Concept Learning, Perceptual Fluency, and Expert Classification

A dissertation presented to
the faculty of
the College of Arts and Sciences of Ohio University

In partial fulfillment
of the requirements for the degree
Doctor of Philosophy

Derek E. Zeigler
August 2016
© 2016 Derek E. Zeigler. All Rights Reserved.
This dissertation titled
Concept Learning, Perceptual Fluency, and Expert Classification

by
DEREK E. ZEIGLER

has been approved for
the Department of Psychology
and the College of Arts and Sciences by

Ronaldo Vigo
Associate Professor of Psychology

Robert Frank
Dean, College of Arts and Sciences
Abstract

ZEIGLER, DEREK E., Ph.D., August 2016, Experimental Psychology

Concept Learning, Perceptual Fluency, and Expert Classification

Director of Dissertation: Ronaldo Vigo

The way in which category specific knowledge is acquired over time has been a longstanding central topic in the cognitive and perceptual sciences. Accordingly, the influence of training and experience on learning has been the focus of much empirical work. This research often involves accounting for the results of concept learning tasks that necessitate classifying category members and non-members. Studies in this area explore questions like the following. Can different concepts be ordered by their relative learning difficulty? Does repeated exposure to a concept result in perceptual expertise and/or expert classification? Is concept acquisition inherently easier for some individuals? The relative difficulty between categories tells us something fundamental about the conceptual system by revealing which relational structures humans are most sensitive. As such, concept learning difficulty orderings for categorical stimuli form an important part of the empirical foundation of concept learning research. However, it is rare that the stability of such orderings is tested over a period of extended learning. Further, this research rarely explores dependent variables beyond classification accuracy that may also indicate relative learning difficulty. Accordingly, this investigation explores the relationship between accuracy and response times (RTs) when practice is gained over multiple category learning sessions. Of particular interest is the extent to which the relative learning difficulty between categories remains stable over sessions of learning.
Of additional interest are measures of perceptual fluency (classification RTs) that might reflect category difficulty. Learning difficulty orderings in terms of classification RTs provide an alternative to the conventional approach that construes difficulty solely in terms of mean proportion of correct/incorrect responses. In light of recent empirical support for an invariance–based structural account of conceptual representations (Vigo, 2011a; 2013; 2014), the acquired data is interpreted in the context of generalized invariance structure theory (GIST; Vigo, 2013, 2014) in order to reveal how task experience influences the way concepts are learned and represented over time.
Dedication

Dedicated to my parents for providing the support

And to Annie for providing the motivation
Acknowledgments

A number of individuals have supported this project and have made contributions which have led to the final draft of this manuscript. The current work was made possible thanks to the combined effort of these individuals.

First and foremost, I would like to thank my advisor and mentor Dr. Ronaldo Vigo. For many years now, Dr. Vigo has been a continual source of inspiration. His passion for science motivated me to pursue a doctoral degree and his close mentorship has molded my approach to research as well as life in general. He invests so much time and energy into mentoring his students and I have benefited greatly from his dedication. Thank you, Dr. Vigo, for always believing in me and encouraging me to realize my potential—for this, I am forever indebted to you.

Additionally, thank you to those who provided technical support on the current project. Dr. Ronaldo Vigo provided the core experimental paradigm as well as the program for the computer-based task and Basawaraj Basawaraj made modifications to this program. Likewise, I have appreciated the daily support of SCOPE lab members, both past and present.

I would also like to provide a special thanks to Dr. Peggy Zoccola, Dr. Keith Markman, Dr. Chao-Yang Lee, and Dr. Robert Briscoe. This document is been significantly improved thanks to their willingness to serve as dissertation committee members. Because of these individuals, the process has been very pleasurable—I am very grateful to all of you.


<table>
<thead>
<tr>
<th>Table of Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
</tr>
<tr>
<td>Dedication</td>
</tr>
<tr>
<td>Acknowledgments</td>
</tr>
<tr>
<td>List of Tables</td>
</tr>
<tr>
<td>List of Figures</td>
</tr>
<tr>
<td>Chapter 1: Introduction</td>
</tr>
<tr>
<td>Chapter 2: Category Learning</td>
</tr>
<tr>
<td>Category Structures</td>
</tr>
<tr>
<td>Integral and Separable Dimension Stimuli</td>
</tr>
<tr>
<td>Serioinformative and Parainformative Tasks</td>
</tr>
<tr>
<td>Learning Difficulty Orderings</td>
</tr>
<tr>
<td>Chapter Summary</td>
</tr>
<tr>
<td>Chapter 3: Concept Representations</td>
</tr>
<tr>
<td>The Classic Representational Paradigms</td>
</tr>
<tr>
<td>Invariance-Based Representations in Ideotype Space</td>
</tr>
<tr>
<td>Chapter Summary</td>
</tr>
<tr>
<td>Chapter 4: Perceptual Fluency and Expertise during Learning</td>
</tr>
<tr>
<td>Expert Knowledge</td>
</tr>
<tr>
<td>Expertise and Perceptual Fluency</td>
</tr>
<tr>
<td>Response Times and Practice</td>
</tr>
<tr>
<td>Classification Response Times and Practice</td>
</tr>
<tr>
<td>Chapter Summary</td>
</tr>
<tr>
<td>Chapter 5: Overview of the Current Empirical Work</td>
</tr>
<tr>
<td>Chapter 6: Repeated Exposure to a Single Structure Type</td>
</tr>
<tr>
<td>Method</td>
</tr>
<tr>
<td>Stimuli and materials</td>
</tr>
<tr>
<td>Participants</td>
</tr>
<tr>
<td>Procedure</td>
</tr>
</tbody>
</table>
Chapter 9: Discussion ..................................................................................................... 115
General Discussion ..................................................................................................... 115
Limitations and Future Directions .............................................................................. 128
References....................................................................................................................... 132
List of Tables

