbers further from 0 represent locations further from center

All				Omission					Comission				Excluded
Variation	Name	Sex	Age	0-Bit	1-Bit	2-Bit	3-Bit	Total	1-Bit	2-Bit	3-Bit	Total	
Fixed	Wenchang	F	0.5	6	2	2	1	11	0	1	0	1	12
Fixed	Odin	M	1	27	11	5	9	52	0	1	1	2	54
Fixed	Luigi	M	3	2	1	4	1	8	4	14	15	33	41
Fixed	Wario	M	3	11	4	1	5	21	0	0	2	2	23
Fixed	Goodall	F	11	0	0	0	0	0	3	0	2	5	5
Fixed	Darwin	F	12	3	2	0	1	6	7	1	3	11	12
Fixed	Mean			8.17	0.83	0.50	0.50	16.33	2.33	2.83	3.83	9.00	12.67
Random	Athena	F	0.5	0	0	0	0	0	0	0	0	0	0
Random	Thoth	M	1	7	7	9	3	26	0	0	0	0	26
Random	Bowser	M	3	0	0	0	0	0	0	0	1	1	1
Random	Waluigi	F	3	0	0	0	1	1	0	1	1	2	3
Random	Vonnegut	M	17	1	2	0	3	6	0	2	5	7	13
Random	Hawthorne	M	17	16	9	9	10	44	2	5	12	19	63
Random	Dickinson	F	18	14	8	16	19	57	1	2	1	4	61
Random	Mean			5.43	3.71	4.86	5.14	19.14	0.43	1.43	2.86	4.71	11.93
Subtle	Itzamná	M	2	2	1	2	1	6	3	2	8	13	19
Subtle	Yoshi	M	4	16	18	12	24	70	8	21	43	72	142
Subtle	Herriot	M	12	0	0	0	0	0	0	1	0	1	1
Subtle	Cousteau	M	13	10	11	7	12	40	1	2	4	7	47
Subtle	Durrell	F	13	0	0	4	4	8	1	3	10	14	22
Subtle	Jubilee	F	17	2	2	3	5	12	3	5	10	18	30
Subtle	Mean			5.00	5.33	4.67	7.67	22.67	2.67	5.67	12.50	20.83	21.75

Table 6.1. The number of omission and commission errors for each subject for each bit condition and the total number of trials excluded over all training sessions. The age of the subject is in years.

All				
Variation	Name	Slope Coefficient	ANOVA	Adjusted R^2
Fixed	Wenchang	-0.06	$F(1, 786) = 34.95$, $p < .001$	0.041
Fixed	Odin	-0.10	$F(1, 744) = 69.37$, $p < .001$	0.084
Fixed	Luigi	-0.05	$F(1, 757) = 24.21$, $p < .001$	0.030
Fixed	Wario	-0.08	$F(1, 775) = 48.44$, $p < .001$	0.058
Fixed	Goodall	-0.02	$F(1, 792) = 18.29$, $p < .001$	0.021
Fixed	Darwin	-0.04	$F(1, 778) = 11.71$, $p < .001$	0.014
Fixed	Mean	-0.06		0.041
Random	Athena	0.00	$F(1, 862) < 1$, $p = .929$	-0.001
Random	Thoth	-0.01	$F(1, 835) < 1$, $p = .422$	0.000
Random	Bowser	0.00	$F(1, 861) < 1$, $p = .883$	-0.001
Random	Waluigi	0.00	$F(1, 859) < 1$, $p = .574$	-0.001
Random	Vonnegut	-0.02	$F(1, 849) = 2.45$, $p = .118$	0.002
Random	Hawthorne	-0.02	$F(1, 799) = 2.05$, $p = .153$	0.001
Random	Dickinson	0.01	$F(1, 801) < 1$, $p = .557$	-0.001
Random	Mean	-0.01		0.000
Subtle	Itzamná	0.02	$F(1, 843) = 3.54$, $p = .06$	0.003
Subtle	Herriot	0.01	$F(1, 861) = 9.84$, $p = .002$	0.010
Subtle	Cousteau	-0.01	$F(1, 815) < 1$, $p = .395$	0.000
Subtle	Durrell	0.00	$F(1, 840) < 1$, $p = .987$	-0.001
Subtle	Jubilee	0.02	$F(1, 832) = 3.27$, $p = .07$	0.003
Subtle	Mean	0.01		0.003

Table 6.2. Results from the simple linear regression of bit condition on median RT for all subjects collapsed across all training sessions.

Chapter 7: A Cognitive Test Battery to Assess General Intelligence in the Pigeon (*Columba livia*)

Abstract

A well replicated result in humans is that performance positively correlates across a wide variety of tasks. Factor analysis consistently extracts one factor that can account for approximately half of the variance in performance. This factor is termed *g* and all cognitive tasks positively load onto this factor. Some neurobiological correlates of *g* have been identified in humans, but causal experiments are not yet possible. Causal neural manipulations are possible in animal models and recently, the potential for *g* in animals has been investigated. When mice and some avian species are assessed with cognitive test batteries, performance positively correlates and the first component extracted has similar properties to *g*. There are some limitations to the species tested thus far, including comparability in the cognitive domains assessed across species and homogeneous samples. The pigeon is an ideal subject to overcome these issues since pigeons, humans, and other primates are frequently given similar tasks and many neural correlates of performance have been identified in the pigeon. We created a test battery that assessed different domains, including associative learning, memory, cognitive flexibility, and reaction time. Yet we did not consistently extract a *g* like factor. Analyses indicated a two-component structure and with differential task loadings. The components seemed to reflect an associative learning/memorization versus general rule task demand. Reasons and implications for this two-component structure are discussed.

Introduction

Does performance on a vocabulary test give any meaningful indication on how someone will perform on a math test? Or how quickly they react to a change in an array of stimuli? Or

how well they can mentally rotate an image? Surprisingly, the answer is yes. Performance across all of these tasks shows a positive correlation; if a person performs well on one task, they are likely to perform well in another. This effect has been replicated many times, but the most compelling results are from full scale intelligence quotient (FSIQ) tests because of the number and variety of tasks used. The exact number and type included vary from test to test, but they typically include 11-17 measures that assess memory, basic math, spatial reasoning, and analogical reasoning (Carroll, 1993; Johnson et al., 2004). Despite differences in the test batteries and variety of tasks used, a positive correlation matrix is found (Carroll, 1993; Johnson et al., 2004). When variable-reducing techniques, like principal component analysis (PCA) or factor analysis, are applied to this positive correlation matrix, one factor is consistently extracted that can account for approximately half of the variance (Carroll, 19993; Deary, 2000). All cognitive tasks positively load onto this factor, meaning the factor can account for variance in performance in the task (Carroll, 1993; Deary, 2000). Because this factor is seemingly related to all cognitive abilities, it is referred to as *g* (Spearman, 1904).

g is extracted with a variety of test batteries in a variety of samples, making it one of the most well-replicated results in psychology (Carroll, 1993; Deary, 2000; Johnson et al., 2004). Despite the ubiquity of *g*, there are still important parameters to consider when creating a test or test battery to extract this factor. As mentioned earlier, almost all cognitive tasks load onto *g*, but some tasks have a higher loading than others. The highest loadings on the *g* factor will be found when tasks are complex, novel, and require reasoning, irrespective of the task content or method of delivery (Jensen, 1992; Quiroga et al., 2019). Even though we know what kinds of tasks load highly onto *g*, no task is a perfect or pure measure of any specific cognitive construct. All cognitive tasks assess *g* to some degree, but they also assess more specific abilities (Gignac,

2015). Using a large and diverse battery of tests will attenuate task-specific variance, resulting in a more accurate *g* factor (Major et al., 2011). Another issue with all measures is random error, variables unrelated to the construct of interest that impact performance on the task. Random error can cause performance to be different when participants complete the measure at different time points or respond differently to theoretically similar items. If a measure produces similar results, despite random error, it is referred to as a reliable measure (John & Benet-Martinez, 2000). Reliable measures are crucial because more of the variance in performance across individuals is due to differences in the actual cognitive ability the task is measuring as opposed to differences caused by random error (Bray et al., 1998). Reliability also impacts the correlation matrix. Less reliable measures will artificially lower the correlations, which impacts the extraction of the *g* factor (John & Benet-Martinez, 2000). While it is important that measures are reliable, they also need to be sensitive enough to detect individual differences across people. The *g* factor accounts for variance in performance across people, therefore the tasks used should show variability based on true differences in cognitive ability (Hedge et al., 2018). Extracting a robust *g* factor depends on using appropriate tasks *and* on the sample that is being assessed. Highly homogeneous samples of human participants may not have true differences in the construct of interest which reduces variance and attenuates the subsequent *g* factor (Sackett & Yang, 2000). It is also best to test many participants due to how correlational and factor analyses are conducted. While a sufficient sample size to detect a reliable correlation depends on a variety of parameters, the sample size required can be in the hundreds (Bonett & Wright, 2000), and for factor analysis a sufficient sample size ranges from as few as 75 to as many as 1,200 participants (Mundfrom et al., 2005, but see de Winter et al., 2009). To summarize, to extract the strongest *g* factor, a large,

heterogeneous sample of people should be given a large variety of cognitive tasks, that are reliable and sensitive to individual differences.

While *g* is consistently replicated, it is still not clear what exactly *g* is. It is tempting to use intelligence and *g* interchangeably, since *g* can be extracted using FSIQ tests and is related to a wide variety of cognitive tasks. Yet *g* only accounts for half of the variance in performance on FSIQ, which means other factors besides *g* are related to performance. In addition, the amount of variance *g* can account for relies on the strength of the correlation matrix (Jensen, 1998). Individuals who perform better on intelligence tests tend to have *weaker* correlations between tasks, meaning that higher FSIQ score comes with increased differentiation of abilities (Blum & Holling, 2017). For higher performing individuals, *g* explains less of the variance in performance (Jensen, 1998). Therefore, it would be incorrect to say that more intelligent people have more *g* (Detterman, 1991). These results indicate that when we refer to intelligence in a colloquial sense, we are referring to more than *g*, even though the two concepts are closely related (Jung & Haier, 2007; Stankov, 2017).

With that distinction stated, understanding *g* is still important given the consistent pattern of correlations across tasks, even among high ability individuals (Blum & Holling, 2017). Even though *g* is a single factor or component, it does not mean it is a single causal entity. Instead, *g* is commonly theorized to be composed of more specific cognitive processes like working memory (WM), short term memory (STM), processing speed, attention, and associative learning (Conway et al., 2002; Deary, 2000; Jensen, 1998; Kaufman et al., 2009; Sheppard & Vernon, 2008). It is likely that the tasks included in FSIQ tests, particularly complex tasks that load highly onto *g*, require support from multiple cognitive domains (Chuderski, 2013). Therefore, *g* could reflect individual differences in how many processes are required to solve a task (Chuderski, 2013).

Another theory suggests that differences in one of these abilities could act as a bottleneck, constraining and weakening the ability of all other cognitive domains to function (Kovacs & Conway, 2019). With this theory, *g* is primarily reflecting differences in one cognitive ability, but it is unclear which cognitive ability. These theories are helpful for understanding the more specific cognitive processes that are involved with intelligence tests and how those processes are used across a large number of tasks (Conway et al., 2002; Deary, 2000; Jensen, 1998; Kaufman et al., 2009; Sheppard & Vernon, 2008). Future research is still needed, however, to fully understand if there is a relationship between these cognitive processes that could impact the positive correlation matrix (Frischkorn et al., 2019).

At the psychological construct level, *g* is related to a variety of cognitive processes. Similarly, *g* and intelligence are correlated with a variety of neurobiological mechanisms, processes, and features, (Deary et al., 2010). Thus far there have been two major lines of research. One focuses on what makes individuals different, for example comparing people who have high IQ scores to people who do not. The most robust result from this line of research is the modest positive correlation between brain size and measures of intelligence (Pietschnig et al., 2015). The other line of research focuses on why performance is positively correlated across tasks, irrespective of individual performance. A variety of methods, including functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and lesions due to accident or stroke, have indicated the importance of the frontal cortex in a wide variety of tasks (Jung & Haier, 2007). The dorsolateral prefrontal cortex in particular is active during a variety of WM and reasoning tasks, though similar patterns of activations in different areas of the frontal cortex for other types of tasks have also been identified (Colom et al., 2013). Yet, brain areas do not function in isolation; rather, different areas are connected, forming functional networks (van

den Heuvel & Sporns, 2017). Similar performance across tasks may be partially due to how whole networks are activated by tasks instead of discrete regions. A network connecting the frontal and parietal cortex is implicated (Jung & Haier, 2007; Zanto & Gazzaley, 2013). Thus, most research has indicated that consistent performance could be due to activation of the frontal cortex over a wide variety of tasks (Colom et al., 2013; Jung & Haier, 2007)

Important neurobiological correlates of *g* and intelligence have been identified in humans, but the techniques used thus far fail to support causal interpretations. Nonhuman animal models (hereafter animals) would be ideal to explore causal manipulations, but it first needs to be established that animals have a *g* factor similar to what is seen in humans (Matzel et al., 2013). Investigations over the past 20 years have generated promising results that are described in more detail elsewhere (Burkart et al., 2017; Flaim & Blaisdell, 2020; Shaw & Schmelz, 2017), but some key results from mice and avian species will be briefly reviewed here. For mice, Matzel and colleagues in particular have been consistently exploring a general factor using a cognitive test battery that targets different domains of learning (Matzel et al., 2003). Briefly, the test battery includes five tasks and measures non-spatial navigation (Lashley III maze), spatial navigation (Morris water maze), suppression of exploratory behavior to avoid an aversive audiovisual stimulus (passive avoidance), using odor to guide a response (odor discrimination), and using an auditory cue to predict an aversive shock (associative fear learning). Multiple experiments found that performance was positively correlated across all tasks, and the first factor extracted could account for 38-43% of the variance in performance (Kolata et al., 2005, 2007; Matzel et al., 2003). For these individual experiments however, the number of subjects ranged from 21-56, which is smaller than what is typically used or recommended in human studies (Mundfrom et al., 2005, but see de Winter et al., 2009). When the results from multiple

experiments were combined to have a total of 241 subjects, the result was replicated, providing robust evidence for a *g* like factor in mice (Kolata et al., 2008). Subsequent experiments have shown that performance on this cognitive test battery is positively correlated with measures of WM, similar to what is seen in humans (Kolata et al., 2005). Investigations with avian species have also yielded interesting results. Cognitive test batteries that typically include motor learning, color discrimination, reversal learning, spatial memory, and inhibitory control, have been administered to robins (Shaw et al., 2015), spotted bowerbirds (Isden et al., 2013), magpies (Ashton et al., 2018), and song sparrows (Anderson et al., 2017; Boogert et al., 2011). For robins and spotted bowerbirds, performance across the tasks was mostly positively, though not significantly, correlated, and a factor that could explain 34% and 44% of the variance in performance, respectively, was found (Isden et al., 2013; Shaw et al., 2015). This result should be treated with some caution since a small number of subjects, 16 robins and 14 bowerbirds, were assessed. More robust results have been obtained with magpies, which assessed 56 subjects. The subsequent correlation matrix was uniformly and significantly positive and the subsequent factor extracted accounted for 64% of the variance (Ashton et al., 2018). Yet, similar results were not found in song sparrows, even though 52 (Boogert et al., 2011) and 41 (Anderson et al., 2017) birds were assessed using the same test battery. Across both experiments, two factors were extracted, and not all tasks loaded onto the first factor extracted. This may be due to the low reliability in performance across years on cognitive tasks in song sparrows (Soha et al., 2019). While the results from animals thus far are interesting and promising, there are some difficulties in comparing *g* across species.

