

Modeling the Dynamics and Representations of Real-World Visual Expertise

By

Jianhong Shen

Dissertation

Submitted to the Faculty of the  
Graduate School of Vanderbilt University  
in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

May 11, 2018

Nashville, Tennessee

Approved:

Thomas J. Palmeri, Ph.D.

Frank Tong, Ph.D.

Isabel Gauthier, Ph.D.

Sun-Joo Cho, Ph.D.

To my parents and to Chenxu.



## ACKNOWLEDGMENTS

The Ph.D. has been a challenging yet fulfilling journey for me, which I could not get through without the love, support, guidance, and encouragement of my mentor, family, colleagues, and friends.

I would first like to thank my mentor, Dr. Thomas J. Palmeri, for his unwavering support and guidance throughout my Ph.D. career. I came to the United States in 2012, as naive and blunt as a fresh-out-of-college kid can be, in addition to the cultural differences. He was the one who guided me with care, patience, and scientific vigor along the way. I am truly grateful for his support, guidance, and friendship both in the scientific arena and on a personal level.

I am honored to have Dr. Isabel Gauthier, Dr. Sun-Joo Cho, and Dr. Frank Tong as my doctoral committee members. They have been generous with their time and expertise to give me valuable feedback and advice throughout my Ph.D. career. In particular, I would like to thank Dr. Isabel Gauthier and Dr. Sun-Joo Cho for being wonderful exemplars as female scientists for me. There were numerous times when I looked up to them for inspiration throughout my time at Vanderbilt.

I have been fortunate and grateful to be surrounded by a group of talented, friendly, and hard-working colleagues from the Category Lab and the Object Perception Lab at Vanderbilt. In particular, I would like to thank Isabel Gauthier, Magen Speegle, Jeff Annis, Michael Mack, Brent Miller, Ting-Yun Chang, Kao-Wei Chua, Rankin McGugin, Katie Ryan, Jackie Floyd, David Ross, Mackenzie Sunday, Ana Van Gulick, and Jennifer Richler. Also, I heartily appreciate the dedicated help from Laura Stelianou, Ethan Schmerling, and Minhee Jo during experiment preparation and data collection.

The Princeton review says that Vanderbilt has the happiest students, which I absolutely agree. I would like to thank the Dean and staff members in various offices across Vanderbilt who create such a warm and supportive environment for students, especially Dean Mark

Wallace, Dr. Ruth Schemmer, Liz Leis, Sheri Kimble, Linda Harris, Dr. Ashley Brady, Jerry Hager, Angel Gaither, Jennifer Lass, Dan Stewart, and Cris Zerface.

I would like to thank my parents, Liren Shen and Xuemei Zhao, for their unconditional love, support, and encouragement. Thank you for being always there for me. It is from them that I learned patience, gratitude, compassion, humbleness, persistence, and love, characteristics that I will cherish for the rest of my life. I would like to thank Chenxu Wen for his company during my time at Vanderbilt. I appreciate his care, advice, counseling, support, and all the heated debates and discussions. It has been fun enduring and surviving graduate school together. I look forward to the journey ahead.

In addition, I would like to thank the wonderful colleagues I met at axialHealthcare as an Intern, especially Elizabeth Ann Stringer, John Donahue, Lindsey Morris, Chad You, Mathilde Granke, Meridith Peratikos, Ray Pasek, Christopher Vandeventer, Emanuel Villa, Ruthie Harding, Matt Shoaf, and Bethany Bedford. Thank you for making me at home and bringing the best out of me for this worthy mission of combating the opioid crisis.

I am fortunate to have friends outside of academic life. Thank you members of *GaiQun* and *Run, Vandy* group, in particular, Jie Yang, Yu Zhang, Yin Guo, Xiang Zhang, Bing Han, Dongqing Zhang, Donglai Zhang, Qiujia Guo, Ying Ji, Yang Zhang, Jizhou Liu, Shuhai Zhang, Sichang Lu, Meng Su, Le Han, Xuebiao Yuchi, Zhiwei Zhang, Shuangxi Han, Chao Ma, Yao Pan, Zheng Huang, Qiao Xu, Xiaojie Gong, Mingfeng Bai, and Kristin Conniff. Thank you for the hugs and laughs throughout the 5 years.

Finally, I would like to acknowledge my funding agencies. This work was funded by an NSF grant to the Temporal Dynamics of Learning Center and the Lisa Quesenberry Scholarship from the Community Foundation of Louisville.

## TABLE OF CONTENTS

	Page
DEDICATION . . . . .	ii
ACKNOWLEDGMENTS . . . . .	iii
LIST OF TABLES . . . . .	vii
LIST OF FIGURES . . . . .	viii
Chapter . . . . .	1
1 Introduction . . . . .	1
1.1 Background . . . . .	1
1.2 Outline . . . . .	10
2 Measuring Perceptual Expertise . . . . .	12
2.1 Introduction . . . . .	12
2.2 Methods . . . . .	16
2.3 Results . . . . .	21
2.4 Discussion . . . . .	30
3 Modeling the Dynamics of Expertise . . . . .	33
3.1 Introduction . . . . .	33
3.2 Method . . . . .	41
3.3 Behavioral Results . . . . .	45
3.4 Modeling Results . . . . .	50
3.5 Discussion . . . . .	57
4 Modeling the Representations of Expertise . . . . .	59
4.1 Introduction . . . . .	59
4.2 Methods . . . . .	62
4.3 Results . . . . .	67

4.4 Discussion . . . . .	78
5 Conclusion . . . . .	81
5.1 Summary of Findings . . . . .	81
5.2 General Discussion . . . . .	84
Appendix A The Questionnaire before the Bird Expertise Test . . . . .	86
Appendix B The Bird Expertise Test . . . . .	90
Appendix C The Coverage Map for the Bird Expertise Test . . . . .	121
BIBLIOGRAPHY . . . . .	125

## LIST OF TABLES

Table	Page
2.1 Coefficients estimates from the latent regression LLTM . . . . .	28
3.1 The dummy variable coding of the design matrix in the model . . . . .	47
3.2 A summary of the posterior estimates . . . . .	56

## LIST OF FIGURES

Figure	Page
2.1 The construct map of the bird expertise test . . . . .	18
2.2 The Eigenvalues of the first 10 factors using exploratory factor analysis . . .	23
2.3 The density plots for the parameter estimates from the three-parameter model	23
2.4 The Wright map of IRT results from the three-parameter model . . . . .	25
2.5 A scatter plot of estimated birding expertise over self-rated birding expertise	30
3.1 An illustration of the basic-first hypothesis and the differentiation hypothesis	35
3.2 An illustration of the diffusion decision process for a choice between word and nonword . . . . .	38
3.3 A demonstration of the hierarchical diffusion model . . . . .	40
3.4 Some example stimuli used in the speeded category-verification task . . . .	42
3.5 An example trial in the speeded category-verification task . . . . .	44
3.6 The mean response time patterns for the "novice" and "expert" groups . . .	46
3.7 The density and histogram of the parameter estimates for $\xi_{Ip}$ and $\zeta_{Ip}$ . . . .	48
3.8 The correlation between the expertise index and the parameter estimates for $\xi_{Ip}$ and $\zeta_{Ip}$ . . . . .	49
3.9 The correlations between the data and the posterior predictions from the HDDM model . . . . .	53
3.10 The traces of the parameter estimates for the key parameters from the HDDM model . . . . .	54
3.11 The density of the parameter estimates for the key parameters from the HDDM model . . . . .	55
4.1 The 10 Warblers species and 10 Sparrows species used in the study . . . . .	64
4.2 An example trial in the similarity-ratings task . . . . .	66

4.3	An example trial in the bird identification task . . . . .	68
4.4	WAIC for the MDS models . . . . .	70
4.5	Averaged Kruskal’s Stress and correlation coefficient for the MDS models .	71
4.6	The ordering of the 20 bird species along each of the 9 dimensions in the representational space . . . . .	72
4.7	The correlation between birds’ dimensional locations and their family in- formation . . . . .	73
4.8	The correlation between the dimensional weightings in MDS and the ex- pertise index . . . . .	74
4.9	The correlation between the dimensional weightings in identification and the expertise index . . . . .	75
4.10	The correlation between the biases toward each bird species and the exper- tise index . . . . .	76
4.11	The sensitivity parameter estimates across the expertise index . . . . .	77

# Chapter 1

## Introduction

### 1.1 Background

Given an X-ray image, expert radiologists can diagnose Alzheimer's from a single look, while novices can barely guess what the image entails. This ability that expert radiologists exhibit, the exceptional ability to make judgments with images, is called real-world visual expertise. Beyond radiology, visual experts play important roles in many domains of our society, such as airport baggage screening and forensic fingerprint identification. Understanding real-world visual expertise has profound theoretical and practical implications. Theoretically, studying expertise as an exemplary performance of human perception can provide insights about the general mechanisms of perception. Practically, understanding expertise can inform us about potential techniques to enhance the development of expertise in workplaces. Given the significance of this topic, how and why perceptual experts differ from novices has always been a topic of great interest to cognitive scientists.

Expertise manifests itself in various behaviors. Experts can learn to identify and categorize new objects in their expertise domain more quickly than novices (Gauthier & Tarr, 1997, 2002; Tanaka, Curran, & Sheinberg, 2005). They can also identify caricatured objects faster and recognize them with a higher accuracy (Rhodes & McLean, 1990). A hallmark of expertise often studied is a phenomenon called the entry-level shift, that novices are slower in categorizing things at a subordinate level (e.g., *Robin* or *Beagle*) than the so-called basic level (e.g., *Bird* or *Dog*), while experts are equally fast in categorizing at these two levels (K. E. Johnson & Mervis, 1997; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Tanaka & Taylor, 1991). As shown in various expertise hallmarks, including the entry-level shift phenomenon, visual expertise is mostly revealed at the subordinate level, e.g., drastic differences were observed between bird experts and novices distinguish-



ing Nashville Warbler vs. Tennessee Warbler rather than distinguishing bird vs. dog.

Visual expertise is a complex construct. Individuals can perform differently in a visual task for a host of reasons, including domain-specific experience as well as domain-general traits such as visual acuity and intelligence quotient (IQ). Given the complexity, visual expertise has often been measured and studied from various angles. The choice of the measurement tool often depends on the nature of the study.

One approach to understand expertise is to measure its components and to examine the relationships among them to uncover the source of performance differences (Richler, Wilmer, & Gauthier, 2017; Sunday, Donnelly, & Gauthier, 2017; Van Gulick, McGugin, & Gauthier, 2016). For example, Van Gulick et al. (2016) used non-visual semantic knowledge as a proxy for measure experience and to study how experience contributes to individual differences in object recognition. In such studies, it is desirable to have measures such as a semantic knowledge test that would tap into a specific component or source of expertise.

Another approach is to understand expertise by modeling the behavior that defines expertise itself, rather than relating it to other measured constructs. The idea is to describe and understand how people vary in cognitive mechanisms by instantiating their behaviors in cognitive models. One example is to use cognitive or computational models to mimic the identification, categorization, or recognition behaviors of people with varying levels of expertise. Although there are successful algorithms outside of Psychology, especially in Computer Science, to mimic or even to excel human perception (e.g., Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015), the mechanisms of these algorithms were often left unaddressed, leaving the mechanisms of perceptual expertise untapped. In the field of psychology, although the computational approach has rarely been applied to real-world expertise, it has been studied extensively using artificial stimuli with “experts” trained in the laboratory (e.g., Nosofsky, 1992a, 1992b; Nosofsky & Palmeri, 1997; Palmeri, 1997). For example, Nosofsky and Palmeri (1997) showed that the

exemplar-based random walk model (EBRW), a model that detailed the cognitive processes behind speeded multidimensional perceptual classification, could accurately depict classification responses over the course of training for several tasks. To apply cognitive models such as EBRW to expertise, a direct measure of the performance on the kind of behavior that defines expertise is necessary.

Expertise has been measured indirectly using different methods. Some tasks focus on the perceptual part of expertise by having participants judge consecutive images in a domain as the same or different (e.g., Gauthier & Tarr, 2002; Gauthier, Skudlarski, Gore, & Anderson, 2000; Hagen, Vuong, Scott, Curran, & Tanaka, 2014; Maurer, Blau, Yoncheva, & McCandliss, 2010). Some tasks focus on the memory part of expertise by having participants recognize objects presented earlier (Duchaine & Nakayama, 2006; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012). To be able to model the behavior that defines expertise, and to complement other developed measures, a direct expertise measure was developed in this project.

The key questions in expertise modeling are how and why people with varying levels of expertise differ from each other in behavioral performances such as identification and categorization. A natural approach to these fundamental questions is to use cognitive models to understand a hallmark of visual expertise, especially the entry-level shift phenomenon. Surprisingly, to my knowledge, no work has been done to directly model the cognitive processes of this expertise hallmark in a real-world setting.

The lack of work on this direction is understandable for several reasons (Shen & Palmeri, 2015). Modeling performance differences seen in real-world expertise has always been challenging, due to the complexity of cognitive models and the extra challenge when incorporating individual differences in the models (M. D. Lee & Wagenmakers, 2014). Such modeling often requires a big sample of participants, which causes logistical difficulty in lab settings, where participants are often recruited locally. Also, the behavioral task that yields the entry-level shift phenomenon, the speeded category verification task (K.

E. Johnson & Mervis, 1997; Rosch et al., 1976; J. W. Tanaka & Taylor, 1991), does not immediately relate to any cognitive models.

In this project, participants with varying levels of expertise were recruited online across the U.S. The experiments were also conducted through interactive online web pages. Online experiments can easily solve the logistical difficulty of arranging many participants. The only caveat is that it is challenging to engage participants for longer than 10 to 20 minutes, while they are in potentially distracting real-world settings such as at home or at work. In lab settings, participants often have a dedicated period of time, sometimes even hours, in a minimally distracting environment. This inevitably changes the nature of the experimental data set from rich data with few participants to sparse data with many participants.

Sparse data with many participants are challenging to fit with complex cognitive models. A well-known dilemma in such situations is that modeling only group averages can fail to detect meaningful individual differences, while individual data may be too noisy to provide sufficient power for modeling. In this project, recently developed hierarchical modeling techniques were employed within a Bayesian framework to resolve this issue, in which both group and individual differences were modeled simultaneously (B. Carpenter et al., 2016; Gelman & Hill, 2007; M. D. Lee & Webb, 2005; Okada & Lee, 2016; Rouder & Lu, 2005; Wiecki, Sofer, & Frank, 2013). In the hierarchical modeling approach, individuals are assumed as samples from a certain population, with model parameters defined at both group and individual level. Hierarchical modeling combines the strength of group modeling and individual modeling. It could cancel meaningless noise while preserving meaningful individual differences, resolving the power issue by sharing information across multiple levels (I. G. Kreft, Kreft, & Leeuw, 1998; M. D. Lee, 2011; Shiffrin, Lee, Kim, & Wagenmakers, 2008). The behavioral differences were thus modeled using Bayesian hierarchical techniques.

Many models have been developed for categorization (Nosofsky, 1984, 1986; Nosofsky

& Palmeri, 1997; J. D. Smith & Minda, 2000). However, no cognitive models have been specifically developed for the cognitive process in the speeded category verification task. The task is not to categorize an object, but to verify whether an image contains a given object (K. E. Johnson & Mervis, 1997; Rosch et al., 1976; Tanaka & Taylor, 1991). However, the task is a speeded decision task in nature, the responses and response times (RT) of which have been successfully depicted using the sequential sampling models, a class of models that has been widely applied to understand speeded decision-making behaviors (S. D. Brown & Heathcote, 2005; Nosofsky & Palmeri, 1997; Ratcliff, 1978; Ratcliff & Smith, 2004; Ratcliff & Starns, 2013). The sequential sampling models were thus used in this project to model the speeded category-verification behavior and to deconstruct the underlying cognitive process into psychological components. In particular, the Drift Diffusion model was used given its excellence in describing the decision processes in a host of cognitive tasks (Ratcliff & Smith, 2004; Ratcliff & Starns, 2013). This fresh perspective of treating speeded category verification as speeded decision-making brought the strength of decision-making models to the field of expertise. In particular, performance differences in the task can be deconstructed into differences in psychological processes such as bias, evidence accumulation rate, and decision boundary.

The sequential sampling models can offer insights about the differences in psychological processes that give rise to expertise behaviors. For example, experts could be faster than novices because of faster evidence accumulation, shorter non-decision time, or a combination of both. The question then is: what kinds of individual differences can give rise to these psychological differences? Are participants with higher level of expertise faster in evidence accumulation because of more differentiated representations, or because of higher sensitivity to subtle differences between the representations of two similar objects? To delve deeper into the question, it is necessary to describe the representations underlying these psychological differences by mapping out the representations of participants along the continuum of real-world expertise. The representations can then be used to feed

into cognitive models to examine whether differences in representations alone can explain behavioral differences.

Representation is defined differently depending on traditions in the field (Shen & Palmeri, 2016). Models in the field of traditional object recognition often fully describe the hierarchy of visual processing, starting from the raw images on the retina, to representation transformations along the visual pathway, to eventual categorization decisions (e.g., Cottrell & Nguyen, 2005; Palmeri & Cottrell, 2009 for reviews; see Palmeri, Wong, & Gauthier, 2004; Riesenhuber & Poggio, 2000; Serre, Oliva, & Poggio, 2007; Tong, Joyce, & Cottrell, 2005). This contrasts with most categorization models that focus on category representations and decision process, where objects are simply represented as points in a multidimensional space, leaving out the details of deriving the space dimensions and object locations from the retinal images (e.g., Hintzman, 1986; Nosofsky, 1988, 1992a; Shepard, 1980).

The representational space might be based on the dimensions that are explicitly manipulated by the experimenter (e.g., Hintzman, 1986; Nosofsky, 1988), or they might be derived using techniques like multidimensional scaling (MDS; e.g., Nosofsky, 1992a; Shepard, 1980). In MDS-based models, it is assumed that in our mind's eye, we represent objects as points located in a multidimensional space. For example, we might represent birds as points in a bird space with some easy-to-define dimensions such as the color of the feather, the size of the bird, the length of the beak, as well as some abstract, hard-to-define dimensions.

The mental representations, i.e., the locations of a set of objects in the multidimensional space, would influence the similarities one observes among the objects (Shepard, 1987). Using the logic, the mental representations of participants with varying levels of expertise can be modeled by gathering their similarity ratings among a set of similar-looking objects in their domain of expertise. The MDS approach has been widely used to uncover hidden perceptual representations, and the derived mental representation has been successful in explaining various cognitive behaviors such as identification, categorization, and recognition (e.g., Borg & Groenen, 2005; T. F. Cox & Cox, 2000; Kruskal, 1964; Nosofsky, 1992a;

Shepard, 1962, 1987). But so far the success has been mostly limited to artificial stimuli.

There has been limited effort to understand the representations underlying real-world expertise using the MDS approach, due to both logistical difficulty and technical challenges (Bocci & Vichi, 2011; Carroll & Chang, 1970; Okada & Lee, 2016; Takane, Young, & De Leeuw, 1977). Some work has used this approach to understand how superordinate-level categories are structured and how those vary with different kinds of expertise (Boster & Johnson, 1989; Medin, Lynch, Coley, & Atran, 1997). For example, Boster and Johnson (1989) examined how fish experts and novices represented fish species differently in their perceptual space and found that experts represented not only morphological, but also functional information, while novices represented only morphological information of the fish species. Some recent work used MDS to examine the representations of rocks. Although real-world stimuli were used, these studies looked at novices learning those objects, not experts (Meagher, Carvalho, Goldstone, & Nosofsky, 2017; Nosofsky, Sanders, Gerdman, Douglas, & McDaniel, 2017a; Nosofsky, Sanders, Meagher, & Douglas, 2017b).

While it is interesting to understand representation differences across individuals at the superordinate level, expertise takes action mostly at the subordinate level, as exemplified in the entry-level shift phenomenon. This calls for a focus on the subordinate level to understand representation differences along the continuum of real-world expertise, i.e., the microstructure of expertise.

Unlike many models in cognitive science (Estes & Maddox, 2005; M. D. Lee & Webb, 2005; Shen & Palmeri, 2016), MDS-based models have long embraced the notion of individual differences (Carroll & Chang, 1970; Takane et al., 1977). In earlier work, MDS-based models could allow individuals to weight dimensions differently in the representational space, but individuals were limited to share a single representational space (Carroll & Chang, 1970; Takane et al., 1977). It was not until recently that MDS-based models can account for not only quantitative, but also qualitative individual differences (Bocci & Vichi, 2011; Okada & Lee, 2016). One technique newly-introduced to these models is

the latent-mixture component, which can be used to identify latent groups of participants who use qualitatively different cognitive processes, i.e., different representational spaces in this case (Bocci & Vichi, 2011). Another important advancement is the introduction of the Bayesian framework, which allows stable parameter estimation in complex models with a latent-mixture component (Okada & Lee, 2016). These advancements laid the groundwork for us to study how representations vary along the continuum of real-world visual expertise, enabling us to test whether individuals differ from each other quantitatively in attention weights, qualitatively in representational space, or both.

Representation mediates the decisions we make with regard to objects, such as identification, categorization, and recognition. Cognitive models with the assumption that objects are represented in a multidimensional space have been successful in explaining various behavioral patterns. For example, identification, a commonly studied behavior in which participants assign a unique response to each stimulus, has been successfully explained using the MDS-choice model (Nosofsky, 1985; Shepard, 1957, 1958), a special case of the classic similarity choice model (SCM) proposed by Shepard (1957) and Luce (1963). However, representation differences alone may not be able to explain performance differences. Nosofsky and Palmeri (1997) trained participants to identify or categorize a set of arbitrarily grouped color chips that varied in saturation and brightness. Based on the MDS solutions derived from identification data, the generalized context model (GCM; Nosofsky, 1986), a popular similarity-based model of categorization, can nicely describe the categorization data. It was found that other parameters in GCM were also necessary to explain the behavior, such as increased discriminability as well as decreased guessing.

Some researchers in the field of face perception also investigated expertise using an MDS framework. Valentine (Valentine, 1991) proposed the hypothesis that a unified MDS-based framework can account for a variety of face perception phenomena, which was verified in several studies (e.g., K. Lee, Byatt, & Rhodes, 2000; Papesh & Goldinger, 2010). One phenomenon in face perception that relates to expertise is the other-race effect, that

is, we recognize own-race faces more easily than faces of a less familiar race. This better performance with own-race faces can be viewed as a form of visual expertise (Tanaka & Farah, 2003). Using MDS analysis, Byatt and Rhodes (2004) found that for Caucasian participants, Chinese faces were more densely clustered in the multidimensional face space than Caucasian faces. Importantly, using the GCM as the cognitive process model, they predicted the behavioral other-race effect from the differential representations of the Chinese and Caucasian faces.

These examples in face perception pointed out a promising direction but fell short of depicting real-world expertise for several reasons. The biggest issue is that individual differences were not considered in any of the studies. Only coarse group differences were addressed, assuming that experts and novices were homogeneous within each group. However, many researchers have highlighted the dangers of omitting individual differences (Ashby, Maddox, & Lee, 1994; Estes, 1956; Estes & Maddox, 2005; Shen & Palmeri, 2016). Often only one group of experts were contrasted with one group of novices, studying only two or a few groups, without a discussion of the fine gradient along the expertise continuum. In addition, the range of experience/expertise is often rather limited within the expertise continuum. In the Byatt and Rhodes (2004) study, Caucasian participants were assumed “experts” in identifying Caucasian faces and “novices” in identifying Chinese faces. While many Caucasian participants might be “experts” in identifying Caucasian faces given our remarkable ability in face recognition, they may not be real “novices” in identifying Chinese faces given the high similarity between Caucasian and Chinese faces. These issues were addressed by studying experts with a wide range of expertise and modeling their expertise behaviors like identification and categorization within a Bayesian hierarchical framework that takes into account both group and individual differences.



## 1.2 Outline

This dissertation describes a series of experimental and modeling work designed to investigate the mechanisms underlying real-world visual expertise. I approached these aims by studying the behaviors of bird watchers with varying levels of birding expertise. Bird watching was chosen as the domain to study for several reasons (Shen, Mack, & Palmeri, 2014). There are numerous bird watchers who have a keen interest in science and are willing to participate in research. This makes it easy to recruit bird watchers with a wide range of expertise. Second, birding is by its very nature an identification and categorization task. Expert birders stand out from novices because of their ability to accurately and rapidly categorize birds seen at a glance at the subordinate level. Thus birding is a domain in which scientific significance and real world interest perfectly overlap. Last but not least, birding domain has been widely used to inform research on expertise and object recognition (Gauthier et al., 2000; K. E. Johnson & Mervis, 1997; Tanaka & Taylor, 1991). These seminal studies laid groundwork for us to understand visual expertise theoretically.

Three key aims were addressed. First, individuals were lined up along a continuum from novice to intermediate to expert by establishing a direct measure of expertise. Second, the dynamics behind an expertise hallmark, the entry-level shift phenomenon, were investigated using the sequential sampling models to uncover the underlying psychological components. Third, the underlying representation of expertise was modeled using the MDS technique in a Bayesian framework to map out representation differences along the expertise continuum. The role of identification was examined by modeling the behavior based on the MDS solutions.