Table 1: Mean classification errors for $3_2[4]$ types in parainformative tasks ..................31
Table 2: Estimated alpha values in Vigo (2013) .............................................................63
Table 3: $3_2[4]$ structural manifold values .....................................................................65
Table 4: Study 1 errors and response times per epoch ....................................................72
Table 5: Study 1 alpha values for errors .........................................................................76
Table 6: Study 1 alpha values for response times ...........................................................77
Table 7: Study 2 alpha values for errors .........................................................................97
Table 8: Study 2 alpha values for response times ...........................................................98
Table 9: Study 1 power and exponential fits per type ..................................................108
Table 10: Study 2 power and exponential fits per type ................................................114
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Boolean cubes for the $3_2[4]$ structure types</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Sample stimuli depicting the $3_2[4]$ structure types</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>Structural manifold transformations</td>
<td>41</td>
</tr>
<tr>
<td>4</td>
<td>Ideotypes points for the $3_2[4]$ category structures</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>Power and exponential curves for two values of the learning rate parameter</td>
<td>55</td>
</tr>
<tr>
<td>6</td>
<td>Power function fit to observed response times reported by Nosofsky and Palmeri (1997)</td>
<td>57</td>
</tr>
<tr>
<td>7</td>
<td>Study 1 mean errors and response times per structure across types</td>
<td>70</td>
</tr>
<tr>
<td>8</td>
<td>Study 2 difference scores for errors and response times</td>
<td>71</td>
</tr>
<tr>
<td>9</td>
<td>Study 1 error and response time line graphs</td>
<td>73</td>
</tr>
<tr>
<td>10</td>
<td>Study 1 ideotypes per type across sessions</td>
<td>79</td>
</tr>
<tr>
<td>11</td>
<td>Study 2 mean errors and response times per structure across types</td>
<td>93</td>
</tr>
<tr>
<td>12</td>
<td>Study 2 difference scores for errors and response times</td>
<td>94</td>
</tr>
<tr>
<td>13</td>
<td>Study 2 ideotypes per type across sessions</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>Cumulative $k$ values for errors</td>
<td>102</td>
</tr>
<tr>
<td>15</td>
<td>Cumulative $k$ values for response times</td>
<td>103</td>
</tr>
<tr>
<td>16</td>
<td>Study 1 power and exponential fits on all type data</td>
<td>107</td>
</tr>
<tr>
<td>17</td>
<td>Study 1 power and exponential function graphs</td>
<td>109</td>
</tr>
</tbody>
</table>
Figure 18: Study 1 response time distributions.............................................................110
Figure 19: Study 2 power and exponential fits on all type data.................................112
Figure 20: Study 2 power and exponential function graphs ......................................113
Figure 21: Study 1 and Study 2 comparison of errors ..............................................119
Figure 22: Study 1 and Study 2 comparison of RTs ................................................120
Figure 23: Study 1 and Study 2 comparison of Session 1 ideotypes .........................125
Figure 24: Study 1 and Study 2 comparison of Session 2 ideotypes .........................126
Figure 25: Study 1 and Study 2 comparison of Session 3 ideotypes .........................127
Chapter 1: Introduction

An expert is often thought of as an individual who has acquired a depth of knowledge in a given area that surpasses that of the average person or novice. As such, if one sets out to become an expert in a given area, extensive explicit learning is undoubtedly required (knowledge of facts through rote memorization). The accumulation of facts in a certain domain is undoubtedly necessary for expertise in that area. However, the development of expertise goes considerably beyond the accumulation of vast amounts of knowledge. That is, in addition to the learning effects that commonly result from experience or training, more complex and accurate perceptual representations also emerge over time. That is, increased task performance and more accurate perceptual representations are typically associated with a concurrent effect of processing fluency marked by quicker response times as well as a reduction in the cognitive and attentional resources dedicated to the task.

In the following, the mechanisms of expertise are operationalized as a normative outcome of the learning process. In this way, the development of expertise (in terms of its behavioral markers) can be characterized accurately and precisely by observing changes in knowledge representation that result from learning. Such benefits have led many researchers in the field to operationalize expertise in a similar fashion (Palmeri & Cottrell, 2009).

Viewing expertise as the endpoint of learning affords the use of formal concept learning models as well as their associated representational theories as a theoretical base for the study of expertise development. The primary aim of these models is to account for
performance on classification tasks by fitting mean classification errors across
categorization problems of varying difficulty. Additionally, theories and models may
attempt to explain rates of learning by predicting decreases in classification errors over
time. Also of interest is how more complex and accurate perceptual representations
emerge over time due to experience.\(^1\) As such, the typical concept learning theory
proposes the structure of concept representation and, if the formal model implementation
of a theory is sufficiently flexible, it can be used to test the extent to which the theory can
account for the influence of practice and training on performance accuracy over time (e.g.
Nosofsky & Palmeri, 1997).

Often neglected in this area of research are models that account for the time
course of classification by predicting mean classification response time (RT) and/or the
structure of RT distributions. Further, empirical work often rarely explores the decreases
in RTs that result from learning and theoretical work typically fails to account for such
decreases in classification RT. It is integral for empirical and theoretical work to move in
this direction since RTs provide an additional behavioral marker which may reveal the
nature of conceptual representation beyond measures of accuracy. RTs may also indicate
ease of processing fluency even after an individual has achieved near perfect

\(^1\) The use of general terms like “experience” and/or “training” should be understood to refer to experiential
learning in particular. Experiential learning (i.e., experience-based learning) is defined by a relatively active
learning process that takes place over an extended period of learning (Kellman, 2002). Research on
experience-based learning is unique with respect to its chosen methodology which involves testing
individual performance over multiple experimental sessions. For instance, previous empirical work within
the concept learning literature has operationalized experiential learning in terms of the incorporation of
multiple experimental sessions in which participants are repeatedly exposed to the same domain specific
object stimuli and are required to provide classification decisions in response to individually presented
stimuli (Nosofsky & Palmeri, 1997; Nosofsky and Stanton, 2005; Nosofsky, Little, Donkin, and Fific,
2011). The resulting data is meant to reveal the unique influence of “experience” and/or “training” on
conceptual knowledge (Nosofsky & Palmeri, 1997).
classification accuracy. Accordingly, the current work repeatedly exposes participants to category instances in order to obtain measures of accuracy as well as RT data over three experimental sessions on consecutive days.