Research with many species thus far indicates that *g* can be found beyond humans, but it is not clear exactly how similar *g* is across species. This ambiguity is partially due to the

differences in test batteries across species. In humans, *g* has been heavily investigated in relationship to processing speed, where more intelligent individuals are consistently faster on simple tasks (Sheppard & Vernon, 2008), yet this has not been investigated or replicated with animal test batteries (see discussion by Flaim & Blaisdell, 2020). In contrast, the relationship between response inhibition and *g* has rarely been investigated in humans, but response inhibition tasks are almost always included in avian cognitive test batteries (Flaim & Blaisdell, 2020). Even when the cognitive domain does overlap, there are differences in the procedures used for humans versus non humans that can impede comparisons. Taking associative learning as an example, in humans an initial investigation using a simple associative learning task, where children had to learn which picture was associated with a reward, was not related to IQ scores (Plenderleith, 1956). More recent investigations have used the word-pairs task, where participants learn up to ten arbitrary pairs of words, like cat-pie, or the three-term contingency task where one word serves as a cue and the participant must learn three response words (Kaufman et al., 2009). These more complex associative learning tasks show a positive relationship to *g* that scales with complexity, where the more complex task, the three-term contingency, has a stronger relationship with *g* (Tamez et al., 2008; Williams & Pearlberg, 2006; but see Kaufman et al., 2009). In contrast, for mice and birds, a simple associative learning task, such as learning how to discriminate one cue from another to obtain a food reward, is related to the *g* like factor extracted in these species (Flaim & Blaisdell, 2020). The finding that associative learning is related to *g* across species, but different levels of difficulty are needed to reveal such a relationship, may be related to the experience of the subject. *g* is related to complexity, but it is also related to novelty, where novel tasks tend to have a high *g* loading (Carroll, 1993; Sternberg & Gastel, 1989). If animal subjects are naïve to any highly artificial experimental stimuli and

procedures, the task may be sufficiently novel to explain why performance is related to *g*, despite the apparent simplicity. In contrast, when many humans are assessed, they have had years of experience in an educational setting with similar materials and task demands as the word-pairs and three-term contingency tasks. Therefore, for humans, task difficulty may be a more important factor for investigating associative learning and *g*. These results could indicate that task loading onto *g* is related to novelty, complexity, and associative learning across species, but further research is necessary to determine if there is a similar relationship between complexity and *g* in animals.

While animals may be relatively naïve to cognitive assessments compared to most human samples, there are other issues when comparing across species. In nonhuman research on *g*, the sample of animals tested is often homogeneous in some way (Shaw & Schmelz, 2017). In mice thus far, only male subjects have been used and all subjects have the same 'home' environment (Kolata et al., 2008), while for the wild bird subjects, like robins, collection is biased towards males and bold individuals (Shaw & Schmelz, 2017). If *g* is a robust phenomenon in animals, then it should replicate across all members of the species, but this has yet to be shown. In addition, most experiments assess a small number of subjects. This can be overcome by using a consistent test battery, which makes it possible to pool results from multiple experiments, as demonstrated by Kolata et al. (2008). Utilizing a species that is more commonly investigated, either in the lab or across field sites, could also increase the number of subjects if multiple labs are willing to work together (Shaw & Schmelz, 2017). Thus, there could be improvements in both the test battery and sample characteristics, particularly for avian species. Tasks that assess clear cognitive domains, facilitate cross species comparisons, and have identified neural

correlates should be favored. Species for which it is possible to obtain a large and diverse sample should also be favored, at least in these preliminary investigations of *g* in animals.

Given these arguments, it is surprising that pigeons have not been given a comprehensive test battery, given their long history as an animal model in psychology. Pigeons have excellent visual acuity and readily learn to peck visual stimuli in a touchscreen operant chamber, similar to procedures used to assess human and nonhuman primates (Wright et al., 2018; Zentall, 2020). Investigations of matching, timing, reaction time, memory, and many other cognitive domains show there are similarities in performance across pigeons and primates that indicate similar underlying mechanism at the psychological and neurobiological level (Colombo & Scarf, 2020; Güntürkün, 2005; Vickrey & Neuringer, 2000; Zentall, 2020). Methods for investigating memory, associative learning, and cognitive flexibility in particular have been well established, and the neural mechanisms supporting performance have been identified on some level. Similar to humans, performance on many cognitive tasks seems to depend on nidopallium caudolaterale (NCL) which is the avian equivalent to the mammalian prefrontal cortex (Güntürkün, 2005). For example, when assessing STM in the pigeon by requiring pigeons to remember a stimulus over a short delay to guide choice behavior, there is sustained neural activity in the NCL that relies on the neurotransmitter dopamine, similar to results found in nonhuman primates (Johnston et al., 2017; Karakuyu et al., 2007). Given this rich history, there are many tasks that could be included in a cognitive test battery for pigeons, but a few were selected as ideal.

The tasks in the battery developed here were selected according to how well they assessed a specific cognitive domain, if the task facilitates cross-species comparisons, and if the neural substrates of performance had been identified (Diekamp et al., 2000; Flaim & Blaisdell, 2020; Izquierdo et al., 2017; Johnston et al., 2017; Karakuyu et al., 2007; Lissek et al., 2002;

Vickrey & Neuringer, 2000). Ultimately, the pigeon cognitive test battery was designed to assess associative learning, cognitive flexibility, memory, and processing speed. Specifically, there were five tasks, matrix displays, symbolic match to sample (SMTS), serial reversal learning, delayed match to sample (DMTS), and a reaction time (RT) task. While the matrix displays task was intended to assess abstract learning ability, results from a transfer test did not indicate that any subject had learned an abstract rule, but rather had memorized which stimuli were associated with food. Therefore, the matrix displays task was considered an additional assessment of associative learning. All the tasks were sufficiently sensitive to detect individual differences in performance, and all subjects completed at least two tasks in the battery (Table 7.1). Surprisingly, the correlation matrix from the test battery was not uniformly positive, and PCAs did not consistently yield a component similar to *g*. Potential procedural issues, the influence of age and experience, and the possibility that these results reflect a genuine difference between pigeons and other species are discussed.

Method

Subjects

Twenty-three pigeons served as subjects. The age at the start of the test battery ranged from 0.5-17 years old and ten were female. Subjects varied in how much experience they had with other cognitive tasks (Table 7.1). Subjects were individually housed in steel home cages with metal wire mesh floors in a vivarium. They were maintained at 80% of their free-feeding weight, but were allowed free access to water and grit while in their home cages. Testing occurred at approximately the midpoint of the light portion of the 12-hour light-dark cycle. All procedures were approved by the UCLA Institutional Review Board.

Apparatus

Testing was conducted in a flat-black Plexiglas chamber (38 cm wide x 36 cm deep x 38 cm high). All stimuli were presented by computer on a color LCD monitor (NEC MultiSync LCD1550M) visible through a 23.2 x 30.5 cm viewing window in the middle of the front panel of the chamber. The bottom edge of the viewing window is 13 cm above the chamber floor. Pecks to the monitor were detected by an infrared touchscreen (Carroll Touch, Elotouch Systems, Fremont, CA) mounted on the front panel. A custom-built food hopper (Pololu, Robotics and Electronics, Las Vegas, NV) was located in the center of the front panel, its access hole flush with the floor. The food hopper contained a mixture of leach grain pigeon pellets and seed (Leach Grain and Milling). All experimental events were controlled and recorded with a Pentium III-class computer (Intel, Santa Clara, California). Stimuli were presented using the 2.7.11 version of Python with the psychopy toolbox, version 3.0.3 (Peirce, 2007).

Procedure

All tasks have been described in detail elsewhere, but are described briefly here to emphasize the features that are relevant to the dependent measures ultimately included in the battery. Subjects did not receive the test battery in the same order and it was not possible to fully counterbalance for order effects. Additionally, the time between tasks was not consistent across subjects. Subjects received one session per day, 3-7 days a week. All tasks were appetitive and used 3-s of access to a mixture of grain and seed as a reward.

Matrix displays.

The goal of the matrix displays task was to detect differences in the ability to use a relational size change rule that could be flexibly applied to novel stimuli. Eleven subjects were

trained to discriminate between displays that showed a change in size from displays that did not. Specifically, the display was a 2x2 matrix that had a pair of shapes in the row or column of the matrix. Reinforced displays had the same shape in the same color, but one of the shapes was a different size (Chapter 2, Figure 2.2d). Nonreinforced displays either had an identical pair of shapes that were the same size or had a set that differed in both shape and color, but were the same size (Chapter 2, Figure 2.2e, f). The shapes could be a rectangle or triangle, in red or blue, in the row or column of the matrix. The key difference is the reinforced displays have a change in size, while the nonreinforced displays do not. During the task, subjects were presented with two displays to the left and right of the midline. One display was always reinforced, presented equally often on the left and right side of the screen and equally often with each type of nonreinforced display. Subjects had to make four consecutive pecks (FR4) to one of the displays to end the trial. Completing the peck requirement to the reinforced display resulted in a food reward, while pecks to the nonreinforced display simply ended the trial. Subjects were trained until they reached criterion, 80% accurate on two consecutive sessions on both types of nonreinforced displays, or until they had trained for 100 sessions. Once they reached criterion, they received probe trials where the display could contain a novel shape, novel color, or both. No subject transferred to the novel displays, indicating that subjects did not learn an abstract rule that could be flexibly applied to novel stimuli. The similar performance across all subjects also indicates that transfer performance is not sufficiently sensitive to detect individual differences. The number of sessions to reach criterion, however, was variable across subjects and was included in the battery as a measure of associative learning (Table 7.1).

Symbolic match to sample.

The SMTS was another measure of associative learning and was conceptually based on the arbitrary word-pairs task given to humans (Kaufman et al., 2009). Instead of words, 18 subjects were presented with pictures of foods and animals which were paired together through reinforcement. Subjects were trained on four pairs of pictures, where one of the pictures was always a food item while the other picture was always an animal, obtained from the food-pics database (Blechert et al., 2014). To associate pairs of pictures, trials had two phases, a sample phase and a choice phase. In the sample phase, one picture was shown in the center of the screen and is referred to as the sample. When subjects completed an observing response to the sample, pecking the picture ten times (FR10), the choice phase began. In the choice phase, the sample remained on the screen, and two comparison stimuli were presented below the sample on the left and right of the screen (Chapter 3, Figure 3.2). If subjects pecked the correct comparison once (FR1), they received a food reward. If they pecked the incorrect comparison, a correction procedure was used where the trial repeated starting at the sample phase. Correction trials were not included in the analysis. The correct comparison was presented equally often on the left and right side of the screen and equally often with the three other incorrect comparison stimuli. Subjects were trained with this procedure until they were 80% accurate on each pair in a single session or until they had trained for 35 sessions. The number of sessions to reach criterion was used to measure associative learning ability (Table 7.1).

Serial reversal learning.

The serial reversal learning task was used to assess cognitive flexibility, being able to update behavior to reflect changes in environment (Izquierdo et al., 2017). Twenty-three subjects were trained with two stimuli, a blue or yellow circle, presented on the left and right side of the

screen. One of the stimuli was always followed by a food reward (S+), while the other was not (S-). Subjects had to peck the stimulus three times (FR3) to indicate their choice and end the trial. When subjects were selecting the reinforced stimulus with 90% accuracy on two consecutive sessions, on the next session the contingency was reversed. Now the stimulus that was the S-, was now the S+ and vice versa. Subjects were trained on five reversals and performance on the first session of each reversal was used as to measure cognitive flexibility (Table 7.1). Due to computer errors, performance on the first and second reversal was only measured for 21 subjects and performance on the fifth reversal was only measured for 22 subjects.

Delayed match to sample.

The DMTS task was used to assess memory, or more specifically the ability to maintain a memory of a stimulus over a short delay (Kangas et al., 2011). Eighteen subjects were trained using a procedure similar to the SMTS task. There are three key differences between the DMTS and the SMTS task, the size of the stimulus set, what determined the correct comparison, and the delay period. Subjects were trained with two stimuli, a red circle and a green circle. First subjects were trained without a delay so the rule determining which comparison was correct could be learned. Similar to the SMTS, during this initial training each trial had two phases, a sample phase and a choice phase. During the sample phase, one stimulus was presented in the center location. After subjects completed the observing response (FR10) to the sample, the choice phase began and the two comparison stimuli were presented. One of the comparisons matched the sample, while the other did not. If subjects pecked (FR1) the matching comparison, they received a food reward. If they pecked the nonmatching comparison, a correction procedure was used where the trial repeated starting at the sample phase. The correction trials were not included in

the analysis. Each stimulus served as the sample an equal number of times and the correct comparison was presented equally often on the left and right side of the screen. When subjects reached criterion on the initial training, 80% correct on two consecutive sessions, they stated the DMTS.

In the DMTS, each trial had three phases, sample, delay, and choice (Chapter 5, Figure 5.1). The sample phase is the same as the training described above. Once subjects had completed the observing response, the sample was removed from the screen, and the delay phase began. The delay could be 0, 2, 4, or 8 s, and each delay length was presented an equal number of times with each comparison. After the delay had elapsed, only the comparison stimuli were presented. Again, if subjects pecked the matching comparison, they received a food reward. If subjects pecked the nonmatching comparison, the trial ended without reinforcement. The correction procedure was not used during the DMTS. Subjects were trained on this procedure for 30 sessions. Accuracy over the last three sessions for all delay lengths was used to assess performance.

Reaction time task.

This task was used to assess speed with a procedure that relied on detecting a change in the stimulus display (Sheppard & Vernon, 2008). Twenty-two subjects were trained and took part in four different experiments with slight procedural differences overall (n = 4 for one experiment, n = 6 for the remaining experiments), but the procedure described here was the same across experiments. The stimulus display consisted of nine circular stimuli that could either be a white outline or completely filled with white. One of the stimuli was in the center of the screen, near the bottom of the viewing window. This stimulus was the home key. The remaining eight stimuli were arranged in a semi-circle around the home key (Chapter 6, Figure 6.1h) and were

potential targets (PTs). Trials had two phases, a home key phase and a choice phase. During the home key phase, the home key was filled with white and subjects had to peck the home key three times on average (VR3 +/- 2). When subjects completed the peck requirement to the home key, the choice phase began. During the choice phase, the home key became a white outline and one of the PTs was filled with white and became the target (Chapter 6, Figure 6.1i). If subjects pecked (FR1) the target, they received a food reward. If subjects pecked a PT it was counted as an error of commission, but if subjects did not peck the target or a PT within 5 s, it was counted as an error of omission. Only trials where subjects successfully pecked the target were included in analysis. The target appeared in each location an equal number of times. Each session either had 20 (n = 10) or 24 trials (n = 12) with eight PTs, and subjects were trained for 10 or 9 sessions respectively. Median reaction time (RT) collapsed across all sessions was used as the dependent measure. The number of trials included in the analysis for each subject ranged from 184-216 depending on how many trials had to be excluded due to errors.

Data Analysis

All dependent measures from the tasks were selected because they reflected differences in cognitive ability, but the tasks do not indicate these differences in the same direction or on the same scale. For the serial reversal learning and DMTS, better performance is indicated with a larger number and worse performance is indicated with a smaller number. For the matrix displays, SMTS, and RT time task, performance is coded in the opposite direction, where a larger number indicates worse performance. For these three tasks, the data was reverse coded, so all results were in the same direction where a larger number indicates better performance. Additionally, the scale of the dependent measure differs across tasks. For example, the matrix displays task ranges from 18-100, while the serial reversal learning and DMTS tasks are bound

between 0-1. Since the range of the data can impact correlations, the data were normalized (observed value – mean value / the standard deviation) so the range was more similar across tasks.

For the DMTS, where performance was collapsed across three sessions, the test-retest reliability of the performance was assessed with a Pearson correlation. Similarly, a Pearson correlation was also used to assess performance over different conditions within a task. This analysis was used to compare performance for the different delay lengths in the DMTS task and the different reversals in the serial reversal learning task. To assess reliability for the RT time task, where performance was collapsed across 9 or 10 sessions, a Pearson correlation was used to compare performance at the beginning and end of training. For the RT time task, because subjects experienced different procedures, an ANOVA was also used to determine if there were any significant differences in median RT across the experiments.

A Pearson correlation with a Bonferroni correction was used to compare performance across tasks. Age and experience were included as potential variables that were related to performance (Table 7.1). PCAs with an unrotated factor solution was used to determine if variance in performance in the test battery could be accounted for with a single component. All statistical analyses were performed using SPSS version 27.

Results

Individual Cognitive Tasks

Matrix Displays

The number of sessions need to reach criterion, 80% accuracy on two consecutive sessions, was used as the dependent measure. The number of sessions subjects required ranged

from 18-100 (*M* = 48.27, *SD* = 24). The data were reverse coded so a larger number indicated better performance and then normalized.

Symbolic Match to Sample

The number of sessions need to reach criterion, 80% accuracy on all four pairs in a single session, was used as the dependent measure. The number of sessions required ranged from 8-35 (*M* = 20.5, *SD* = 9.13). The data were reverse coded so a larger number indicated better performance and then normalized.