All participants were recruited and tested online, which resolved the logistical difficulty of recruiting real-world experts in lab settings. Once participants were recruited, their expertise levels were measured using a bird expertise test that was newly developed based on the kind of task that defines bird expertise – bird identification. Participants were then invited to complete the speeded category-verification task, in which they verified the cat-

egories of objects at different levels of abstraction. The response time and choice data were modeled using the sequential sampling models to deconstruct the performance differences into differences in constituent psychological processes. Participants then completed a similarity judgment task, in which they were asked to judge the similarity of bird pairs. The similarity ratings were used to derive participants' mental representations of the bird species. I analyzed how their representations varied along the continuum of expertise, either quantitatively, qualitatively, or both. Next, participants were invited to participate in further behavioral tasks, in this case an identification task. Representations derived from the MDS framework were used to predict participants' performances, to test whether their representational differences could alone account for their behavioral differences, or other psychological factors like overall sensitivity were also necessary in the prediction.

In summary, I modeled the dynamics and representations underlying perceptual expertise, focusing on where expertise takes action, the subordinate level. In modeling the dynamics of expertise, I examined how individual performance differs along the expertise continuum, and what psychological components give rise to the performance differences. In modeling the representations of expertise, I investigated how individuals differ from each other in representing objects at the subordinate level, and how their representations relate to their performance. Together, these modeling work provided insights into the dynamics and representations of expertise, proposing a coherent computational framework of real-world visual expertise.

## Chapter 2

### Measuring Perceptual Expertise

#### 2.1 Introduction

A prerequisite for almost any expertise-relevant study is a valid and reliable measure of the expertise under study. In perceptual expertise related research, participants' identification ability has been extensively used to predict changes in behavioral or neural markers, such as response time, accuracy, and brain activation (e.g., Gauthier & Tarr, 2002; Gauthier et al., 2000; Hagen et al., 2014). Various methods such as self-report and indirect quantifications of expertise have been used in these studies to differentiate expert and novice participants. However, there have been limited attempts to develop a valid and reliable direct measure of the identification ability.

Some earlier studies used self-report or peer nomination (Tanaka & Taylor, 1991). These measures are easy to carry out. But self-report is known to be unreliable (Ericsson, 2006, 2009), and no numeric quantification on an interval scale could be obtained from such measures. In more recent studies, different tasks have been developed to quantify perceptual expertise with higher precision, focusing on various aspects of skilled performance. One class of tasks, matching tasks, focuses on the "perceptual" part of perceptual expertise. Such tasks include the same-different task, the one-back task etc. (e.g., Gauthier & Tarr, 2002; Gauthier et al., 2000; Hagen et al., 2014; Maurer et al., 2010). In the same-different task, participants judge whether two sequentially presented images are of the same category or not. In the one-back task, images are presented one at a time and participants are asked to judge whether consecutive images represent the same category or not. In these matching tasks, experts perform better on images from their domain of expertise than novices. But someone without domain knowledge can also perform well simply because of a keen eye for details.

Another genre of measures focuses on perceptual memory as an index of perceptual expertise. For example, modeled after the Cambridge Face Memory task (Duchaine & Nakayama, 2006), the Vanderbilt Expertise Task (VET; McGugin et al., 2012) uses eight different object categories to measure visual expertise. On each block of trials, participants are asked to study one exemplar from each of six different species/models. In the case of bird expertise, they first study six different bird species. On each following test trial within that block, the task is to select the target species from image triplets, including images of two distractor species and one target species that is either the same or different exemplar of one of the six studied species. For example, they might need to pick the *Eastern Bluebird* because that was one of the six studied bird species from among distractors like a *Stellar's Jay* or a *Blue Jay*. Therefore, the relative memory for the domain is deemed as a measure of expertise.

Perceptual expertise in a domain can lead to improved perceptual discriminability, as measured by the perceptual matching tasks, and improved perceptual memory, as measured by the VET. However, these two behavioral markers may or may not be a direct quantification of the “expert” part of perceptual expertise. An expert radiologist quickly and accurately identifies particular types of malignant or benign growths in a medical image, but may not be able to score high on the matching tasks or the VET. Accurately judging images as the same or similar as a previously studied image could be markers but not measures of domain expertise per se. To study the consequences of increased domain expertise, experts should also be quantitatively assessed based on the kind of tasks that distinguish them as experts in a domain (e.g., Ericsson, 2006, 2009).

The choice of expertise measure for a study can be dictated by the type of problems being addressed. Some studies seek to understand the source of expertise by relating it to factors such as experience, IQ, demographic information, and other skill measures (e.g. Gauthier et al., 2013; Van Gulick et al., 2016). Some studies focus on the impact of expertise by using expertise to predict behavioral or neural changes such as the holistic

processing and brain activity in the Fusiform face area (FFA, Gauthier & Tarr, 2002). In such studies, expertise measures with a clear focus might be preferred over a conglomerate measure so that the source of variance can be clearly delineated. A different line of studies, often in lab training settings, examines the identification or categorization performance that defines expertise, e.g., bird identification for bird experts (Nosofsky, 1987; Nosofsky, Clark, & Shin, 1989). In such studies, methods like cognitive modeling are often used to understand expertise. The goal of these studies is not to parse out the variance in the expertise construct by relating expertise to external factors (those factors other than the skilled identification or categorization behavior that defines expertise). Rather, the goal is to understand expertise internally by mapping out how individuals vary in their underlying psychological processes when performing identification or categorization in their domain of expertise. To expand such studies from lab training to real-world settings, the prerequisite is to develop a measure to directly quantify participants' identification or categorization performance in the domain.

This project complements the perceptual expertise research by providing a direct quantification of the “expert” part of perceptual expertise. The measure was developed using bird expertise as an example domain, because bird expertise is a commonly used and accessible domain to study perceptual expertise, unlike fields like radiology and forensics. The bird expertise construct was defined as the ability to quickly and accurately identify various birds when presented with bird images, because it is the kind of task that distinguishes people as experts in a domain, i.e., an expert birder quickly and accurately identifies and categorizes birds perceptually while a novice oftentimes cannot.

The famous “Four Building Blocks” approach (Wilson, 2004) using the item response theory (IRT; Lord & Novick, 1968), a fundamental paradigm in psychometrics for designing, analyzing, and scoring measurement instruments, was used to develop and optimize the test. The first building block, the construct map, requires a theoretical definition of the construct to be mapped out along a range from one extreme to the other. The construct

map provides motivation and structure for the test. In this case, bird expertise, defined as the ability to quickly and accurately identify birds, was mapped out along the range from novice to expert.

The second block, the item design, concerns the realizations of the construct. Given that the construct was defined as bird identification ability, the measure was implemented as a bird identification test, where participants identified birds based on the images provided. Bird identification questions with varying levels of difficulty were designed to tap into different levels of expertise. The third block, the outcome space, defines how the responses are scored. In this case, the question response by participants were categorized into “correct” or “wrong” based on the scoring rubric. Then the last block, the measurement tool, in the form of psychometric or statistical model, evaluates the scores from the item responses and determines the validity and reliability of the measurement.

The “Four Building Blocks” approach provides a systematic paradigm to design measurement instruments through iterative processes. Following this guideline, a bird expertise test was developed, evaluated, and revised through several iterations. Participants’ demographic information and self-report on their expertise were also collected using a questionnaire before the test. The responses were used to explain and validate the test scores.

Equipped with this direct measure and recruiting tool, one can now seek to understand the consequences of increased domain performance on the underlying psychological processes. In one way, the resulting expertise scores can be used as a predictor to explain individual differences in behavioral or neural markers of expertise like response time (RT), accuracy, and brain activation. For example, faster RT in many behavioral tasks is a hallmark of perceptual expertise (Palmeri & Cottrell, 2009; Palmeri et al., 2004), one can use the expertise scores to study how much variance in RT differences can be explained by birding skills. In addition, one can seek to understand the individual differences theoretically by using the expertise scores to predict variance in cognitive process parameters, like the drift-rate parameter in the diffusion model (Smith & Ratcliff, 2004; Vandekerckhove,

Tuerlinckx, & Lee, 2011; Wagenmakers, 2009) and the lognormal race model (Rouder, Province, Morey, Gomez, & Heathcote, 2014). The drift-rate parameter in these models denotes the evidence quality, which was shown to increase with practice (Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009). This bird expertise test can also serve as a recruiting tool for online participants, since many bird watchers are very interested in testing their bird knowledge and taking part in bird identification challenges.

## 2.2 Methods

**Participants.** Bird watchers with all levels of expertise were recruited online. Our lab contacted birding societies across North America and established a participant pool for current and future research on expertise. So far, our lab has gathered contact information for 489 birding societies in North America. For the current project, contact permission was obtained from each society to be able to email their members for participant recruitment. Bird watchers participated in the studies voluntarily or received a modest amount of monetary compensation. Participants completed all experiment sessions online on our website (<http://expertise.psy.vanderbilt.edu/>). Informed consents were obtained prior to participation in accordance with the Institutional Review Board at Vanderbilt University. The questionnaire and the bird expertise test were completed by 741 participants (407 female, 55 missing due to technical issues) aged between 18 and 85 (56 missing, mean = 47.93, SD = 17.23).

**Test Design.** The bird expertise construct, defined as the ability to accurately and speedily identify birds, was mapped out along the range from novice to expert (Figure 2.1). In each question, the task is to identify the bird species out of four numbered label choices given the bird image presented. Participants responded by pressing the number key that corresponded to the bird species in the image. The image was removed after five seconds to ensure that participants did not refer to external sources like birding books or the Internet for consultation. The four choices were presented on the screen until a response was

made. Only visual presentation was used because the goal was to quantify birder watchers' visual ability to identify birds rather than to measure their ability to identify birds using all available information.

The initial set of test items was selected from published references including birding guide books (Dunn & Alderfer, 2011; Floyd, 2008; Kaufman, 2011) and online resources such as [www.aba.org](http://www.aba.org) and [www.allaboutbirds.com](http://www.allaboutbirds.com). Each test item consists of one target bird and three distractor/foil birds that were chosen based on physical and taxonomical similarity to the target bird. Lists of common backyard birds were used for the easy and intermediate identifications; noted bird mis-identifications were used for the more difficult items, as recommended by expert birders and published books (Dunn & Alderfer, 2011; Floyd, 2008; Kaufman, 2011).

Certain structure was built into the test to achieve varying levels of difficulty for the questions. Beginner through intermediate questions have one foil from the same taxonomic family, one foil from a similar family, and one foil from a dissimilar family; while advanced through expert questions have one similar foil from the same family, one less similar foil from the same family, and one foil from a similar family. Beyond these rules, foils with similar color names were picked to avoid color names being the only cue for correct identification. For instance, even novices without any knowledge of birds can correctly identify a *Yellow-Breasted Chat* by observing the yellow breast of the bird when given foils like *Blue Jay*, *Red-Winged Blackbird*, and *White-Breasted Nuthatch*. But if *Yellow Warbler* is included in the foils, some knowledge about the bird would be required to correctly identify the target bird. Birds that are geographically widespread across North America were picked so that participants across the U.S. can perform equally on the test despite their residence in different areas of the country. Detailed coverage map is listed in the Appendix.

The initial test contains 85 test questions. Easy questions appear relatively early in the test, while difficult questions appear relatively late in the test, to engage participants throughout the test. Test questions range from easy identifications, like *Blue Jay* or *Rock Pi-*








Levels		Respondents	Responses to Items	Examples (choice A is the correct answer for all examples)
Level 1	Novice	Novices who rarely or never do bird watching, have no formal coursework in ornithology, do not travel primarily for birding, less skilled than most birders	correctly identify everyday birds of North America	a) Blue Jay b) Clark's Nutcracker c) Mountain Bluebird d) Carolina Wren 
Level 2	Beginner	Recognized among birders as beginners, they may do occasional bird watching, have one or more formal workshops or community courses in ornithology, occasionally travel primarily for birding, less skilled than many birders	correctly identify common birdfeeder birds of North America	a) Steller's Jay b) Blue Jay c) Blue Grosbeak d) Hermit Warbler 
Level 3	Intermediate	Recognized among birders as intermediate birders, they may do bird watching every 2-3 weeks, have one or more college-level courses in ornithology, travel primarily for birding every other year	correctly identify less common, but relatively distinctive birds	a) Cactus Wren b) Carolina Wren, c) Abert's Towhee d) Fox Sparrow 
Level 4	Advanced	Recognized among birders as advanced birder, they may do bird watching once a week, have Master or Bachelor's degree in ornithology, travel primarily for birding once a year, can lead birding tour groups professionally	correctly identify relatively uncommon confusable birds	a) Acorn Woodpecker b) Red-headed Woodpecker c) Gila Woodpecker d) Yellow-bellied Sapsucker 
Level 5	Expert	Recognized among birders as experts, they may do bird watching two or more times per weeks, have PhD degree in ornithology or related disciplines, travel primarily for birding more than once a year or even regularly	correctly identify the most difficult bird identification challenges	a) Nashville Warbler b) Orange-crowned Warbler c) Tennessee Warbler d) Palm Warbler 

Figure 2.1: The construct map of the bird expertise test

*geon*, common birds throughout North America, to birds most beginners know, like *Tufted Titmouse* or *Carolina Wren*, common to parks and many backyard bird feeders, to distinctive yet far less common birds, like *Pileated Woodpecker* or *Great Kiskadee*, ranging up to quite difficult identifications that even fairly expert bird watchers can find difficult, like discriminating *Bohemian Waxwing* from *Cedar Waxwing*, *Hairy Woodpecker* from *Downy Woodpecker*, or correctly identifying the many extremely similar Warblers, Sparrows, or Flycatchers.

**Measurement Tool.** The IRT models were used to evaluate the test items. One central question in IRT models concerns the dimensionality assumption. In this test, the underlying construct is assumed to be one dimension, i.e., birders are seen as progressing from little or none bird identification ability as a novice to having more such ability as an expert. Experts can identify various birds more quickly and accurately than novices simply because of the one-dimensional bird identification ability, rather than a high-dimensional composite of abilities. However, in some tests, performance is indeed assessed on multiple dimensions. For example, mathematical questions can be measuring both reading comprehension and mathematical skills if the questions are not written in plain English. The dimensionality of a test can be assessed using an exploratory factor analysis (EFA; e.g., Thompson, 2004), a widely-used statistical technique for dimensionality reduction.

Once the dimensionality is determined, the IRT models can be used to assess the test items by relating the scored responses to person latent ability and item attributes such as difficulty, discriminability, and guessing rate. Among the item attributes, discriminability and guessing rate are optional parameters. The best-fitting model can be selected based on model selection criteria such as Akaike's information criterion (AIC; Akaike, 1973), Bayesian information criterion (BIC; Schwarz, 1978), and the likelihood ratio test (LRT).

The simplest IRT model is the Rasch model, or one-parameter logistic (1PL) item response model. In this model, the log odds (logit) of the probability of one person scoring in a given question is the linear function of that person's ability minus the difficulty of the

question, as shown in the equation below:

$$\log \left( \frac{\pi_{pi}}{1 - \pi_{pi}} \right) = \theta_p - \beta_i. \quad (2.1)$$

In this equation,  $\pi_{pi}$  represents person  $p$ 's probability of answering question  $i$  correctly,  $\theta_p$  represents the ability of person  $p$ , and  $\beta_i$  represents the difficulty of item  $i$ . Ideally, the bird expertise test should have questions with a range of difficulty spanning the whole expertise continuum, since the goal is to measure expertise equally well along the whole continuum. The second model is the two-parameter logistic (2PL) item response model, which includes an additional item discriminability parameter compared to the Rasch model, as shown in the equation below:

$$\log \left( \frac{\pi_{pi}}{1 - \pi_{pi}} \right) = \alpha_i(\theta_p - \beta_i). \quad (2.2)$$

The item discriminability parameter,  $\alpha_i$ , represents item  $i$ 's ability to distinguish high-ability participants from low-ability participants, with higher  $\alpha$  representing higher discriminability of the question. Since the goal of a test is to tell people apart, i.e., to discriminate, higher  $\alpha$  is thus more desirable. Negative  $\alpha$  would suggest confusing items for which novices get correct while experts get wrong, which should be removed. The three-parameter logistic (3PL) item response model includes an additional item guessing parameter compared to the 2PL model, as shown below:

$$\pi_{pi} = \gamma_i + (1 - \gamma_i) \frac{\exp[\alpha_i(\theta_p - \beta_i)]}{1 + \exp[\alpha_i(\theta_p - \beta_i)]}. \quad (2.3)$$

The item guessing parameter,  $\gamma_i$ , represents the probability of participants answering question  $i$  correctly by pure guessing. Since the bird expertise test consists of four-alternative forced-choice questions, it is reasonable to consider the inclusion of an item guessing parameter.

The parameter estimates for items were used to guide item selection. For example,

item discriminability is expected to be positive, which means that experts should have a higher chance of scoring the items than novices. Negative discriminability could suggest a confusing question. The distribution of item difficulty was examined to determine whether participants along the whole expertise continuum were measured equally well. The guessing rate is expected to be around 0.25, while any estimates far off might indicate poor item design.

**Questionnaire.** The questionnaire collected participants' demographic information, self-reported bird watching experience, and self-rated birding expertise. First, participants answered questions about their age, gender, and whether they had any neurological conditions that might affect their vision, hearing, memory, or thinking. Next, they answered questions about their bird watching experience, including the length of their bird watching experience, relevant training, bird watching frequency, and relevant travel frequency. They were also asked to rate on a scale of 1 (novice) to 5 (expert) their overall expertise level as well as specific expertise for nine birding regions covering North America, including: Eastern US and Eastern Canada, Western US and Western Canada, Arctic, Pelagic (Atlantic and Pacific), South Texas, Southeast Arizona, South Florida, Mexico and Central America, and Caribbean. The complete questionnaire can be found in the Appendix.

### 2.3 Results

The questions were evaluated and revised using IRT through an iterative process. After constructing the initial test, the test was implemented in JavaScript as a web page and administered online to bird watchers across the U.S. For the first round, responses from 332 participants (186 female) for a test of 85 questions were analyzed. Eigenvalues estimated in an EFA using the EQSIRT software (Bentler & Wu, 2012) suggested one general factor: the eigenvalues for the first 10 factors were 41.50, 3.91, 3.05, 2.39, 2.12, 1.94, 1.82, 1.77, 1.68, and 1.62, respectively. The ratio of the first to second eigenvalue was 10.618, indicating a dominant general factor, i.e., one dimensionality in the test.

Estimates for the difficulty and discriminability parameter were used to select items. Item discriminability parameters for all items were greater than 0, which suggests that the items discriminate in favor of the expert group as desired. One item had an extreme difficulty parameter estimate at -45.91 (“item 2”), which means the item was too easy and was thus removed. For the other items, difficulty parameter estimates had a mean of -0.56. On the person side, the participants’ ability was assumed to have a mean of 0. This suggests that the initial test was outperformed by the participants. In particular, very few questions tapped into the upper end of the expertise continuum. To cover the entire range of participants’ ability, more questions were then added. Data from the second version of the test were collected and modeled using IRT again to select questions. Eventually, 93 questions were kept in the current version of the test (see Appendix). The results below were based on analysis of responses to the current version of the test.

**Model Selection.** The dimensionality of the test was assessed using an EFA in the EQSIRT software. The Eigenvalues suggested one general factor: the eigenvalues for the first 10 factors were 37.66, 5.76, 2.88, 2.38, 2.10, 1.98, 1.88, 1.71, 1.60, and 1.54, respectively Figure 2.2. The ratio of the first to second eigenvalue was 6.54, indicating a dominant general factor (Reise, Moore, & Haviland, 2010). Thus it is safe to assume unidimensionality in the test, suggesting that the test was measuring one major construct along a single dimension.

Second, given that one dominant dimension was sufficient to explain item variances, the three variations of unidimensional item response models were considered to determine the final model to be used to measure participants’ expertise scores and items’ difficulty scores.

The three versions of IRT models were fitted to response data using the R package *ltm* (Version 1.0-0, Rizopoulos, 2006). The likelihood ratio test (LRT) result indicated that the two-parameter unidimensional item response model fit better than the Rasch unidimensional model (Chi-square value = 1750.93,  $df = 93$ ,  $p < 0.05$ ), and the three-parameter

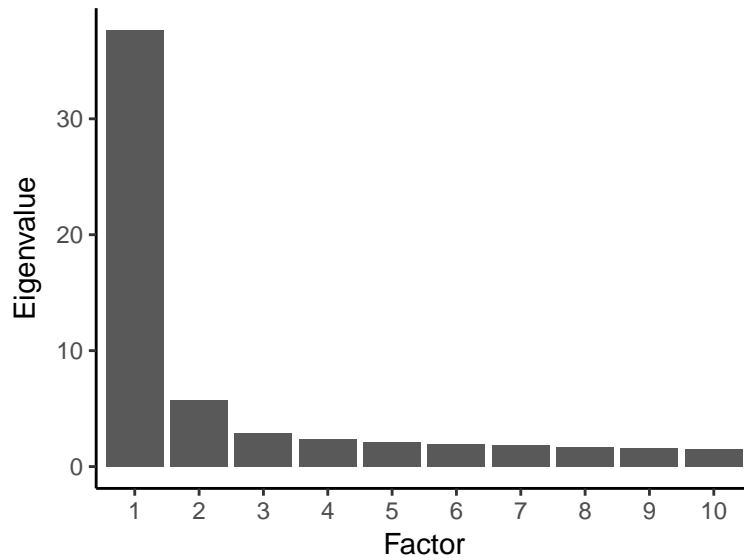


Figure 2.2: The Eigenvalues of the first 10 factors using exploratory factor analysis. The ratio of the first to second eigenvalue was 6.54, indicating a dominant general factor.

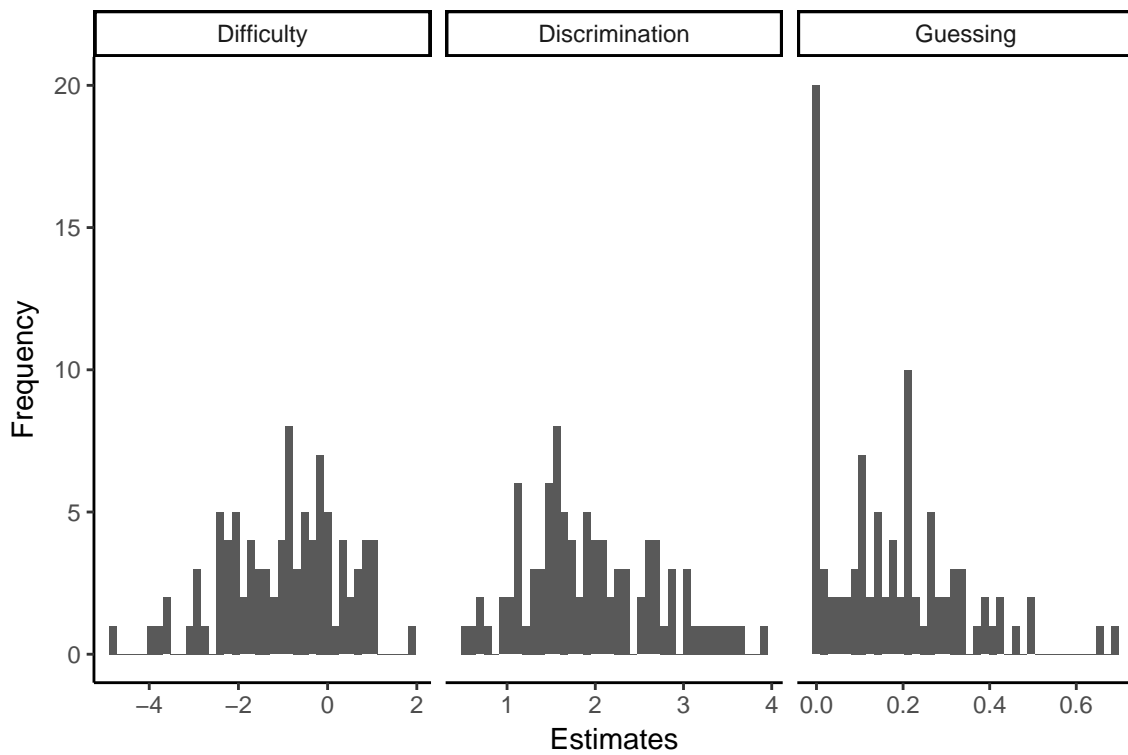


Figure 2.3: The density plots for the parameter estimates including difficulty, discriminability, and guessing parameter from the three-parameter model.

unidimensional item response model fit better than the two-parameter model (Chi-square value = 779.24,  $df = 94$ ,  $p < 0.05$ ). Therefore, based on available LRT and item fit results, a three-parameter unidimensional item response model fit the data the best. This model was thus used to estimate the participants' ability and the items' difficulty. Information criteria supported this result, as the Akaike's information criterion (AIC) and Bayesian information criterion (BIC) for the three-parameter unidimensional model (AIC = 55491.84; BIC = 56791.30) were smaller than those of the Rasch unidimensional model (AIC = 57648.01; BIC = 58085.77) and the two-parameter unidimensional model (AIC = 56083.08; BIC = 56949.39).