Of additional importance in the concept learning literature are empirically robust learning difficulty orderings for categorical stimuli. The relative difficulty between such categories tells us something fundamental about the conceptual system. Notably, it tells us which relational structures humans are most sensitive. Further, since models should account for the relative difficulty between categories, such empirical orderings are of fundamental importance in model comparisons. Accordingly, participants in the current studies repeatedly learn category structures belonging to a category family that has repeatedly shown a robust learning difficulty ordering. The conventional approach seeks to establish learning difficulty orderings in terms of mean classification accuracy. However, the time course of classification decisions may also be used to reveal relative difficulty in terms of processing fluency. Accordingly, the current experiments are designed to establish orderings for both accuracy and RTs across stages of learning.

Results reveal an interesting relationship between accuracy and fluency throughout training. In sum, the current empirical results demonstrate that classification RTs are a good indicator of relative learning difficulty. In addition, stable difficulty orderings in terms of accuracy are only revealed when category structures are presented in the context of one another, rather than in isolation. Specifically, when instances of a single category structure are presented to participants in isolation, the relative learning difficulty ordering disappears across sessions. However, the learning difficulty ordering
remains stable over multiple experimental sessions (both in terms of proportion of errors and RTs) when multiple category structures are presented in the context of one another throughout learning. Further, the current results show evidence that conceptual expertise is domain specific and accompanied by a speedup in processing marked by quicker RTs.

To our knowledge, this is the first empirical work that has explored the relationship between learning difficulty orderings and corresponding classification RTs over multiple training sessions.

The obtained results are interpret primarily in the context of a recent theory of classification (Generalized Invariance Structure Theory or GIST; Vigo, 2013, 2014) that has recently gained considerable empirical support (Vigo, 2013, 2014; Vigo & Basawaraj, 2013; Vigo, Zeigler, & Halsey, 2013; Vigo & Doan, 2015) but has not yet been used to explain classification learning over multiple sessions. This theory proposes a very different notion of conceptual representation than previous theories. Given the success of the core models of the theory in accurately predicting a wide range of classification data and on describing the nature of conceptual representation, it is now important to consider how these models may account for changes in concept learning difficulty over time (and due to experience). The associated formal mathematical model, the Generalized Invariance Structure Theory Model (GISTM), is used to account for the current results.

Specifically, the current work demonstrates how the psychologically meaningful parameters included the parametric variant of the GISTM advance our understanding of concept representations over the course of learning. Further, these data support
underlying principles of the model. For instance, observers place disproportionate emphasis on fully diagnostic and fully redundant dimensions of categorical stimuli. In addition, structures with comparable degrees of diagnosticity are similarly represented and additional experience with these structures results in progressively similar representations.

Beyond the concept learning literature, this work applies to fields concerned with the acquisition of perceptual skills and researchers interested in the development of perceptual expertise. Also, a growing amount of applied research has explored performance in jobs that require expert classification, such as radiology and air traffic control. Understanding the mechanisms of classification learning over time may elucidate the difference between experts and novices in these areas. If so, the current study might have applications for job training procedures and managerial decisions.

In summary, the current experimental tasks require individuals to repeatedly classify categorical stimuli associated with a robust learning difficulty ordering. As a result, accuracy and RT orderings can be established across learning sessions. Further, classification data is collected on three consecutive days in order to reveal the nature of concept learning beyond measures of accuracy by exploring the decreases in RTs that result from learning over time.

The dissertation proceeds as follows. First, Chapter 2 provides an overview of important methodological considerations by reviewing key empirical work in the concept learning literature. Chapter 3 provides a gentle introduction to representational theories before explaining GIST as well as the details of the formal model. Chapter 4 discusses
the relationship between response times and task training by surveying classic results in the perceptual learning literature. Chapter 5 introduces the two experiments that are the focus of the current dissertation while Chapter 6 and 7 present the methodological details and results of these studies. Chapter 8 provides response time analyses of both studies. Finally, Chapter 9 presents a detailed discussion of the current work along with an outline of future research directions.
Chapter 2: Category Learning

Being human partly means being able to intelligently interact with the environment. An essential aspect of this intelligent interaction is the ability to form accurate representations so that appropriate actions are made in response to environmental stimuli. Prior experiences with such stimuli form a rich base of knowledge used to identify, predict, and reason about familiar as well as novel stimuli. As such, the acquisition of knowledge can be described as the process of learning categories through the formation of appropriate conceptual representations of these categorical stimuli. This process of forming concept representations based on categorical stimuli has been the focus of much research in the cognitive sciences. In this chapter we discuss important methodological considerations by reviewing key empirical work in the concept learning literature.

First, however, it is important to make a few comments regarding some terminology that will become critical in the following discussion. Formal terminology will be defined as necessary throughout this thesis; however, certain informal terminology should be addressed from the start. Such practical notes concerning definitions are similarly presented by Vigo (2014, pg. 15) and Murphy (2002). First, we must distinguish between “concepts” and “categories.” In short, “concepts” are the internal mental representations of “categories” in the world. That is, “concepts” are assumed to correspond to (or represent) an objective “category” composed of a set of objects. These categories are said to be “dimensionally definable” if the set of objects can be characterized in terms of a shared number of fixed features. Moreover, we will refer to
such a dimensionally definable set as a “categorical stimulus.” Additionally, we operationalize what will be referred to as “concept learnability” or “concept difficulty” in terms of categorization/classification performance. That is, the observable behaviors under consideration will be the proportion of classification errors and the response times (RTs) for classification decisions. The discussion of specific experiments to come will likely serve to clarify this terminology.