Serial Reversal Learning

Performance on the first session of the initial discrimination and each reversal was used as the dependent measure. Performance on the initial discrimination and the first and second reversals had weak and nonsignificant correlations with all measures. Performance on the third and fifth and fourth and fifth reversals were significantly, positively correlated with each other. While performance on the third and fourth reversal was also strongly, positively correlated, this did not survive a Bonferroni correction (Table 7.2). This indicated that the initial discrimination and the first and second reversals were not assessing the same ability or cognitive domain as the third, fourth, and fifth reversals. In addition, previous research (Chapter 4) indicated that the third, fourth, and fifth reversals were more sensitive to individual differences. Due to the strong correlations between performance on the fourth and fifth reversals, an aggregate measure of performance was created by collapsing across the two conditions. While performance on the third and fifth reversals also had a strong correlation, this may have been due to both conditions having the same reinforced stimulus.

Delayed Match to Sample

Performance collapsed over the last three sessions (28, 29, and 30) of training was used as the dependent measure. For one subject, Goodall, the 27, 28, and 29th sessions were used due to a computer error in the 30th session. Performance with the 2-second delay was the most reliable with an average correlation of .809 across the last three sessions. Performance with the 4-second delay was also fairly reliable, with an average correlation of .713. Performance at the 0- and 8-second delays were not as reliable, with an average correlation of .442 and .554 respectively. Performance on the 2-second and 4-second delay were also significantly, positively correlated with each other (r = .783, n = 18, p < .001), thus an aggregate measure of performance was collected by collapsing across the two conditions.

Reaction Time

Performance collapsed over all sessions of training was used as the dependent measure. Only trials where subjects correctly pecked the target were used in the analysis. Reliability was assessed by comparing the mean median RT in the first and last 3 sessions of training. Performance was very reliable, with a correlation of .922 and previous analyses indicated there were no significant differences in performance due to training (Chapter 6). Therefore, the median RT collapsed across sessions of training were used in the analyses. A one-way ANOVA confirmed there were no statistically significant differences in median RT based on which experiment a subject experienced (F < 1).

Cognitive Test Battery

Correlation matrix

A Pearson correlation was used to determine if there was any relationship in performance across the different cognitive tasks. Before the analysis, the data from the matrix displays, RT,

and SMTS tasks were reversed coded so for all tasks a larger number indicated better performance. The data were also normalized so there would be a similar range in scale across all tasks. Tasks with multiple dependent measures were analyzed in two ways, with the aggregate measure where performance was collapsed across highly correlated conditions as described above, and with the individual task performance that composed the aggregate measure. For the DMTS, performance on the 2- and 4-second delay were included, and for the reversal learning task, performance on the fourth and fifth reversal were included in the analyses. Finally, age and experience were also included in the analysis since they could have a potential relationship with performance.

The correlation matrix with the aggregate measures was not uniformly positive across the cognitive tasks (Table 7.3). Reversal learning was positively correlated with the matrix displays and DMTS tasks, yet the matrix display task and DMTS were not correlated with each other. Performance on the SMTS and RT were positively correlated, but had weak to negative correlations with the other tasks. None of these correlations survived correction ($\alpha = .003$). The average correlation between the cognitive tasks was .224.

Age and experience had a strong, significant positive correlation with each other and thus had a similar relationship with the cognitive tasks. Age and experience were negatively correlated with almost all of the cognitive tasks, though only the correlation between age and the SMTS was significant at the conventional level. Only the DMTS had no relationship with age or experience. A partial correlation controlling for age had little effect on the correlations between the cognitive tasks (Table 7.3).

Similar results were obtained when using the individual measures of performance for the reversal learning and DTMS tasks (Table 7.4). Performance on the fifth reversal had a stronger

correlation with the matrix displays task and 2- and 4-second delay from the DMTS task compared to the fourth reversal. Performance on the 4-second delay had more positive correlations with the other cognitive measures compared to the 2-second delay. The average correlation between tasks across the matrix was .258. None of the correlations *across* tasks were significant after a Bonferroni correction ($\alpha = .002$). Similar to the analysis with the aggregate measures, age and experience were negatively correlated with all cognitive measures, except for the 4-second delay condition. Controlling for age did not substantially alter the correlations between the cognitive tasks (Table 7.4).

Principal component analysis

To maximize the number of subjects in the PCA, performance on the matrix displays task was not included in the analysis. A total of ten PCAs were conducted. To investigate the effect of missing data, PCAs were performed using pairwise and listwise deletion (Dray & Josse, 2015). For listwise deletion, this meant that only the subjects that had completed all four tasks ($n = 15$, Table 7.1), were included in the analyses. Since some of the tasks had multiple dependent measures or an aggregate measure, five PCAs were conducted to ensure that all tasks were equally represented and to determine if the component structure was robust. Finally, because the data were normalized, the PCA was based on the covariance matrix (Jolliffe & Cadima, 2016) and was unrotated so the first component could account for the maximum amount of variance. By convention, only Eigenvalues larger than 1 were retained.

For all the PCAs, two components with Eigenvalues larger than 1 were extracted (Table 7.5). For the analyses using pairwise deletion, the task loadings depended on which measure of reversal learning was included in the analysis. When the aggregate measure and fifth reversal was included, all tasks positively loaded onto the first component, which could account for 41.93

– 44.28% of the variance in performance. While the aggregate measure of reversal learning, fifth reversal, and DMTS positively loaded onto the second component, the SMTS and RT negatively loaded onto this component, which could account for 31.64 - 35.3% of the variance in performance. In contrast, when the fourth reversal was included, the DMTS measures only had a positive loading on the second component. Interestingly, when the fourth reversal and 4-second delay were included in the analysis, the first and second components could account for a similar amount of variance in performance (38.33 and 37.41% respectively), and all tasks positively loaded onto the second component.

For the analyses using listwise deletion, the five PCAs showed a similar pattern, where performance on the SMTS and RT positively loaded onto the first component, which could account for 39.87-44.5% of the variance, while performance on the reversal learning and DMTS positively loaded onto the second component, which could account for 33.52-40.76% of the variance. While the DMTS never positively loaded onto the first component, there was more variation with the measures of reversal learning. The fifth reversal had positive loadings onto the first component, though this was weaker than the loading on the second component. In contrast, the fourth reversal never positively loaded onto the first component (Table 7.5).

Discussion

This was the first time that cognitive performance in the pigeon has been systematically investigated using a test battery. The battery was created with the intention of assessing different cognitive domains, including associative learning, cognitive flexibility, STM, and RT. We predicted that performance would positively correlate across tasks, and that the matrix displays and SMTS tasks may show a stronger correlation with each other since they seemed to rely on associative learning. In contrast to our predictions, the correlation matrix was not uniformly

positive and there were indications that tasks were forming two clusters. The matrix displays, reversal learning, and DMTS tasks showed stronger positive correlations, while performance on the SMTS and RT tasks were only positively correlated with each other. While age and experience were negatively related to almost all cognitive tasks, controlling for age did not have a substantial impact on the overall pattern of correlations between the cognitive tasks.

The PCAs only partially confirmed the clustering seen in the correlation matrix. When pairwise deletion was used to handle missing data, all tasks had a positive loading on the same factor in four out of the five PCAs, which could be interpreted as evidence for a *g* like factor. Yet even within these analyses, there is evidence for a divide between the SMTS and RT tasks and the reversal learning and DMTS tasks. This is clear when listwise deletion was used, where there was no evidence for *g*, instead the cognitive tasks differentially loaded onto the two components. The SMTS and RT tasks always had a positive loading onto the first component and no, or a negative loading on the second component, with two exceptions. The reversal learning and DMTS measures had variable loadings on the first component, but always positively loaded onto the second component. A negative loading indicates that these tasks capture the opposite of what the second component represents (Bro & Smilde, 2014). For the DMTS, performance always had strong positive loading onto the second component. When the DMTS had a positive loading on the first component, it was always weaker than the loading on the second component, and it was more common for the DMTS to not load onto the first component or have a negative loading. For the serial reversal learning task, the loading was more variable depending on if performance on the fourth or fifth reversal was included in the analyses. The fourth reversal had a variable loading on the first component, ranging from weakly positive to weakly negative. The fifth reversal always had a positive loading onto the first component, but which component had a

stronger loading depended on how the missing data were handled. For the pairwise deletion, there was a stronger loading on the first component, but with the listwise deletion, there was a stronger loading on the second component. Listwise deletion intensifies the divide between the SMTS and RT tasks primarily reflecting one component while the reversal learning and DMTS tasks reflected another. This divide is present to a smaller degree when pairwise deletion is used. While listwise deletion is viewed as less reliable since the analyses are conducted with a smaller sample size, for these data set there is not an extreme difference in sample size for listwise (n = 15) compared to pairwise (n = 16-21) deletion (Van Ginkel et al., 2014). In addition, the correlation matrix supports separate clusters of cognitive tasks. Therefore, despite some evidence for a *g* like factor using pairwise deletion, it is worthwhile to explore what these two components represent.

Why do the tasks in the battery show these patterns of results? It could be due to an overall difference in strategy that could be used for the two sets of tasks. In the matrix displays, serial reversal learning, and DMTS tasks, it was possible to use a more general rule-based strategy. For the matrix displays task, all of the reinforced displays had a change in size, for serial reversal learning subjects could use a 'win-stay, lose-shift' rule, and for DMTS subjects could, as the name implies, use a matching rule. For the SMTS and RT tasks, it was not possible to use a rule-based strategy. In the SMTS task, the pairs were selected because they did not have any consistent perceptual feature that could be used to guide choice behavior. In the RT task, the location of the target was pseudorandomized and it would have been difficult for subjects to predict where it would appear. Therefore, one interpretation of the components would be tasks that rely on memorization or associative learning (component 1) and tasks where it is possible to use a rule-based strategy (component 2). Yet, this is not wholly satisfactory given the difference

in loadings for the fourth and fifth reversals. While the fourth and fifth reversal positively loaded onto the second 'rule-based' component, the fifth reversal always loaded positively on the first 'memorization/associative learning' component as well. In contrast, the fourth reversal rarely positively loaded onto this first component. According to the proposed interpretation, this would indicate that by the fifth reversal, a general rule has less control over performance. Instead of component 2 reflecting a rule-based strategy, it could reflect inhibitory control. Arguably, subjects would perform better on the choice RT task with poor inhibitory control since they could react faster to a change in their environment. Yet poor inhibitory control would likely lead to worse performance on the serial reversal learning task since subjects would need to inhibit their peck response to the previously reinforced stimulus when presented with a new contingency.

It is also possible these two components reflect differences in automaticity. Human research has indicated that skill or task learning occurs in three phases, where the initial learning phase requires effortful cognitive processes while the final phase relies on automated responses (Ackerman, 1988). Progression though these phases partially depends on the complexity and consistency of the task (Ackerman, 1988). Component 1 could reflect tasks that have become automized, while component 2 reflects tasks that are not yet automized. It is consistent with the human literature that the choice RT task would quickly become automated due to its high level of stimulus-response compatibility and consistent task demands. This could also account for the difference between the fourth and fifth reversal in the serial reversal learning task. As training continues, performance is more likely to rely on automatic responses (Ackerman, 1988).

Finally, these components could also reflect differences in which tasks are sensitive to age-related declines. Component 1 reflects tasks that show age related declines in performance,

while component 2 reflects tasks where performance does not change with age. Performance on the SMTS and choice RT tasks was negatively correlated with age, which is consistent with the human literature. Research with humans has demonstrated that associative learning, as measured with the arbitrary word pairs task, is related to processing speed and that performance declines with age (Rast & Zimprich, 2009; Salthouse, 1994). This indicates that pigeons may have similar age-related changes in cognition compared to humans, but additional research is needed. Even though these interpretations are presented individually, they are not mutually exclusive. It is likely that more specialized cognitive abilities, like inhibitory control, influence rate of automation or are also influenced by age-related changes in performance.

While these interpretations are still speculative, the most striking result from the test battery is that there is more evidence for a two-component structure rather than a *g*-like factor. The potential two-component structure contradicts our predictions and many of the previous results with other species (Flaim & Blaisdell, 2020). While the results from this test battery are not a definitive conclusion on the structure of pigeon cognition, it would be helpful to identify which features of the sample or test battery led to these results. There are three primary issues that will be discussed. The first and most obvious issue is that the statistical analyses are underpowered (Bonett & Wright, 2000; Mundfrom et al., 2005). Nevertheless, a *g* like factor has been found in samples of similar size (Isden et al., 2013; Shaw et al., 2015). Thus, while this is an issue across a few animal cognitive test batteries, it doesn't necessarily preclude finding the results we predicted.

The second issue is that the tasks in this cognitive battery required more sessions of training compared to the cognitive test batteries given to other species, including humans (Johnson et al., 2004). For example, it took robins an average of 17 days to complete a five-task

battery (Shaw et al., 2015). In contrast, subjects needed an average of 84.45 days (*SD* = 12.78) to complete all of the following tasks, serial reversal learning, SMTS, DMTS, and RT tasks, not including preliminary training. It could be that the amount of training and time each subject spent on the task causes this factor structure, instead of a genuine difference between pigeons and other species. Yet it is not clear how differences in training could cause this specific cognitive structure or prevent a *g* like factor from being found, particularly when the tasks were specifically investigated for sensitivity to individual differences. Even with the length of training, performance on tasks in the battery did not show ceiling or floor effects (Schubiger et al., 2020; Völter et al., 2018). When possible, the data were also analyzed for reliability and the results indicated that the measures used in analyses were moderately stable over a short time frame. Similarly, positive correlations between multiple measures from the same task indicated that the task was consistently measuring the same underlying construct. While noncognitive traits like persistence were not assessed, many of these tasks were selected because previous research indicated that they were cognitively demanding (Izquierdo et al., 2017; Kaufman et al., 2009; Zentall & Smith, 2016). The least cognitively demanding task, RT, was selected because of the consistent relationship between g and RT in humans (Sheppard & Vernon, 2008). Therefore, it seems unlikely that either the length of training or the tasks themselves could prevent a *g* like factor from being consistently extracted from a PCA.

The third issue is the difference in experience with cognitive tasks that our pigeons have compared to the other species primarily discussed thus far. The mice and avian cognitive test batteries have used experimentally-naïve subjects, which would make the testing apparatus and subsequent experimental conditions very novel. The subjects in this test battery, like many other pigeon research laboratories (and also similar to nonhuman primate research), had different

amounts of experience with other cognitive tasks in the touchscreen operant chamber. This meant that all subjects had been exposed to different stimuli, peck requirements, and reinforcement rates, which reduces the novelty of the tests in the battery (Table 7.1). As mentioned in the introduction, novelty is a factor for why some tasks show a strong *g* loading, but this is not true for all tasks used in human research. For example, vocabulary tests also have a strong *g* loading and primarily rely on retrieving previously learned knowledge under familiar conditions (Colom et al., 2002; Gignac, 2015). In addition, human FSIQ are typically administered using formats commonly found in Western educational settings (Clark et al., 2016), such that the apparatus and general procedure are familiar to most of the participants, similar to the pigeon subjects in this experiment. So, if the prior experience with the apparatus and general experimental procedure prevented a *g* like factor in pigeons, it would have major implications when comparing *g* across humans as well. Ultimately, however, the difference between naïve and experienced subjects is an unlikely explanation for these differences since experience with other cognitive tasks had a relatively weak relationship with performance on the cognitive test battery (Table 7.2, 7.3).

Differences in the test battery and experimental history of the subjects are not satisfactory explanations for why we did not extract a robust *g* like factor. Is there a satisfactory explanation at the species level? While previous research has demonstrated striking similarities in performance between pigeons and primates, the parameters for each species were different (Colombo & Scarf, 2020). Pigeons need more stimuli or examples to show evidence for general rule learning compared to primates. For example, Wright and Katz (2006) have investigated same/different concept learning in Rhesus and capuchin monkeys and pigeons, using similar procedures and identical stimuli. Their criterion for full abstract concept learning is that

performance with completely novel stimuli must match performance with previously trained stimuli. Initially, all subjects were trained with a set of eight pictures until they were 80% accurate, then they were tested with novel pictures. The set size of the pictures doubled if subjects failed to show full concept learning at test. Rhesus and capuchin monkeys had full concept learning with a set size of 128 pictures, whereas most pigeons needed a set size 256 pictures, and one pigeon needed 1024! It should be noted that *none* of the species here showed any evidence for concept learning with the initial set of 8 pictures. So, it is possible that all subjects initially learned the task by memorizing each initial item pair then all gradually transferred to using an abstract matching concept, but pigeons were the slowest to transfer. The obvious question is why are pigeons slower to transfer! The most likely answer is because pigeons are birds, but the type of bird may be an important qualifier. Subsequent investigations have shown that nutcrackers and black-billed magpies, members of the corvid family, show full concept learning with 128 pictures, the same as the monkeys (Wright et al., 2018). It has been stated that pigeon's 'preferred strategy' is the concrete one, where subjects will use each unique stimulus configuration to guide behavior as opposed to a more flexible, stimulus independent rule (Wright, 1997, p. 119). This seems to apply to a wide variety of tasks, but the goal of these experiments is to figure out how to break the concrete strategy (Colombo & Scarf, 2020; Wright, 1997; Wright & Katz, 2006). It would be helpful to understand why pigeons tend to use a concrete strategy over an abstract one, as it may be related to the two-component structure extracted in this experiment.