The Wright map from IRT analysis showcases the participants' ability and the items' difficulty distribution side by side, which provides a nice visualization to determine whether the test is difficult enough for the sample population. As can be seen in Figure 2.4 (created using the R package *WrightMap*, Irribarra & Freund, 2014), the estimated item difficulty lined up with designed difficulty really well, with a Pearson rank order correlation of 0.77. Also, the person ability distribution sat a little higher than the item difficulty distribution along the y axis, suggesting that there were a few experts who were better than the test. Since the person ability distribution was slightly negatively skewed but overall normal, the test was not concerned as being too easy for the whole sample. Overall, the test could discriminate individuals at different expertise levels and cover a wide range of the expertise continuum.

**Explanatory Modeling.** To understand the testing scores on both item and person sides, person and item covariates were added to the modeling process consecutively<sup>1</sup>. Instead of adding all covariates in the model at once, I started with a basic IRT model with person random effects and item random effects. Then the item covariates were added, con-

---

<sup>1</sup>Here, the item difficulty was explained using several item covariates, without considering the item guessing and discrimination parameters. This could be problematic if the item difficulty was distorted from the true estimates without the constraint of the guessing and discrimination parameters. This was not an issue in this sample, since the item difficulty parameter estimates from the 1PL and 3PL models were highly correlated ( $r = 0.99$ ,  $p < 0.05$ ).

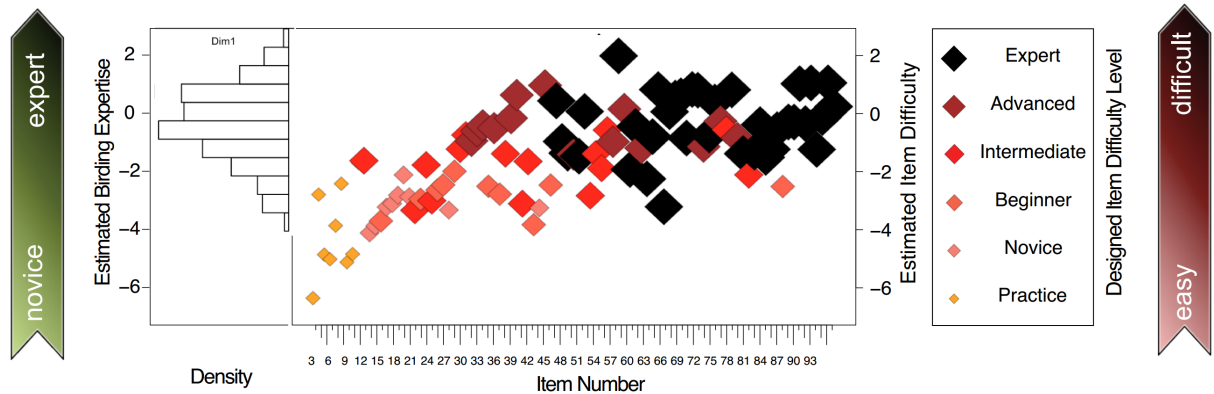


Figure 2.4: The Wright map of IRT results from the three-parameter model. The histogram on the left side shows the distribution of person ability along the standardized scale. On the right side, each diamond represents one question. The color of the diamonds represents designed difficulty level of questions. The locations of the diamonds along y axis represent their estimated difficulty.

stituting the linear logistic test model (LLTM). Lastly, the person covariates were added, constituting the latent regression LLTM. The model was run in R 3.3.3 using R package *lme4* (Version 1.1-12, Bates, Mchler, Bolker, & Walker, 2015) and *optimx* (Nash, 2014; Version 2013.8.7, Nash & Varadhan, 2011).

Several factors were considered as item covariates, including the commonness of the target bird in the question (commonness rating from the American Birding Association, the values are 1-*common* or 2-*less common* among the birds used in the test), the gender of the target bird in the testing picture (*female* dummy coded as 1, *male* as 0), and three covariates on foil design. There were three foils/distractors in each question. Each of the three foils could be bird species labels from different orders, different families, different genera, or the same genus as the target bird. Thus, the foil design yielded four variables that could be used as covariates: the number of foils from different orders, different families, different genera, and the same genus as the target bird. Since these four covariates add up to three, the total number of foils, the last covariate, number of foils from the same genus, was dropped to avoid perfect collinearity in the covariates. With item covariates but not person covariates, the model was called the LLTM. The likelihood ratio test indicated that this model fit better



than the basic Rasch model (Chi-square value = 112.20,  $df = 6$ ,  $p < 0.05$ ).

Several hypotheses were made about the effect of the item covariates on item difficulty. It was reasonable to assume that less common birds and female birds were harder to identify, thus positive coefficients were expected for target bird commonness and the gender of the bird picture in predicting item difficulty. Given the taxonomic tree (order  $\rightarrow$  family  $\rightarrow$  genus  $\rightarrow$  species), birds from different orders are more different from each other in visual appearance than birds from different families but the same order, which are then more different than birds from different genera but the same family. Birds from the same genus are the most similar to each other, making the questions the most difficult. But those from different orders, families, and genera should all make the question easier. Thus it was hypothesized that the number of foils from different orders, families, and genera all correlated negatively with item difficulty.

Finally, I added person covariates in the model as well, including person birding frequency, birding-relevant training, birding experience, birders' gender, and their medical conditions (whether they had any neurological conditions that might affect their vision, hearing, memory, or thinking or not). With both item and person side covariates, the model was called the latent regression LLTM. The likelihood ratio test result indicated that the latent regression LLTM fit better than the LLTM (Chi-square value = 470.86,  $df = 18$ ,  $p < 0.05$ ). The coefficient estimates from latent regression LLTM are shown in Table 2.1

For the five item factors analyzed, only the factors that were relevant to foil structures were significant. Questions having one more foils from different orders tended to be 0.86 unit lower on the logit scale ( $z = -6.08$ ,  $p < 0.05$ ), suggesting decreased difficulty. Questions having one more foils from the same order, but different families were 0.78 unit lower on the logit scale ( $z = -5.11$ ,  $p < 0.05$ ). Questions having one more foils from the same order, the same family, but different genera were 0.39 unit lower on the logit scale ( $z = -2.98$ ,  $p < 0.05$ ). Thus among these three variables, number of foils from different orders had the biggest effect in making the items easy, while the number of foils from the same

order, the same family, but different genera had the smallest effect in making the item easy. These results were consistent with the predictions, suggesting that this foil design was very effective. In contrast, the commonness of the target bird ( $z = 0.36, p = 0.13$ ) and using female bird image did not affect item difficulty significantly ( $z = 0.16, p = 0.88$ ). This was likely due to a sampling issue. In the test, there were only 4 items that were less common and 8 items that used female bird images, among all 93 items. With such few items and a wide range of difficulty for the other items, these few items were less likely to be detected as significant in affecting item difficulty.

Table 2.1: Coefficients estimates from the latent regression LLTM

	Estimate	Std. Error	z value	Pr(> z )
Rare Target Bird	0.05	0.12	-0.38	0.70
Female Bird Image	0.03	0.12	-0.22	0.83
Number of Foils - Different Orders	-0.86	0.14	6.18	0.00
Number of Foils - Different Families	-0.78	0.14	5.48	0.00
Number of Foils - Different Genera	-0.39	0.13	2.87	0.00
Frequency - Rarely	1.70	0.36	4.74	0.00
Frequency - Occasionally	2.21	0.28	7.77	0.00
Frequency - Every Two Weeks	2.54	0.29	8.78	0.00
Frequency - Once per Week	2.70	0.28	9.62	0.00
Frequency - Twice per Week	3.08	0.28	11.01	0.00
Experience in Year	0.12	0.04	2.98	0.00
Female Birder	-0.24	0.04	-6.43	0.00
Training - None	-0.60	0.24	-2.52	0.01
Training - One Workshop	-0.63	0.25	-2.53	0.01
Training - One Course	-0.31	0.25	-1.22	0.22
Training - Two Courses	-0.28	0.27	-1.02	0.31
Training - Masters	-0.12	0.26	-0.45	0.65
Travel - Every Few Years	-0.31	0.13	-2.39	0.02
Travel - More Than Once Yearly	0.11	0.11	1.00	0.32
Travel - Every Other Year	0.27	0.19	1.42	0.16
Travel - Regularly as A Professional	0.40	0.17	2.40	0.02
Travel - Rarely	-0.56	0.11	-4.92	0.00
Medical Condition	-0.03	0.04	-0.72	0.47

Three of the person covariates, *birding frequency*, *birding-relevant training*, and *birding-relevant travel*, were recorded as ordinal variables in the data set. Each level of these covariates was dummy coded, so that the effect of each level of the covariates was estimated. It was found that all levels of birding frequency were significant, with higher birding frequency associated with a stronger increase in the person ability. Among all levels of training, only the levels *no training* and *one workshop* were significant, with these two levels negatively associated with person ability. Among all travel levels, professionals who reported traveling regularly for birding purposes had significantly higher ability estimates than the average, while people who reported traveling infrequently (once every few years or rarely) had significantly lower ability estimates. These results were consistent with the expectations, in that people with more experience and more training would have higher birding expertise. Overall, more training was associated with higher person abilities. For birding frequency, every level of this ordinal variable differed significantly from the intercept, suggesting that birding frequency predicted the birding skill significantly.

All other person covariates had a significant effect, except for the medical condition factor. Person ability estimates increased by 0.12 on the logit unit per year of experience increases ( $z = 2.98, p < 0.05$ ). Note that the age covariate was not added because age correlated moderately with experience ( $t = 7.66, df = 678, p < 0.05$ ) but not with ability estimates ( $t = 0.30, df = 683, p = 0.76$ ). Gender was another significant factor in that female birders had a lower ability estimates than male birders by 0.38 unit on the logit scale ( $z = -5.05, p < 0.05$ ). The factor medical condition, on the other hand, did not have a significant effect on person ability ( $z = 0.15, p = 0.88$ ). This is likely due to a sampling issue. Since there were only 25 participants reported as having had some medical conditions, they were not likely to have a wide range of birding expertise (-1.96,1.61), whereas the other 662 healthy participants did have a wide range of birding expertise (-2.87,2.48). Thus it was less likely that the self-reported patients would be detected as different from the healthy participants in their birding ability estimates. I also analyzed how well birders' self-rated

expertise level correlate with their estimated expertise. Interestingly, the correlation is decently high with a Pearson correlation of 0.59 (Figure 2.5).

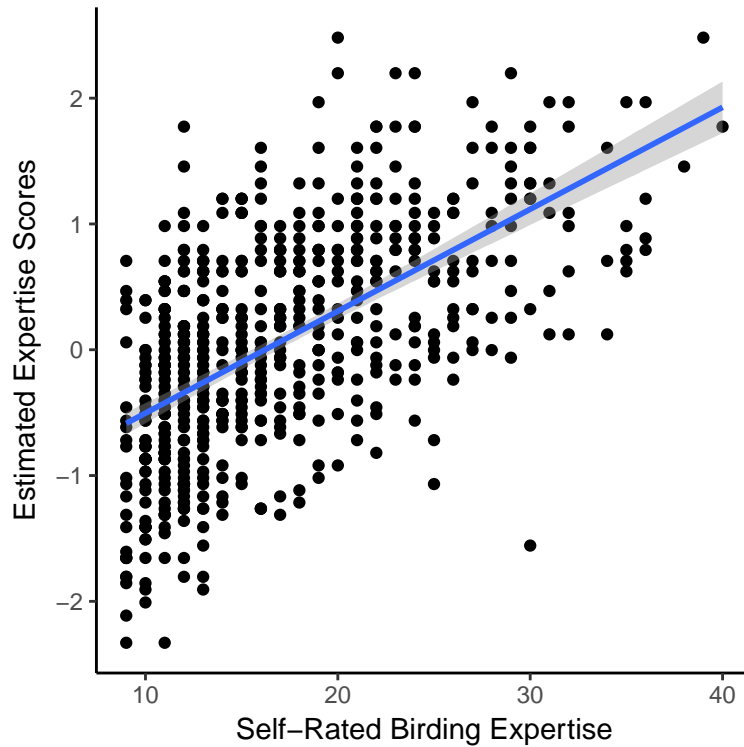


Figure 2.5: A scatter plot of estimated birding expertise over self-rated birding expertise

## 2.4 Discussion

In summary, the bird expertise test was shown to provide a valid and reliable measure of the birding expertise of the participants. The ability estimates from the bird expertise test using IRT correlated reasonably well with participants' self-rated ability and their performance on the specific bird expertise tests.

Using latent regression LLTM, I explored the effect of person covariates and item covariates on person ability and item difficulty estimates. On the person side, it was found that birding frequency, gender, experience, and training all associated with person birding skill significantly. The factor medical condition did not correlate with person ability estimates, potentially due to the overly restricted sample size of the patient population. On the item

side, it was found that the item design factors associated with the item difficulty estimates in the expected direction. In particular, foil structure was very effective in determining item difficulty. The other two factors, commonness of the target bird and female/male bird image, did not affect item difficulty significantly. However, the direction of influence was as expected.

As mentioned in the introduction, there exist different measures of perceptual expertise, such as the matching tests and the VET. The expertise scores can be used to explore the covariance with perceptual discriminability and perceptual memory, as measured by perceptual matching tasks and the VET. There could be important structure in the patterns of covariance across these three different measures of perceptual expertise. While it is possible that all three emerge from a single latent variable of perceptual expertise, it is also possible that the results from these different measures are not well-correlated. For example, some participants could score high on the perceptual matching tasks, but score low on the bird identification test.

It would be interesting to test the same population on these different kinds of test to measure the correlations among these tests and their convergent validity. Unfortunately this task falls out of the scope of the current project due to the necessity to develop special versions of the matching test and the VET. The previous matching tests was developed for local participants, testing their knowledge with local bird species. To administer such a test, one would first need to replace the bird species with birds that are local to the target population. Then several iterations of test evaluation and revision would be necessary to eventually establish a valid and reliable test. The situation was slightly different for the VET – the performance of the birder population was at ceiling for the VET-bird (Van Gulick et al., 2016), a subset of the VET that uses bird as the testing category. Using the VET-bird again would require iterations of evaluation and revision, particularly adding difficult items for this special birder population. The re-development of these two tests is a good future direction to pursue to understand the measurement of perceptual expertise.

Also, further research is necessary to parse out the variability in perceptual expertise using various measures such as IQ and demographic information, as well as various exploratory and confirmatory data analysis techniques such as clustering, structural equation modeling, principal components analysis.

## Chapter 3

### Modeling the Dynamics of Expertise

#### 3.1 Introduction

Many people spend years to become experts in categorizing and identifying specialized image sets, such as airport baggage screening, forensic fingerprint identification, imaging-based diagnosis, bird identification etc. An important hallmark of such visual expertise is the entry-level shift phenomenon: while it takes novices significantly longer to categorize at the subordinate level (e.g., *Blue Jay* or *Eastern Bluebird* in the case of bird identification) than at the basic level (e.g., *Bird* or *Dog*), experts are equally fast at the two abstraction levels.

What gives rise to the entry-level shift phenomenon? Why do people with varying levels of expertise differ from each other in visual category-verification? This question was first tackled by addressing how novices make category decisions. In their seminal 1976 paper, Rosch and colleagues pioneered in investigating people's visual categorization of natural objects (Mervis & Rosch, 1981; Rosch et al., 1976). They defined the basic category as "the most general and inclusive level at which categories can delineate real-world correlational structures" (p. 384). Generally, basic categories are intermediate level categories (e.g., *bird* vs. *dog*), as opposed to superordinate categories (e.g., *animal* vs. *vehicle*) and subordinate categories (e.g., *Blue Jay* vs. *Eastern Bluebird*). The authors identified multi-fold advantages of basic categories. The one of most interest to the current project is the entry-level shift phenomenon, which was observed from a paradigm called the speeded category verification task. This task is now widely used as a staple task to study expertise, with entry-level shift phenomenon deemed a hallmark of visual expertise. In this task, a trial begins with a category label, followed some time (500~1000ms) later by an image containing an object. Participants then verify whether the object is a member of



the category presented. The category labels are from three levels of abstraction, including the superordinate level, the intermediate level, and the subordinate levels. Among the three levels, participants were fastest in the basic-level category verification.

What gives rise to this basic-level advantage? One theory viewed categorization as a level dependent process (Grill-Spector & Kanwisher, 2005; Jolicoeur, Gluck, & Kosslyn, 1984; Rosch et al., 1976). That is, basic-level categorization is fastest because it is the first stage of categorization, after which further conceptual/perceptual analysis is required to make superordinate/subordinate-level categorizations. Basic level was later termed the *entry level* to emphasize that many objects make contact with semantic representations first at the basic level (Jolicoeur et al., 1984). In other words, the subordinate-level categorization ends later because of later onset of the process, i.e., one categorizes an object as *Blue Jay* only after determining that it is a *bird* (but see Mack & Palmeri, 2011; Mack, Wong, Gauthier, Tanaka, & Palmeri, 2009). Contrary to this basic-first hypothesis, Murphy and Brownell's (1985) differentiation hypothesis does not assume basic-level categorization as a prerequisite before further category decisions. Instead, categorizations at different levels are independent and are influenced by category accessibility. They argued that basic categories are most accessible, which leads to the fastest category verification responses. In other words, the subordinate-level categorization ends later because of slower accumulation of evidence for the category decisions, not because of later onset of the categorization process (see Figure 3.1 for illustration). This view was later extended as the parallel distributed processing theory to explain the basic-level advantage (see Rogers & Patterson, 2007).

The question then is: how do experts categorize differently from novices? Using the same speeded category verification paradigm as used to identify the basic-level advantage, researchers observed that for experts, subordinate-level categories are verified as fast as basic-level categories (K. E. Johnson & Mervis, 1997; Tanaka & Taylor, 1991). The term *entry-level shift* was thus coined to denote that for experts, the subordinate level functions

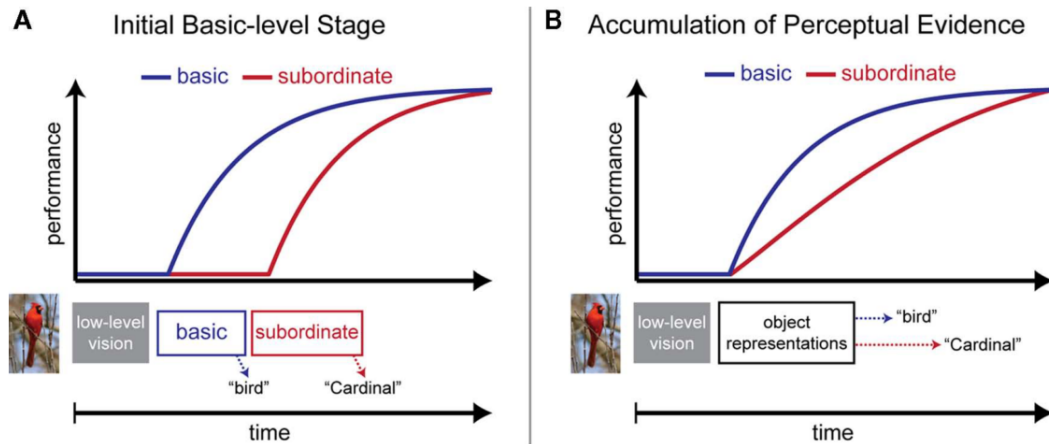


Figure 3.1: An illustration of the basic-first hypothesis (A) and the differentiation hypothesis (B). A: In the basic-first hypothesis, the decisions made at the subordinate level are slower because of a delayed onset of the decision process. B: In the differentiation hypothesis, the decisions at the subordinate level are slower because of slower accumulation of evidence over time. Adapted from Mack & Palmeri (2011).

just like the basic level, i.e. an entry level.

What gives rise to this expertise-related response time (RT) change? In line with the basic-first hypothesis for novices' categorization behavior, it is possible that in addition to having the basic level as an entry level, experts also have the subordinate level as an extra entry level into conceptual knowledge. Thus categorizations at the basic and the subordinate levels are equally fast. In other words, the shorter RT for the subordinate-level categorization are from earlier onset of the process. In contrast, the differentiation hypothesis predicts that as expertise increases, the subordinate category becomes as accessible as the basic category. Thus expertise is associated with accelerated accumulation of evidence for decision-making, which makes categorizations at subordinate levels faster (Palmeri et al., 2004).

Few attempts have been made to theoretically differentiate these two explanations. Using a signal-to-response technique, Mack et al. (2009) systematically varied the time participant were given to process a task and mapped out the onset and evolvement of categorization performance over time. They found that for novices, there is no difference in the

onset of basic vs. subordinate-level categorizations. This finding suggests that novices categorize faster at basic level because of faster information processing speed than that of the subordinate level, rather than an earlier onset of the process. Curby and Gauthier (2009) used a backward masking paradigm and found that experience influences the availability of information early in processing, supporting the differentiation hypothesis. Novices and experts were often sampled as two or three points along the continuum of expertise development (K. E. Johnson & Mervis, 1997; Tanaka & Taylor, 1991), essentially treating expertise as binary or ternary rather than continuous. While it is interesting to contrast two or three drastically different groups, information about the whole trajectory is largely missing.

To formally differentiate the two competing hypotheses, this project seeks to describe differences in individual categorization behavior along the continuum of expertise and to explain the differences theoretically using process models. These are important yet difficult questions, requiring a combination of psychometric and cognitive models to be fully addressed. Cognitive models can provide a detailed account of the underlying psychological processes, but lack the strength of incorporating individual differences, mostly due to the complexity of most cognitive models. Thus in cognitive modeling, oftentimes only group averages are modeled to test specific mechanistic hypotheses (e.g., Nosofsky & Palmeri, 1997; Ratcliff, 1981; Viken, Treat, Nosofsky, McFall, & Palmeri, 2002). Psychometric models could fill in the gap by incorporating individual differences. Yet these models often lack the power to test mechanistic hypotheses and to explain the underlying cognitive processes (for a review, see Shen & Palmeri, 2016). In this project, cognitive and psychometric models were combined to provide a detailed account of the psychological processes underlying the RT differences along the continuum of expertise.

The sequential sampling models provide ideal tools to understand the RT differences observed in various speeded two-choice tasks. This class of models includes three main players (for a review of this class of models, see Ratcliff, Gomez, & McKoon, 2004):

the diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008), the linear ballistic model (S. D. Brown & Heathcote, 2005, S. D. Brown & Heathcote (2008)), and the random walk model (Nosofsky & Palmeri, 1997; Nosofsky & Stanton, 2005). In such models, it is assumed that evidence accumulates over time from a starting point toward two or more response boundaries. A decision is made when one of the boundaries is crossed. The choice and RT for each trial is mainly decided by four process parameters, including the starting point ( $b$ , representing priori bias), the separation of the two boundaries ( $A$ , representing response caution), the drift rate of evidence accumulation ( $v$ , representing rate of information processing), and non-decision time ( $\tau$ , representing time spent on processes other than evidence accumulation). These parameters could vary across conditions, i.e., they are potential candidates of the psychological differences underlying the entry-level shift phenomenon (Figure 3.2).

The sequential sampling models were chosen to model categorization decisions for several reasons. The psychological interpretations of its parameters have been justified by many studies (Ratcliff & Rouder, 1998; Voss, Rothermund, & Voss, 2004). For example, higher drift rate was found for easier stimuli, while wider boundary separation was found when accuracy was emphasized over speed (Voss et al., 2004). Also, these models consider RT distribution and accuracy data simultaneously and thus naturally account for speed-accuracy trade-off. In addition, among its broad applications to psychological tasks in various fields, it has been successful in explaining the learning process (for a review, see Ratcliff et al., 2004). For example, by fitting the diffusion model to RT data collected in a 10,000-trial lexical decision task, Dutilh and colleagues (2009) found that the practice effect consists of several components, including increased evidence quality, decreased non-decision time, and decreased response caution. Last but not least, in addition to statistically deconstructing the expertise effect (see also Petrov, Van Horn, & Ratcliff, 2011; Ratcliff, Thapar, & McKoon, 2006), the model also shows the potential for incorporating mechanistic components to fully explain individual differences in expertise. Notably, the

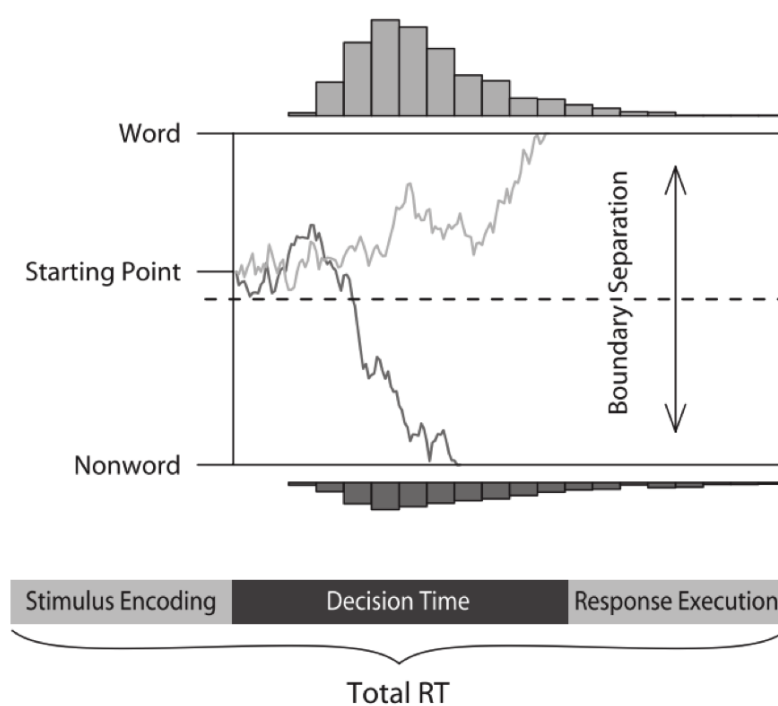


Figure 3.2: An illustration of the diffusion decision process for a choice between word and nonword. From a starting point, evidence accumulates over time toward the two response boundaries: *word* or *nonword*. A decision is made when one of the boundaries is crossed. The light gray trace is an example of a decision-making process when *word* is chosen, while the darker gray trace shows a trace of *nonword* choice. Adapted from Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers (2009).

exemplar-based random walk model (EBRW; Nosofsky & Palmeri, 1997) – a special version of the diffusion model in which evidence accumulates discretely in small steps rather than continuously – successfully explained the drift rate change as a result of an efficient retrieval of exemplars in memory for expert perceptual classification. Given these considerations, the diffusion model provides an ideal technique for us to deconstruct expertise effect in speeded two-choice tasks into its constituent psychological processes.