Category Structures

Models of conceptual behavior are generally tested in terms of classification performance. Specifically, researchers may obtain average learning times (time taken to learn a concept) or average learning accuracies (percentage of correct responses) in terms of correct object classifications for categories of varying learnability and then fit model predictions to the aggregate response data across participants. Research with learning times typically uses a pre-specified learning criterion and learning proceeds until the criterion is met. In these tasks, the number of trials/blocks to criterion is the primary dependent variable (e.g. Nosofsky, Gluck, Palmeri, Mckinley, & Glauthier, 1994). Alternatively, research using learning accuracies typically takes mean proportion of errors as the primary dependent variable (e.g., Feldman, 2000; Vigo, 2013; Vigo, Zeigler, & Halsey, 2013). Additionally, the process by which concepts are learned can be studied by obtaining the rate of learning throughout an experimental session and models can be subsequently compared by quantitatively predicting block-by-block changes in performance, providing a more comprehensive comparison than measures of aggregate
performance. This approach has greatly advanced our understanding of concept formation.

When studying conceptual behavior, researchers often use well-defined category structures that adhere to logical rules, where membership is determined by a combination of features. The advantages of using such well-defined category structures when studying conceptual behavior can be enumerated as follows: (1) the objects provide clear quantities that are comparable across experiments; (2) they represent the simplest cases which models of concept learning should be able to account for first and foremost, before accounting for more ill-defined dimensional gradients; (3) despite their apparent simplicity, they are generalizable to a large amount of real world stimuli—in the way that many real world categories can be defined over a few constituent dimensions; (4) and, as mentioned above, they help researchers understand the role individual features play in category learning, aiding in the development and proper comparison of quantitative models.

The main goal of concept learning models has been to first and foremost make predictions on classification tasks using well-defined category structures. Models can be characterized in terms of their breadth and depth of performance when it comes to accounting for classification tasks. That is, a model might be very good at accounting for particular kinds of category structures (depth in a certain area) but poor at accounting for performance across a range of category structures (breadth of performance).
**Integral and Separable Dimension Stimuli**

The main goal Classification tasks either use integral or separable dimension stimuli. Integral dimension stimuli are those that are encoded and represented as single, unitary wholes. The features of integral dimension stimuli cannot be selectively attended. Color patches are the canonical example of an integral dimension stimulus (i.e., Nosofsky & Palmeri, 1997; Nosofsky & Zaki, 2003; Nosofsky, 1987; 1988). Such stimuli might vary according to hue, brightness, and saturation. The values on one of these features cannot be separated from the values of the other features. That is, single feature values cannot be changed in isolation. Changing the brightness of a color patch, for instance, results in a necessary and corresponding change in hue and saturation. Hue, brightness, and saturation cannot be attended to in isolation.

Alternatively, separable stimulus dimensions can be selectively attended and such stimuli are encoded and represented in terms of their separate dimensions. For instance, stimuli may differ in terms of the dimensions of size and shape. Each individual stimulus can thus be processed in terms of its relative size in a way that is isolated from the shape of the stimulus. That is, size can be attended to in isolation from other dimension.

This thesis deals almost exclusively with separable dimensions stimuli. However, a particularly relevant study (Nosofsky & Palmeri, 1997) using integral dimension stimuli is discussed in Chapter 4. The majority of recent work has used separable dimension stimuli (e.g., Feldman, 2000; Kurtz, Levering, Stanton, Romero, & Morris, 2013; Vigo, 2013, 2014; Vigo & Basawaraj, 2013; Vigo, Zeigler, Halsey, 2013; Vigo & Doan, 2015). The current empirical work uses the separable dimensions of size, shape, and color.
Serioinformative and Parainformative Tasks

Experimental procedures in the classification literature conform to one of two general paradigms which can be characterized as either serioinformative or parainformative (these terms were originally introduced and defined by Vigo, 2014). In serioinformative tasks, subjects classify members and non-members of categorical stimuli serially, one at a time, and with no prior experience with the categories. In parainformative tasks, however, subjects are presented with all the members of categorical stimuli simultaneously during a learning phase before individual stimuli are presented for classification. Henceforth, these labels will be used to describe experimental situations under consideration.

Historically, serioinformative tasks (Kruschke, 1992; Nosofsky, Gluck, Palmeri, & McKinley, 1994; Love & Medin, 1998) outnumber parainformative tasks (Haygood & Bourne, 1965; Posner & Keele, 1968). However, a number of recent studies have investigated learning in parainformative tasks (Feldman, 2000; Vigo, 2013; Vigo, Owens, & Evans, 2014; Vigo, Zeigler, & Halsey, 2013). It’s reasonable to believe that models should first account for data from parainformative tasks, which represent the simpler case then serioinformative tasks (Vigo, person. comm., 2015). The more inductive nature of serioinformative tasks places strain on memory, introducing a host of additional variables. Models should thus aim to first account for the results of parainformative tasks where all the information is presented for participants and classification decisions can be made in a more deductive manner. The recent focus on parainformative tasks reflects the primary importance of accounting for these data over and above data from more difficult
classification tasks. As such, the current empirical work uses a parainformative task where subjects are presented category members and non-members in a learning phase. Object stimuli are then presented one-at-a-time in a classification phase.

Classification tasks can also vary with respect to the amount of performance feedback given to participants. Tasks may show classification accuracy per trial by displaying “correct” / “incorrect” labels after each response. Such corrective feedback can be used as an additional aid throughout learning. Corrective feedback is nearly always used in serioinformative tasks being that such feedback serves as the only cue regarding category membership. Although corrective feedback can be used in both paradigms, is not typically used in parainformative tasks. One would expect improved accuracy with the inclusion of feedback in parainformative tasks and decreases in accuracy if feedback is not given during serioinformative tasks. However, this is an empirical question that remains to be tested. Currently, the extent to which corrective feedback impacts concept formation remains an open question (Vigo, 2014).