Further speculations on why pigeons have a different cognitive structure than other species should be resisted until these results are replicated. While there are clear strengths to the tasks used in this battery, there are also weaknesses. Future, stronger test batteries should assess

a broader array of cognitive domains. Spatial reasoning in particular would be an excellent addition since there have been investigations directly comparing the abilities of pigeons and humans in a variety of paradigms (Hollard & Delius, 1982; Spetch et al., 1996). Non-cognitive factors, like persistence and neophobia, should also be assessed. While it seems unlikely to explain the results obtained here, non-cognitive factors have been shown to differentially impact performance on seemingly cognitive tasks and should be explicitly accounted for (Carere & Locurto, 2011; Isden et al., 2013; Shaw & Schmelz, 2017). Finally, and on a more practical note, ideally the entire test battery should take far less time to complete. The amount of time to train and test subjects limits the number of tasks that can be included and how feasible it is for other labs to replicate the results. Investigating *g* across a variety of species could help determine if there are consistent neuroanatomical features present in species that exhibit a *g* factor compared to species that do not. Ultimately this test battery is an interesting step towards understanding the general cognitive abilities of the pigeon. Future investigations are sure to yield insights about the structure of general cognitive abilities across species.

Test Battery			Tasks					
Age	Experience	Name	Matrix Displays	Symbolic Match to Sample	Serial Reversal Learning	Delayed Match to Sample	Reaction Time	Total
17	10	Vonnegut	1	1	1	1	1	5
17	9	Dickinson	1	1	1	1	1	5
3	2	Bowser	1	1	1	1	1	5
3	1	Peach	1	1	1	1	1	5
3	1	Waluigi	1	1	1	1	1	5
3	0	Luigi	1	1	1	1	1	5
3	0	Mario	1	1	1	1	1	5
3	0	Shy Guy	1	1	1	1	1	5
3	0	Wario	1	1	1	1	1	5
17	6	Estelle	0	1	1	1	1	4
16	9	Jubilee	0	1	1	1	1	4
11	7	Herriot	0	1	1	1	1	4
17	9	Hawthorne	1	1	1	0	1	4
11	6	Goodall	0	1	1	1	1	4
0.5	0	Athena	0	1	1	1	1	4
0.5	0	Wenchang	0	1	1	1	1	4
16	9	Gambit	0	1	1	1	0	3
12	11	Darwin	0	0	1	1	1	3
12	6	Durrell	0	1	1	0	1	3
12	5	Cousteau	0	0	1	1	1	3
3	2	Yoshi	1	0	1	0	1	3
1	2	Itzamná	0	0	1	0	1	2
1	1	Odin	0	0	1	0	1	2
		Total	11	18	23	18	22	

Table 7.1. All subjects in the test battery and which tasks they completed, where 1 signifies they completed the task and 0 signifies they did not. Experience refers to the number of cognitive tasks completed before or between tasks in the cognitive test battery.

Reversal Learning						
		Initial	First	Second	Third	Fourth
First	*r*	0.013 (20)	--			
	p	0.957				
Second	*r*	0.031 (22)	0.189 (21)	--		
	p	0.891	0.413			
Third	*r*	-0.05 (22)	0.04 (21)	0.154 (23)	--	
	p	0.825	0.863	0.483		
Fourth	*r*	0.308 (22)	0.083 (21)	0.001 (23)	*.572 (23)*	--
	p	0.162	0.72	0.995	*0.004*	
Fifth	*r*	0.108 (21)	0.001 (20)	0.101 (22)	**.712 (22)**	**.623 (22)**
	p	0.641	0.996	0.656	>**0.001**	**0.002**

Table 7.2. Correlation matrix between the measures of the serial reversal learning task. The number inside the parenthesis is the sample size. Italicized values indicate the result was significant before correcting for multiple comparisons, while bolded values indicate the result was significant after a Bonferroni correction.

Pearson Correlation							
Variable		Matrix Displays	SMTS	Reversal Learning	DMTS	RT	Experience
Matrix Displays	*r*		0.126 (7)	**0.748 (8)**	-.105 (6)	-.207 (8)	-0.256 (8)
	p		0.747	**0.013**	0.804	0.566	0.475
SMTS	*r*	0.31 (10)		.111 (15)	.092 (13)	*.424 (14)*	0.157 (15)
	p	0.383		0.672	0.745	*0.101*	0.548
Reversal Learning	*r*	**0.753 (11)**	0.201 (18)		**0.51 (14)**	.092 (18)	0.309 (19)
	p	**0.007**	0.423		**0.044**	0.7	0.173
DMTS	*r*	-0.1 (9)	0.074 (16)	**0.495 (17)**		-.174 (14)	0.111 (15)
	p	0.798	0.785	**0.043**		0.52	0.67
RT	*r*	-0.016 (11)	**0.524 (17)**	0.163 (21)	-0.166 (17)		0.267 (19)
	p	0.962	**0.031**	0.479	0.525		0.241
Experience	*r*	-0.485 (11)	-0.404 (18)	-0.089 (22)	0.052 (18)	-0.247 (22)	
	p	0.131	0.096	0.694	0.836	0.269	
Age	*r*	-0.43 (11)	**-0.491 (18)**	-0.218 (22)	0.011 (18)	-0.367 (22)	**0.927 (23)**
	p	0.187	**0.038**	0.33	0.964	0.093	**< .001**

Table 7.3. Correlation matrix between the measures of the cognitive test battery, age, and experience. An aggregate measure was used for the reversal learning and delayed match to sample (DMTS) tasks. The values in the lower half of the triangle are zero-order correlations and the number inside the parenthesis is the sample size. The values in the upper half of the triangle are partial correlations controlling for age and the number inside the parenthesis are the degrees of freedom. Bolded values indicate the result was significant before correcting for multiple comparisons and italicized values indicate results that were no longer significant after controlling for age.

Variables		Matrix Displays	SMTS	Fourth Reversal	Fifth Reversal	2 Sec Delay	4 Sec Delay	RT	Experience
Matrix Displays	*r*		0.126 (7)	0.55 (8)	**0.744 (8)**	-0.403 (6)	0.171 (6)	-0.207 (8)	-0.256 (8)
	p		0.747	0.1	**0.014**	0.322	0.686	0.566	0.475
SMTS	*r*	0.31 (10)		0.037 (15)	0.165 (15)	-0.005 (13)	0.179 (13)	*0.424 (14)*	0.157 (15)
	p	0.383		0.889	0.528	0.987	0.523	*0.101*	0.548
Fourth Reversal	*r*	0.522 (11)	0.062 (18)		**0.644 (19)**	0.291 (15)	0.463 (15)	0.017 (19)	0.314 (20)
	p	0.1	0.806		**0.002**	0.257	0.061	0.943	0.155
Fifth Reversal	*r*	**0.779 (11)**	0.307 (18)	**0.623 (22)**		0.428 (14)	**0.55 (14)**	0.113 (18)	0.25 (19)
	p	**0.005**	0.216	**0.002**		0.098	**0.027**	0.634	0.275
2 Sec Delay	*r*	-0.337 (9)	0.026 (16)	0.294 (18)	0.421 (17)		**0.792 (15)**	-0.09 (14)	0.017 (15)
	p	0.375	0.925	0.237	0.092		**< .001**	0.74	0.948
4 Sec Delay	*r*	0.119 (9)	0.116 (16)	0.456 (18)	**0.485 (17)**	**0.783 (18)**		-0.24 (14)	0.191 (15)
	p	0.76	0.669	0.057	**0.048**	**< .001**		0.371	0.462
RT	*r*	-0.016 (11)	**0.524 (17)**	0.038 (22)	0.227 (21)	-0.062 (17)	-0.252 (17)		0.267 (19)
	p	0.962	**0.031**	0.867	0.322	0.814	0.329		0.241
Experience	*r*	-0.485 (11)	-0.404 (18)	0.06 (23)	-0.237 (22)	-0.05 (18)	0.146 (18)	-0.247 (22)	
	p	0.131	0.096	0.785	0.288	0.845	0.564	0.269	
Age	*r*	-0.43 (11)	**-0.491 (18)**	-0.062 (23)	-0.351 (22)	-0.061 (18)	0.08 (18)	-0.367 (22)	**0.927 (23)**
	p	0.187	**0.038**	0.78	0.11	0.811	0.752	0.093	**< .001**

Table 7.4. Correlation matrix between the measures of the cognitive test battery, age, and experience. The values in the lower half of the triangle are zero-order correlations and the number inside the parenthesis is the sample size. The values in the upper half of the triangle are partial correlations controlling for age and the number inside the parenthesis are the degrees of freedom. Bolded values indicate the result was significant before correcting for multiple comparisons and italicized values indicate results that were no longer significant after controlling for age. Bolded and underlined values indicate the measures were from the same task

Principal Component Analysis				
	Pairwise deletion		Listwise deletion (n = 15)	
Task	PC1	PC2	PC1	PC2
SMTS	0.76	-0.39	0.87	
RT	0.65	-0.62	0.92	
Reversal Learning	0.70	0.51		0.57
DMTS	0.44	0.79		0.97
Eigenvalue	1.67	1.41	1.64	1.29
% Variance Explained	41.93	35.30	44.53	35.18
SMTS	0.87		0.87	
RT	0.86		0.91	
Fourth Reversal	0.22	0.81		0.68
2-Sec Delay		0.79		0.93
Eigenvalue	1.54	1.32	1.63	1.37
% Variance Explained	37.96	32.66	39.87	33.52
SMTS	0.62	0.61	0.80	0.41
RT	0.82	0.37	0.92	
Fourth Reversal	-0.40	0.76	-0.23	0.75
4-Sec Delay	-0.58	0.65	-0.35	0.76
Eigenvalue	1.55	1.51	1.66	1.32
% Variance Explained	38.33	37.41	42.76	33.83
SMTS	0.78	-0.36	0.87	
RT	0.72	-0.50	0.90	
Fifth Reversal	0.72	0.47	0.29	0.42
2-Sec Delay	0.36	0.82		1.04
Eigenvalue	1.77	1.27	1.67	1.28
% Variance Explained	44.29	31.64	45.03	35.49
SMTS	0.80	-0.29	0.86	
RT	0.64	-0.66	0.93	
Fifth Reversal	0.75	0.43	0.23	0.48
4-Sec Delay	0.39	0.84		0.93
Eigenvalue	1.77	1.40	1.68	1.17
% Variance Explained	44.18	35.08	47.64	33.15

Table 7.5. The loadings and percentage of variance explained for each principal component, with two different methods of handling missing data and different dependent measures. The top row of analyses used an aggregate measure from the delay match to sample (DMTS) and reversal learning task, while the subsequent analyses use the individual measures from the tasks. For the symbolic match to sample (SMTS) and RT (reaction time) tasks, the same measures are used for all analyses.

Chapter 8: Conclusion

The goal of this dissertation was to understand the general cognitive abilities of pigeons. This was inspired by the general intelligence research conducted with humans, where participants are given a diverse battery of cognitive tests. Performance is positively correlated across all of the different tests, resulting in a uniformly positive correlation matrix. Subsequent factor analysis on the positive correlation matrix extracts one factor that accounts for approximately half of the variance in performance. This factor is termed *g* (Carroll, 1993; Deary, 2000; Jensen, 1998). My goal was to use a similar methodology, creating a diverse cognitive test battery, to determine if pigeons also have a *g* factor.

To create this test battery, in the first chapter I reviewed how *g* is assessed in humans and nonhuman animals (hereafter animals). For research with humans, it was emphasized that *g* is a robust finding, related to retrieving knowledge and novel problem solving, and can be extracted irrespective of test format or exact test battery contents (Johnson et al., 2004). In addition, the relationship between *g* and other cognitive abilities was also discussed. Working memory (WM), short-term memory (STM), associative learning, and processing speed have a strong or consistent relationship with *g*, while inhibition has a weak to nonexistent relationship (Flaim & Blaisdell, 2020). For research with animals, I focused on nonhuman primates, mice, and avian species due to the large amount of research that had been conducted thus far. While there have been failures to replicate *g* in chimpanzees (Herrmann et al., 2010), mice (Locurto et al., 2003, 2006), and song sparrows (Boogert et al., Anderson et al., 2017), a majority of species assessed do show evidence for *g* (Table 1.1). It may be more accurate, however, to state that animals show a *g like* factor, due to differences in cognitive test battery construction. Some test batteries given to primates include social problems (Kaufman et al., 2019; Herrmann et al., 2010; Hopkins et al.,

2014), while avian test batteries always include a measure of inhibition (Anderson et al., 2017; Ashton et al., 2018; Boogert et al., 2011; Isden et al., 2013; Shaw et al., 2015). This means that animal *g* may be related to social abilities and inhibition, unlike the *g* extracted in humans. Additionally, even though the *g* factor found in humans has shown a consistent relationship to processing speed, no similar measure has been included in animal test batteries. Even when the same cognitive abilities are purportedly being investigated across species, the differences in the methodology employed could result in the tasks relying on different cognitive abilities. Using associative learning for example, in humans this is assessed by presenting unrelated word pairs and tracking how accuracy improves over subsequent study and test blocks. For avian species and mice, associative learning is assessed by training animals to associate one stimulus, either a color or smell, with an outcome like food. It is unclear if these two tasks, with their differing levels of complexity, truly assess the same underlying abilities. So, while there is some evidence that humans and animals have a *g* factor, there is also evidence that this *g* factor may differ across species, but it is difficult to determine due to the differences in test batteries.

The pigeon cognitive test battery was created to facilitate comparisons between human and animal cognitive test batteries. This battery was administered to 23 pigeons that ranged in age from 6 months to 18 years old. It ultimately included a matrix displays task and a symbolic match to sample (STMS) task to assess associative learning, serial reversal learning to assess inhibition, a delayed match to sample (DMTS) to assess STM, and a choice reaction time (RT) task to assess processing speed. In chapters 2-6, I describe the procedure and results of each task in depth to justify which dependent measure was included in the battery analysis. Measures were included if they were sensitive to individual differences, but were also reliable measure of performance. The measures included in the battery were the sessions to criterion in the matrix

displays task, sessions to criterion in the SMTS, first session performance on the fourth and fifth reversal for the serial reversal learning task, accuracy on the 2 and 4 second delay collapsed across the last three sessions of the training for the DMTS, and the mean median RT collapsed across 9 or 10 sessions of training for the choice RT task. Performance on the serial reversal learning and DMTS was also investigated with aggregate measures for each task that collapsed across the two individual measures described above. The data were further processed before analysis to enhance comparisons across tasks. The sessions to criterion and mean median RT data were reversed coded so better performance was always indicated by a higher number. Then all data were normalized (observed value – mean value/standard deviation) to create a similar range and variance across all measures. Finally, age and experience, measured as the number of other cognitive tasks experienced before or during completion of the test battery, were also included in the analyses when appropriate.

Once the data were processed, the relationships between tasks was investigated using a Pearson correlation. In contrast to the results obtained with humans and many other species, the correlation matrix was not uniformly positive, instead there was evidence that the cognitive tasks formed two clusters. Performance on the SMTS and choice RT tasks were significantly, positively correlated with each other, forming the first cluster. The aggregate measure for the serial reversal learning task was significantly, positively correlated with the matrix displays and aggregate measure for the DMTS task, forming the second cluster. Similar results were obtained when the individual measures, performance on the 2 and 4 second delays for the DMTS and performance on the fourth and fifth reversal, were used. The main difference when using the individual measures was between the fourth and fifth reversal. Performance on the fifth reversal had stronger, positive correlations with all other cognitive tasks compared to the fourth reversal.

Age and experience were also included in the correlation analyses, but, since they were strongly correlated with each other (r = .93), only age will be discussed here. There was a significant, negative correlation between age and the SMTS task. While this was the only significant correlation, choice RT and matrix displays also showed negative correlations. Performance on the 2 and 4 second delays had zero-order correlations with age. Finally, even though the fourth reversal had a zero-order correlation with age, the fifth reversal showed a more substantial negative correlation. With these correlation matrices, there is an indication that the serial reversal learning tasks relies on different underlying processes as the task progresses, which may not be equally impacted by age. In addition, there was no compelling evidence for *g* since the correlation matrices were not uniformly positive. Instead, there was preliminary evidence that these cognitive tasks form two separable clusters.