For subordinate-level categorizations, expertise related faster RT could be due to either a faster processing speed (differentiation hypothesis) or an earlier onset of the process (basic-first hypothesis). The sequential sampling models could easily translate these two possibilities into the changes in its process parameters. To be specific, faster processing speed can be interpreted as higher drift rate  $\nu$ , as the drift rate naturally accounts for accumulation of evidence for the decision-making process; while earlier onset of the process can be seen as shorter non-decision time  $\tau$ , as the basic-level identification process is now removed from non-decision processes. Therefore,  $\nu$  for the expertise condition can be used to differentiate the two possible mechanisms of the expertise effect in categorization. If the basic-first hypothesis is true, it is expected that  $\nu$  does not correlate with expertise. In contrast, if the differentiation hypothesis is true, one would expect a positive correlation between  $\nu$  and expertise.

The diffusion model was used to deconstruct the entry-level shift phenomenon into its constituent psychological constructs. In order to map out the differences associated with expertise, the diffusion model was applied to the behavioral data of participants with varying levels of expertise. The expertise score was used to predict individual differences in model parameters. This could be done in two ways. One is to fit the diffusion model to individual data and then make inferences based on individual parameters. The other is to view individuals' parameters as randomly drawn from a population distribution, so all data can be analyzed simultaneously to find the population mean and the variability for each parameter (a population/hierarchical model, Figure 3.3). The second approach is better

because it allows more reliable parameter estimates, while requiring fewer data points per participant (Linden, 2007; Vandekerckhove et al., 2011). It can also be used to explain individual differences in parameters from a predictor (e.g., expertise score in this case) within the model specification.

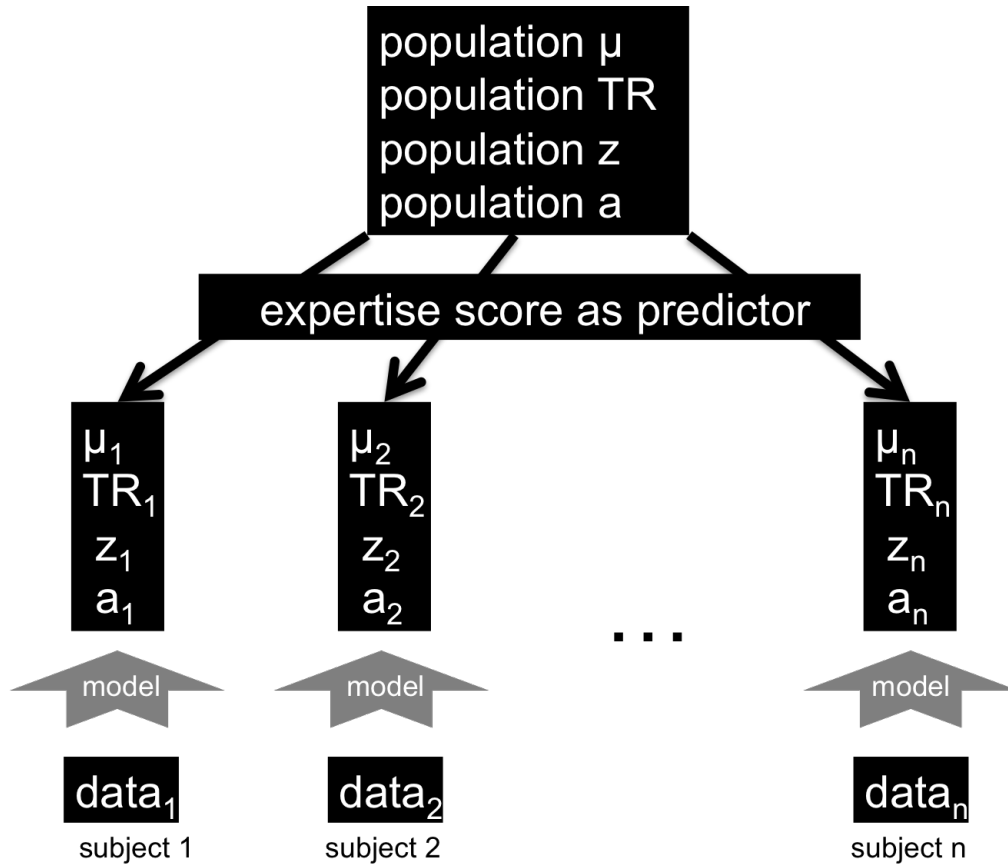


Figure 3.3: A demonstration of the hierarchical diffusion model. Individuals parameters are assumed to be randomly drawn from a population distribution. Expertise score is used as a predictor for the parameters in certain conditions (see the modeling section in Results).

Yet the hierarchical approach is technically challenging, especially given the complexity of the diffusion model. Thanks to the work of Vandekerckhove and others (Dutilh et al., 2009; Vandekerckhove et al., 2011; Wabersich & Vandekerckhove, 2014; Wiecki et al., 2013), the hierarchical diffusion model (HDM) was made possible to implement in free software like JAGS (Plummer, 2003) and WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). Therefore, I chose the hierarchical diffusion model to fit choice RT data from a population with varying levels of expertise.

### 3.2 Method

**Participants.** Bird watchers with all levels of expertise were recruited online. To solicit participation, members of 59 birding societies across the U.S. were contacted. During the initial recruitment, 314 bird watchers (113 female) participated voluntarily in the bird expertise test as described in Chapter 2. Next, invitations were sent through email to all 314 participants who finished the bird expertise test, 63 of which (26 female) participated in the behavioral task and were entered into a poll to receive Amazon gift card. Informed consent was obtained prior to participation in accordance with Vanderbilt University's Institutional Review Board.

**Stimuli.** Test stimuli were 256 pictures of common birds and dogs. Bird stimuli consisted of pictures of eight geographically wide-spread and common bird species in the U.S. according to the Cornell Lab of Ornithology (<http://www.allaboutbirds.org>), including *Cardinal*, *Blue Jay*, *Crow*, *Hawk*, *Oriole*, *Pigeon*, *Robin*, and *Sparrow*. Dog stimuli consisted of pictures of eight most popular dog breeds in the U.S. according to the American Kennel Club (<http://www.akc.org/>), including *Labrador Retriever*, *German Shepherd*, *Yorkshire Terrier*, *Beagle*, *Golden Retriever*, *Bulldog*, *Boxer*, and *Dachshund*. Participants' familiarity with these dog breeds and bird species were confirmed before the experiment. I collected 24 pictures for each bird or dog species from online resources and from friends (12 for basic-level categorizations and 12 for subordinate-level categorizations). The objects in the images were of various orientations and various background contexts. The images were cropped so that the objects were centered and prominent, limiting the influence of background. No image was repeated during the entire experiment. Figure 3.4 shows some example stimuli.

**Procedure.** After participants finished the bird expertise test, they were invited via email to complete the speeded category-verification task online. I adapted the classic paradigm used in previous research (K. E. Johnson & Mervis, 1997; Rosch et al., 1976) in several ways to better fit the online testing setting and the modeling purpose. First, in





Figure 3.4: Some example stimuli used in the speeded category-verification task

online testing settings, participants can be easily distracted by their surroundings and other applications on their computer, critically limiting their engagement time in experiment sessions. To minimize the attrition rate due to a lengthy session, the task was made as short as possible by including only relevant conditions. In the classic paradigm, participants were tested at three different levels of abstraction, including the superordinate, the basic, and the subordinate levels. Since it was known from previous studies and the pilot study that performance at the superordinate level does not correlate with expertise (K. E. Johnson & Mervis, 1997; Rosch et al., 1976), I only included basic and subordinate levels in this experiment. Second, the sequential sampling models require some error responses in each condition to be properly constrained (Donkin, Averell, Brown, & Heathcote, 2009; Ratcliff & Childers, 2015). Several pilot studies were done to ensure that participants could make enough errors. Eventually, I added noise to the images and reduced the image presentation time from unlimited to 200 ms. In the pilot study, images with and without noises were both tested. It was found that these adaptations made the task harder, causing more errors in participants' responses. This effect was seen uniformly across all conditions. The overall result pattern was consistent with previous research, as can be seen in the result section.

Three factors were combined in the experimental design, including Levels of Abstraction (*basic* or *subordinate*), Object Category (*expert*: bird or *novice*: dog), and Trial Type (*match* or *mismatch*), which created 8 within-subject conditions. In a *match* trial, the test image contained the object specified in the preceding label, while the image in a *mismatch* trial did not. Images for false trials were always from the same level of abstraction, for example, a *western bluebird* image for a label *BLUE JAY*, or a *golden retriever* image for a label *BIRD*. The test consisted of 256 trials, 32 for each condition. Each trial began with a fixation cross of 500 ms. Then participants saw a category label of various levels of abstraction (e.g., *DOG* for the basic level, *DACHSHUND* for the subordinate level). The category label was displayed for 2,000 ms, so that participants had enough time to understand the label and got prepared for the upcoming image. The label was followed by a fixation cross

with varied length of display (0~1500ms). The fixation period was jittered to minimize effects of anticipation and habituation. Then the test image was presented on the screen for 200 ms. This presentation time was intentionally reduced compared to the classic paradigm in order to collect more error responses. Participants were instructed to decide whether the object in the image matched the label or not. They responded by pressing the “D” key for a “no” response and the “K” key for a “yes” response .

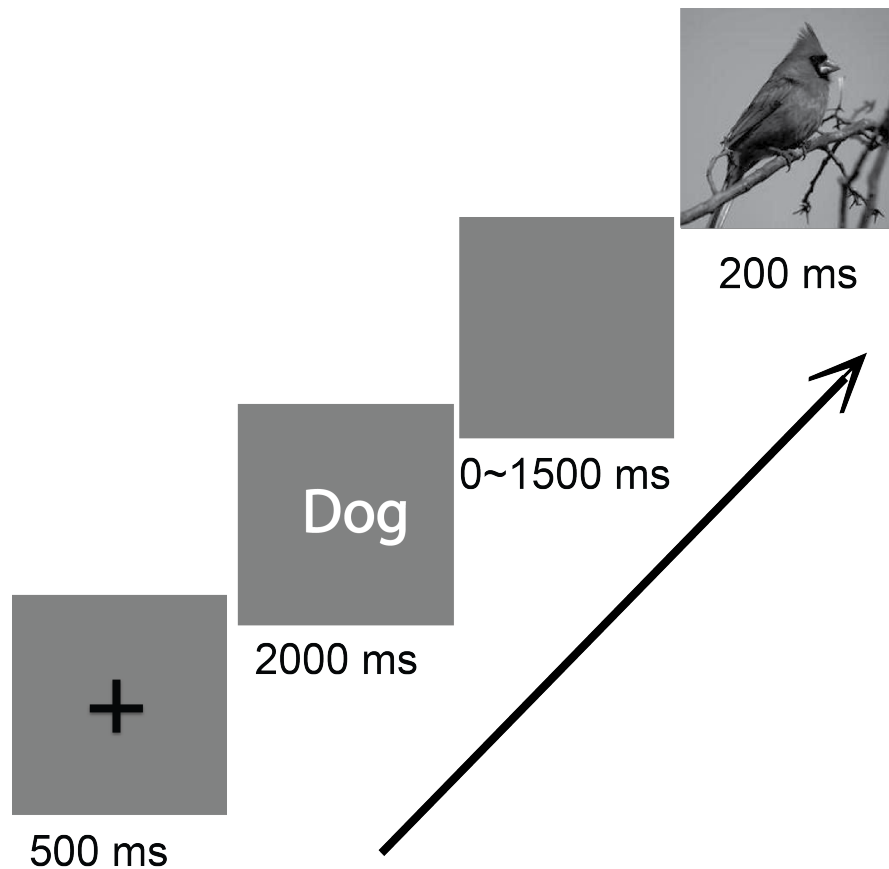


Figure 3.5: An example trial in the speeded category-verification task

The trial sequence was randomized but was limited to one of two randomly-generated sequences for all participants. Since participants’ performances in this task were expected to be different due to their different birding expertise, it was undesirable to include the randomness in trial sequences as a potential source for differences in their performance,

thus the trial sequence was intentionally controlled to be the same. But to check that order did not play a role in causing performance differences, participants were randomly assigned to one of the two randomly-generated sequences to examine the effect of trial sequence. After 8 practice trials, participants went through 2 sessions of 128 trials. Participants were instructed to respond as fast and accurate as possible. If the participant did not respond within 1,500 ms, a warning message appeared on the screen, prompting the participant to respond quickly and accurately. The warning message would remain on the screen until a response was made. The entire session lasted for about 15 minutes.

### 3.3 Behavioral Results

In previous studies, the entry-level shift phenomenon was defined as experts being equally fast at the basic and the subordinate level when categorizing objects from their domain of expertise, but much slower at the subordinate level when the objects are from a novice domain. It was thus expected that experts would show this pattern of interaction between the level of abstraction and the domain of expertise, but not the novices. A similar pattern was observed in this study. Figure 3.6 shows the response time patterns after a median split of the participants based on their expertise scores. Note that participants in the "expert" group were not truly experts but intermediate- to expert-level bird watchers. For this reason, the interaction was not significant for this group, but one can still observe a consistent general pattern with the classic entry-level shift phenomenon.

Since participants in this project scattered along the continuum of expertise rather than being labeled experts/novices in a binary fashion, the interaction magnitude was expected to be continuous rather than binary, with the expertise index as a potential predictor of the interaction magnitude. Instead of using the customary *analysis of variance* (ANOVA) or *analysis of covariance* (ANCOVA) approach, the behavioral results were analyzed using using a *general linear model* (GLM) approach because GLM was incorporated into the Bayesian Hierarchical modeling to differentiate hypotheses. The definition of the in-

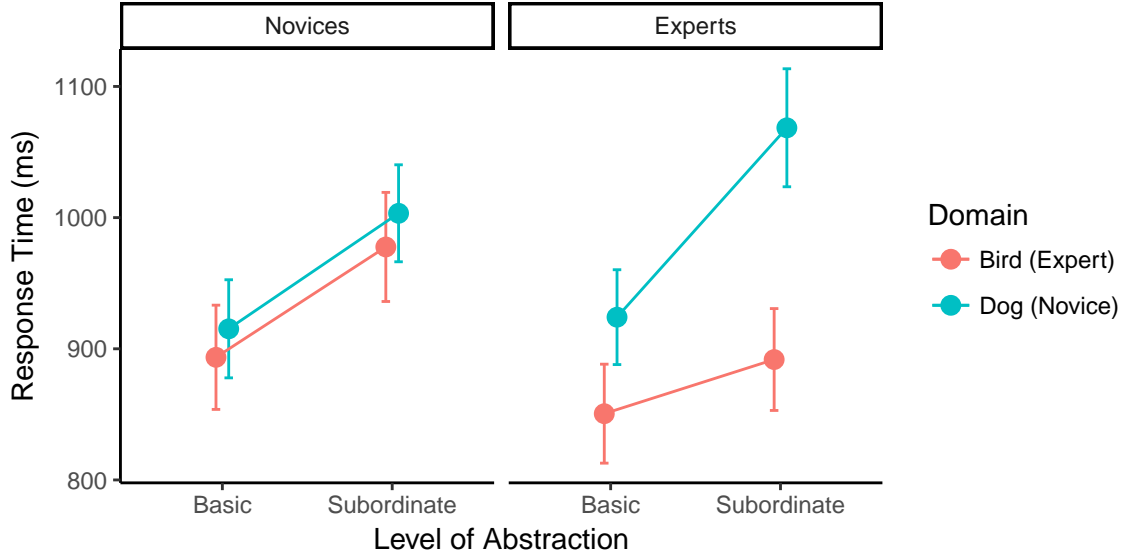


Figure 3.6: The mean response times across levels of abstraction and domains of expertise for the "novice" and "expert" group. Participants were assigned into the "novice" or "expert" group based on a median split of their expertise scores. The general pattern is consistent with the classic entry-level shift phenomenon.

teraction index was thus derived from the traditional two-way ANOVA analysis, which is equivalent to the GLM assumption as in the following equation:

$$RT_{ig} = \zeta_0 + \zeta_L L_i + \zeta_D D_i + \zeta_{Lg} L_i D_i + \varepsilon_i \quad (3.1)$$

The response time of each trial was assumed to be the sum of an intercept ( $\zeta_0$ ), the effect of the first factor, *level of expertise* ( $\zeta_L L_i$ ), the effect of the second factor, *domain of expertise* ( $\zeta_D D_i$ ), the effect of the interaction term ( $\zeta_{Lg} L_i D_i$ ), and an error term ( $\varepsilon_i$ ). The level of abstraction,  $L_i$ , and the domain of expertise,  $D_i$ , of each trial was dummy coded<sup>1</sup> as in Table 3.1. Given that  $L_i$  is 0 for the basic level and 1 for the subordinate level, a positive  $\zeta_L$  would indicate longer mean RT for the subordinate level compared to the basic level. Similarly, with  $D_i$  defined as 0 for the novice domain and 1 for the expertise domain, a positive  $\zeta_D$  would indicate longer mean RT for the expertise domain (bird) compared to the

<sup>1</sup>Note that the effects that are estimated based on dummy coding are referred to as "simple effects", i.e., an effect of a factor while holding other factors at one level. These are different from the main effects and interactions according to the classical definition (Hardy, 1993).

novice domain (dog). The  $\zeta_{I_g}$  concerns the interaction magnitude for each group  $g$ . The classic entry-level shift phenomenon entails a  $\zeta_{I_g}$  not different from 0 for the novice group but one smaller than 0 for the expert group.

Table 3.1: The dummy variable coding of the design matrix in the model

Level of Abstraction	L	Domain of Expertise	D
Basic	0	Novice	0
Subordinate	1	Novice	0
Basic	0	Expert	1
Subordinate	1	Expert	1

Note that in Equation 3.1, the interaction magnitude  $\zeta_{I_g}$  was assumed to be binary, 0 for the novice group and negative for the expert group. To expand the binary assumption to a continuous one to take into account individual differences in the interaction magnitude, a new definition is laid out as below, with individual subscript for  $\zeta_I$ . This enables us to estimate a separate interaction magnitude for every individual.

$$RT_{ip} = \zeta_0 + \zeta_L L_i + \zeta_D D_i + \zeta_{I_p} L_i D_i + \varepsilon_i \quad (3.2)$$

The R package *lme4* (version 1.1-12) was used to estimate the individual  $\zeta_{I_p}$  with both the RT and the accuracy data. A linear mixed-effect model was fit with the RT data. The right panel of Figure 3.7. shows the density and histogram of the parameter estimates for  $\zeta_{I_p}$ . One-sample  $t$ -test for the  $\zeta_{I_p}$  showed that the interaction magnitude on average was not significantly different from zero ( $\zeta_{I_p}$ ,  $t(62) = -1.99$ ,  $p = 0.05$ ). Participants on average had a greater-than-zero RT for the novice domain at the basic level ( $\zeta_0$ ,  $t(64.70) = 37.76$ ,  $p < 0.05$ ). The mean RT increased significantly for the subordinate level compared to the basic level ( $\zeta_L$ ,  $t(798.00) = 15.63$ ,  $p < 0.05$ ), while significantly decreased for the expert domain compared to the novice domain ( $\zeta_D$ ,  $t(816.50) = -9.41$ ,  $p < 0.05$ ). These results

suggest that the sample population was slower at the species level than the animal level, was faster in categorizing bird images than dog images, and did not show the entry-level shift phenomenon on average. These are not surprising given that the sample population is inherently interested in bird watching, but have varying levels of expertise. Only some participants showing the interaction could result in no interaction on average.

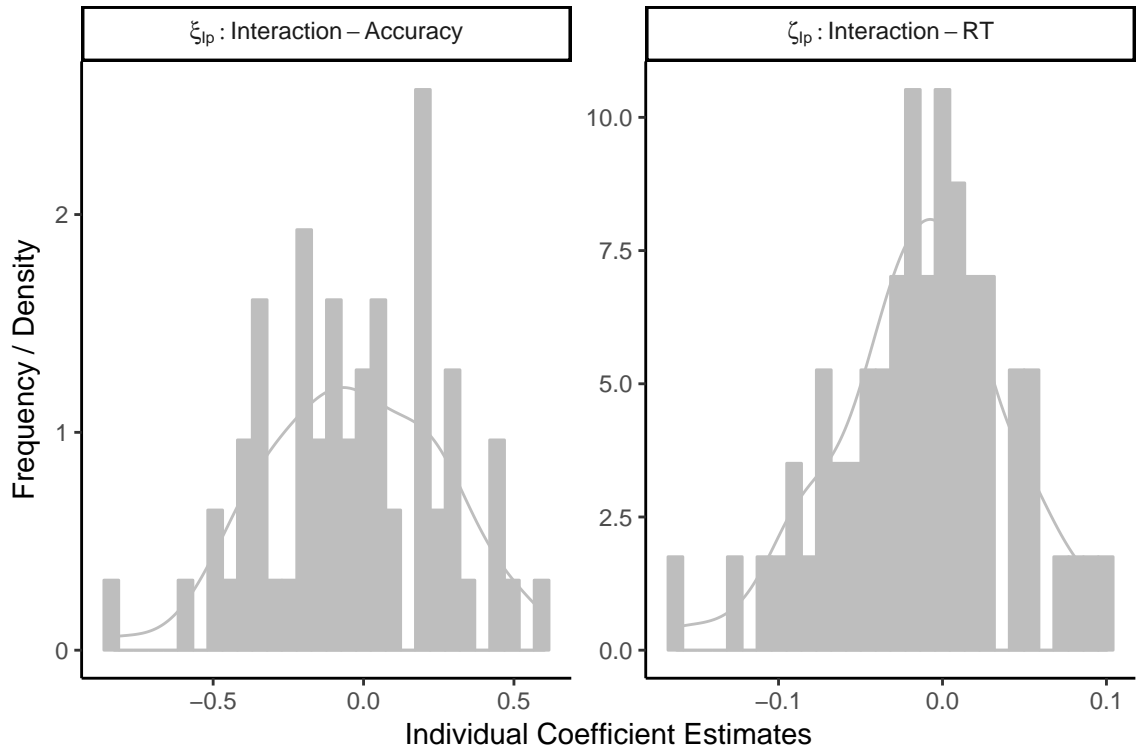


Figure 3.7: The density and histogram of the parameter estimates for  $\xi_{I_p}$  and  $\zeta_{I_p}$ .  $\zeta_{I_p}$  represents individual interaction coefficient for participant  $p$  from the accuracy data, for which experts are expected to have positive values.  $\xi_{I_p}$  represents individual interaction coefficient for participant  $p$  from the RT data, for which experts are expected to have negative values.

A necessary next step is to estimate the correlation between the expertise and the individual coefficient estimates (Figure 3.8). Expertise was defined as the expertise index estimated using IRT modeling with responses from the bird expertise test (Chapter 2). T-test on Pearson’s product-moment correlation showed that expertise correlated significantly with  $\zeta_{I_p}$  ( $r = -0.50, p < 0.05$ ). This result suggests that increased expertise is associated with larger interaction magnitude observed in RT.