**Learning Difficulty Orderings**

Can concepts be ordered by their relative learning difficulty? Indeed, for certain categorical stimuli, research has shown that robust learning difficulty orderings can be established. Such relative difficulty between categories tells us something fundamental about the conceptual system. Notably, relative classification accuracy can be used to reveal which relational structures humans are most sensitive. Sensitivity to certain structures likely results in better classification of its members whereas categorical stimuli
with structures in which we are relatively less sensitive likely results a corresponding decrease in classification accuracy.

Concept learning difficulty orderings for categorical stimuli form the empirical foundation of concept learning research. Classification performance for sets of eight stimuli that vary along three binary-valued separable dimensions has been the focus of much empirical work. In classification tasks, the set is divided into two categories of four objects. For the eight possible unique stimuli, there are 70 distinct ways to partition them into two groups of four. These 70 partitions result in six structurally equivalent types in that any two instances of a type can be transformed into one another as a result of dimensional reassignment. The 70 partitions are distributed among the six types as such: [I (6); II (6); III (24); IV (8); V (24); VI (2)]. When participants are asked to learn the six structural relationships that can be partitioned from the $3_2^4$ family the following difficulty ordering emerges for the six structures: $I<II<III,IV,V<VI$, where Type I is the least difficult to learn, followed by Type II. Types III, IV and V are approximately equivalent in difficulty, and type VI is the most difficult.

The structure of the $3_2^4$ family types is of particular interest because of the dimensional size and cardinality of the categories (Vigo, pers. comm., 2013a). That is, three dimensions defining four positive examples turns out to be quite tractable for humans. This tractability seems to be due in part to the $3_2^4$ family’s cardinality being within the limited boundaries of short term memory capacity, which is now believed to be around four chunks of information (Luck & Vogel, 1997; Cowen, 2001). Accordingly, learning a category of four objects, each of which are composed of three dimensions,
seems an ideal test of category learning based on the established limits of short term memory. In addition, despite varying degrees of categorization performance observed within this category, the $3_2[4]$ category types are all manageable with respect to their learnability. Due to this privileged dimensional configuration, the error rates for these category types have been interpreted as revealing key psychological principles underlying the process of concept formation, including how object features influence category learning (Love, Medin, & Gureckis, 2004; Nosofsky, Palmeri, & McKinley, 1994; Vigo, 2011a, 2013) as well as how observers distribute attention along object features (Kruschke, 1992; Rehder & Hoffman, 2005). As such, the learning difficulty ordering for this category family serves as a particularly relevant benchmark for models of categorization behavior and concept formation.

A convenient method of showing category membership for categorical stimuli involves the use of Boolean cubes (e.g. Feldman, 2003). Members and non-members can be visualized in the form of a cube when stimulus dimensions take the value of 1 or 0 (i.e., true or false) in a $D$-dimensional space. Figure 1 depicts each of the $3_2[4]$ structures using Boolean cubes. Sides of the cube in Figure 1 represent dimensions. Also, corners with circles represent positive examples and empty corners represent negative examples of the category. The red lines are meant to show simple relations between object stimuli. Notice the four simple relations in Type I whereas Type VI has no simple relations. Figure 1 also shows an instance of each category structure type using the object stimuli used in the current experiments. These stimuli have been used in several previous studies (Vigo, 2013, 2014; Vigo, Zeigler, & Halsey, 2013; Vigo, Evans, & Owens, 2014). Using
stimuli that have been used in previous work provides the benefit of a common frame of reference in which meaningful comparisons can be made between studies. For instance, if our obtained classification data differs from previous work, we need not be concerned with the possibility that the chosen object stimuli resulted in differential processing. In short, we can expect participants to process and classify these object stimuli in a way that is consistent with previous work. It should be noted that the $3_2[4]$ structures have also been studied using different types of stimuli, including insect stimuli and clock stimuli (Vigo & Doan, 2015).

Figure 2 shows a Boolean cube using example stimuli where the sides of the cube represent the dimensions of size, shape, and color. Corresponding dimensional changes occur when shifting along the sides of the cube. Figure 2 also shows example instances of each category type using the stimuli depicted in the cube. Try learning the members of each type. Which structures seem to be more difficult?
Figure 1. Boolean cubes for the $3_2[4]$ structure types. Members of the $3_2[4]$ structure types are denoted by the corners of a cube. Sides of the cube represent dimensions. Corners with circles represent positive examples and empty corners represent negative examples of the category. The red lines show simple relations between object stimuli.
Figure 2. Sample stimuli depicting the $3_2[4]$ structure types. The Boolean cube depicted above is shown using stimuli from the current study. Sides of the cube represent the size, shape, and color. Moving along the sides of the cube results in a corresponding dimensional change. Example instances of each category type are given for these dimensions. These are the same stimuli used in a number of previous studies (Vigo, 2013, 2014; Vigo, Zeigler, & Halsey, 2013; Vigo, Evans, & Owens, 2014).

Shepard, Hovland, and Jenkins (1961) were the first to establish the learning difficulty ordering for the $3_2[4]$ structures in a series of experiments where classification tasks were preceded by identification tasks where unique responses were learned for each stimulus. For instance, Shepard et al. used a serioinformative task with corrective feedback after each classification in Experiment 1. Subjects classified members and non-members of categorical stimuli until a learning criterion of 32 consecutive correct responses was obtained. All participants were tested on Types I, II, and VI but only one
of either Types III, IV, & V resulting in a total of four types per subject. Five consecutive instances of each of the four types were presented to participants. Spatially separated stimuli were presented and features varied in that they were either a screw or a bolt, a candle or light bulb, and either a violin or trumpet. The ordering was also observed using geometric objects in Experiment 2.