Principal component analysis (PCA) was used to further investigate the results of the correlation matrices. Due to different methods of handling missing data and the nuances of the individual measures, 10 PCAs were conducted using the measures from the SMTS, choice RT, serial reversal learning, and DMTS tasks to maximize the number of subjects analyzed. The results of the PCAs replicated the clusters seen in the correlation matrices a majority of the time, where the SMTS and choice RT tasks loaded onto the first component and the serial reversal learning and DMTS tasks loaded onto the second component.

While these results are preliminary, it would be helpful for future investigations to speculate on what these two components could reflect. I theorized that these components could reflect how these tasks rely on different underlying cognitive abilities, like memorization and inhibitory control. These components could also reflect differences in how automatic a task could be. Finally, these components could reflect differences in age related declines on cognitive

performance. These theorized components are not mutually exclusive since age could impact inhibitory control or how quickly a task becomes ‘automatic’. Additional research, particularly on avian aging, is needed before these components can be interpreted with more confidence.

Ultimately, this dissertation was successful at creating a cognitive test battery for pigeons. It assessed a wide range of cognitive domains, was sensitive to individual differences, and had reliable measures of performance. Despite this, I did not find evidence for a *g* factor in pigeons as seen in humans and many other species. Instead of a uniform positive correlation matrix, there was evidence for two components, which was reflected in the PCAs. While I was able to speculate on what these components could reflect, further research is necessary before drawing strong conclusions. These results are a sound first step in the novel investigation of pigeon intelligence.

References

Aben, B., Stapert, S., & Blokland, A. (2012). About the distinction between working memory and short-term memory. *Frontiers in Psychology*, *3*, 301.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*(3), 288.

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied*, *6*(4), 259 –290. https://doi.org/10.1037/1076-898X.6.4.259.

Adams, E. J., Nguyen, A. T., & Cowan, N. (2018). Theories of working memory: Differences in definition, degree of modularity, role of attention, and purpose. *Language, Speech, and Hearing Services in Schools*, *49*(3), 340-355.

Adolphs, R. (1999). Social cognition and the human brain. *Trends in Cognitive Sciences*, *3*(12), 469-479.

Aiello, L. C., & Wheeler, P. (1995). The expensive-tissue hypothesis: The brain and the digestive system in human and primate evolution. *Current Anthropology*, *36*(2), 199-221.

Alexander, J. R. M., & Smales, S. (1997). Intelligence, learning and long-term memory. *Personality and Individual Differences*, *23*(5), 815-825.

Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., ... & Ferreira, T. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832-838.

Allen, K. L., & Kay, R. F. (2012). Dietary quality and encephalization in platyrrhine primates. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1729), 715-721.

Anderson, B. (1993). Evidence from the rat for a general factor that underlies cognitive performance and that relates to brain size: Intelligence?. *Neuroscience Letters*, *153*(1), 98-102.

Anderson, C., & Colombo, M. (2019). Matching-to-sample: Comparative overview. In J. Vonk & T.K. Shackelford (Eds.), *Encyclopedia of Animal Cognition and Behavior. Springer International Publishing AG*. Springer, Cham. doi.org/10.1007/978-3-319-47829-6_1708-1

Anderson, R. C., Searcy, W. A., Peters, S., Hughes, M., DuBois, A. L., & Nowicki, S. (2017). Song learning and cognitive ability are not consistently related in a songbird. *Animal Cognition*, *20*(2), 309-320.

Ardila, A., Pineda, D., & Rosselli, M. (2000). Correlation between intelligence test scores and executive function measures. *Archives of Clinical Neuropsychology*, *15*(1), 31-36.

Ashton, B. J., Ridley, A. R., Edwards, E. K., & Thornton, A. (2018). Cognitive performance is linked to group size and affects fitness in Australian magpies. *Nature*, *554*(7692), 364-367.

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence*, *33*(4), 431-444.

Astur, R. S., Tropp, J., Sava, S., Constable, R. T., & Markus, E. J. (2004). Sex differences and correlations in a virtual Morris water task, a virtual radial arm maze, and mental rotation. *Behavioural Brain Research*, *151*(1-2), 103-115.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *22*(2), 366-377.

Audet, J. N., & Lefebvre, L. (2017). What's flexible in behavioral flexibility?. *Behavioral Ecology*, *28*(4), 943-947.

Auersperg, A. M., Von Bayern, A. M., Gajdon, G. K., Huber, L., & Kacelnik, A. (2011). Flexibility in problem solving and tool use of kea and New Caledonian crows in a multi access box paradigm. *PLoS One*, *6*(6).

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829-839.

Baddeley, A. D. (2002). Is working memory still working?. *European Psychologist*, *7*(2), 85.

Bailey, R. C. (1997). Hereditarian scientific fallacies. *Genetica*, *99*(2-3), 125-133.

Baker, D. P., Eslinger, P. J., Benavides, M., Peters, E., Dieckmann, N. F., & Leon, J. (2015). The cognitive impact of the education revolution: A possible cause of the Flynn Effect on population IQ. *Intelligence*, *49*, 144-158.

Balda, R. P., & Kamil, A. C. (1992). Long-term spatial memory in Clark's nutcracker, *Nucifraga columbiana*. *Animal Behaviour*, *44*(4), 761-769.

Banerjee, K., Chabris, C. F., Johnson, V. E., Lee, J. J., Tsao, F., & Hauser, M. D. (2009). General intelligence in another primate: individual differences across cognitive task performance in a New World monkey (*Saguinus oedipus*). *PLoS One*, *4*(6), e5883.

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). A new lease of life for Thomson's bonds model of intelligence. *Psychological Review*, *116*(3), 567-579.

Basile, B. M., & Hampton, R. R. (2013). Dissociation of active working memory and passive recognition in rhesus monkeys. *Cognition*, 126(3), 391-396.

Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. D. (2003). The complexities of complex span: Explaining individual differences in working memory in children and adults. *Journal of Experimental Psychology: General*, *132*(1), 71 – 92. https://doi.org/10.1037/0096-3445.132.1.71.

Beauchamp, J. P., Cesarini, D., Johannesson, M., van der Loos, M. J., Koellinger, P. D., Groenen, P. J., ... & Christakis, N. A. (2011). Molecular genetics and economics. *Journal of Economic Perspectives*, *25*(4), 57-82.

Behroozi, M., Helluy, X., Ströckens, F., Gao, M., Pusch, R., Tabrik, S., Tegenthoff, M., Otto, N., Axmacher, N., Kumsta, R., Moser, D., Genc, E., & Güntürkün, O. (2020). Event-related functional MRI of awake behaving pigeons at 7T. *Nature Communications*, *11*(1), 1-12.

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What does the WAIS-IV measure?. *Psychological Assessment*, *22*(1), 121-130.

Bickley, P. G., Keith, T. Z., & Wolfle, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence*, *20*(3), 309-328.

Bitterman, M. E. (1965). The evolution of intelligence. *Scientific American*, *212*(1), 92-101.

Bizon, J. L., Foster, T. C., Alexander, G. E., & Glisky, E. L. (2012). Characterizing cognitive aging of working memory and executive function in animal models. *Frontiers in Aging Neuroscience*, 4, 19.

Blaga, O. M., Shaddy, D. J., Anderson, C. J., Kannass, K. N., Little, T. D., & Colombo, J. (2009). Structure and continuity of intellectual development in early childhood. *Intelligence*, *37*(1), 106-113.

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences*, *29*(2), 109-125.

Blaisdell, A. P., & Cook, R. G. (2005). Two-item Same-different concept learning in pigeons. *Animal Learning & Behavior*, *33*(1), 67-77.

Blasco, R., Finlayson, C., Rosell, J., Marco, A. S., Finlayson, S., Finlayson, G., Negro, J. J., Pacheco, F. G., & Vidal, J. R. (2014). The earliest pigeon fanciers. *Scientific Reports*, *4*(1), 1-7.

Blechert, J., Meule, A., Busch, N. A., & Ohla, K. (2014). Food-pics: an image database for experimental research on eating and appetite. *Frontiers in Psychology*, *5*, 617.

Blough, D. S. (1979). Effects of the number and form of stimuli on visual search in the pigeon. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*(3), 211.

Blum, D., & Holling, H. (2017). Spearman's law of diminishing returns. A meta-analysis. *Intelligence*, *65*, 60-66.

Boesch, C. (2007). What makes us human (*Homo sapiens*)? The challenge of cognitive cross-species comparison. *Journal of Comparative Psychology*, *121*(3), 227 – 240. https://doi.org/10.1037/0735-7036.121.3.227

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*, *65*(1), 23-28.

Boogert, N. J., Anderson, R. C., Peters, S., Searcy, W. A., & Nowicki, S. (2011). Song repertoire size in male song sparrows correlates with detour reaching, but not with other cognitive measures. *Animal Behaviour*, *81*(6), 1209-1216.

Bornstein, M. H., Hahn, C. S., Bell, C., Haynes, O. M., Slater, A., Golding, J., ... & ALSPAC Study Team. (2006). Stability in cognition across early childhood: A developmental cascade. *Psychological Science*, *17*(2), 151-158.

Bray, M. A., Kehle, T. J., & Hintze, J. M. (1998). Profile analysis with the Wechsler Scales: Why does it persist?. *School Psychology International*, *19*(3), 209-220.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812-2831.

Broadbent, D. E., & Gregory, M. (1965). On the interaction of S-R compatibility with other variables affecting reaction time. *British Journal of Psychology*, *56*(1), 61-67.

Brodnick, R. J., & Ree, M. J. (1995). A structural model of academic performance, socioeconomic status, and Spearman's g. *Educational and Psychological Measurement*, *55*(4), 583-594.

Brooks-Gunn, J., Klebanov, P. K., & Duncan, G. J. (1996). Ethnic differences in children's intelligence test scores: Role of economic deprivation, home environment, and maternal characteristics. *Child Development*, *67*(2), 396-408.

Brown, E. K., & Hampton, R. R. (2020). Cognitive control of working memory but not familiarity in rhesus monkeys (*Macaca mulatta*). *Learning & Behavior*, 48(4), 444-452.

Brown, K. G., Le, H., & Schmidt, F. L. (2006). Specific aptitude theory revisited: Is there incremental validity for training performance?. *International Journal of Selection and Assessment*, *14*(2), 87-100.

Brown, P. L., & Jenkins, H. M. (1968). Auto-shaping of the pigeon’s key peck. *Journal of the Experimental Analysis of Behavior*, *11*(1), 1-8.

Bunting, M. (2006). Proactive interference and item similarity in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 183.

Burgoyne, A. P., Hambrick, D. Z., & Altmann, E. M. (2019). Is working memory capacity a causal factor in fluid intelligence?. *Psychonomic Bulletin & Review*, *26*(4), 1333-1339.

Burkart, J. M., Schubiger, M. N., & van Schaik, C. P. (2017). The evolution of general intelligence. *Behavioral and Brain Sciences*, *40*.

Byrne, R. W., & Bates, L. A. (2007). Sociality, evolution and cognition. *Current Biology*, *17*(16), R714-R723.

Cahan, S., & Noyman, A. (2001). The Kaufman ability battery for children mental processing scale: A valid measure of "pure" intelligence?. *Educational and Psychological Measurement*, *61*(5), 827-840.

Carere, C., & Locurto, C. (2011). Interaction between animal personality and animal cognition. *Current Zoology*, *57*(4), 491-498.

Carey, J. R., & Judge, D. S. (2000). *Life spans of mammals, birds, amphibians, reptiles, and fish*. University Press of Southern Denmark.

Carlson, J. S., Jensen, C. M., & Widaman, K. F. (1983). Reaction time, intelligence, and attention. *Intelligence*, *7*(4), 329-344.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404 –431. https://doi.org/10.1037/0033-295X.97.3.404

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In *The Scientific Study of General Intelligence* (pp. 5-21). Pergamon.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Elsevier.

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement*, *15*(3), 139-164.

Cauchoix, M., Chow, P. K. Y., van Horik, J. O., Atance, C. M., Barbeau, E. J., Barragan-Jason, G., Bize, P., Boussard, A., Buechel, S. D., Cabirol, A., Cauchard, L., Claidière, N., Dalesman, S., Devaud, J. M., Didic, M., Doligez, B., Fagot, J., Fichtel, C., Henke-von der Malsburd, J., ... Cauchard, L. (2018). The repeatability of cognitive performance: A meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170281.

Cavojová, V., Mikusková, E. B., & Hanák, R. (2013). Do you have to be smart to know what the others are thinking?. *Studia Psychologica*, *55*(1), 61-66.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology*, *27*(5), 703 – 722. https://doi.org/10.1037/0012-1649.27.5.703

Ceci, S. J. (1996) *On Intelligence: A Biological Treatise on Intellectual Development*. Harvard University Press.

Chabris, C. F. (2007). Cognitive and neurobiological mechanisms of the Law of General Intelligence. In M. J. Roberts (Ed.), *Integrating the mind: Domain specific versus domain general processes in higher cognition* (pp. 449–491). Hove, UK: Psychology Press.

Chabris, C. F., Hebert, B. M., Benjamin, D. J., Beauchamp, J., Cesarini, D., Van der Loos, M., ... & Freese, J. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science*, *23*(11), 1314-1323.

Chen, J., Zou, Y., Sun, Y. H., & Ten Cate, C. (2019). Problem-solving males become more attractive to female budgerigars. *Science*, *363*(6423), 166-167.

Chittka, L., & Niven, J. (2009). Are bigger brains better?. *Current Biology*, *19*(21), R995-R1008.

Chooi, W. T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*(6), 531-542.

Chuderski, A. (2013). When are fluid intelligence and working memory isomorphic and when are they not?. *Intelligence*, *41*(4), 244-262.

Chuderski, A., Taraday, M., Nęcka, E., & Smoleń, T. (2012). Storage capacity explains fluid intelligence but executive control does not. *Intelligence*, *40*(3), 278-295.

Clarin, T. M., Ruczyński, I., Page, R. A., & Siemers, B. M. (2013). Foraging ecology predicts learning performance in insectivorous bats. *PloS One*, *8*(6).

Clark, C. M., Lawlor-Savage, L., & Goghari, V. M. (2016). The Flynn effect: A quantitative commentary on modernity and human intelligence. *Measurement: Interdisciplinary Research and Perspectives*, *14*(2), 39-53.

Cole, E. F., & Quinn, J. L. (2012). Personality and problem-solving performance explain competitive ability in the wild. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1731), 1168-1175.

Colom, R., Abad, F. J., Garcıa, L. F., & Juan-Espinosa, M. (2002). Education, Wechsler's full scale IQ, and g. *Intelligence*, *30*(5), 449-462.

Colom, R., Abad, F. J., Quiroga, M. Á., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why?. *Intelligence*, *36*(6), 584-606.

Colom, R., Burgaleta, M., Román, F. J., Karama, S., Álvarez-Linera, J., Abad, F. J., Martínez, K., Quiroga, M. A., & Haier, R. J. (2013). Neuroanatomic overlap between intelligence and cognitive factors: Morphometry methods provide support for the key role of the frontal lobes. *Neuroimage, 72*, 143-152.

Colom, R., García, L. F., Juan-Espinosa, M., & Abad, F. J. (2002). Null sex differences in general intelligence: Evidence from the WAIS-III. *The Spanish Journal of Psychology*, *5*(1), 29-35.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition, 34*(1), 158-171.

Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., ... & Karama, S. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence*, *41*(5), 712-727.

Colombo, M., & Broadbent, N. (2000). Is the avian hippocampus a functional homologue of the mammalian hippocampus?. *Neuroscience & Biobehavioral Reviews*, *24*(4), 465-484.

Colombo, M., & Scarf, D. (2012). Neurophysiological studies of learning and memory in pigeons. *Comparative Cognition & Behavior Reviews, 7*, 23-43.

Colombo, M., & Scarf, D. (2020). Are There Differences in" Intelligence" Between Nonhuman Species? The Role of Contextual Variables. *Frontiers in Psychology*, *11*, 2072-2072.

Conway, A. R., & Kovacs, K. (2015). New and emerging models of human intelligence. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(5), 419-426

Conway, A. R., & Kovacs, K. (2015). New and emerging models of human intelligence. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(5), 419-426.

Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic bulletin & review*, *8*(2), 331-335.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163-183.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7*(12), 547-552.

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.

Cook, R. G., & Blaisdell, A. P. (2006). Short-term item memory in successive same–different discriminations. *Behavioural Processes*, *72*(3), 255-264.

Cook, R. G., Katz, J. S., & Blaisdell, A. P. (2012). Temporal properties of visual search in pigeon target localization. *Journal of Experimental Psychology: Animal Behavior Processes, 38*(2), 209-216.

Cook, R. G., Kelly, D. M., & Katz, J. S. (2003). Successive two-item same-different discrimination and concept learning by pigeons. *Behavioural Processes, 62*(1-3), 125-144.

Cook, R. G., Levison, D. G., Gillett, S. R., & Blaisdell, A. P. (2005). Capacity and limits of associative memory in pigeons. *Psychonomic Bulletin & Review*, *12*(2), 350-358.

Coppola, V. J., Flaim, M. E., Carney, S. N., & Bingman, V. P. (2015). An age-related deficit in spatial–feature reference memory in homing pigeons (Columba livia). *Behavioural Brain Research*, *280*, 1-5.

Coppola, V. J., Flaim, M. E., Carney, S. N., & Bingman, V. P. (2015). An age-related deficit in spatial–feature reference memory in homing pigeons (*Columba livia*). *Behavioural Brain Research*, *280*, 1-5.

Coppola, V. J., Hough, G., & Bingman, V. P. (2014). Age-related spatial working memory deficits in homing pigeons (Columba livia). *Behavioral Neuroscience*, *128*(6), 666.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(1), 7.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87-114.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory?. *Progress in Brain Research*, *169*, 323-338.

Coyle, T. R., Pillow, D. R., Snyder, A. C., & Kochunov, P. (2011). Processing speed mediates the development of general intelligence (g) in adolescence. *Psychological Science*, *22*(10), 1265-1269.

Crabtree, D. A., & Antrim, L. R. (1988). Guidelines for measuring reaction time. *Perceptual and Motor Skills*, *66*(2), 363-370.

Damerius, L. A., Burkart, J. M., van Noordwijk, M. A., Haun, D. B., Kosonen, Z. K., Galdikas, B. M., Sarawati, Y., Kurniawn, D. & van Schaik, C. P. (2018). General cognitive abilities in orangutans (*Pongo abelii* and *Pongo pygmaeus*). *Intelligence, 74*, 3-11.

de Winter*, J. C., Dodou*, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44*(2), 147-181.

De Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, *18*(1), 10.

Deary IJ. *Looking down on human intelligence: From psychometrics to the brain*. Oxford, England: Oxford University Press; 2000.

Deary, I. J., & Brett, C. E. (2015). Predicting and retrodicting intelligence between childhood and old age in the 6-Day Sample of the Scottish Mental Survey 1947. *Intelligence, 50*, 1-9.

Deary, I. J., Pattie, A., & Starr, J. M. (2013). The stability of intelligence from age 11 to age 90 years: the Lothian birth cohort of 1921. *Psychological Science, 24*(12), 2361-2368.

Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, *11*(3), 201-211.

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish mental surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, *86*(1), 130-147.

DeCasien, A. R., Williams, S. A., & Higham, J. P. (2017). Primate brain size is predicted by diet but not sociality. *Nature Ecology & Evolution*, *1*(5), 1-7.

Demetriou, A., Makris, N., Spanoudis, G., Kazi, S., Shayer, M., & Kazali, E. (2018). Mapping the dimensions of general intelligence: An integrated differential-developmental theory. *Human Development*, *61*(1), 4-42.

Dempster, F. N. (1991). Inhibitory processes: A neglected dimension of intelligence. *Intelligence*, *15*(2), 157-173.

Derksen, J., Kramer, I., & Katzko, M. (2002). Does a self-report measure for emotional intelligence assess something different than general intelligence?. *Personality and Individual Differences*, *32*(1), 37-48.

Detterman, D. K. (1991). Reply to Deary and Pagliari: Is g intelligence or stupidity?. *Intelligence*, *15*(2), 251-255.

DeYoung, C. G., & Clark, R. (2012). The gene in its natural habitat: The importance of gene–trait interactions. *Development and Psychopathology*, *24*(4), 1307-1318.

Diekamp, B., Kalt, T., Ruhm, A., Koch, M., & Güntürkün, O. (2000). Impairment in a discrimination reversal task after D1 receptor blockade in the pigeon" prefrontal cortex". *Behavioral Neuroscience*, *114*(6), 1145.

Divac, I., Mogensen, J., & Björklund, A. (1985). The prefrontal 'cortex' in the pigeon. Biochemical evidence. *Brain Research*, *332*(2), 365-368.

Doebler, P., & Scheffler, B. (2016). The relationship of choice reaction time variability and intelligence: A meta-analysis. *Learning and Individual Differences*, *52*, 157-166.

Dray, S., & Josse, J. (2015). Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*, *216*(5), 657-667.

Dudchenko, P. A. (2004). An overview of the tasks used to test working memory in rodents. *Neuroscience & Biobehavioral Reviews*, *28*(7), 699-709.

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., ... & Emslie, H. (2000). A neural basis for general intelligence. *Science*, *289*(5478), 457-460.

Durlach, P. J., & Mackintosh, N. J. (1986). Transfer of serial reversal learning in the pigeon. *The Quarterly Journal of Experimental Psychology*, *38*(1), 81-95.

Eagle, D. M., Bari, A., & Robbins, T. W. (2008). The neuropsychopharmacology of action inhibition: Cross-species translation of the stop-signal and go/no-go tasks. *Psychopharmacology*, *199*(3), 439-456.

Edhouse, W. V., & White, K. G. (1988). Cumulative proactive interference in animal memory. *Animal Learning & Behavior*, *16*(4), 461-467.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309-331.

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J., & Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence*, *50*, 93-99.

Fagan, J. F., Holland, C. R., & Wheeler, K. (2007). The prediction, from infancy, of adult IQ and achievement. *Intelligence*, *35*(3), 225-231.

Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement*, *63*(6), 915-930.

Farabaugh, S. M., Brown, E. D., & Hughes, J. M. (1992). Cooperative territorial defense in the Australian Magpie, *Gymnorhina tibicen* (Passeriformes, Cracticidae), a group-living songbird. *Ethology*, *92*(4), 283-292.

Finn, P. G., & Hughes, J. M. (2001). Helping behaviour in Australian magpies, *Gymnorhina tibicen*. *Emu*, *101*(1), 57-63.

Flaim, M., & Blaisdell, A. (2021, April 6). A Procedure for the Assessment of Individual Differences in Relational Discrimination Behavior in the Pigeon (Columba livia). PsyArXiv. https://doi.org/10.31234/osf.io/x7by3

Flaim, M., & Blaisdell, A. P. (2020). The comparative analysis of intelligence. *Psychological Bulletin*, *146*(12), 1174.

Flinn, M. V., Geary, D. C., & Ward, C. V. (2005). Ecological dominance, social competition, and coalitionary arms races: Why humans evolved extraordinary intelligence. *Evolution and Human Behavior*, *26*(1), 10-46.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 1030.

Frank, A. J., & Wasserman, E. A. (2005). Associative symmetry in the pigeon after successive matching-to-sample training. *Journal of the Experimental Analysis of Behavior*, *84*(2), 147-165.

Frank, R. (2015). Back to the future? The emergence of a geneticized conceptualization of race in sociology. *The ANNALS of the American Academy of Political and Social Science, 661*(1), 51-64.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science, 17*(2), 172-179.

Frischkorn, G. T., Schubert, A. L., & Hagemann, D. (2019). Processing speed, working memory, and executive functions: Independent or inter-related predictors of general intelligence. *Intelligence, 75*, 95-110.

Galsworthy, M. J., Paya-Cano, J. L., Liu, L., Monleon, S., Gregoryan, G., Fernandes, C., Schalkwyk, L.C., & Plomin, R. (2005). Assessing reliability, heritability and general cognitive ability in a battery of cognitive tasks for laboratory mice. *Behavior Genetics, 35*(5), 675-692.

Galsworthy, M. J., Paya-Cano, J. L., Monleon, S., & Plomin, R. (2002). Evidence for general cognitive ability (g) in heterogeneous stock mice and an analysis of potential confounds. *Genes, Brain and Behavior, 1*(2), 88-95.

Garlick, D. (2002). Understanding the nature of the general factor of intelligence: The role of individual differences in neural plasticity as an explanatory mechanism. *Psychological Review, 109*(1), 116-136.

Garlick, D., Fountain, S. B., & Blaisdell, A. P. (2017). Serial pattern learning in pigeons: Rule-based or associative?. *Journal of Experimental Psychology: Animal Learning and Cognition, 43*(1), 30-47.

Gignac, G. E. (2014). Dynamic mutualism versus g factor theory: An empirical test. *Intelligence, 42*, 89-97.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of *g*. *Intelligence*, *52*, 71-79.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance*, *15*(1-2), 25-46.

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence predicts health and longevity, but why?. *Current Directions in Psychological Science*, *13*(1), 1-4.

Gow, A. J., Johnson, W., Pattie, A., Brett, C. E., Roberts, B., Starr, J. M., & Deary, I. J. (2011). Stability and change in intelligence from age 11 to ages 70, 79, and 87: the Lothian Birth Cohorts of 1921 and 1936. *Psychology and Aging*, *26*(1), 232-240.

Güntürkün, O. (2005). The avian 'prefrontal cortex' and cognition. *Current Opinion in Neurobiology*, *15*(6), 686-693.

Güntürkün, O., & Bugnyar, T. (2016). Cognition without cortex. *Trends in Cognitive Sciences, 20*(4), 291-303.

Güntürkün, O., Ströckens, F., Scarf, D., & Colombo, M. (2017). Apes, feathered apes, and pigeons: differences and similarities. *Current Opinion in Behavioral Sciences, 16*, 35-40.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*(3), 179-203.

Gustafsson, J. E. (2003). Measurement from a Hierarchical Point of View. In Braun, H. I., Jackson, D. N., and Wiley, D. E. (Eds.), *The Role of Constructs in Psychological and Educational Measurement.* Educational Testing Service.

Gutman, L. M., & Schoon, I. (2013). The impact of non-cognitive skills on outcomes for young people. *Education Endowment Foundation*, *59*(22.2), 4-57.

Hackman, D. A., Farah, M. J., & Meaney, M. J. (2010). Socioeconomic status and the brain: Mechanistic insights from human and animal research. *Nature Reviews Neuroscience*, *11*(9), 651-659.

Hakstian, A. R., & Cattell, R. B. (1978). Higher-stratum ability structures on a basis of twenty primary abilities. *Journal of Educational Psychology, 70*(5), 657-669. doi:http://dx.doi.org/10.1037/0022-0663.70.5.657

Harootunian, B. (1966). Intelligence and the ability to learn. *The Journal of Educational Research*, *59*(5), 211-214.

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

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review, 103*(6), 2052-86.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166-1186.

Herbranson, W. T., & Shimp, C. P. (2008). Artificial grammar learning in pigeons. *Learning & Behavior*, *36*(2), 116-137.

Herman, L. M. (2010). What laboratory research has told us about dolphin cognition. *International Journal of Comparative Psychology, 23*(3).

Herndon, J. G., Moss, M. B., Rosene, D. L., & Killiany, R. J. (1997). Patterns of cognitive decline in aged rhesus monkeys. *Behavioural Brain Research*, *87*(1), 25-34.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science, 317*(5843), 1360-1366.

Hertwig, R., & Todd, P. M. (2003). More is not always better: The benefits of cognitive limits. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making*, (pp. 213-231). Wiley.

Heyes, C. (2012). Simple minds: a qualified defence of associative learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1603), 2695-2703.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4*(1), 11-26.

Hogan, D. E., Edwards, C. A., & Zentall, T. R. (1981). Delayed matching in the pigeon: Interference produced by the prior delayed matching trial. *Animal Learning & Behavior*, *9*(3), 395-400.

Holekamp, K. E. (2007). Questioning the social intelligence hypothesis. *Trends in Cognitive Sciences, 11*(2), 65-69.

Hollard, V. D., & Delius, J. D. (1982). Rotational invariance in visual pattern recognition by pigeons and humans. *Science*, *218*(4574), 804-806.

Hopkins, W. D., Li, X., & Roberts, N. (2019). More intelligent chimpanzees (Pan troglodytes) have larger brains and increased cortical thickness. *Intelligence*, *74*, 18-24.

Hopkins, W. D., Russell, J. L., & Schaeffer, J. (2014). Chimpanzee intelligence is heritable. *Current Biology*, *24*(14), 1649-1652.

Horn, N. R., Dolan, M., Elliott, R., Deakin, J. F. W., & Woodruff, P. W. R. (2003). Response inhibition and impulsivity: An fMRI study. *Neuropsychologia*, *41*(14), 1959-1966.

Hoyer, W. J., Stawski, R. S., Wasylyshyn, C., & Verhaeghen, P. (2004). Adult age and digit symbol substitution performance: a meta-analysis. *Psychology and Aging*, *19*(1), 211-214.

Isden, J., Panayi, C., Dingle, C., & Madden, J. (2013). Performance in cognitive and problem-solving tasks in male spotted bowerbirds does not correlate with mating success. *Animal Behaviour*, *86*(4), 829-838.

Isler, K., & van Schaik, C. (2006). Costs of encephalization: The energy trade-off hypothesis tested on birds. *Journal of Human Evolution*, *51*(3), 228-243.

Iwaniuk, A. N., & Nelson, J. E. (2003). Developmental differences are correlated with relative brain size in birds: A comparative analysis. *Canadian Journal of Zoology*, *81*(12), 1913-1928.

Izquierdo, A., & Jentsch, J. D. (2012). Reversal learning as a measure of impulsive and compulsive behavior in addictions. *Psychopharmacology, 219*(2), 607-620

Izquierdo, A., Brigman, J. L., Radke, A. K., Rudebeck, P. H., & Holmes, A. (2017). The neural basis of reversal learning: An updated perspective. *Neuroscience, 345*, 12-26.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, 105*(19), 6829-6833.

Jarrold, C., Tam, H., Baddeley, A. D., & Harvey, C. E. (2011). How does processing affect storage in working memory tasks? Evidence for both domain-general and domain-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 688 -705. https://doi.org/10.1037/a0022527

Jasnow, A. M., Cullen, P. K., & Riccio, D. C. (2012). Remembering another aspect of forgetting. *Frontiers in Psychology, 3*, 175.

JASP Team (2020). *JASP* (Version 0.14.1). JASP Team. https://jasp-stats.org/

Jensen, A. R. (1992). Commentary: Vehicles of g. *Psychological Science, 3*(5), 275-279.

Jensen, A. R. (1998). *The g factor: The science of mental ability* (Vol. 648). Westport, CT: Praeger.

Jensen, A. R., & Weng, L. J. (1994). What is a good g?. *Intelligence, 18*, 231-258

John, O. P., & Benet-Martinez, V. (2000). Measurement: reliability, construct validation, and scale construction. In H.T. Reis & C. M. Judd (Eds) *Handbook of research methods in social and personality psychology* (pp. 339-39). Cambridge University Press.

Johnson, W., & Bouchard Jr, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*(4), 393-416.

Johnson, W., Bouchard Jr, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence*, *32*(1), 95-107.

Johnson, W., te Nijenhuis, J., & Bouchard Jr, T. J. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence*, *36*(1), 81-95.

Johnston, M., Anderson, C., & Colombo, M. (2017). Neural correlates of sample-coding and reward-coding in the delay activity of neurons in the entopallium and nidopallium caudolaterale of pigeons (Columba livia). *Behavioural Brain Research*, *317*, 382-392.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Jung, R. E., & Haier, R. J. (2007). The Parieto-Frontal Integration Theory (P-FIT) of intelligence: converging neuroimaging evidence. *Behavioral and Brain Sciences*, *30*(2), 135.

Kabadayi, C., Bobrowicz, K., & Osvath, M. (2018). The detour paradigm in animal cognition. *Animal Cognition*, *21*(1), 21-35.

Kan, K. J., Kievit, R. A., Dolan, C., & van der Maas, H. (2011). On the interpretation of the CHC factor Gc. *Intelligence*, *39*(5), 292-302.

Kan, K. J., Wicherts, J. M., Dolan, C. V., & van der Maas, H. L. (2013). On the nature and nurture of intelligence and specific cognitive abilities: The more heritable, the more culture dependent. *Psychological Science*, *24*(12), 2420-2428.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615-622.