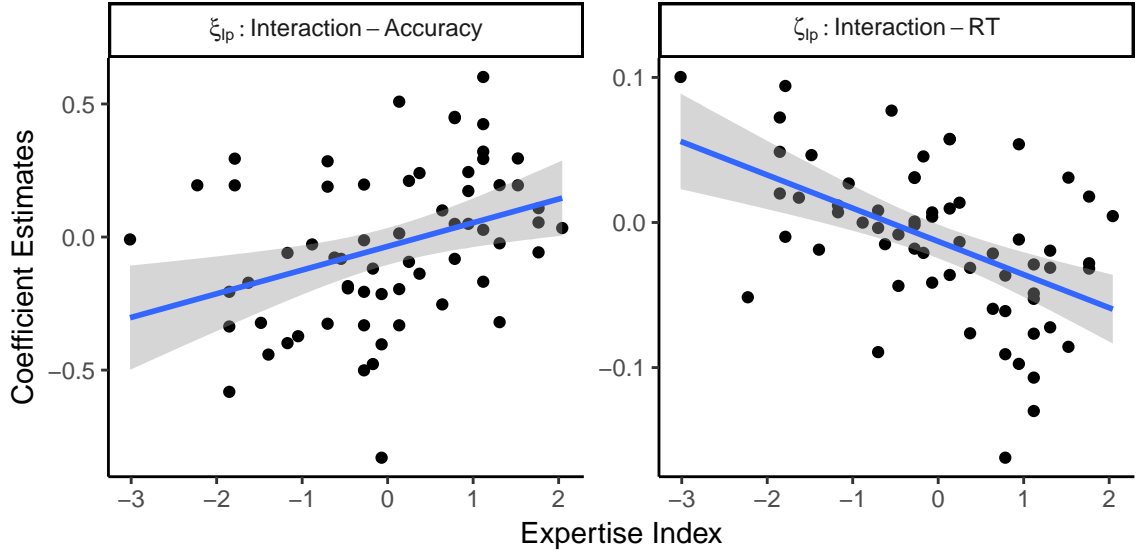


Figure 3.8: The correlation between the expertise index and the parameter estimates for  $\xi_{Ip}$  and  $\zeta_{Ip}$ .  $\zeta_{Ip}$  represents individual interaction coefficient for participant  $p$  from the accuracy data, for which experts are expected to have positive values.  $\xi_{Ip}$  represents individual interaction coefficient for participant  $p$  from the RT data, for which experts are expected to have negative values.

Accuracy data was analyzed in a similar fashion, except that a logit link function was added, assuming that the log odds ( $\log \frac{P}{1-P}$ ) of a participant scoring a question correctly comes from the same linear function that defined RT.

$$\log \frac{P_{pi}(Y = 1)}{1 - P_{pi}(Y = 1)} = \xi_0 + \xi_L L_i + \xi_D D_i + \xi_{Ip} L_i D_i + \varepsilon_i \quad (3.3)$$

The left panel of Figure 3.7 shows the density and histogram of the parameter estimates for all individual  $\xi_{Ip}$ s. One-sample  $t$ -test for the  $\xi_{Ip}$  showed that the interaction magnitude on average was not significantly different from zero ( $\xi_{Ip}$ ,  $t(62) = -0.93$ ,  $p = 0.36$ ). Participants on average had a greater-than-zero log odds for the novice domain at the basic level ( $\xi_0$ ,  $z = 25.94$ ,  $p < 0.05$ ), suggesting better-than-chance performance in this condition. The mean log odds decreased significantly for the subordinate level compared to the basic level ( $\xi_L$ ,  $z = -13.97$ ,  $p < 0.05$ ), while significantly increased for the expert domain compared to the novice domain ( $\xi_D$ ,  $z = 7.51$ ,  $p < 0.05$ ). These results suggest that the



sample population was less accurate at the species level than the animal level, was more accurate in categorizing bird images than dog images, and did not show the entry-level shift phenomenon on average. These results are consistent with the RT results, suggesting that the sample population was not only more accurate but also faster in categorizing in their expertise domain, although moving from the basic level to the subordinate level did cost them both speed and accuracy.

### 3.4 Modeling Results

**Model Assumptions.** A hierarchical diffusion model with linear regression component was run using the HDDM package (Wiecki et al., 2013) to understand the data. At the core of the model, each pair of RT and response choice data, i.e.,  $(t_{(pi)}, \chi_{(pi)})$  for person  $p$  and trial  $i$ , collected from a button press is assumed to be generated by a diffusion process.  $W$  represents the probability density function of the Wiener diffusion process (see Figure 3.2 for an illustration of the diffusion process). The four process parameters, including drift rate  $v$ , non-decision time  $\tau$ , bias  $b$ , and decision boundary  $A$  is allowed to vary across each person-by-item combination.

$$(t_{(pi)}, \chi_{(pi)}) \sim W(A_{(pi)}, b_{(pi)}, \tau_{(pi)}, v_{(pi)}). \quad (3.4)$$

The second layer of the model concerns the population distribution. Each of the four process parameters is assumed as random samples from a population distribution. The mean of the population distribution is assumed to differ depending on the person's expertise level and the condition of the trial. Formally,

$$v_{(pi)} \sim N(\mu_{v(pi)}, \sigma_v^2), \quad (3.5a)$$

$$\tau_{(pi)} \sim N(\mu_{\tau(pi)}, \sigma_\tau^2), \quad (3.5b)$$

$$b_{(pi)} \sim N(\mu_{b(pi)}, \sigma_b^2), \quad (3.5c)$$

$$A_{(pi)} \sim N(\mu_{A(pi)}, \sigma_A^2). \quad (3.5d)$$

Thus far, the model is descriptive in nature in that the variations across person and item is described as random samples from different population distributions. The last layer of the model attempts to explain the variations by defining the population mean for each person-by-item combination as the result of a linear component. For each parameter, the mean is defined as an intercept term (e.g.,  $\beta_{0v}$  for drift rate  $v$ ) plus the effect of the level of abstraction (e.g.,  $\beta_{Lv}L_{(i)}$  for drift rate  $v$ , with the beta coefficient  $\beta_{Lv}$  times the dummy variable  $L_{(i)}$  that represents the level of abstraction for that trial), plus the effect of the domain of expertise (e.g.,  $\beta_{Dv}D_{(i)}$  for drift rate  $v$ , with the beta coefficient  $\beta_{Dv}$  times the dummy variable  $D_{(i)}$  that represents the domain of expertise for that trial), plus the effect of the interaction term  $\beta_{Iv}\theta_{(p)}L_{(i)}D_{(i)}$  and an error term  $\varepsilon_i$ .

The interaction term is less intuitive, defined as the beta coefficient  $\beta_{Iv}$  times the dummy variable  $L_{(i)}$  that represents the level of abstraction for that trial, times the dummy variable  $D_{(i)}$  that represents the domain of expertise for that trial (see Table 3.1 for the coding of the two dummy variables), times the ability estimates for that person  $\theta_{(p)}$ . The magnitude of the interaction is assumed to be influenced by person ability in a linear fashion. An estimated  $\beta_{Iv}$  greater or smaller than 0 would suggest a significant positive/negative linear relationship between expertise and the interaction magnitude. Formally, the linear components are defined as:

$$\mu_{v(pi)} = \beta_{0v} + \beta_{Lv}L(i) + \beta_{Dv}D(i) + \beta_{Iv}\theta_{(p)}L(i)D(i) + \varepsilon_i, \quad (3.6a)$$

$$\mu_{\tau(pi)} = \beta_{0\tau} + \beta_{L\tau}L(i) + \beta_{D\tau}D(i) + \beta_{I\tau}\theta_{(p)}L(i)D(i) + \varepsilon_i, \quad (3.6b)$$

$$\mu_{b(pi)} = \beta_{0b} + \beta_{Lb}L(i) + \beta_{Db}D(i) + \beta_{Ib}\theta_{(p)}L(i)D(i) + \varepsilon_i, \quad (3.6c)$$

$$\mu_{A(pi)} = \beta_{0A} + \beta_{LA}L(i) + \beta_{DA}D(i) + \beta_{IA}\theta_{(p)}L(i)D(i) + \varepsilon_i. \quad (3.6d)$$

Based on the equations above and the dummy variable coding in Table 3.1, one can observe that  $\beta_0$  always stands for the intercept term;  $\beta_L$  stands for the effect of the level of abstraction;  $\beta_D$  stands for the effect of the domain of expertise; and  $\beta_I$  stands for the effect of the expertise index on the magnitude of the interaction.

Note that the interaction term here includes an extra expertise covariate, to test the effect of expertise on the magnitude of the interaction within the model itself, essentially doing hypothesis testing within the parameter space (a one-step approach). In contrast, in the behavioral models, I defined the traditional interaction term without the extra expertise covariate and then tested the correlation between expertise and interaction magnitude after obtaining estimates for the interaction magnitude (a two-step approach). Both approaches are acceptable practices. The one-step approach is more powerful and thus preferable when the data are sparse or the models are complex, because uncertainty could propagate from the entire data set to the parameter estimates at all levels of the model, allowing information to be shared across participants and conditions during model estimation. These advantages are inherent to hierarchical models (Gelman & Hill, 2007; M. D. Lee, 2011).

The model was fit using the HDDM package (Wiecki et al., 2013), specifically the regressor component. Three chains were run, with 6000 iteration for each chain. The first 3000 iterations were discarded as burn-in period. The posterior predictions from the model described the data fairly well, with strong correlations between the data and the posterior predictions for both the mean RT ( $r = 0.95$ ,  $p < 0.05$ , Figure 3.9) and the mean accuracy

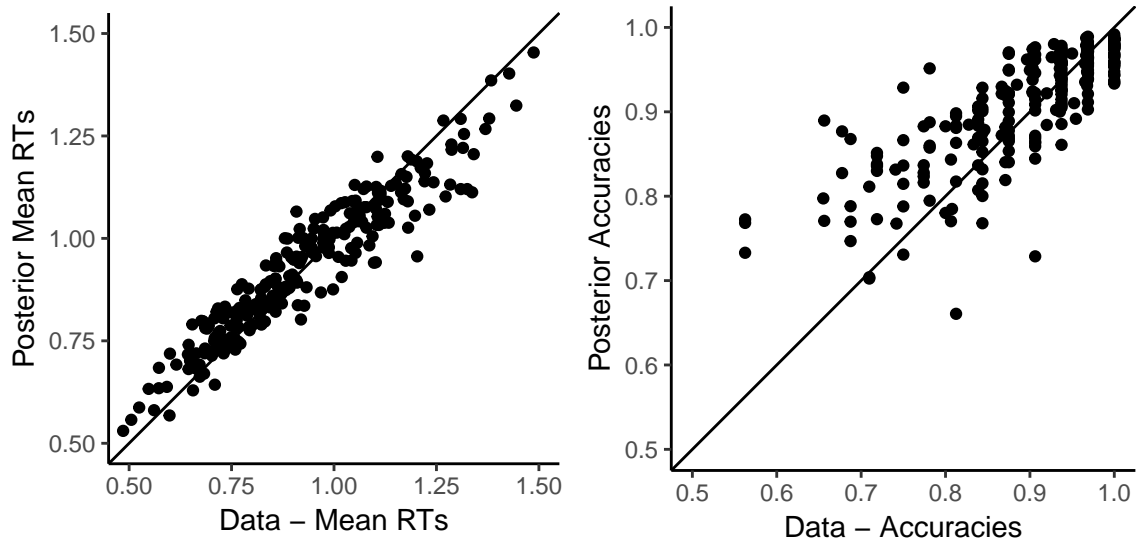


Figure 3.9: The correlations between the data and the posterior predictions from the HDDM model. Left panel shows the mean RT for each person-by-condition combination. Right panel shows the accuracy for each person-by-condition combination.

( $r = 0.82$ ,  $p < 0.05$ , Figure 3.9). The correlation for the mean accuracy is slightly less than ideal, potentially because of the low error rate.

The convergence of the chains was assessed using the *potential scale reduction factor*,  $\hat{R}$  (Gelman & Rubin, 1992). All  $\hat{R}$  is between 1.00 and 1.03, suggesting that convergence has been achieved among the three chains (a rule of thumb is that  $\hat{R}$  less than 1.10 indicates convergence). The key results of the model fitting concern the  $\beta$  parameters. Figure 3.10 illustrates the traces during the parameter estimation for these key parameters after the burn-in period (iteration number 3001 through 6000). The trace plots suggest good convergence for these parameters by showing the perfect overlap among the three chains. The traces cover a fixed range along the y axis evenly across iterations, which suggests that the fit is satisfactory.

The posterior probability density of the  $\beta$  parameters are displayed in Figure 3.11, with the two vertical lines in each subplot indicating the 95% Bayesian credibility interval. A credibility interval excluding 0 would suggest a significant  $\beta$  coefficient.

The  $\beta_L$  parameter is significant for all four parameters, with  $\beta_{LV}$  and  $\beta_{LA}$  estimated to

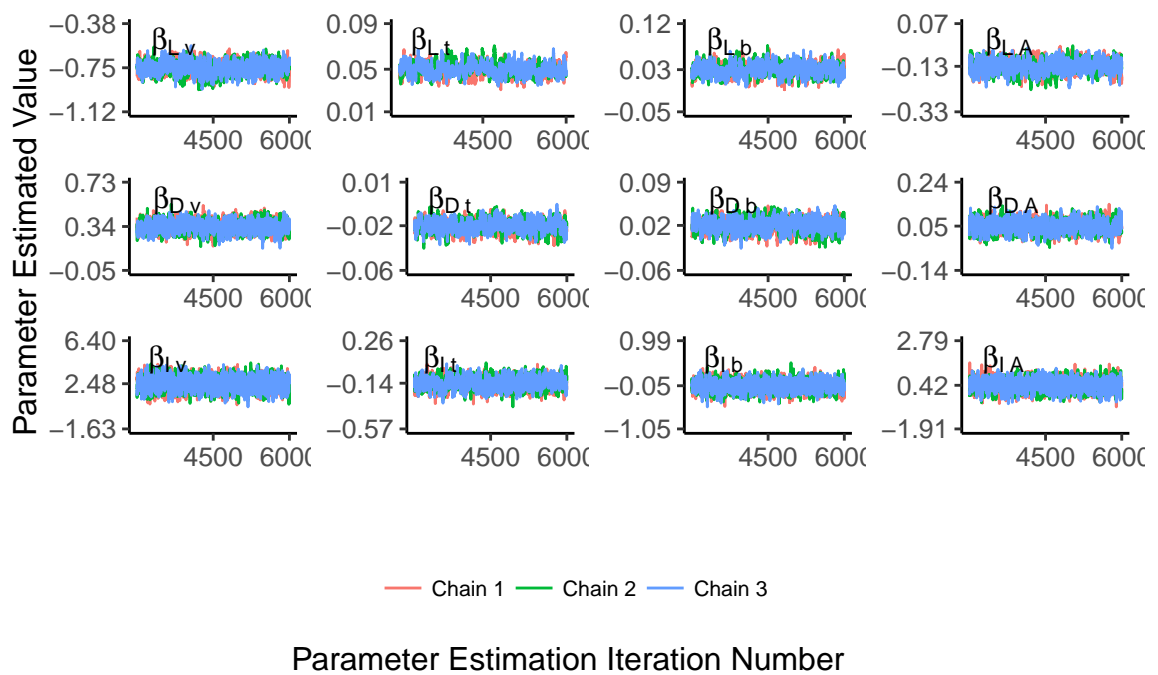
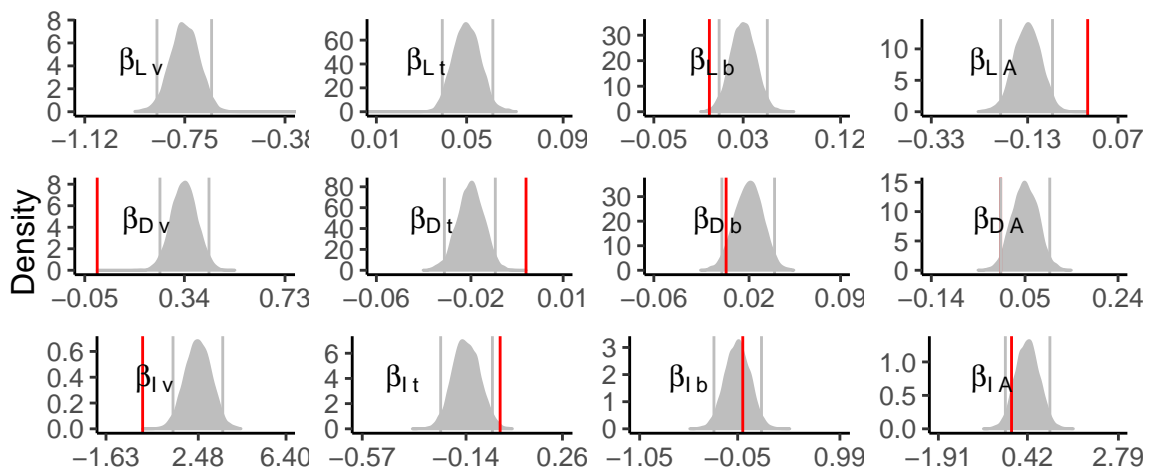


Figure 3.10: The traces of the parameter estimates for the key parameters from the HDDM model – all of the  $\beta$  parameters – after the burn-in period (iteration number 3001 through 6000). The traces cover a fixed range along the y axis evenly across iterations, which suggests that the fit is satisfactory.



Parameter Estimates

Figure 3.11: The density of the parameter estimates for the key parameters from the HDDM model – all of the  $\beta$  parameters – after the burn-in period (iteration number 3001 through 6000). The two gray vertical lines in each subplot indicate the 95 percent Bayesian credibility interval. The red vertical line indicates 0. A credibility interval excluding 0 suggests a significant  $\beta$  coefficient.

be smaller than 0 and  $\beta_{L\tau}$  and  $\beta_{Lb}$  to be greater than 0. Since the level of abstraction was dummy coded with 1 for the subordinate level and 0 for the basic level, this suggests that compared to the basic level, the subordinate level has decreased drift rate and decision boundary, but increased non-decision time and bias.

The  $\beta_D$  parameter is significant for the  $\nu$  and  $\tau$  parameters, with  $\beta_{D\nu}$  estimated to be smaller than 0 and  $\beta_{D\tau}$  to be greater than 0. Since the domain of expertise was dummy coded with 1 for the expert domain and 0 for the novice domain, this suggests that compared to the novice domain, the expert domain has an increased drift rate and a reduced non-decision time.

Table 3.2: A summary of the posterior estimates

	$\beta_\nu$	$\beta_\tau$	$\beta_b$	$\beta_A$
	Drift Rate	Non-Decision Time	Bias	Decision Boundary
$\beta_L$				
Level of Abstraction	-	+	+	-
$\beta_D$				
Domain of Expertise	+	-	.	.
$\beta_I$				
Expertise on Interaction	+	-	.	.

The  $\beta_I$  parameter is also significant for the  $\nu$  and  $\tau$  parameters, with  $\beta_{I\nu}$  estimated to be greater than 0 and  $\beta_{I\tau}$  to be smaller than 0. Since  $\beta_I$  represents the coefficient for the linear association between expertise and the magnitude of interaction, a positive coefficient for  $\nu$  suggests that increased expertise is associated with increased interaction magnitude in drift rate, while a negative coefficient for  $\tau$  suggests that increased expertise is associated with decreased interaction magnitude in non-decision time. The entry-level shift phenomenon boils down to different magnitude in the interaction, with experts showing interaction but

not the novices. The modeling results shows that the difference in interaction magnitude can be explained by both an increased evidence accumulation rate and a reduced non-decision time.

### 3.5 Discussion

In this work, I mapped out real-world visual expertise along the expertise continuum by modeling the category-verification behavior of participants with varying levels of expertise. Participants' expertise levels were measured using a psychometrically verified bird expertise test. The entry-level shift phenomenon was replicated across the expertise spectrum using an adapted version of the speeded category-verification task online. Using the hierarchical diffusion model, the behavioral differences exhibited by participants, i.e., varying magnitude of interaction, were explained by differences in their drift rate as well as a slight change in their non-decision time. The results suggest that experts are faster than novices not only because they have better evidence in the decision accumulation, but also because they are slightly faster in processes other than the decision accumulation, including but not limited to the processes for encoding the stimuli and for executing the responses.

The modeling results provide important insights about the underlying theory of visual expertise. I observed a positive  $\beta_{IV}$  from the modeling, which denotes the linear regression coefficient for the expertise index in predicting the interaction magnitude. Per the differentiation hypothesis, one would expect a positive correlation between the interaction magnitude of  $v$  and the expertise index. But no such correlation is expected per the basic-first hypothesis. Thus the modeling results support the differentiation hypothesis. This result provides another example of how formal models can be powerful in differentiating verbal theories (Ross, Deroche, & Palmeri, 2014).

This work bridged the gap between cognitive and psychometric models by applying process models to participants with varying levels of expertise in a hierarchical Bayesian framework. Traditional psychometric models have been used to predict performance from



expertise scores and thus to locate the source of individual differences, but usually cannot account for detailed psychological processes, while cognitive models usually lack the component to account for individual differences. This work illustrates that combining cognitive models and psychometric models not only helps us to locate the source of individual differences in behavior, but also to understand the underlying processes that give rise to the behavioral differences (Cronbach, 1975; Vandekerckhove, 2014; Vandekerckhove et al., 2011).

While visual expertise can be explained by faster evidence accumulation and faster non-decision processes, further questions arise as to what gives rise to the better evidence and shorter non-decision time. The exemplar-based random walk (EBRW) model (Nosofsky & Palmeri, 1997) provides an excellent example in explaining categorization behavior at the individual level. Individual categorization behavior was mathematically specified as a decision making process in which evidence accumulates over time. The accumulation speed is determined by the number and quality of domain examples that the participant has in memory, with more high-quality exemplars leading to faster and more accurate decision-making. The following chapter pursues this matter further by modeling the individual mental representations along the expertise continuum.

## Chapter 4

### Modeling the Representations of Expertise

#### 4.1 Introduction

In the previous chapter, real-world visual expertise was deconstructed into individual differences in underlying psychological processes such as evidence-accumulation speed in the decision-making process and non-decision time. The question is: what kinds of individual differences in the mental representations can give rise to such differences in these psychological processes and further to their behavioral differences? This chapter delves deeper into the mechanisms of visual expertise by modeling the individual differences in mental representations that eventually yield the individual differences in behaviors.

As shown in various expertise hallmarks, especially the entry-level shift phenomenon, visual expertise is mostly revealed at the subordinate level, e.g., drastic differences were observed between bird experts and novices distinguishing Nashville Warbler vs. Tennessee Warbler rather than distinguishing bird vs. dog. Thus I focused on the subordinate level to examine individual differences in representations, which is here called the microstructure of expertise.

To examine this microstructure, I estimated how people with different levels of expertise represent a set of similar objects in their expertise domain at the subordinate level, by using the multidimensional scaling (MDS) approach. In such techniques, it is assumed that objects are represented as points in a multidimensional psychological space. For example, we might represent birds as points in a bird space with some explicable dimensions such as spottiness of the feather color, size of the bird, length of the tail, as well as some incomprehensible dimensions. The distance between two objects determines their similarity, with shorter distance corresponding to object pairs more similar to each other. Understandably, two birds locating close to each other in the bird space would appear similar to the eye.

While the bird space might be abstract and unmeasurable, similarity can be approximated by people's similarity ratings of the objects pairs. Then the underlying representations can be modeled using the observable similarity data.

The MDS approach is a widely-used technique to uncover hidden perceptual representations (Borg & Groenen, 2005; T. F. Cox & Cox, 2000; Kruskal, 1964; Nosofsky, 1992a; Shepard, 1962, 1987). Importantly, the mental representation derived from MDS has been successful in explaining a host of cognitive behaviors such as identification, categorization, and recognition. Such a computational approach to these cognitive behaviors has been studied extensively using artificial stimuli with "experts" trained in the laboratory (e.g., Nosofsky, 1992a, 1992b; Nosofsky & Palmeri, 1997; Palmeri, 1997). In these work, the mental representations in a multidimensional space have been used in conjunction with cognitive models such as the generalized context model (Nosofsky, 1986) model to accurately describe the cognitive processes. Admittedly, lab-acquired experience with artificial objects cannot fully describe the complexities in real-world expertise, but such well-controlled studies provided important insights that a multidimensional space framework is viable to characterize various expertise behaviors.

Surprisingly, there has been no previous attempt to use MDS to understand the microstructure of real-world expertise. Some past work has used similarity scaling (via MDS or clustering techniques) to understand how relatively superordinate-level categories are structured and how those vary with different kinds of expertise, but not how more subordinate-level categories are structured and how they vary with expertise (Boster & Johnson, 1989; Medin et al., 1997). There has been theoretical work and laboratory-based training studies (Nosofsky, 1986, 1992a, 1992c; Palmeri, 1997) using MDS (both as a construct and as a measurement tool) to explain changes in learning and expertise using MDS (the construct, not the measurement tool), but those did not examine real-world perceptual expertise. Some recent work used MDS to examine the similarity structure of rocks, but while using real-world stimuli, these work looked at novices learning those objects, not experts (Meagher et

al., 2017; Nosofsky et al., 2017a, 2017b).

The MDS technique has not been used to tackle the micro-structure of real-world expertise for several reasons. Logistically, it has been difficult to recruit a big sample of real-world participants in traditional laboratory settings (Shen et al., 2014). In this project, this problem is solved by recruiting bird watchers with varying levels of expertise across the country for online participation. In addition to logistical difficulty, it has been challenging to model individual differences in the representational space using the MDS technique.