This learning difficulty ordering has remained relevant because of multiple subsequent replications. For instance, these structures were studied further by Nosofsky, Gluck, Palmeri, Mckinley and Glauthier (1994) who replicated the results obtained by Shepard et al. while providing block-by-block learning curves suitable for quantitative fitting. The resulting learning curves conform to the difficulty ordering in terms of percentage of correct responses, showing a high level representation of the difficulty of these types in terms of learning category membership across time. Accordingly, models of classification learning were able to be compared by fitting their predictions to these data (ALCOVE, Kruschke, 1992; the rational model, Anderson, 1991; the configural-cue model, Gluck, & Bower, 1988).

In addition to these replications of serioinformative tasks, the ordering has also been replicated using parainformative tasks (Vigo, 2013; Vigo, Zeigler, & Halsey, 2013) where participants are able to study all category members and non-members in a learning phase before classifying objects one at a time. Table 1 shows the reported mean classification errors per type in studies that used a parainformative task without feedback. In all three tasks, participants studied category members and non-members in a 20 second learning phase. Following the learning phase, participants classified serially presented
stimuli. Vigo, Evans, and Owens (2014) gathered classification performance from adults, adolescents, and adolescents with ADHD. The data presented in Table 1 corresponds to the adult sample. These studies are of particular interest because the current work uses the same experimental design.

Table 1

<table>
<thead>
<tr>
<th>Study</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vigo (2013)</td>
<td>.02</td>
<td>.07</td>
<td>.19</td>
<td>.18</td>
<td>.17</td>
<td>.32</td>
</tr>
<tr>
<td>Vigo, Zeigler, &amp; Halsey (2013)</td>
<td>.01</td>
<td>.13</td>
<td>.19</td>
<td>.18</td>
<td>.19</td>
<td>.32</td>
</tr>
<tr>
<td>Vigo, Evans, &amp; Owens (2013)</td>
<td>.02</td>
<td>.12</td>
<td>.17</td>
<td>.19</td>
<td>.20</td>
<td>.30</td>
</tr>
</tbody>
</table>

Note: Reported mean classification errors in studies using a parainformative task without feedback.

Additionally, Vigo (2013, Experiment 3) has demonstrated that upon a brief visual inspection, participants intuitively rank the difficulty of these structures in accordance with the empirical ordering in terms of correct classifications. Participants gave judgments of learning difficulty on a scale of 1 through 10 for instances of the $3_2[4]$ structures which were simultaneously presented on each trial. Mean magnitude judgments for each of the six $3_2[4]$ structures were reported as follows: I (1.7) < II (4.3) < [III (5.5), IV (5.6), V (5.5)] < VI (6.3). Vigo (2013) interprets this result as evidence for the capacity to form metajudgments about concept learnability. Further, such metajudgments are accurate in terms of classification performance.
It should be noted that there are instances where the classic ordering was not observed (Love & Markman, 2003; Lewandowsky, 2011; Feldman, 2000, Nosofsky & Palmeri, 1996; Love, 2002; Rehder & Hoffman, 2005) leading Kurtz et al. (2012) to argue that the classic Type II advantage over Type IV is achieved as a result of (i) the particular interplay between stimulus material and dimensional assignment, and (ii) instructional language that places emphasis on “rules” and “features.” However, the alternative orderings are likely due to experimental designs that do not sample from all possible structure instances (Vigo, pers. comm., 2013a, 2014). The learning difficulty of these types is meant to be an aggregated average proportion of performance across all instances of the type and a limited subset of these instances will quite likely result in deviations from the true difficulty ordering.

In summary, classification performance on the six $3_2[4]$ structure types is significant because it provides an empirically validated difficulty ordering by which concept learning models can be compared. These results gain new significance light of recent developments in the way that few models can predict the ordering precisely (Nosofsky, 1984; ALCOVE, Kruschke, 1992; SUSTAIN, Love, Medin, & Gureckis, 2004; GISM-SE, Vigo, 2013) and recent accounts (Feldman, 2000; Vigo, 2006; Goodwin & Johnson-Laird, 2011; Kemp, 2012) do not predict the learning difficulty ordering.

While it is important for models to account for difficulty orderings for well-known category structures (e.g. $3_2[4]$ and 5–4 structures) it is also important that the generality of the model be tested to the extent that it accounts for the classification performance across a large number of category structures. For example, Vigo (2013)
conducted a large-scale category learning experiment using well-defined categories by obtaining the learning difficulty, in terms of percentage of correct object classifications, for 84 category structures across 10 category families. Associated with these 84 structures is a pool of 5100 categorical stimuli that were randomly selected during the study.

The categorical invariance model (CIM, ECIM; Vigo, 2009; 2011a) as well as its generalization and model variants (GIST-NP, GISTM, GISTM-SE; Vigo, 2013) provide the best fit to data from these large-scale classification experiments as well as several other performance benchmarks (see Table 4, Vigo, 2013). These data were also fit to recent structural accounts which make predictions based on a category’s minimal description (MinC; Feldman, 2000; Vigo, 2006), algebraic complexity via spectral decomposition (ACM; Feldman, 2006), and number of mental models (NOMM; Goodwin and Johnson-Laird, 2011). Exemplar – based approaches have been compared unfavorably with respect to these more recent models (Goodwin & Johnson-Laird, 2011; Vigo, 2013; 2014).

Chapter Summary

This chapter provided a summary of the methods used in category learning research. In doing so, we covered the kind of stimuli used as well as the methodological variations found in the literature. To start, we discussed the advantages of using well-defined category structures as the stimuli in category learning tasks. Next, we highlighted a key distinction between stimuli whose dimensions can be selectively attended (separable) and those that cannot (integral). Further, the two general paradigms in the
classification literature were discussed. Parainformative tasks include a learning phase where category members are presented in parallel whereas serioinformative tasks require that category membership be learned during the classification of individually presented stimuli. Additionally, these tasks may or may not use corrective feedback to indicate the accuracy of classification responses per trial.