Kangas, B. D., Berry, M. S., & Branch, M. N. (2011). On the development and mechanics of delayed matching-to-sample performance. *Journal of the experimental analysis of behavior*, *95*(2), 221-236.

Karakuyu, D., Herold, C., Güntürkün, O., & Diekamp, B. (2007). Differential increase of extracellular dopamine and serotonin in the 'prefrontal cortex' and striatum of pigeons during working memory. *European Journal of Neuroscience*, 26(8), 2293-2302.

Katz, J. S., & Wright, A. A. (2006). Same/different abstract-concept learning by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *32*(1), 80.

Kaufman, A. B., Reynolds, M. R., & Kaufman, A. S. (2019). The structure of ape (hominoidea) intelligence. *Journal of Comparative Psychology*, *133*(1), 92-105.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Brown, J., & Mackintosh, N. (2009). Associative learning predicts intelligence above and beyond working memory and processing speed. *Intelligence*, *37*(4), 374-382.

Keagy, J., Savard, J. F., & Borgia, G. (2011). Complex relationship between multiple measures of cognitive ability and male mating success in satin bowerbirds, *Ptilonorhynchus violaceus*. *Animal Behaviour*, *81*(5), 1063-1070.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, *47*(7), 635-650.

Kent, P. (2017). Fluid intelligence: A brief history. *Applied Neuropsychology: Child*, *6*(3), 193-203.

Klein, C., Rauh, R., & Biscaldi, M. (2010). Cognitive correlates of anti-saccade task performance. *Experimental Brain Research*, *203*(4), 759-764.

Kolata, S., Light, K., & Matzel, L. D. (2008). Domain-specific and domain-general learning factors are expressed in genetically heterogeneous CD-1 mice. *Intelligence*, *36*(6), 619-629.

Kolata, S., Light, K., Grossman, H. C., Hale, G., & Matzel, L. D. (2007). Selective attention is a primary determinant of the relationship between working memory and general learning ability in outbred mice. *Learning & Memory*, *14*(1-2), 22-28.

Kolata, S., Light, K., Townsend, D. A., Hale, G., Grossman, H. C., & Matzel, L. D. (2005). Variations in working memory capacity predict individual differences in general learning abilities among genetically diverse mice. *Neurobiology of Learning and Memory*, *84*(3), 241-246.

Kotrschal, A., Buechel, S. D., Zala, S. M., Corral-Lopez, A., Penn, D. J., & Kolm, N. (2015). Brain size affects female but not male survival under predation threat. *Ecology Letters*, *18*(7), 646-652.

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*(3), 151-177.

Kovacs, K., & Conway, A. R. (2019). What is iq? life beyond “general intelligence”. *Current Directions in Psychological Science, 28*(2), 189-194.

Krasheninnikova, A., Berardi, R., Lind, M. A., O’Neill, L., & von Bayern, A. M. (2019). Primate cognition test battery in parrots. *Behaviour*, *156*(5-8), 721-761.

Krimsky, S., & Sloan, K. (Eds.). (2011). *Race and the genetic revolution: Science, myth, and culture*. Columbia University Press.

Kvist, A. V., & Gustafsson, J. E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's Investment theory. *Intelligence, 36*(5), 422-436.

Lamar, M., & Resnick, S. M. (2004). Aging and prefrontal functions: dissociating orbitofrontal and dorsolateral abilities. *Neurobiology of Aging*, 25(4), 553-558.

Lefebvre, L., & Sol, D. (2008). Brains, lifestyles and cognition: Are there general trends?. *Brain, Behavior and Evolution*, *72*(2), 135-144.

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*(4), 391-444.

Lien, M. C., & Proctor, R. W. (2002). Stimulus-response compatibility and psychological refractory period effects: Implications for response selection. *Psychonomic Bulletin & Review*, *9*(2), 212-238.

Light, K. R., Kolata, S., Wass, C., Denman-Brice, A., Zagalsky, R., & Matzel, L. D. (2010). Working memory training promotes general cognitive abilities in genetically heterogeneous mice. *Current Biology*, *20*(8), 777-782.

Lind, J., Enquist, M., & Ghirlanda, S. (2015). Animal memory: A review of delayed matching-to-sample data. *Behavioural Processes, 117*, 52-58.

Lissek, S., Diekamp, B., & Güntürkün, O. (2002). Impaired learning of a color reversal task after NMDA receptor blockade in the pigeon (Columbia livia) associative forebrain (Neostriatum Caudolaterale). *Behavioral Neuroscience, 116*(4), 523.

Locurto, C., & Scanlon, C. (1998). Individual differences and a spatial learning factor in two strains of mice (*Mus musculus*). *Journal of Comparative Psychology, 112*(4), 344-352.

Locurto, C., Benoit, A., Crowley, C., & Miele, A. (2006). The structure of individual differences in batteries of rapid acquisition tasks in mice. *Journal of Comparative Psychology, 120*(4), 378.

Locurto, C., Fortin, E., & Sullivan, R. (2003). The structure of individual differences in heterogeneous stock mice across problem types and motivational systems. *Genes, Brain and Behavior, 2*(1), 40-55.

Longstreth, L. E. (1984). Jensen's reaction-time investigations of intelligence: A critique. *Intelligence, 8*(2), 139-160.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General intelligence,' objectively determined and measured". *Journal of Personality and Social Psychology, 86*(1), 96-111.

Mackintosh, N. J., & Cauty, A. (1971). Spatial reversal learning in rats, pigeons, and goldfish. *Psychonomic Science, 22*(5), 281-282.

Macphail, E. M. (1987). The comparative psychology of intelligence. *Behavioral and Brain Sciences, 10*(4), 645-656.

Madden, J. R. (2008). Do bowerbirds exhibit cultures?. *Animal Cognition*, *11*(1), 1-12.

Major, J. T., Johnson, W., & Bouchard Jr, T. J. (2011). The dependability of the general factor of intelligence: Why small, single-factor models do not adequately represent g. *Intelligence*, *39*(5), 418-433.

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science*, *341*(6149), 976-980.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*(2), 107-127.

Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, M. Á., & Colom, R. (2011). Can fluid intelligence be reduced to 'simple' short-term storage?. *Intelligence*, *39*(6), 473-480.

Matzel, L. D., & Kolata, S. (2010). Selective attention, working memory, and animal intelligence. *Neuroscience & Biobehavioral Reviews*, *34*(1), 23-30.

Matzel, L. D., Han, Y. R., Grossman, H., Karnik, M. S., Patel, D., Scott, N., Specht, S. M., & Gandhi, C. C. (2003). Individual differences in the expression of a "general" learning ability in mice. *Journal of Neuroscience*, *23*(16), 6423-6433.

Matzel, L. D., Sauce, B., & Wass, C. (2013). The architecture of intelligence: Converging evidence from studies of humans and animals. *Current Directions in Psychological Science*, *22*(5), 342-348.

Matzel, L. D., Townsend, D. A., Grossman, H., Han, Y. R., Hale, G., Zappulla, M., Light. K., & Kolata, S. (2006). Exploration in outbred mice covaries with general learning abilities

irrespective of stress reactivity, emotionality, and physical attributes. *Neurobiology of Learning and Memory, 86*(2), 228-240.

Mazza, V., Eccard, J. A., Zaccaroni, M., Jacob, J., & Dammhahn, M. (2018). The fast and the flexible: Cognitive style drives individual variation in cognition in a small mammal. *Animal Behaviour, 137*, 119-132.

Mettke-Hofmann, C. (2014). Cognitive ecology: ecological factors, life-styles, and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(3), 345-360.

Mettke-Hofmann, C., & Gwinner, E. (2003). Long-term memory for a life on the move. *Proceedings of the National Academy of Sciences, 100*(10), 5863-5866.

Mogle, J. A., Lovett, B. J., Stawski, R. S., & Sliwinski, M. J. (2008). What's so special about working memory? An examination of the relationships among working memory, secondary memory, and fluid intelligence. *Psychological Science, 19*(11), 1071-1077.

Moran, R., Zehetleitner, M., Liesefeld, H. R., Müller, H. J., & Usher, M. (2016). Serial vs. parallel models of attention in visual search: accounting for benchmark RT-distributions. *Psychonomic Bulletin & Review, 23*(5), 1300-1315.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*(2), 159-168.

Murray, C., Johnson, W., Wolf, M. S., & Deary, I. J. (2011). The association between cognitive ability across the lifespan and health literacy in old age: The Lothian Birth Cohort 1936. *Intelligence, 39*(4), 178-187.

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology, 53*(1), 53-81.

Neubauer, A. C. (1991). Intelligence and RT: A modified Hick paradigm and a new RT paradigm. *Intelligence*, *15*(2), 175-192.

Neubauer, A. C., & Bucik, V. (1996). The mental speed—IQ relationship: unitary or modular?. *Intelligence*, *22*(1), 23-48.

Neumann, C. G., Murphy, S. P., Gewa, C., Grillenberger, M., & Bwibo, N. O. (2007). Meat supplementation improves growth, cognitive, and behavioral outcomes in Kenyan children. *The Journal of Nutrition*, *137*(4), 1119-1123.

Nisbett, R. E. (2009). *Intelligence and how to get it: Why schools and cultures count*. WW Norton & Company.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779-819.

Oberauer, K., Süß, H. M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—facets of a cognitive ability construct. *Personality and Individual Differences*, *29*(6), 1017-1045.

Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: a meta-analysis. *Psychology and Aging*, *23*(1), 104.

Osvath, M., Kabadayi, C., & Jacobs, I. (2014). Independent evolution of similar complex cognitive skills: The importance of embodied degrees of freedom. *Animal Behavior and Cognition*, *1*(3), 249-264.

Papini, M. R. (2008). *Comparative psychology: Evolution and development of behavior*. Psychology Press.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1-2), 8-13.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*(2), 109-130.

Pepperberg, I. M. (2018). Alex the Parrot. In *Encyclopedia of Animal Cognition and Behavior* (pp. 1-10). Springer, Cham.

Phelan, J. C., Link, B. G., & Feldman, N. M. (2013). The genomic revolution and beliefs about essential racial differences: A backdoor to eugenics?. *American Sociological Review, 78*(2), 167-191.

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909–2013). *Perspectives on Psychological Science, 10*(3), 282-306.

Pietschnig, J., Penke, L., Wicherts, J. M., Zeiler, M., & Voracek, M. (2015). Meta-analysis of associations between human brain volume and intelligence differences: How strong are they and what do they mean?. *Neuroscience & Biobehavioral Reviews, 57,* 411-432.

Pins, D., & Bonnet, C. (1996). On the relation between stimulus intensity and processing time: Piéron's law and choice reaction time. *Perception & Psychophysics, 58*(3), 390-400.

Plenderleith, M. (1956). Discrimination learning and discrimination reversal learning in normal and feebleminded children. *The Journal of Genetic Psychology, 88*(1), 107-112.

Plomin, R. (2001). The genetics of g in human and mouse. *Nature Reviews Neuroscience, 2*(2), 136-141.

Plomin, R., & von Stumm, S. (2018). The new genetics of intelligence. *Nature Reviews Genetics*, *19*(3), 148-159.

Polderman, T. J., de Geus, E. J., Hoekstra, R. A., Bartels, M., van Leeuwen, M., Verhulst, F. C., Posthuma, D. & Boomsma, D. I. (2009). Attention problems, inhibitory control, and intelligence index overlapping genetic factors: A study in 9-, 12-, and 18-year-old twins. *Neuropsychology*, *23*(3), 381.

Pollen, A. A., Dobberfuhl, A. P., Scace, J., Igulu, M. M., Renn, S. C., Shumway, C. A., & Hofmann, H. A. (2007). Environmental complexity and social organization sculpt the brain in Lake Tanganyikan cichlid fish. *Brain, Behavior and Evolution*, *70*(1), 21-39.

Posner, M. I., & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review*, *74*(5), 392.

Postlethwaite, B. E. (2011). *Fluid ability, crystallized ability, and performance across multiple domains: A meta-analysis* [Unpublished doctoral dissertation]. University of Iowa.

Proctor, R. W., & Schneider, D. W. (2018). Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology*, *71*(6), 1281-1299.

Quiroga, M. A., Diaz, A., Román, F. J., Privado, J., & Colom, R. (2019). Intelligence and video games: Beyond "brain-games". *Intelligence*, *75*, 85-94.

Racey, D., Young, M. E., Garlick, D., Pham, J. N-M., & Blaisdell, A. P. (2011). Pigeon and human performance in a multi-armed bandit task in response to changes in variable interval schedules. *Learning & Behavior*, *39*, 245-258. PMID: 21380732

Rajkumar, A. P., Yovan, S., Raveendran, A. L., & Russell, P. S. S. (2008). Can only intelligent children do mind reading: The relationship between intelligence and theory of mind in 8 to 11 years old. *Behavioral and Brain Functions*, *4*(1), 51.

Rast, P., & Zimprich, D. (2009). Individual differences and reliability of paired associates learning in younger and older adults. *Psychology and Aging*, *24*(4), 1001.

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General*, *140*(3), 464.

Raven, J. (2003). Raven progressive matrices. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 223-237). Springer.

Raven, J. (2008). The Raven progressive matrices tests: their theoretical basis and measurement model. *Uses and abuses of intelligence. Studies advancing Spearman and Raven's quest for non-arbitrary metrics*, 17-68.

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, *19*(1), 137-150.

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *20*(6), 1102-1113.

Redick, T.S., Shipstead, Z., Harrison, T.L., Hicks, K.L., Fried, D.E., Hambrick, D.Z., Kane, M.J., & Engle, R.W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, *142*(2), 359- 379.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science*, *1*(3), 86-89.

Reynolds, M. R., & Turek, J. J. (2012). A dynamic developmental link between verbal comprehension-knowledge (Gc) and reading comprehension: Verbal comprehension-knowledge drives positive change in reading comprehension. *Journal of School Psychology*, *50*(6), 841-863.

Reynolds, M. R., Floyd, R. G., & Niileksela, C. R. (2013). How well is psychometric g indexed by global composites? Evidence from three popular intelligence tests. *Psychological Assessment*, *25*(4), 1314-1321.

Roberts, W. A. (1972). Short-term memory in the pigeon: Effects of repetition and spacing. *Journal of Experimental Psychology*, *94*(1), 74.

Roberts, W. A., & Kraemer, P. J. (1982). Some observations of the effects of intertrial interval and delay on delayed matching to sample in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *8*(4), 342.

Rodewald, H. K. (1974). Symbolic matching-to-sample by pigeons. *Psychological Reports*, *34*(3), 987-990.

Rodriguez, J. S., & Paule, M. G. (2009). Working memory delayed response tasks in monkeys. In J. J. Buccafusco (Ed.), *Methods of behavior analysis in neuroscience* (2nd ed., Ch. 12). Taylor & Francis.

Rolfhus, E. L., & Ackerman, P. L. (1999). Assessing individual differences in knowledge: Knowledge, intelligence, and related traits. *Journal of Educational Psychology*, *91*(3), 511-526. http://dx.doi.org/10.1037/0022-0663.91.3.511

Rose, S. A., Feldman, J. F., Jankowski, J. J., & Van Rossem, R. (2008). A cognitive cascade in infancy: Pathways from prematurity to later mental development. *Intelligence, 36*(4), 367-378.

Roth, G., & Dicke, U. (2005). Evolution of the brain and intelligence. *Trends in Cognitive Sciences*, *9*(5), 250-257.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273-1283.

Rubenstein, D. R., & Lovette, I. J. (2007). Temporal environmental variability drives the evolution of cooperative breeding in birds. *Current Biology*, *17*(16), 1414-1419.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*(2), 235 – 294. https://doi.org/10.1037/1076-8971.11.2.235

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: an expanded typology. *Journal of Applied Psychology*, *85*(1), 112.

Salthouse, T. A. (1994). Aging associations: influence of speed on adult age differences in associative learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1486.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403.

Sauce, B., & Matzel, L. D. (2018). The paradox of intelligence: Heritability and malleability coexist in hidden gene-environment interplay. *Psychological Bulletin*, *144*(1), 26-47.

Sauce, B., Bendrath, S., Herzfeld, M., Siegel, D., Style, C., Rab, S., Korabelnikov, J. & Matzel, L. D. (2018). The impact of environmental interventions among mouse siblings on the heritability and malleability of general cognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170289.

Sauce, B., Wass, C., Smith, A., Kwan, S., & Matzel, L. D. (2014). The external–internal loop of interference: Two types of attention and their influence on the learning abilities of mice. *Neurobiology of Learning and Memory*, *116*, 181-192.