One important extension of the MDS model, the individual differences scaling model (INDSCAL, Carroll & Chang, 1970; Takane et al., 1977), include individual differences in the weights on each dimension of the representational space, but still assumes a single representational space, i.e., no group differences. To allow group differences in the representations, latent-mixture modeling is required (M. D. Lee & Vanpaemel, 2008; Winsberg & De Soete, 1993), in which each mixture/group has their own spatial representations. However, it was not feasible until recent developments to model both group and individual differences. The recent K-INDSCAL (Bocci & Vichi, 2011) model incorporates both the individual weights and latent-mixture features to infer group differences. But the authors relied on least-squares optimization, a limiting optimization method for such complex models with both group and individual differences. Okada and Lee (2016) further extended the K-INDSCAL model to a Bayesian framework using Stan, a probabilistic programming language for Bayesian inference and optimization (Gelman, Lee, & Guo, 2015), which allows for more stability and tolerance for complexity in the models. In this project, the Bayesian K-INDSCAL model was used to understand group as well as individual differences in the representations.

Either quantitative or qualitative differences could be observed among the participants. Quantitative difference means that participants with different levels of expertise share the same psychological space but weight dimensions differently; qualitative difference means that different groups (e.g., serious participants vs. contaminants, novices vs. experts) have

different psychological spaces with different dimensions. Because the techniques to be used are situated in a Bayesian framework, strong statistical tests, including BIC (Bayesian Information Criterion), can be used to select the best model, among models with different numbers of groups and different numbers of dimensions.

Individual differences in mental representations may or may not alone explain behavioral differences. People's mental representations estimated from MDS were used as input to cognitive models to understand people's performances in behavioral tasks, in this case bird identification. In identification tasks, participants assign a unique response to each stimulus. This behavior was chosen because it is a commonly studied behavior in perception. More importantly, identification patterns have been successfully explained by classic cognitive models that relied on a multidimensional psychological representation, especially the MDS-choice model (Nosofsky, 1985; Shepard, 1957, 1958), a special case of the classic similarity choice model (SCM) proposed by Shepard (1957) and Luce (1963).

Different hypotheses on the mechanisms of expertise were tested in the modeling process. Specifically, the MDS-choice model was used to explain people's behavioral differences. I investigated whether representational differences alone can explain behavioral differences, or other factors like perceptual sensitivity also contribute to perceptual expertise.

## 4.2 Methods

**Participants.** There were 130 participants who completed the similarity ratings task, while 95 completed the identification task. 54 participants (28 female) aged between 21 and 73 (mean = 46.15, SD = 15.04) completed the similarity ratings task, the identification task, and the bird expertise test (Chapter 2).

**Stimuli.** Given my focus on the microstructure of expertise and that expertise is mostly revealed at the subordinate level, I chose birds at the subordinate (i.e., species) level, especially birds that look similar within a taxonomic family. To maximize the differences

between participants with varying levels of expertise, I chose two bird families instead of just one. Potentially, expertise could affect the visual similarities because participants had varying levels of knowledge about the bird families. *Wood-Warblers* (Order: *Passeriformes*, Family: *Parulidae*, called Warblers thereafter) and *New World Sparrows and Allies* (Order: *Passeriformes*, Family: *Emberizidae*, called Sparrows thereafter) were chosen as the two families because among all bird families in North America, these two are the most populous, common, and geographically widespread. These birds are also highly confusable with each other (Figure 4.1), e.g., *Nashville Warbler* vs. *Tennessee Warbler*, which makes them ideal stimuli to investigate expertise. According to the Cornell Lab of Ornithology (<http://www.allaboutbirds.org>), while many bird families have a small number of bird species in North America, the Warblers family has 47 species, while the Sparrows family has 38 species, thus these two families provide abundant bird species and images to be used as stimuli.



Figure 4.1: The 10 Warblers species and 10 Sparrows species used in the study. The top two rows are Warblers, while the bottom two rows are Sparrows.

Test stimuli consisted of pictures of 10 Warblers species (Order: *Passeriformes*: Family: *Parulidae*) and 10 Sparrows (Order: *Passeriformes*, Family: *Emberizidae*). The Warblers were *Black-and-White Warbler*, *Blackburnian Warbler*, *Magnolia Warbler*, *Nashville Warbler*, *Northern Waterthrush*, *Orange-Crowned Warbler*, *Ovenbird*, *Tennessee Warbler*, *Townsend's Warbler*, and *Yellow Warbler*. The Sparrows were *American Tree Sparrow*, *Chipping Sparrow*, *Fox Sparrow*, *Lark Sparrow*, *Lincoln's Sparrow*, *Song Sparrow*, *Swamp Sparrow*, *Vesper Sparrow*, *White-Crowned Sparrow*, and *White-Throated Sparrow* (see Figure 4.1 for all these 20 bird species). Among all Warblers and Sparrows, these species were chosen because they are geographically wide-spread and relatively common in the U.S. according to the Cornell Lab of Ornithology (<http://www.allaboutbirds.org>).

The bird images are from the NABirds dataset, a collection of 48,000 annotated pho-

tographs of the 400 species of common birds in North America (Van Horn et al., 2015). The birds in the images were of various orientations and various background contexts. The images were cropped so that the birds were centered and prominent, limiting the influence of the background. The color images were converted to gray-scale to keep consistency with the previous chapters. This ensured consistent stimuli in all experiment tasks in this project. Since color perception plays important roles in perceptual expertise (Hagen et al., 2014), it would be interesting to test with color images in the future. Image did not repeat during each experiment task.


**Procedure.** After participants finished the bird expertise test, they were invited via email to complete the similarity ratings task online. The experiment was programmed in JavaScript and implemented as interactive web pages. Participants completed the tasks within their web browsers. On their computer, they received the following on-screen instructions:

“In this experiment, you will be asked to judge the similarity of bird species. You will be presented with two bird images at a time and a slider scale from ‘Most dissimilar’ to ‘Most similar’. Please make your judgment based on the visual similarity of the two bird species. Ignore superficial characteristics like image sizes or bird orientations. Note that we remove the pictures after 3 seconds, so you can focus on the species, rather than the pictures themselves.”

Next, participants went through three demo trials, with additional instructions after each trial, to ensure that they rated the similarities based on their own criteria, and focused on the bird species rather than superficial characteristics of pictures such as backgrounds and orientations of the birds. Next, participant were given a collage of the 20 bird pictures that would appear during the experiment session, one picture for each species. They were given enough time to study all bird species, so they could understand the range of dissimilarity to appear in the following task. This was to encourage participants to give consistent ratings across the session.



How visually similar are these bird species? Progress: 3/210



Most dissimilar    Somewhat dissimilar    Somewhat similar    Most similar

Figure 4.2: An example trial in the similarity-ratings task

Out of the 20 bird species, there were 190 different-species pairs. Pictures did not repeat within this task, thus 19 pictures were used for each species. Figure 4.2 shows an example trial. The choice of pictures for each pair was randomized for each participant, to ensure that participants judge the similarities of bird species, rather than idiosyncrasies of specific bird pictures. E.g., on the same trial that contains a *Nashville Warbler*, one participant might see a different image of the species from another due to the randomization. Participants viewed each pair of pictures and rated them along an interval scale. Seven ticks were shown along the slider scale, with four labels placed at tick 1, 3, 5, and 7, to encourage participants to use the full scale during ratings. The four labels were “Most dissimilar”, “Somewhat dissimilar”, “Somewhat similar”, and “Most similar” from left to right. The pairs of bird species were presented in a random order. This random order was set as the same for all participants, to keep the random order from being a source of individual differences in the results.

After participants finished the similarity ratings task, they were invited via email to complete the identification task online. In this task, they identified each bird species 10 times, resulting in 200 trials in total. In each trial, one bird image would appear, along with all 20 species labels (Figure 4.3). Participants were asked to use their mouse to click on the correct species label. The bird image was removed after 5 seconds, to avoid participants to refer to external sources like bird books or the Internet. Images did not repeat during the session. The order of the trials was randomized but kept the same for all participants, to eliminate order as a potential source of individual differences.

### 4.3 Results

**Multidimensional Scaling.** The data from the similarity ratings task were analyzed using an extended version of the Bayesian K-INDSCAL model (Okada & Lee, 2016). I implemented the models using the open-source software Stan (B. Carpenter et al., 2016) and its interface to R (Stan Development Team, 2016), as did Okada and Lee (2016). Stan

What is the species of this bird?



Progress: 3/300

- American Tree Sparrow
- Black-and-white Warbler
- Blackburnian Warbler
- Chipping Sparrow
- Fox Sparrow
- Lark Sparrow
- Lincoln's Sparrow
- Magnolia Warbler
- Nashville Warbler
- Northern Waterthrush
- Orange-crowned Warbler
- Ovenbird
- Song Sparrow
- Swamp Sparrow
- Tennessee Warbler
- Townsend's Warbler
- Vesper Sparrow
- White-crowned Sparrow
- White-throated Sparrow
- Yellow Warbler

Next

Figure 4.3: An example trial in the bird identification task

provides full Bayesian inference for continuous-variable models through Markov chain Monte Carlo methods using the so-called No U-Turn Sampler (NUTS, Hoffman & Gelman, 2014). This fairly new software quickly gained popularity because it can flexibly fit complex models efficiently using NUTS.

In the model, the observed dissimilarities are assumed to be the sum of the true distance in the psychological space and the measurement error. Thus the dissimilarity between bird species  $i$  and  $j$ , rated by participant  $p$ , denoted as  $y_{ijp}$ , follows a normal distribution, with the true distance  $d_{ijp}$  as the mean and the measurement error  $\sigma$  as the standard deviation of the distribution:

$$y_{ijp} \sim N_{[0,\infty)}(d_{ijp}, \sigma^2) \quad (4.1)$$

In the equation,  $N_{[0,\infty)}$  represents truncated normal, since the ratings cannot have negative values. There are  $K$  classes of participants assumed, which is always smaller than the number of participants  $P$ . The true distance  $d_{ijp}$  is defined as the Euclidean distance weighted by attention weights  $w_{pm}$ , assuming that there are  $M$  dimensions in total:

$$d_{ijp} = \sqrt{\left( \sum_{m=1}^M w_{pm} (X_{imk} - X_{jmk}) \right)} \quad (4.2)$$

The individual weight ( $w$ ), class membership for each participant, and representation coordinates ( $X$ ) for each class were estimated. Participants can differ from each other in two ways. Quantitatively, participants might have different weights ( $w$ ) on different dimensions. Qualitatively, participants might belong to different classes, i.e., they have different configurations of the psychological space ( $X$ ) and/or different weights ( $w$ ).

I relied on three criteria to select the number of classes  $K$  and the number of dimensions  $J$ . The first criterion is Kruskal's stress (Kruskal, 1964), a common classical measure of normalized squared errors in MDS. The second criterion is a form of posterior predictive checking (Rubin & others, 1984), the correlation coefficient between the observed dissim-

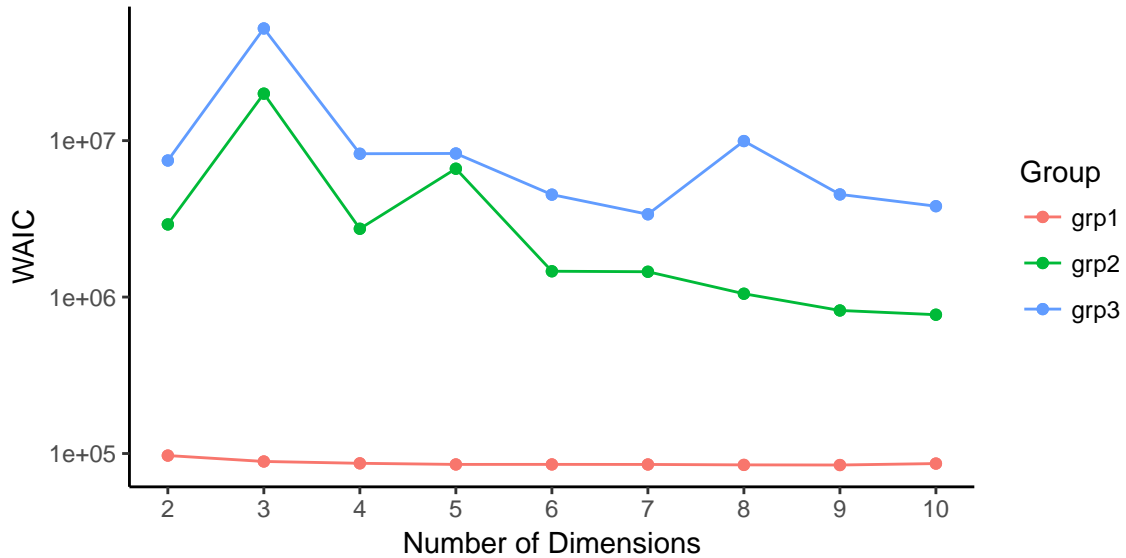


Figure 4.4: WAIC across MDS models with different number of dimensions and groups. The models with 1 group are all better than models with 2 or 3 groups, with much smaller WAIC. Across dimensions, the gain in WAIC becomes relatively marginal as the number of dimensions approaches 9.

ilarity and its mean posterior predicted values. For each combination of the number of groups ( $K = 1, 2, 3$ ) and the number of dimensions ( $W = 2, 3, 4, 5, 6, 7, 8, 9$ ), the Kruskal's Stress and the correlation coefficient were calculated for each individual and then averaged across individuals. The third criterion is WAIC (the widely applicable or Watanabe-Akaike information criterion, Vehtari, Gelman, & Gabry, 2016, 2017; Watanabe, 2010), which is a predictive information criterion for Bayesian models. This criterion estimates point-wise out-of-sample prediction accuracy from a fitted Bayesian model. It concerns not only the current data set, but also the generalizability of the model. The WAIC was calculated for each group-dimension combination .

The results are shown in Figure 4.4 and Figure 4.5. Clearly, with higher number of dimensions, the Kruskal's stress decreases while the correlation increases. This is intuitive because the space with higher number of dimensions can encompass more complexity in the spacing of objects and can better approximate the similarity ratings. The gain in the goodness-of-fit, in this case higher correlation and lower Stress, becomes relatively

marginal as the number of dimensions approaches 9, across all number of groups. A similar pattern was observed using the WAIC metric. These results suggest that  $J = 9$  dimensions provides an adequate representation of the space.

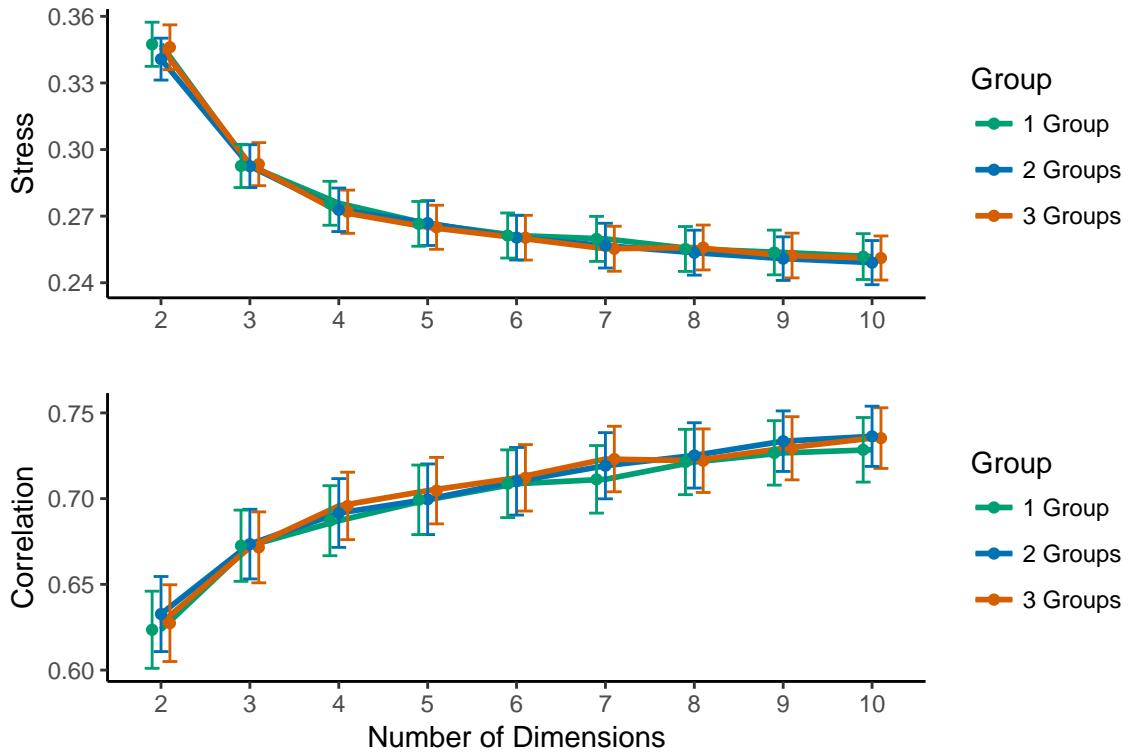


Figure 4.5: Averaged Kruskal’s Stress and correlation coefficient between the observed dissimilarity and its mean posterior predicted values across MDS models with different number of dimensions and groups. The models with different number of groups are almost indistinguishable from each based on both criteria. Across dimensions, the gain in both criteria becomes relatively marginal as the number of dimensions approaches 9.

For the number of groups, Stress and correlation are almost indistinguishable across the three models with 1, 2, and 3 groups respectively. Overall, the models with 3 groups have higher correlation and lower Stress. Understandably, the most complex model can fit the data better because of more parameters. However, WAIC results strongly support the 1-group models, with significantly lower WAIC values for 1-group models across all dimensions, compared to 2-group and 3-group models. Together, these three criteria suggest that the best model is the one with  $K = 1$  group and  $J = 9$  dimensions.

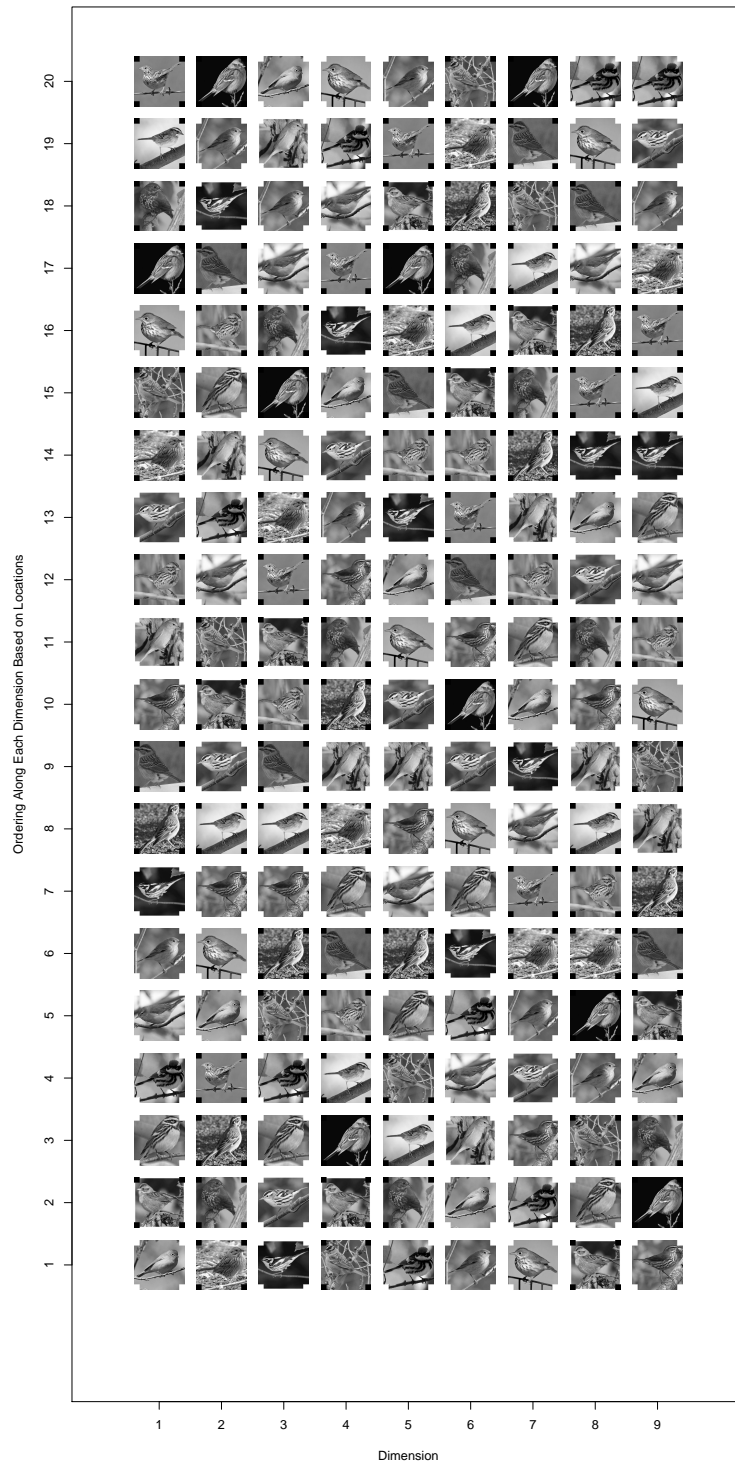


Figure 4.6: The ordering of the 20 bird species along each of the 9 dimensions in the representational space. The 9 dimensions are mapped out along the  $x$  axis. The 20 birds species are ordered based on their locations along each dimensions, with 1 along the  $y$  axis indicating the smallest value and 20 indicating the largest value along the corresponding dimension.

The posterior spatial representations of the birds is shown in Figure 4.6. Given the complexity of the 9-dimensional space, only the ordering of the 20 bird species along each dimension is shown. The Sparrow images are flagged with four black corners, while the Warblers are flagged with four white corners. As the 20 bird species belong to two different families, it is likely that some of the 9 dimensions might be relevant to the bird family information. The relationship between dimensional locations of the bird species and their family information (0 for Sparrow and 1 for Warbler) was tested and shown in Figure 4.7. Dimensional locations along 4 of the 9 dimensions, dimension 1, 4, 6, and 7, have a significant relationship with their family information, which suggests that these 4 dimensions represents bird families.

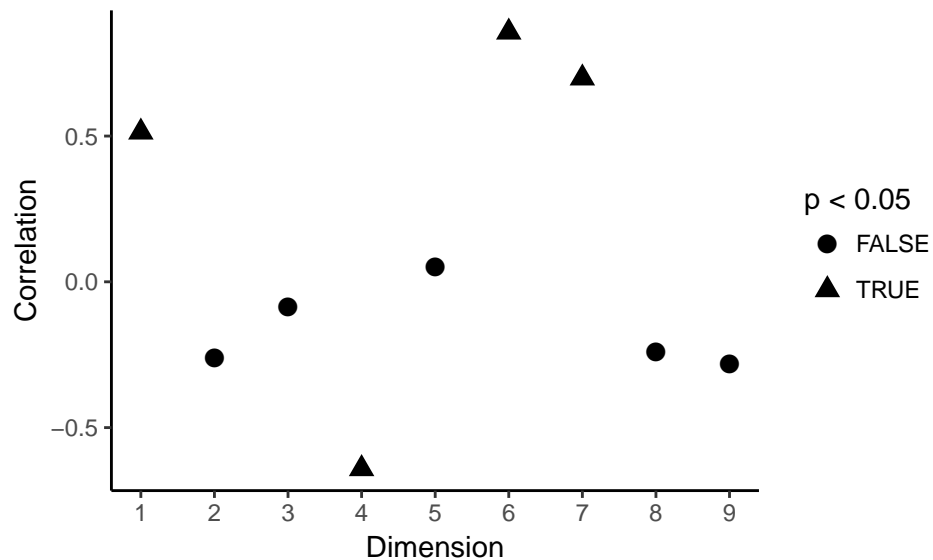


Figure 4.7: The correlation between the dimensional locations of the 20 birds species along each of the 9 dimensions and their family information (Warbler or Sparrow). It is shown that some of the dimensions, including dimension 1, 4, 6, and 7, are family-related.

Since all participants share one space, the individual difference in their representations lie solely in the weightings of the dimensions. The relationship between the dimensional weights and the expertise index was tested and shown in Figure 4.8. The expertise index was defined as the expertise index estimated from IRT modeling with responses from the bird expertise test (Chapter 2). The weights on 6 of the 9 dimensions correlate significantly



with the expertise index, with 4 positive and 2 negative relationships. These results suggest that experts weighted 4 of the 9 dimensions more and 2 of the 9 dimensions less in their similarity ratings.