The significance of the $3_2[4]$ category family was also discussed. Specifically, this category provides an empirically validated difficulty ordering for the six structurally equivalent types. Since models should account for the observed difficulty of the six types, the $3_2[4]$ category family can be used to compare models. Also, since it is additionally important that classification performance be tested across a large number of category structures, we reviewed a large-scale study that tested classification across 84 category structures (Vigo, 2013).

The work discussed in this chapter is particularly relevant in the context of the current project. For instance, the current work tests classification performance on the $3_2[4]$ category structure types in a parainformative task that does not include corrective feedback.
Chapter 3: Concept Representations

Central to the research described above are the theories and models that have been used to describe the form of the concept learning process. Accordingly, in this section we introduce the main representational paradigms to provide a context for the empirical work to follow. More importantly, we introduce these paradigms to distinguish them from the core theory and model that will be the focus of this report.

The classic representational paradigms reviewed in the next section can be considered “process models” of concept learning in that these accounts describe how concepts are acquired in terms of mental processes (i.e., attention, similarity assessment, and memory). Further, exemplar and prototype representations are based on similarity assessment and their associated formal models transform similarity scores into probability scores so that predictions can be made in terms of the probability that any one member of a category will be correctly classified. The ability to make predictions of this kind is a critical component of process models.

In contrast, “structural models” aim to account for classification behavior by measuring the complexity of categorical stimuli. For instance, recent structural accounts (Feldman, 2000, 2006; Vigo, 2006; Goodwin & Johnson-Laird, 2011) make predictions based predominantly on the dimensions of categorical stimuli. Accordingly, while structural models are stimulus-oriented, process models are relatively observer-oriented (Vigo, 2014).
The Classic Representational Paradigms

Concept learning theories and their associated models generally specify three mental phenomena: (i) the structure of the categorical representation, (ii) the process of associating target stimuli to stored representations in memory, and (iii) the process of classification (i.e., placing target stimuli into their respective categories; Kruschke, 2008). In this way, models aim to account for classification performance while suggesting how categorical information might be organized in the mind. Ultimately, however, the predictions of the model as well as the hypothesized organization will depend on the chosen process of association. Models of categorization are typically grouped based on the first of the above criteria—that is—by the structure of the proposed internal knowledge representation. The three paradigms that have historically had the greatest impact on the field are concepts as rules, prototypes and exemplars.

Early theories of category learning had the goal of defining category membership in terms of necessary and sufficient conditions (e.g. Bourne, 1966; Bruner, Goodnow, & Austin, 1956). That is, features are singly necessary in the way that every instance of the concept must have the chosen defining property and jointly sufficient in the way that every entity containing the defining properties must be an instance of the concept. Accordingly, category membership in the rule view (also called the classical or definitional view) is determined by whether or not an object or entity meets the definition of a particular category. The requisite criteria (or membership criteria) that defines the category is based on defining properties of the category members, called the intension of the category. The members of the categories themselves are the extension of the
category. The terms intension and extension are barrowed from mathematical logic and their use highlights the goal of these early pioneers to create a precise and rigorous field of study.

Rather than a common definition, it seems that any instance shares some resemblance to other instances of a given category (Rosch & Mervis, 1975). The idea of family resemblance allows for some variation within a category and suggests that categorization is based on the similarity of an instance to stored representations. A formal description of family resemblance has been given in terms of cue validity (Rosch, Mervis, Gray, & Johnson, 1976) and a set theoretic representation of similarity proposed by Tversky (1977) and Tversky and Gati (1982). The general idea is that the cue validity of cue \( x \) and category \( y \) increases as a result of (i) increased associations between a cue \( x \) and category \( y \) and (ii) increased differentiation from other contrasting categories. The idea is highlighted by Tversky’s (1977) notion of “category resemblance” which is the weighted sum of the common features within a category minus the sum of distinctive features.

The process of assessing similarity is thus emphasized in both the prototype and exemplar approaches. In both views, category learning is explained in terms of attention-weighted similarity to reference points where a novel object is classified based on its similarity to a stored category representation. In prototype models (e.g. Rosch, 1978; Estes, 1986; Nosofsky & Zaki, 2002), a new instance is classified based on its similarity to a prototype, whereas exemplar models (Medin & Schaffer, 1978; Nosofsky, 1984, 1986) propose that such novel instances are classified based on their similarity to
individually stored exemplars. These approaches assume that objects are represented in a multidimensional space where similar items are stored close together in the coordinate space and dissimilar items are far apart in the space.

The prototype view achieves category distinctiveness in the way that concepts consists of features most representative of members and least representative of contrasting category members (Rosch, 1973; 1975). Prototypes can be thought of as abstract representations of a category based on the average or summary representation of many prior experiences. That is, our representation is composed of those dimensional values (i.e. features) that are most frequently encountered in category instances. Many authors interpret prototype representation as the storage of a single prototype (or best example); however, representation in prototype theories should be thought of as a summary representation of the frequency with which features occur in category members.

The exemplar view rejects the notion of a summary representation (via a list of feature frequencies) and instead claims that exemplars, because they are more accessible in memory, play a more central role in categorization (Medin & Schaffer, 1978). A relative lack of an abstraction process is characteristic of the exemplar view in that memories for individually stored exemplars form the representation of a category (Smith & Medin, 1981) rather than an abstract or “ideal” prototype. As such, classification and identification in the exemplar approach depends on the ability of a target stimulus to activate these stored exemplars in memory via a similarity relation between the target and source exemplars in memory.
In summary, prototype and exemplar-based theories propose that concepts are represented in a high-level psychological space. That is, information about a given category is instantiated by a high-level representation that is based on low-level properties in terms of commonly occurring diagnostic features of category members. Objects are classified based on their similarity to stored category representations. Over time and due to experience and/or training, there can be changes in attention allocation to diagnostic dimensions but classification will always occur by comparing an object to a high-level representation in psychological space.