Sayol, F., Maspons, J., Lapiedra, O., Iwaniuk, A. N., Székely, T., & Sol, D. (2016). Environmental variation and the evolution of large brains in birds. *Nature Communications*, *7*(1), 1-8.

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence*, *46*, 156-168.

Schmidt, F. L. (2011). A theory of sex differences in technical aptitude and some supporting evidence. *Perspectives on Psychological Science, 6,* 560 – 573.

Schmidt, F. L. (2014). A general theoretical integrative model of individual differences in interests, abilities, and personality traits, and academic and occupational achievement. *Perspectives on Psychological Science*, *9,* 211 – 218.

Schmidt, F. L. (2017). Beyond questionable research methods: The role of omitted relevant research in the credibility of research. *Archives of Scientific Psychology, 5,* 32 – 41. DOI: http://dx.doi.org/10.1037.arc0000033

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience*, *2*, 27.

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in Psychology*, *5*, 1475.

Schmitt, J. E., Raznahan, A., Liu, S., & Neale, M. C. (2020). The genetics of cortical myelination in young adults and its relationships to cerebral surface area, cortical thickness, and intelligence: A magnetic resonance imaging study of twins and families. *NeuroImage*, *206*, 116319.

Schrank, F. A., & McGrew, K. S. (2001). Woodcock-Johnson® III. *Itasca, IL: Riverside*.

Schubert, A. L., Hagemann, D., Voss, A., Schankin, A., & Bergmann, K. (2015). Decomposing the relationship between mental speed and mental abilities. *Intelligence*, *51*, 28-46.

Schubert, A. L., Nunez, M. D., Hagemann, D., & Vandekerckhove, J. (2019). Individual differences in cortical processing speed predict cognitive abilities: A model-based cognitive neuroscience account. *Computational Brain & Behavior*, *2*(2), 64-84.

Schubiger, M. N., Fichtel, C., & Burkart, J. M. (2020). Validity of cognitive tests for non-human animals: pitfalls and prospects. *Frontiers in Psychology*, *11*, 1835.

Schuck-Paim, C., Alonso, W. J., & Ottoni, E. B. (2008). Cognition in an ever-changing world: Climatic variability is associated with brain size in neotropical parrots. *Brain, Behavior and Evolution*, *71*(3), 200-215.

Schweizer, K. (2007). Investigating the relationship of working memory tasks and fluid intelligence tests by means of the fixed-links model in considering the impurity problem. *Intelligence*, *35*(6), 591-604.

Serpell, R. (2000). *Intelligence and culture.* In R. J. Sternberg (Ed.), *Handbook of Intelligence* (p. 549–577). Cambridge University Press. https://doi.org/10.1017/CBO9780511807947.026

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, *125*(1), 4 –27. https://doi.org/10.1037/0096-3445.125.1.4

Shaw, R. C. (2017). Testing cognition in the wild: Factors affecting performance and individual consistency in two measures of avian cognition. *Behavioural Processes*, *134*, 31-36.

Shaw, R. C., & Schmelz, M. (2017). Cognitive test batteries in animal cognition research: evaluating the past, present and future of comparative psychometrics. *Animal Cognition*, *20*(6), 1003-1018.

Shaw, R. C., Boogert, N. J., Clayton, N. S., & Burns, K. C. (2015). Wild psychometrics: evidence for 'general' cognitive performance in wild New Zealand robins, Petroica longipes. *Animal Behaviour*, *109*, 101-111.

Shelton, J. T., Elliott, E. M., Matthews, R. A., Hill, B. D., & Gouvier, W. M. (2010). The relationships of working memory, secondary memory, and general fluid intelligence: Working memory is special. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 813-820.

Sheppard, L. D., & Vernon, P. A. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences*, *44*(3), 535-551.

Shipstead, Z., & Engle, R. W. (2013). Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 277.

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2015). Working memory capacity and the scope and control of attention. *Attention, Perception, & Psychophysics*, *77*(6), 1863-1880.

Shultz, S., & Dunbar, R. I. M. (2006). Both social and ecological factors predict ungulate brain size. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1583), 207-215.

Sleimen-Malkoun, R., Temprado, J. J., & Berton, E. (2013). Age-related dedifferentiation of cognitive and motor slowing: insight from the comparison of Hick–Hyman and Fitts' laws. *Frontiers in Aging Neuroscience*, *5*, 62.

Snodgrass, J. J., Leonard, W. R., & Robertson, M. L. (2009). The energetics of encephalization in early hominids. In *The evolution of hominin diets* (pp. 15-29). Springer, Dordrecht.

Soha, J. A., Peters, S., Anderson, R. C., Searcy, W. A., & Nowicki, S. (2019). Performance on tests of cognitive ability is not repeatable across years in a songbird. *Animal Behaviour*, *158*, 281-288.

Sol, D., Bacher, S., Reader, S. M., & Lefebvre, L. (2008). Brain size predicts the success of mammal species introduced into novel environments. *The American Naturalist*, *172*(S1), S63-S71.

Sol, D., Sayol, F., Ducatez, S., & Lefebvre, L. (2016). The life-history basis of behavioural innovations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1690), 20150187.

Sol, D., Székely, T., Liker, A., & Lefebvre, L. (2007). Big-brained birds survive better in nature. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1611), 763-769.

Sorato, E., Zidar, J., Garnham, L., Wilson, A., & Løvlie, H. (2018). Heritabilities and co-variation among cognitive traits in red junglefowl. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170285.

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

Spetch, M. L., & Edwards, C. A. (1986). Spatial memory in pigeons (Columba livia) in an open-field feeding environment. *Journal of Comparative Psychology*, *100*(3), 266.

Spetch, M. L., Cheng, K., & MacDonald, S. E. (1996). Learning the configuration of a landmark array: I. Touch-screen studies with pigeons and humans. *Journal of Comparative Psychology*, *110*(1), 55.

Spinath, F. M., Ronald, A., Harlaar, N., Price, T. S., & Plomin, R. (2003). Phenotypic g early in life: On the etiology of general cognitive ability in a large population sample of twin children aged 2–4 years. *Intelligence*, *31*(2), 195-210.

Spritzer, M. D., Meikle, D. B., & Solomon, N. G. (2005a). Female choice based on male spatial ability and aggressiveness among meadow voles. *Animal Behaviour*, *69*(5), 1121-1130.

Spritzer, M. D., Solomon, N. G., & Meikle, D. B. (2005b). Influence of scramble competition for mates upon the spatial ability of male meadow voles. *Animal Behaviour*, *69*(2), 375-386.

St Clair-Thompson, H. L. (2010). Backwards digit recall: A measure of short-term memory or working memory?. *European Journal of Cognitive Psychology*, *22*(2), 286-296.

Stankov, L. (2017). Overemphasized "g". *Journal of Intelligence*, *5*(4), 33.

Stankov, L., & Crawford, J. D. (1993). Ingredients of complexity in fluid intelligence. *Learning and Individual Differences*, *5*(2), 73-111.

Stankov, L., & Roberts, R. D. (1997). Mental speed is not the 'basic' process of intelligence. *Personality and Individual Differences*, *22*(1), 69-84.

Stephenson, C. L., & Halpern, D. F. (2013). Improved matrix reasoning is limited to training on tasks with a visuospatial component. *Intelligence*, *41*(5), 341-357.

Sternberg, R. J., & Gastel, J. (1989). Coping with novelty in human intelligence: An empirical investigation. *Intelligence*, *13*(2), 187-197.

Stevenson, H. W., & Zigler, E. F. (1957). Discrimination learning and rigidity in normal and feebleminded individuals. *Journal of Personality*.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643-662.

Swick, D., Ashley, V., & Turken, U. (2011). Are the neural correlates of stopping and not going identical? Quantitative meta-analysis of two response inhibition tasks. *Neuroimage*, *56*(3), 1655-1665.

Tamez, E., Myerson, J., & Hale, S. (2008). Learning, working memory, and intelligence revisited. *Behavioural Processes*, *78*(2), 240-245.

Teague E.B., Langer K.G., Borod J.C., Bender H.A. (2011) Proactive Interference. In: Kreutzer J.S., DeLuca J., Caplan B. (eds) Encyclopedia of Clinical Neuropsychology. Springer, New York, NY. https://doi.org/10.1007/978-0-387-79948-3_1142

Teichner, W. H., & Krebs, M. J. (1974). Laws of visual choice reaction time. *Psychological Review*, *81*(1), 75.

Theeuwes, J. (1995). Abrupt luminance change pops out; abrupt color change does not. *Perception & Psychophysics*, *57*(5), 637-644.

Thorsen, C., Gustafsson, J. E., & Cliffordson, C. (2014). The influence of fluid and crystallized intelligence on the development of knowledge and skills. *British Journal of Educational Psychology*, *84*(4), 556-570.

Trafimow, D. (2016). The attenuation of correlation coefficients: A statistical literacy issue. *Teaching Statistics*, *38*(1), 25-28.

Tucker-Drob, E. M., & Bates, T. C. (2016). Large cross-national differences in gene × socioeconomic status interaction on intelligence. *Psychological Science*, *27*(2), 138-149.

Turkheimer, E., Haley, A., Waldron, M., d'Onofrio, B., & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science*, *14*(6), 623-628.

Unsworth, N., & Engle, R. W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language*, *54*(1), 68-80.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133*(6), 1038-1066.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009). There's more to the working memory capacity—fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review, 16*(5), 931-937.

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory, 17*(6), 635-654.

Urcuioli, P. J. (2008). Associative symmetry, antisymmetry, and a theory of pigeons' equivalence-class formation. *Journal of the Experimental Analysis of Behavior, 90*(3), 257-282.

Urcuioli, P. J. (2015). A successful search for symmetry (and other derived relations) in the conditional discriminations of pigeons. *Conductual: Revista Internacional de Interconductismo y Analisis de Conducta, 3*(1), 4.

van den Heuvel, M. P., & Sporns, O. (2013). Network hubs in the human brain. *Trends in Cognitive Sciences, 17*(12), 683-696.

van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842-861.

Van Ginkel, J. R., Kroonenberg, P. M., & Kiers, H. A. (2014). Missing data in principal component analysis of questionnaire data: a comparison of methods. *Journal of Statistical Computation and Simulation*, *84*(11), 2298-2315.

van Horik, J. O., & Lea, S. E. (2017). Disentangling learning from knowing: Does associative learning ability underlie performances on cognitive test batteries?. *The Behavioral and Brain Sciences*, *40*, e220.

van Horik, J. O., Langley, E. J., Whiteside, M. A., Laker, P. R., Beardsworth, C. E., & Madden, J. R. (2018). Do detour tasks provide accurate assays of inhibitory control?. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1875), 20180150.

van Horik, J., & Emery, N. J. (2011). Evolution of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 621-633.

van Moorselaar, D., & Slagter, H. A. (2020). Inhibition in selective attention. *Annals of the New York Academy of Sciences*, *1464*(1), 204.

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2014). SAwSu: An integrated model of associative and reinforcement learning. *Cognitive Science*, *38*(3), 580-598.

Velasco, S. M., Huziwara, E. M., Machado, A., & Tomanari, G. Y. (2010). Associative symmetry by pigeons after few-exemplar training. *Journal of the Experimental Analysis of Behavior*, *94*(3), 283-295.

Verbruggen, F., Best, M., Bowditch, W. A., Stevens, T., & McLaren, I. P. (2014). The inhibitory control reflex. *Neuropsychologia*, *65*, 263-278.

Vernon, P. A. (1983). Speed of information processing and general intelligence. *Intelligence*, *7*(1), 53-70.

Vickrey, C., & Neuringer, A. (2000). Pigeon reaction time, Hick's law, and intelligence. *Psychonomic Bulletin & Review*, *7*(2), 284-291.

Vincze, O. (2016). Light enough to travel or wise enough to stay? Brain size evolution and migratory behavior in birds. *Evolution*, *70*(9), 2123-2133.

Viswesvaran, C., Ones, D. S., Schmidt, F. L., Le, H., & Oh, I. S. (2014). Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analyses. *Industrial and Organizational Psychology*, *7*(4), 507-518.

Völter, C. J., Tinklenberg, B., Call, J., & Seed, A. M. (2018). Comparative psychometrics: establishing what differs is central to understanding what evolves. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170283.

Warne, R. T., & Burningham, C. (2019). Spearman's g found in 31 non-Western nations: Strong evidence that g is a universal phenomenon. *Psychological Bulletin*, *145*(3), 237-272.

Wass, C., Pizzo, A., Sauce, B., Kawasumi, Y., Sturzoiu, T., Ree, F., Otto, T., & Matzel, L. D. (2013). Dopamine D1 sensitivity in the prefrontal cortex predicts general cognitive abilities and is modulated by working memory training. *Learning & Memory*, *20*(11), 617-627.

Wass, C., Sauce, B., Pizzo, A., & Matzel, L. D. (2018). Dopamine D1 receptor density in the mPFC responds to cognitive demands and receptor turnover contributes to general cognitive ability in mice. *Scientific Reports*, *8*(1), 1-13.

Wasserman, E. A., & Miller, R. R. (1997). What's elementary about associative learning?. *Annual Review of Psychology*, *48*(1), 573-607.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WAIS-IV and clinical validation of the four-and five-factor interpretative approaches. *Journal of Psychoeducational Assessment, 31*(2), 94-113.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS One*, 11 (3):e0152719.doi:10.1371/journal.pone.0152719

Wicherts, J. M. (2017). Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results). *Intelligence, 60*, 26-38.

Wicherts, J. M., Dolan, C. V., & van der Maas, H. L. (2010). A systematic literature review of the average IQ of sub-Saharan Africans. *Intelligence, 38*(1), 1-20.

Widaman, K. F., & Carlson, J. S. (1989). Procedural effects on performance on the Hick paradigm: Bias in reaction time and movement time parameters. *Intelligence, 13*(1), 63-85.

Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it?. *Frontiers in Psychology*, *4*, 433.

Williams, B. A., & Pearlberg, S. L. (2006). Learning of three-term contingencies correlates with Raven scores, but not with measures of cognitive processing. *Intelligence*, *34*(2), 177-191.

Williams, D. I. (1967). The overtraining reversal effect in the pigeon. *Psychonomic Science, 7*(7), 261-262.

Wills, S., & Mackintosh, N. J. (1999). Relational learning in pigeons? *The Quarterly Journal of Experimental Psychology B*, *52*(1), 31-52.

Woodley of Menie, M. A. W., Fernandes, H. B., & Hopkins, W. D. (2015). The more *g*-loaded, the more heritable, evolvable, and phenotypically variable: Homology with humans in chimpanzee cognitive abilities. *Intelligence, 50*, 159-163.

Woodrow, H. (1946). The ability to learn. *Psychological Review*, *53*(3), 147.

Wright, A. A. (1997). Concept learning and learning strategies. *Psychological Science, 8*(2), 119-123.

Wright, A. A., Katz, J. S., Magnotti, J., Elmore, L. C., & Babb, S. (2010). Testing pigeon memory in a change detection task. *Psychonomic Bulletin & Review*, *17*(2), 243-249.

Wright, A. A., Kelly, D. M., & Katz, J. S. (2018). Comparing cognition by integrating concept learning, proactive interference, and list memory. *Learning & Behavior*, *46*(2), 107-123.

Wright, A. A., Santiago, H. C., Sands, S. F., Kendrick, D. F., & Cook, R. G. (1985). Memory processing of serial lists by pigeons, monkeys, and people. *Science*, *229*(4710), 287-289.

Wright, C. E., Marino, V. F., Belovsky, S. A., & Chubb, C. (2007). Visually guided, aimed movements can be unaffected by stimulus–response uncertainty. *Experimental Brain Research*, *179*(3), 475-496.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9*(2), 79-94.

Zanto, T. P., & Gazzaley, A. (2013). Fronto-parietal network: flexible hub of cognitive control. *Trends in Cognitive Sciences*, *17*(12), 602-603.

Zentall, T. R. (1997). Animal memory: the role of "instructions". *Learning and Motivation*, *28*(2), 280-308.

Zentall, T. R. (2015). When animals misbehave: Analogs of human biases and suboptimal choice. *Behavioural Processes, 112,* 3-13.

Zentall, T. R. (2020). Revisited: pigeons have much cognitive behavior in common with humans. *Frontiers in Psychology*,

Zentall, T. R., & Smith, A. P. (2016). Delayed matching-to-sample: A tool to assess memory and other cognitive processes in pigeons. *Behavioural Processes*, *123*, 26-42.