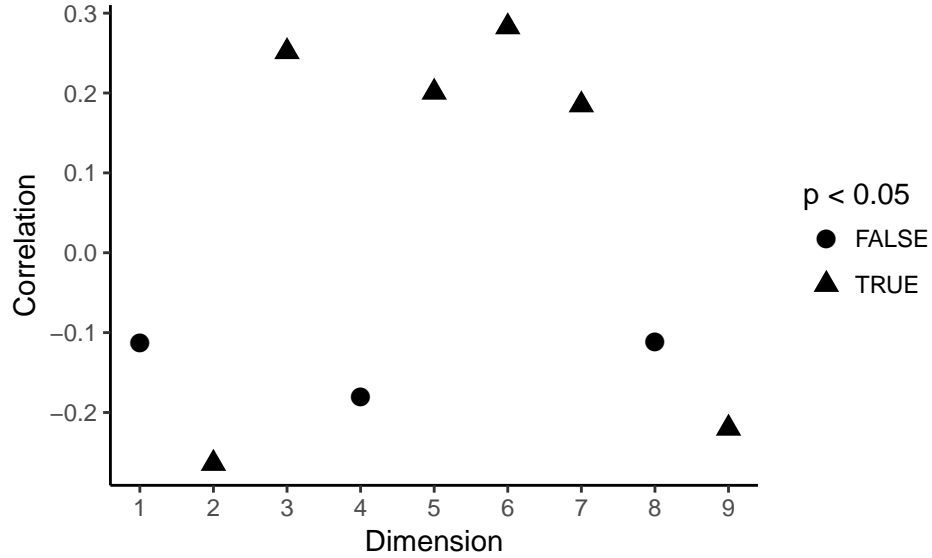


Figure 4.8: The correlation between the dimensional weightings by the participants in the similarity-ratings task and their expertise index. It is shown that some of the dimensions, including dimension 2, 3, 5, 6, 7 and 9, are positively/negatively related to expertise, suggesting that experts weighted some dimensions more and some dimensions less.

**Similarity Choice Model.** The data from the identification task were fit with the similarity-choice model using JAGS (Plummer, 2003) and its interface to R, which has been widely used for Bayesian modeling. The data were analyzed using an extended version of the MDS-Choice Model (Nosofsky, 1985; Shepard, 1957). In its original form, the MDS-choice model is a special case of the aforementioned GCM, in which each stimulus defines its own category. In the MDS-choice model, the probability of participant  $p$  identifying stimulus  $i$  as  $j$  is given by

$$P(R_j|S_i) = \frac{b_j \eta_{ij}}{\sum_{k=1}^n b_k \eta_{ik}} \quad (4.3)$$

where  $0 \leq b \leq 1$ ,  $\sum_{k=1}^n b_k = 1$ ,  $\eta_{ij} = \eta_{ji}$  and  $\eta_{ii} = 1$ .  $b_k$  represents the bias for making response  $R_k$ .  $\eta_{ij}$  represents the similarity between stimuli  $S_i$  and  $S_j$ , which is converted

from the distance defined using the following function:

$$\eta_{ij} = \exp(-c \cdot d_{ij}^h) \quad (4.4)$$

in which  $c$  ( $0 \leq c < \infty$ ) is a sensitivity parameter that represents participants' discriminability in the psychological space. According to previous theoretical work (Ennis, 1988; Nosofsky, 1985; Shepard, 1986, 1987), an exponential decay function ( $h = 1$  in Equation 4.4) can well describe the relationship between similarity and distance for readily discriminable stimuli. In contrast, a Gaussian function ( $h = 2$ ) performs better for highly confusable stimuli.

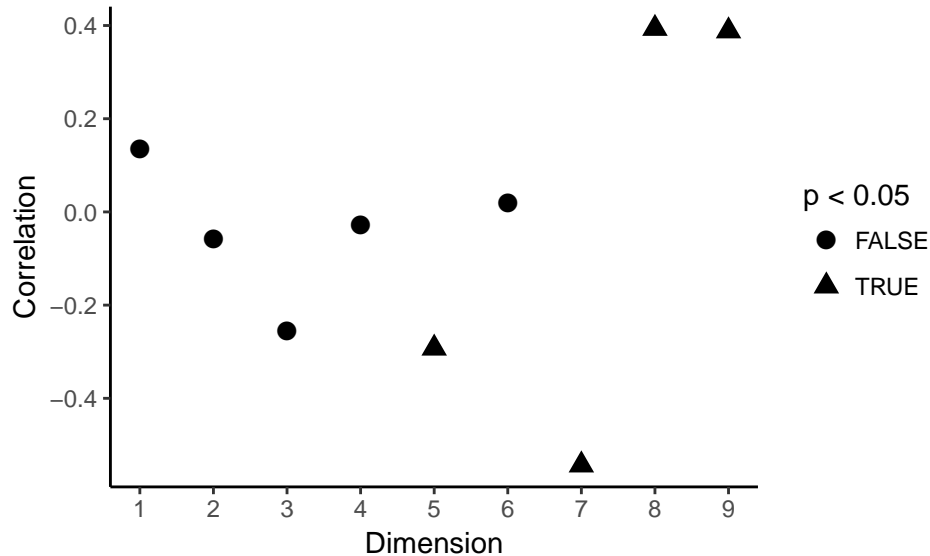


Figure 4.9: The correlation between the dimensional weightings by the participants in the bird identification task and their expertise index. It is shown that some of the dimensions, including dimension 5, 7, 8 and 9, are positively/negatively related to expertise, suggesting that experts weighted some dimensions more and some dimensions less.

In this experiment, since the birds are highly confusable within each family, but less so between the two families, the data were fit with both the exponential and the Gaussian decay function. The final model was chosen based on the Deviance Information Criterion (DIC Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) and the correlation coefficient between the identification data and their posterior predictions. The DIC was much smaller

for the model with  $h = 1$  than  $h = 2$ . The correlation coefficient was 0.95 for  $h = 1$  and 0.89 for  $h = 2$ . Thus both criteria support the exponential decay function for the bird stimuli.

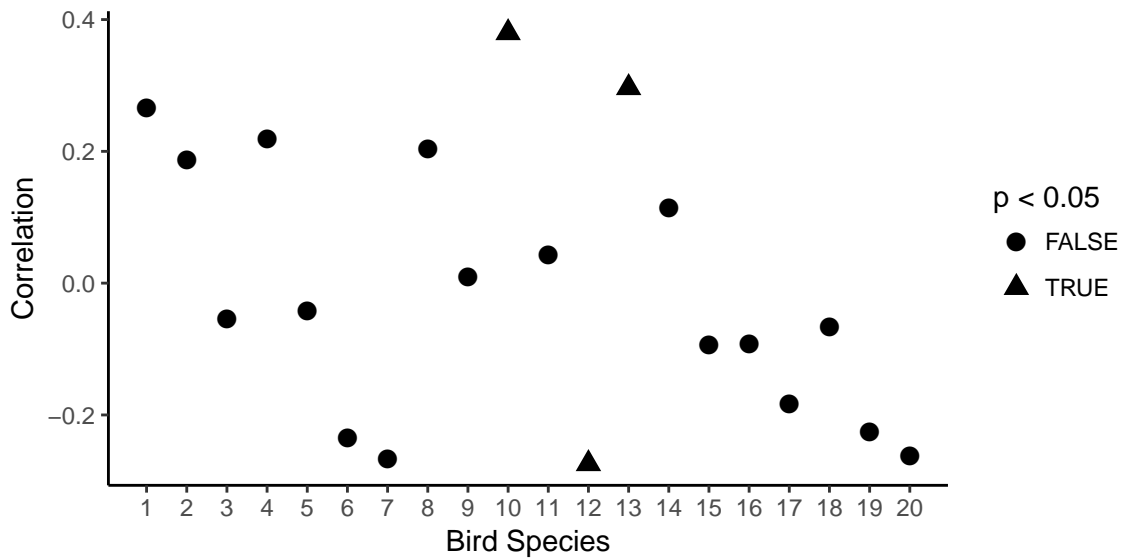


Figure 4.10: The correlation between the biases toward each bird species by the participants in the bird identification task and their expertise indices. It is shown that biases toward a few of the bird species, including species 10, 12 and 13, are positively/negatively related to expertise, suggesting that experts biased toward some species more and one species less. But biases toward most of the species don't correlation with expertise.

The individual distances  $d$  in Equation 4.4 can be determined using the same function (Equation 4.2) as in the MDS model. The coordinate locations  $X$  are best-fit parameter estimates from the MDS modeling. However, the individual weights  $w$  for the identification task might differ from those for the similarity ratings task, as individuals could distribute their attention in one way when rating the similarities between bird pairs, but in another way when identifying birds each at a time. In addition to individual attention weights, it is necessary to assign individual sensitivity parameter  $c$  and bias parameter  $b_j$ . The variance in attention weight parameters, sensitivity parameters, and bias parameters across individuals reflect the individual-difference extension to the MDS-choice model.

The relationship between the expertise index and the dimensional weights was tested. As shown in Figure 4.9, weights on 4 of the 9 dimensions correlate significantly with the expertise index, with 2 positive and 2 negative correlations. This suggests that experts

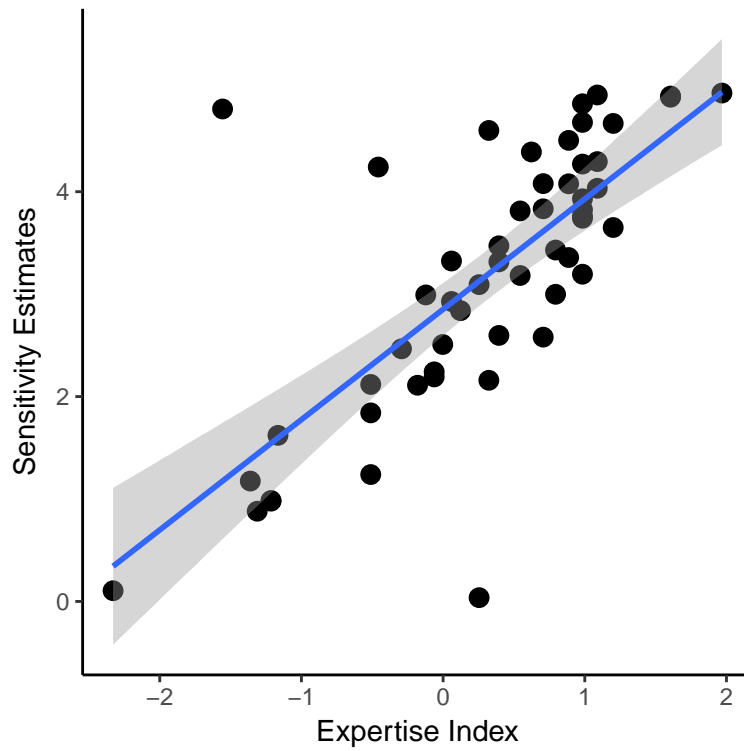


Figure 4.11: The correlation between the sensitivity parameter estimates and the expertise index.

have knowledge about diagnostic dimensions, while novices might be easily distracted by superficial perceptual features. The sensitivity parameter estimates correlates positively (Figure 4.11) with the expertise index ( $r = 0.74$ ,  $t(52) = 7.85$ ,  $p < 0.05$ ). This was expected since the sensitivity parameter reflects the overall discriminability in the psychological space. The correlation between biases toward each bird species and the expertise index was also tested. As shown in Figure 4.10, biases toward the majority of the bird species, 17 out of 20, do not correlate with expertise, with 2 positive and 1 negative correlations for the rest of the 3 bird species.

#### 4.4 Discussion

The representational mechanism of real-world visual expertise has long remained elusive. In this work, I modeled participants' mental representations along the continuum of expertise and related the representations to participants' performances in the identification task. I used bird identification ability as an example to study generic real-world visual expertise, since bird identification is a commonly used and accessible example of visual expertise, unlike radiology and forensics. Participants with varying levels of bird expertise rated the similarities between all possible pairs among 20 similar-looking bird species, including 10 Warblers and 10 Sparrows. Multidimensional scaling with the latent mixture component was used within a Bayesian Hierarchical framework to model the representations based on the similarity-ratings data. The representations were also used to model participants' identification performance.

The three model selection criteria, including the Kruskal's Stress, WAIC, and the correlation coefficient between the observed dissimilarity and its mean posterior predicted values, together suggested that participants along the expertise continuum shared one representation space which has 9 dimensions. However, these 9 dimensions were weighted differently across individuals. In particular, participants with higher level of expertise weighted 4 of the 9 dimensions more and 2 of the 9 dimensions less in their similarity

ratings. These results suggest that rather than having a different space, experts were paying attention to more diagnostic dimensions while novices were distracted by superficial dimensions of the psychological space for these 20 bird species.

When identifying birds, participants weighted the dimensions in ways that were different from when they were rating the similarities between birds. Attention weights on 4 of the 9 dimensions correlate significantly with the expertise index, with 2 positive and 2 negative correlations. Similarly, this suggests that experts were able to capitalize on diagnostic information when identifying the bird species, while novices might be distracted by superficial characteristics of the images. Also, the sensitivity parameter estimates correlate positively with the expertise index, which suggests that experts were more sensitive to differences in the psychological space, i.e., experts can better distinguish two objects that are close to each other in space than novices. Thus participants with higher levels of expertise are better than novices mainly because of their strategic weighting of the dimensions and their sensitivity to subtle differences between similar-looking bird species.

A remaining question about the representation is the meanings of dimensions. This was partly answered by correlating the locations of the 20 bird species along each dimension with the birds' family information (Warbler or Sparrow). It was found that 4 of the 9 dimensions are family-relevant. However, as shown in Figure 4.6, with the high similarity among the birds, the meaning of each dimension was hard to read from their mappings. Bird experts can potentially help to identify and characterize the visual features along which the birds vary. However, it is expected that not all dimensions could be explicitly verbalized, as has been the case in other previous MDS studies with real-world stimuli (Meagher et al., 2017; Nosofsky et al., 2017a, 2017b). Also, the aim here is not to understand “what are the dimensions” of the “bird space” – in a way that might be of great importance in some object domains, like asking “what are the dimensions” of the “face space”. Here the question is whether the space varies qualitatively or quantitatively with expertise and whether the space can be used to predict other aspects of performance known to vary with expertise,

like identification and categorization.

## Chapter 5

### Conclusion

#### 5.1 Summary of Findings

The formal mechanisms of real-world expertise has always been challenging to study, due to the logistical difficulties in recruitment and data collection, as well as the technical complexities in modeling their cognitive processes. The present research explored this topic by recruiting participants online and modeling their behaviors using recently developed Bayesian Hierarchical modeling techniques.

Bird expertise was used as an example domain to study the dynamics and representations of real-world visual expertise, given that the bird expertise domain is commonly used and easily accessible. A bird expertise test was established using the Item Response Theory in this project. The expertise test was face-valid with the bird identification design. The ability estimates from the test correlated reasonably well with participants' self-rated ability. Also, explanatory modelling showed that the participants' birding frequency, experience, and training all associated significantly with their ability estimates, providing validity support for the test.

Test results from the bird expertise test was used as benchmarks to line up participants along the continuum from novice to intermediate to expert. Then behavioral differences along the expertise continuum in category-verification, similarity ratings, and identification were observed and modeled to understand the dynamics and representations of expertise.

The dynamics of expertise is best reflected in the entry-level shift phenomenon, in which novices are faster and more accurate to verify category membership at an intermediate level of abstraction, the so-called the basic or entry level (e.g., "bird"), than a superordinate (e.g., "animal") or a subordinate level (e.g., "Blue Jay"). In contrast, experts are equally fast at the subordinate and the basic level. One explanation for the different re-



response patterns by novices and experts is called the basic-first hypothesis, which suggests that the categorization at the basic level is a prerequisite, i.e., the entry level, for more superordinate and subordinate categorizations - you need to know that it is a *bird* before you can tell whether it is an *animal* or a *Blue Jay*. Experts are faster because their subordinate level becomes an entry level, namely the entry-level shift phenomenon. An alternative explanation, the differentiation hypothesis, suggests that basic-level categorizations are faster because the basic level is more differentiated and informative, not that it happens first. Thus experts are faster at the subordinate level simply because the objects are well differentiated for them at the subordinate level.

The speeded-category verification task was adapted to an online form and completed by a group of remote participants with varying levels of birding expertise ranging from novice to intermediate to expert. The two competing hypotheses were evaluated by fitting the well-known drift-diffusion model of speeded perceptual decisions to the accuracy and response time data collected online. I specifically identified these two hypotheses with differences in process parameters within the diffusion model: variability in only non-decision processing time across category levels would indicate the basic-first hypothesis, whereas variability in drift rate across category levels would support the differentiation hypothesis. The diffusion model was applied within a Bayesian hierarchical framework, which provides a powerful account of individual differences in the model parameters across conditions. Behaviorally, the entry-level shift phenomenon was replicated for participants in online settings. Theoretically, it was found that the variability in category-verification speed patterns across varying levels of expertise was well captured by variability in the drift rate as well as a slight difference in the non-decision processing time. These results support the differentiation hypothesis, providing insights about the psychological processes that give rise to the behavioral pattern in speeded category-verification and informing the understanding of individual differences seen in the entry-level shift phenomenon as a function of expertise.

The representational mechanism of real-world visual expertise was studied using the

latent mixture multidimensional scaling technique, within a Hierarchical Bayesian framework. I modeled participants' mental representations along the continuum of expertise and related the representations to participants' performances. Participants with varying levels of bird expertise rated the similarities between all possible pairs among 20 highly confusable bird species, including 10 Warblers and 10 Sparrows. The modeling results suggest that instead of having different perceptual spaces, participants along the expertise continuum shared one representational space with 9 dimensions but weighted the 9 dimensions differently across individuals. Particularly, participants with more expertise weighted 4 of the 9 dimensions more and 2 of the 9 dimensions less than participants with less expertise. These results suggest that experts and novices were seeing the same amount of information, but experts might be paying more attention to diagnostic dimensions while novices might be distracted by superficial dimensions of the psychological space.

In an attempt to explain behavioral differences from representations, the representations from MDS were used to model identification performance. Participants were recruited to complete an identification task, in which they chose the bird species label out of all 20 possibilities, based on images of the bird. Similar to the multidimensional scaling results, participants with different levels of expertise weighted the dimensions differently when identifying the 20 birds. Weights on 4 of the 9 dimensions correlate significantly with the expertise index, either positively or negatively. Similarly, this is likely due to different distribution of resources to the diagnostic or distracting information when identifying the bird species. Also, experts were shown to be more sensitive to differences in the psychological space. That is, experts can better distinguish two objects that are close to each other in space than novices, likely due to extensive experience. Together, experts are better than novices in bird identification mainly because of their strategic weighting of the dimensions and their sensitivity to subtle differences between bird species.

## 5.2 General Discussion

This work provides insights into the dynamics and representations of real-world visual expertise, yet more work is necessary to gain a full understanding of expertise. Real-world expertise manifests itself in a host of behaviors, such as better recognition as shown in the Vanderbilt Expertise Task, better semantic identification as shown in the Semantic Vanderbilt Expertise Test, holistic effect as shown in the composite task, the inverted task, and the part-whole task etc.

In this project, the MDS-choice model specific to the identification behavior was used to model identification data with representations derived from similarity ratings, while a generic Drift Diffusion model was used to understand the category verification data, without representational assumptions. An intriguing next step would be to understand more expertise-related behaviors by building a generic model that describes a generic decision-making process built on top of individualized representations. The EBRW model (Nosofsky & Palmeri, 1997) provides an excellent starting point, with a generic random walk decision-making process built on top of multidimensional representations that can be easily individualized. If the representation component can be individualized based on individual expertise level or even more individual metrics such as IQ and visual acuity, then ideally the full model can reproduce most of the behaviors at an individual level.

Another question not tapped in this project is the relationship between the parameters of visual expertise and other external variables such as IQ, genotype, experience, and visual acuity. For example, the sensitivity parameter might be a function of visual acuity and IQ, while the weights on diagnostic dimensions can be predicted by experience. Scores from tests such as the Vanderbilt Expertise Test (McGugin et al., 2012), the Semantic Vanderbilt Bird Expertise Test (Van Gulick, 2014; Van Gulick et al., 2016), and the Raven Progressive Matrices Test (P. A. Carpenter, Just, & Shell, 1990) can be used as covariates or predictors in cognitive process models to predict critical parameters such as the sensitivity parameter. Studies in this direction could provide insights into the training of visual expertise.

In this project, cognitive processing was assumed to start from abstract representations of bird species, without considering the processes from the retina to the abstract representations. This poses another interesting challenge, to model the entire cognitive processing from retina to eventual decision-making. Theoretically, it is possible to use the representations derived from an object recognition model (e.g., Cottrell & Nguyen, 2005; Riesenhuber & Poggio, 2000; Serre et al., 2007; Tong et al., 2005) as the input for further cognitive processing in models such as MDS-choice model or EBRW (e.g., Mack & Palmeri, 2010; Ross et al., 2014)<sup>1</sup>.

---

<sup>1</sup>In an effort to promote open science and reproducible research, this document in its entirety was written using *RMarkdown*. The behavioral and modeling data were read in, analyzed, and visualized using R functions within the document. The source code for this document can be requested from the author at [j.shen33@gmail.com](mailto:j.shen33@gmail.com)

## Appendix A

### The Questionnaire before the Bird Expertise Test

#### **Demographic Information**

1. Year of birth (e.g., 1975)
2. Gender: Female / Male / Other
3. Do you have any neurological condition that might affect your vision, hearing, memory, or thinking? Yes / No / Refuse to answer

#### **Birding Experience**

1. Do you consider yourself a birdwatcher? Are you interested in birds? Answer yes if you're someone who merely enjoys identifying birds at your bird feeder or on a hike, someone who regularly does birding as a hobby, or are a professional with expertise in bird identification or ornithology. If yes, answer the remaining questions: Yes / No
2. At what age did you first develop an interest in birds?
3. At what age did you first start birding relatively seriously (e.g., spending time learning bird identifications, going on planned bird walks, joining local Audubon or ornithological societies, etc.)
4. How much formal training and coursework do you have in ornithology?
  - (a) PhD, with a concentration in ornithology or a related discipline
  - (b) Master or Bachelors degree, with a concentration in ornithology or a related discipline

- (c) Two or more college-level courses in ornithology
  - (d) One college-level course in ornithology
  - (e) One or more formal workshops or community courses in ornithology
  - (f) No formal coursework in ornithology
5. How often do you try to go birding outside of your home during peak birding times of the year, on average?
- (a) Two or more times per week
  - (b) Once a week
  - (c) Every 2-3 weeks
  - (d) Occasionally
  - (e) Rarely or never
6. How often have you planned a vacation with a primary intent of birding, on average?
- (a) I am a professional who regularly identifies birds (e.g., ornithological research, photographer, tour leader, educator, wildlife resource manager)
  - (b) More than once a year
  - (c) Once a year
  - (d) Every other year
  - (e) Once every few years
  - (f) Rarely or never
7. How would rate your own bird expertise for birds where you live?
- (a) I am a novice. Nearly all other birders I meet are more skilled than I am.
  - (b) I am a beginner. Most birders I meet are more skilled than I am, but I occasionally meet other beginners like me when out birding.

- (c) I have intermediate birding skills. While there are many birders more skilled than I am, I can identify many birds that beginners cannot.
- (d) I have advanced birding skills. While I am not the most expert birder that I know in my area, I often identify birds quicker and more accurately than others.
- (e) I have expert birding skills. While not a professional, I often lead birding trips for my local birding societies, organize local bird counts, etc.
- (f) I have expert birding skills. While I have met some people who are more expert than I am, I have done things like lead birding tour groups professionally, conduct ornithological research, educate about bird identification and bird conservation, or work in wildlife management.
- (g) I have expert birding skills. I am recognized by my peers in my state, nationally, or internationally as someone other experts would turn to because of my expertise.

8. Rate your expertise on a scale of 1 (novice) to 5 (expert) for the following birding regions:

- (a) Eastern US and Eastern Canada: 1 / 2 / 3 / 4 / 5
- (b) Western US and Western Canada: 1 / 2 / 3 / 4 / 5
- (c) Arctic: 1 / 2 / 3 / 4 / 5
- (d) Pelagic (Atlantic and Pacific): 1 / 2 / 3 / 4 / 5
- (e) South Texas: 1 / 2 / 3 / 4 / 5
- (f) Southeast Arizona: 1 / 2 / 3 / 4 / 5
- (g) South Florida: 1 / 2 / 3 / 4 / 5
- (h) Caribbean: 1 / 2 / 3 / 4 / 5
- (i) Mexico and Central America: 1 / 2 / 3 / 4 / 5

- (j) If you have expertise with any other world regions, please specify:
9. (Optional) Name up to five people you know personally who you would turn to for their expert birding knowledge (e.g., you would want them around to verify a tough bird identification, teach you better birding skills, or help you find new birds to add to your life list):
  10. (Optional) eBird login ID. If you provide your eBird login ID, we may access your eBird records.



## Appendix B

### The Bird Expertise Test

---

Target Bird	Distractor Birds
Turkey Vulture	Northern Harrier Ferruginous Hawk Parasitic Jaeger
Osprey	Golden Eagle Crested Caracara Ruffed Grouse
Canada Goose	Brant Common Loon Sanderling

---

---

**Target Bird****Distractor Birds**

---

Bald Eagle

Golden Eagle

Broad-Winged Hawk

Great Egret

Greater Roadrunner

Black-Billed Cuckoo

Northern Bobwhite

Ruddy Turnstone

Killdeer

Piping Plover

Willet

Herring Gull

---

**Target Bird****Distractor Birds**

---

Barn Owl

Barred Owl

Common Poorwill

American Bittern

Great Blue Heron

Cattle Egret

American Avocet

Anhinga

Wood Duck

Harlequin Duck

Pied-Billed Grebe

Bridled Tern

---

**Target Bird****Distractor Birds**

---

Western Tanager

Hepatic Tanager

Painted Bunting

Varied Bunting

Northern Cardinal

Painted Bunting

Scarlet Tanager

Mountain Chickadee

American Robin

Swainson's Thrush

Lazuli Bunting

Song Sparrow

---

**Target Bird**

---

**Distractor Birds**

---

American Tree Sparrow

Spotted Towhee

Western Meadowlark

Western Scrub-Jay

European Starling

Common Raven

Lark Bunting

Pine Grosbeak

American Goldfinch

House Finch

Yellow Grosbeak

Summer Tanager

---

**Target Bird**

---

**Distractor Birds**

---

Blue Jay

Clark's Nutcracker

Mountain Bluebird

Carolina Wren

Northern Mockingbird

Brown Thrasher

Gray Catbird

Barn Swallow

Rock Pigeon

Mourning Dove

Least Flycatcher

Cape May Warbler

---

**Target Bird****Distractor Birds**

---

Black-Billed Magpie

American Crow

Brewer's Blackbird

Lark Bunting

Steller's Jay

Blue Jay

Blue Grosbeak

Hermit Warbler

Wood Thrush

Varied Thrush

American Pipit

Spragues's Pipit

---

**Target Bird**

---

---

**Distractor Birds**

---

Painted Bunting

Lazuli Bunting

Hooded Oriole

Rose-Breasted Grosbeak

Western Scrub Jay

Steller's Jay

Indigo Bunting

Great Kiskadee

House Finch

Common Redpoll

Flame-Colored Tanager

Crissal Thrasher



---

**Target Bird****Distractor Birds**

---

Baltimore Oriole

Eastern Meadowlark

Blackburnian Warbler

Black-Capped Chickadee

Brown-Headed Cowbird

Baltimore Oriole

Song Sparrow

Cape May Warbler

Cactus Wren

Carolina Wren

Abert's Towhee

Fox Sparrow

---

**Target Bird**

---

**Distractor Birds**

---

Clark's Nutcracker

Pinyon Jay

Bridled Titmouse

Scott's Oriole

Bay-Breasted Warbler

Blackburnian Warbler

Cape May Warbler

Pine Warbler

Summer Tanager

Scarlet Tanager

Hepatic Tanager

Flame-Colored Tanager

---

**Target Bird****Distractor Birds**

---

Blackburnian Warbler

Bay-Breasted Warbler

Magnolia Warbler

Palm Warbler

Northern Flicker

Yellow-Bellied Sapsucker

Brown Thrasher

House Sparrow

Yellow-Breasted Chat

Common Yellowthroat

Wilson's Warbler

Mourning Warbler

---

**Target Bird**

---

---

**Distractor Birds**

---

Bushtit

Verdin

Swamp Sparrow

Painted Bunting

American Redstart

Altamira Oriole

Cape May Warbler

Ovenbird

Bridled Titmouse

Tufted Titmouse

Mountain Chickadee

Black-Crested Titmouse

---

**Target Bird**

---

**Distractor Birds**

---

Scott's Oriole

Audubon's Oriole

Orchard Oriole

Bullock's Oriole

Pileated Woodpecker

Downy Woodpecker

Black-Billed Magpie

Townsend's Solitaire

Cerulean Warbler

Tropical Parula

Black Throated Blue Warbler

Collared Whitestart

---

**Target Bird****Distractor Birds**

---

Red-Winged Blackbird

Common Grackle  
American Redstart  
Scott's Oriole

Mourning Dove

Rock Pigeon  
Loggerhead Shrike  
Lark Sparrow

Bullock's Oriole

Baltimore Oriole  
Audubon's Oriole  
Altamira Oriole

---

**Target Bird**

---

**Distractor Birds**

---

Red-Bellied Woodpecker

Pileated Woodpecker

Vermillion Flycatcher

Northern Mockingbird

Veery

Swainson's Thrush

Wood Thrush

Hermit Thrush

American Dipper

Gray Catbird

Black Phoebe

Gray Jay

---

**Target Bird****Distractor Birds**

---

Black-and-White Warbler

White-Breasted Nuthatch

Red-Breasted Nuthatch

Brown-Headed Nuthatch

Acorn Woodpecker

Red-Headed Woodpecker

Gila Woodpecker

Yellow-Bellied Sapsucker

Chipping Sparrow

American Tree Sparrow

Clay-Colored Sparrow

Field Sparrow



---

**Target Bird**

---

**Distractor Birds**

---

Nashville Warbler

Orange-Crowned Warbler

Tennessee Warbler

Palm Warbler

Belted Kingfisher

Bridled Titmouse

Phainopepla

Great Crested Flycatcher

Great Kiskadee

Say's Phoebe

Yellow-Throated Vireo

Brewer's Sparrow

---

**Target Bird****Distractor Birds**

---

Vermilion Flycatcher

Say's Phoebe

Great Kiskadee

Scarlet Tanager

Blue Grosbeak

Indigo Bunting

Blue Jay

Cerulean Warbler

Indigo Bunting

Blue Grosbeak

Eastern Bluebird

Lazuli Bunting

---

**Target Bird****Distractor Birds**

---

Mountain Bluebird

Eastern Bluebird

Western Bluebird

Blue Grosbeak

Brown Creeper

Canyon Wren

Bewick's Wren

Cactus Wren

Common Redpoll

Hoary Redpoll

Pine Siskin

Cassin's Finch

---

**Target Bird****Distractor Birds**

---

Fox Sparrow

Song Sparrow

Henslow's Sparrow

Sage Sparrow

Bobolink

Lark Bunting

Eastern Meadowlark

Yellow-Headed Blackbird

Eastern Towhee

Spotted Towhee

Lark Sparrow

Black-Headed Grosbeak

---

**Target Bird****Distractor Birds**

---

Blue-Headed Vireo

Warbling Vireo

Hutton's Vireo

Cassin's Vireo

Hermit Warbler

Black-Throated Green Warbler

Golden-Cheeked Warbler

Townsend's Warbler

White-Breasted Nuthatch

Tufted Titmouse

Pygmy Nuthatch

Carolina Chickadee

---

**Target Bird****Distractor Birds**

---

Golden-Fronted Woodpecker

Red-Bellied Woodpecker

Gila Woodpecker

Red-Headed Woodpecker

Warbling Vireo

Hutton's Vireo

Philadelphia Vireo

Bell's Vireo

Townsend's Warbler

Black-Throated Green Warbler

Blackburnian Warbler

Magnolia Warbler

---

**Target Bird****Distractor Birds**

---

Yellow-Bellied Sapsucker

Red-Naped Sapsucker

Golden-Fronted Woodpecker

Ladder-Backed Woodpecker

Pygmy Nuthatch

White-Breasted Nuthatch

Brown-Headed Nuthatch

Bushtit

Hermit Thrush

Swainson's Thrush

Veery

Wood Thrush

---

**Target Bird****Distractor Birds**

---

Prothonotary Warbler

Yellow Warbler

Prairie Warbler

Wilson's Warbler

Wilson's Warbler

Yellow Warbler

Hooded Warbler

Common Yellowthroat

Vesper Sparrow

Song Sparrow

Lincoln's Sparrow

Swamp Sparrow



---

**Target Bird****Distractor Birds**

---

Say's Phoebe

Black Phoebe

Eastern Phoebe

Olive-Sided Flycatcher

Phainopepla

Pyrrhuloxia

Cedar Waxwing

Northern Cardinal

White-Crowned Sparrow

White-Throated Sparrow

Song Sparrow

Swamp Sparrow

---

**Target Bird****Distractor Birds**

---

Common Raven

American Crow

Fish Crow

Gray Jay

Purple Martin

Tree Swallow

Common Grackle

Barn Swallow

Pyrrhuloxia

Northern Cardinal

Scarlet Tanager

Bobolink

---

**Target Bird****Distractor Birds**

---

Northern Rough-Winged Swallow    Tree Swallow

Barn Swallow

Bank Swallow

Evening Grosbeak

American Goldfinch

Pine Grosbeak

Purple Finch

Brown-Headed Nuthatch

Pygmy Nuthatch

Red-Breasted Nuthatch

Brown Creeper

---

**Target Bird****Distractor Birds**

---

White-Winged Dove

Common Ground Dove

Inca Dove

White-Tipped Dove

Townsend's Solitaire

Mountain bluebird

Northern Shrike

Gray-Cheeked Thrush

Carolina Wren

Cactus Wren

Ovenbird

Tufted Titmouse

---

**Target Bird****Distractor Birds**

---

Ovenbird

Northern Waterthrush

Lark Sparrow

Louisiana Waterthrush

Verdin

Lucy's Warbler

Bushtit

Blue-Gray Gnatcatchers

Louisiana Waterthrush

Northern Waterthrush

Ovenbird

Song Sparrow

---

**Target Bird**

---

**Distractor Birds**

---

Lawrence's Goldfinch

Lesser Goldfinch

American Goldfinch

Evening Grosbeak

Canyon Wren

Rock Wren

Winter Wren

Sedge Wren

Varied Thrush

American Robin

Veery

Swainson's Thrush

---

**Target Bird****Distractor Birds**

---

Ruby-Crowned Kinglet

Golden-Crowned Kinglet

Hutton's Vireo

Cassin's Vireo

Western Wood-Pewee

Eastern Phoebe

Greater Pewee

Olive-Sided Flycatcher

Winter Wren

House Wren

Bewick's Wren

Marsh Wren

## Appendix C

### The Coverage Map for the Bird Expertise Test

Target Bird	Expertise Level	Level Number
Turkey Vulture	Practice	1
Osprey	Practice	1
Canada Goose	Practice	1
Bald Eagle	Practice	1
Greater Roadrunner	Practice	1
Killdeer	Practice	1
Barn Owl	Practice	1
Great Blue Heron	Practice	1
Wood Duck	Practice	1
Western Tanager	Intermediate	4
Northern Cardinal	Novice	2
American Robin	Novice	2
American Tree Sparrow	Beginner	3
European Starling	Novice	2
American Goldfinch	Novice	2
Blue Jay	Novice	2
Northern Mockingbird	Novice	2
Rock Pigeon	Novice	2
Black-Billed Magpie	Intermediate	4
Steller's Jay	Beginner	3
Wood Thrush	Intermediate	4



Target Bird	Expertise Level	Level Number
Painted Bunting	Intermediate	4
Western Scrub Jay	Beginner	3
House Finch	Beginner	3
Baltimore Oriole	Novice	2
Brown-Headed Cowbird	Beginner	3
Cactus Wren	Intermediate	4
Clark's Nutcracker	Intermediate	4
Bay-Breasted Warbler	Advanced	5
Summer Tanager	Advanced	5
Blackburnian Warbler	Advanced	5
Northern Flicker	Beginner	3
Yellow-Breasted Chat	Advanced	5
Bushtit	Beginner	3
American Redstart	Intermediate	4
Bridled Titmouse	Advanced	5
Scott's Oriole	Advanced	5
Pileated Woodpecker	Intermediate	4
Cerulean Warbler	Intermediate	4
Red-Winged Blackbird	Beginner	3
Mourning Dove	Novice	2
Bullock's Oriole	Advanced	5
Red-Bellied Woodpecker	Beginner	3
Veery	Expert	6
American Dipper	Expert	6
Black-and-White Warbler	Expert	6

Target Bird	Expertise Level	Level Number
Acorn Woodpecker	Advanced	5
Chipping Sparrow	Expert	6
Nashville Warbler	Expert	6
Belted Kingfisher	Intermediate	4
Great Kiskadee	Intermediate	4
Vermilion Flycatcher	Intermediate	4
Blue Grosbeak	Intermediate	4
Indigo Bunting	Advanced	5
Mountain Bluebird	Expert	6
Brown Creeper	Advanced	5
Common Redpoll	Expert	6
Fox Sparrow	Expert	6
Bobolink	Advanced	5
Eastern Towhee	Expert	6
Blue-Headed Vireo	Expert	6
Hermit Warbler	Expert	6
White-Breasted Nuthatch	Expert	6
Golden-Fronted Woodpecker	Expert	6
Warbling Vireo	Expert	6
Townsend's Warbler	Expert	6
Yellow-Bellied Sapsucker	Expert	6
Pygmy Nuthatch	Expert	6
Hermit Thrush	Expert	6
Prothonotary Warbler	Advanced	5
Wilson's Warbler	Expert	6

Target Bird	Expertise Level	Level Number
Vesper Sparrow	Expert	6
Say's Phoebe	Advanced	5
Phainopepla	Intermediate	4
White-Crowned Sparrow	Expert	6
Common Raven	Advanced	5
Purple Martin	Expert	6
Pyrrhuloxia	Intermediate	4
Northern Rough-Winged Swallow	Expert	6
Evening Grosbeak	Expert	6
Brown-Headed Nuthatch	Expert	6
White-Winged Dove	Expert	6
Townsend's Solitaire	Expert	6
Carolina Wren	Beginner	3
Ovenbird	Expert	6
Verdin	Expert	6
Louisiana Waterthrush	Expert	6
Lawrence's Goldfinch	Expert	6
Canyon Wren	Expert	6
Varied Thrush	Expert	6
Ruby-Crowned Kinglet	Expert	6
Western Wood-Pewee	Expert	6
Winter Wren	Expert	6

## BIBLIOGRAPHY

- Akaike, H. (1973). Information measures and model selection. *Bulletin of the International Statistical Institute*, *50*, 277–290.
- Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). on the Dangers of Averaging Across Subjects When Using Multidimensional Scaling or the Similarity-Choice Model. *Psychological Science*, *5*, 144–151.
- Bates, D., Mchler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.
- Bentler, P. M., & Wu, E. J. (2012). EQSIRT—A comprehensive item response theory program. *Multivariate Software*.
- Bocci, L., & Vichi, M. (2011). The K-INDSCAL model for heterogeneous three-way dissimilarity data. *Psychometrika*, *76*, 691–714.
- Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. New York, NY: Springer Science & Business Media.
- Boster, J. S., & Johnson, J. C. (1989). Form or function: A comparison of expert and novice judgments of similarity among fish. *American Anthropologist*, *91*, 866–889.
- Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*, 117–128.
- Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–78.
- Byatt, G., & Rhodes, G. (2004). Identification of own-race and other-race faces: Implications for the representation of race in face space. *Psychonomic Bulletin & Review*, *11*,

735–741.

- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*, 1–37.
- 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*, 404.
- Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. *Psychometrika*, *35*, 283–319.
- Cottrell, G. W., & Nguyen, N. H. (2005). Owls and wading birds: Generalization gradients in expertise. In *Proceedings of the cognitive science society* (Vol. 27).
- Cox, T. F., & Cox, M. A. (2000). *Multidimensional scaling*. London: Chapman; Hall.
- Cronbach, L. J. (1975). The two disciplines of scientific psychology. *American Psychologist*, *30*, 671–684.
- Curby, K. M., & Gauthier, I. (2009). The temporal advantage for individuating objects of expertise : Perceptual expertise is an early riser. *Journal of Vision*, *9*(6), 1–13.
- Donkin, C., Averell, L., Brown, S., & Heathcote, A. (2009). Getting more from accuracy and response time data: methods for fitting the linear ballistic accumulator. *Behavior Research Methods*, *41*, 1095–1110.
- Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*, 576–585.
- Dunn, J. L., & Alderfer, J. K. (2011). *National geographic field guide to the birds of north*

*america*. Washington, D.C.: National Geographic.

- Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, *16*, 1026–1036.
- Ennis, D. M. (1988). Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General*, *117*, 408–411.
- Ericsson, K. A. (2006). The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 691–698). Cambridge, UK: Cambridge University Press.
- Ericsson, K. A. (2009). Enhancing the Development of Professional Performance: Implications From the Study of Deliberate Practice. In K. A. Ericsson (Ed.), *The development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 412–425). New York, NY: Cambridge University Press.
- Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140.
- Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*, 403–408.
- Floyd, T. (2008). *Smithsonian field guide to the birds of north america*. (P. Hess & G. Scott, Eds.). New York, NY: HarperCollins Publishers.
- Gauthier, I., & Tarr, M. J. (1997). Becoming a “Greeble” expert: Exploring mechanisms

- for face recognition. *Vision Research*, 37, 1673–1682.
- Gauthier, I., & Tarr, M. J. (2002). Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 431–446.
- Gauthier, I., McGugin, R., Richler, J. J., Herzmann, G., Speegle, M., & Van Gulick, A. (2013). Experience with objects moderates the overlap between object and face recognition performance, suggesting a common ability. *Journal of Vision*, 13, 982–982.
- Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3, 191–197.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models* (Vol. 1). Cambridge University Press New York, NY, USA.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457–472.
- Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*.
- Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: as soon as you know it is there, you know what it is. *Psychological Science*, 16, 152–60.
- Hagen, S., Vuong, Q. C., Scott, L. S., Curran, T., & Tanaka, J. W. (2014). The role of color in expert object recognition. *Journal of Vision*, 14, 9–9.
- Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.
- Hintzman, D. L. (1986). “Schema abstraction” in a multiple-trace memory model. *Psychological Review*, 93, 411–428.
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path

- lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.
- Iribarra, D. T., & Freund, R. (2014). *Wright map: IRT item-person map with conquest integration*.
- Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General*, *126*, 248–277.
- Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: making the connection. *Cognitive Psychology*, *16*, 243–275.
- Kaufman, K. (2011). *Kaufman Field Guide to Advanced Birding: Understanding What You See and Hear*. New York, NY: Houghton Mifflin Harcourt.
- Kreft, I. G., Kreft, I., & Leeuw, J. de. (1998). *Introducing multilevel modeling*. Sage.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097-1105).
- Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*, 115–129.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.
- Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science*, *11*, 379–385.
- Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.
- Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities, and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive Science*,



32, 1403–1424.

- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.
- Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*, 605–621.
- Linden, W. J. van der. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York, NY: Wiley.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. J. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.
- Mack, M. L., & Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision*, *10*, 11–11.
- Mack, M. L., & Palmeri, T. J. (2011). The timing of visual object categorization. *Frontiers in Psychology*, *2*, 1–8.
- Mack, M. L., Wong, A. C.-N., Gauthier, I., Tanaka, J. W., & Palmeri, T. J. (2009). Time course of visual object categorization: Fastest does not necessarily mean first. *Vision Research*, *49*, 1961–1968.
- Maurer, U., Blau, V. C., Yoncheva, Y. N., & McCandliss, B. D. (2010). Development of visual expertise for reading: Rapid emergence of visual familiarity for an artificial

- script. *Developmental Neuropsychology*, 35, 404–422.
- McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision Research*, 69, 10–22.
- Meagher, B. J., Carvalho, P. F., Goldstone, R. L., & Nosofsky, R. M. (2017). Organized simultaneous displays facilitate learning of complex natural science categories. *Psychonomic Bulletin & Review*, 1–8.
- Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*, 32, 49–96.
- Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32, 89–115.
- Murphy, G. L., & Brownell, H. H. (1985). Category differentiation in object recognition: typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 70–84.
- Nash, J. C. (2014). On best practice optimization methods in R. *Journal of Statistical Software*, 60, 1–14.
- Nash, J. C., & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, 43, 1–14.
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 104.
- Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, 38, 415–432.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization rela-

- tionship. *Journal of Experimental Psychology: General*, *115*, 39–57.
- Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.
- Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700–708.
- Nosofsky, R. M. (1992a). Exemplar-based approach to relating categorization, identification, and recognition. In G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 363–394). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Nosofsky, R. M. (1992b). Exemplars, prototypes, and similarity rules. In *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 149–167).
- Nosofsky, R. M. (1992c). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*, 25–53.
- Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.
- Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 608–629.
- Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 282–304.
- Nosofsky, R. M., Sanders, C. A., Gedom, A., Douglas, B. J., & McDaniel, M. A. (2017a).

- On learning natural-science categories that violate the family-resemblance principle. *Psychological Science*, 28, 104–114.
- Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2017b). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*.
- Okada, K., & Lee, M. D. (2016). A Bayesian approach to modeling group and individual differences in multidimensional scaling. *Journal of Mathematical Psychology*, 70, 35–44.
- Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 324–54.
- Palmeri, T. J., & Cottrell, G. W. (2009). Modeling Perceptual Expertise. In D. N. Bub, M. J. Tarr, & I. Gauthier (Eds.), *Perceptual expertise: Bridging brain and behavior* (pp. 197–245). New York, NY: Oxford University Press.
- Palmeri, T. J., Wong, A. C.-N., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in Cognitive Sciences*, 8, 378–386.
- Papesh, M. H., & Goldinger, S. D. (2010). A multidimensional scaling analysis of own-and cross-race face spaces. *Cognition*, 116, 283–288.
- Petrov, A. A., Van Horn, N. M., & Ratcliff, R. (2011). Dissociable perceptual-learning mechanisms revealed by diffusion-model analysis. *Psychonomic Bulletin & Review*, 18, 490–497.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statisti-*

*cal Computing (DSC 2003)*, 20–22.

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

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review*, 88, 552–572.

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, 2, 237.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, 20, 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9, 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111, 333–367.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, 120, 697–719.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*.

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review*, 13, 626–635.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92, 544–559.

Rhodes, G., & McLean, I. G. (1990). Distinctiveness and expertise effects with homoge-

- neous stimuli: Towards a model of configural coding. *Perception*, *19*, 773–794.
- Richler, J. J., Wilmer, J. B., & Gauthier, I. (2017). General object recognition is specific: Evidence from novel and familiar objects. *Cognition*, *166*, 42–55.
- Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*, 1199–1204.
- Rizopoulos, D. (2006). Ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25.
- Rogers, T. T., & Patterson, K. (2007). Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General*, *136*, 451–469.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.
- Ross, D. A., Deroche, M., & Palmeri, T. J. (2014). Not just the norm: Exemplar-based models also predict face aftereffects. *Psychonomic Bulletin and Review*, *21*, 47–70.
- Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.
- Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2014). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*.
- Rubin, D. B., & others. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*, 1151–1172.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*,

461–464.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 6424–9.

Shen, J., & Palmeri, T. (2015). Modeling the dynamics of visual object categorization. In *Journal of vision* (Vol. 15, p. 1160). St. Pete Beach, FL.

Shen, J., & Palmeri, T. J. (2016). Modelling individual difference in visual categorization. *Visual Cognition*, *24*, 260–283.

Shen, J., Mack, M. L., & Palmeri, T. J. (2014). Studying real-world perceptual expertise. *Frontiers in Psychology*, *5*, 1–6.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, *55*, 509.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*, 125–140.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390–398.

Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky.

Shepard, R. N. (1987). Universal Law of Generalization for Psychological Science Primacy

- of Generalization Apparent Noninvariance of Generalization. *Science*, 237, 1317–1323.
- Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248–1284.
- Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 3–27.
- Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, 27, 161–168.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.
- Stan Development Team. (2016). RStan: The R interface to Stan.
- Sunday, M. A., Donnelly, E., & Gauthier, I. (2017). Individual differences in perceptual abilities in medical imaging: The Vanderbilt Chest Radiograph Test. *Cognitive Research: Principles and Implications*, 2, 36.
- Takane, Y., Young, F. W., & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42, 7–67.
- Tanaka, J. W., & Farah, M. J. (2003). The Holistic Representation of Faces. In *Perception of faces, objects, and scenes: Analytic and holistic processes* (pp. 53–82).
- Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23, 457–482.
- Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real-



- world perceptual expertise. *Psychological Science*, *16*, 145–151.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association.
- Tong, M. H., Joyce, C. A., & Cottrell, G. W. (2005). Are Greebles special? Or, why the Fusiform Face Area would be recruited for sword expertise (if we had one). In *Proceedings of the 27th annual cognitive science conference*. Mahwah, New Jersey: Lawrence Erlbaum.
- Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *43*, 161–204.
- Van Gulick, A. E. (2014). *Measurement of semantic knowledge and its contribution to object recognition performance* (PhD thesis). Vanderbilt University.
- Van Gulick, A. E., McGugin, R. W., & Gauthier, I. (2016). Measuring nonvisual knowledge about object categories: The Semantic Vanderbilt Expertise Test. *Behavior Research Methods*, *48*, 1178–1196.
- Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., . . . Belongie, S. (2015). Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 595–604).
- Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.
- Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*, 44–62.
- Vehtari, A., Gelman, A., & Gabry, J. (2016). Loo: Efficient leave-one-out cross-validation

and waic for Bayesian models.

- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*, 1413–1432.
- Viken, R. J., Treat, T. A., Nosofsky, R. M., McFall, R. M., & Palmeri, T. J. (2002). Modeling individual differences in perceptual and attentional processes related to bulimic symptoms. *Journal of Abnormal Psychology*, *111*, 598–609.
- Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206–1220.
- Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods*, *46*, 15–28.
- Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.
- Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, *7*.
- Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Winsberg, S., & De Soete, G. (1993). A latent class approach to fitting the weighted Euclidean model, clascal. *Psychometrika*, *58*, 315–330.