**Invariance-Based Representations in Ideotype Space**

In the Generalized Invariance Structure Theory (GIST, Vigo, 2013; 2014), the nature of concept representation is significantly different than the traditional proposals given above. In GIST, “invariance detection” is the specific kind of pattern detection necessary for concept formation. From the principles on which GIST is based, Vigo derives a candidate law of invariance for conceptual behavior. The fully parameterized variant of the law (also known as the generalized invariance structure model or GISTM) is expressed in Equation 1 below:

$$\psi(X) = pe^{-k \Phi^2_{a_j}(X)}$$  \hspace{1cm} (1)

where $\psi$ is the degree of perceived learning difficulty of a continuous or dichotomous category $(X)$, $p$ is the cardinality or size of the categorical stimulus, and $\Phi$ is the degree of perceived categorical invariance. The parameterized version includes a discrimination
parameter $k$ and a sensitivity parameter $\alpha_d$. For a proper explanation of the formal details of the model please refer to Vigo (2013; 2014). For our purposes, however, we will focus on the parameters of the model which reveal changes in the structure of representation across sessions of learning.

Figure 3 summarizes the process of detecting invariants using a simple category structure ($\textit{small black triangle}; \textit{small black circle}; \textit{large white circle}$) consisting of three objects and three dimensions. In the original structural account of the model (Vigo, 2009), an operator generates the degree of partial invariance by perturbing dimensions of categorical stimuli. These perturbations are dimensional transformations that determine the number of invariants per feature. The number of invariants per dimension equals the number of common objects between the original and perturbed categories. Thus, upon the shape transformation in Figure 3 we see that the $\textit{small black circle}$ and the $\textit{small black triangle}$ remain after perturbation. Upon the color and size transformations, however, no objects are common to the original and perturbed sets.
Figure 3. Structural manifold transformations across shape, color, and size for a simple category structure. Determining the structural kernels (SKs) and structural manifold is a straightforward process after accurate identification of invariants (shown in the red squares).

The differential operator is given cognitive meaning as an “invariance detection operator” in Vigo’s (2013) generalization of the model. The invariance detection operator generates one structural kernel (SK) per dimension where SKs are the proportion of invariants to the total number of objects in the category. The structural manifold of the category is found by dividing the number of invariants by the total number of objects given by the SKs per dimension. Structural manifolds are quantitative descriptions of the core representation in GIST: namely, the “ideotype.” In GIST, ideotypes encode the invariance structure information of categories.
In general, the process of SK detection determines how essential a given dimension is in terms of characterizing category membership. Simply, objects either remain or are eliminated after a perturbation. Dimensions with a greater number of eliminated objects after perturbation are more essential for determining category membership. Alternatively, dimensions with a greater number of objects that remain after perturbation are relatively non-essential for determining category membership. Therefore, the structural manifold obtained in our Figure 3 example indicates that color and size are essential for classification, whereas shape is relatively non-essential. Underlying this process of invariance detection are the processes of rapid attention shifting and dimensional binding that result in “partial similarity” assessment between category members. Partial similarity assessment takes place for each dimension resulting in the implicit detection of invariance patterns.

A version of the GISTM without free parameters (GISTM-NP; Vigo, 2013; 2014) uses the deterministic quantitative index $D_0/D$ to characterize this discrimination capacity. $D_0$ stands for the minimum number of dimensions needed to non-trivially describe a category of objects ($D_0 = 2$). $D$ is the number of dimensions that define the categorical stimulus.

GISTM can incorporate a distance scaling parameter $k$ which serves to account for individual differences regarding the ability to extract invariance patterns. That is, by indicating the overall degree of discrimination between differing category structures, $k$ serves as a summary index which reveals how discrimination at the category level contributes to classification performance. Technically, $k$ indicates the overall degree of
discriminability between the ideotypes and the standard ideotype represented by the 0
point in the coordinate space. Larger $k$ values indicate better discrimination and relatively
higher discriminability between distances means relatively higher sensitivity to
invariance patterns.

The ability of the scaling parameter $k$ to reflect group differences is well
demonstrated in recent work reporting differences in classification performance between
populations. Vigo, Evans, and Owens (2014) tested college aged adults, non-ADHD
adolescents, and adolescents with ADHD on a parainformative classification task without
feedback. Participants were tested on four random instances of each of the six structure
types in the $3_2[4]$ family. Error rates for the $3_2(4)$ category structures were highest for the
ADHD participants, followed by the non-ADHD and adult participants while the classic
difficulty ordering for these structures was observed for the adults and non-ADHD
groups but absent for ADHD participants. Of particular interest for our purposes are the
estimated values of $k$ for each group which were reported as follows: adults ($k = .76$);
non-ADHD adolescents ($k = .54$); adolescents with ADHD ($k = .29$). These $k$ values
might be interpreted to indicate relative degrees of overall discrimination, where higher
values represent better overall discrimination.

Sensitivity to structural kernels (SKs) can also vary and such variations can be
captured by the law of invariance with the inclusion of SK sensitivity weights that act as
invariance pattern detection weights corresponding to dimension $d$ (one parameter per
SK, i.e., per dimension). These sensitivity parameters are henceforth denoted $\alpha_d$, where
$\alpha$ is the obtained sensitivity weight corresponding to dimension $d$. For the sake of
exposition, these SK sensitivity weights may also be referred to as “alpha values” or “alpha parameters.” These sensitivity weights reflect relative variability (due to noise, limitation of resources, and other factors) in the number of invariants detected by an observer. These values reflect the mediation of ideotype formation (Vigo, 2013; 2014). Ideotypes contain the essential structural patterns inherent to a categorical stimulus and are described quantitatively by structural manifolds.

The core representation in GIST is the ideotype, which encodes the invariance structure information of categories/concepts. Ideotypes are described quantitatively in terms of the structural manifold. GIST posits that invariance detection underlies all other conceptual representations, including representations in terms of rules and prototypes. Vigo (2014) suggests that ideotypes play a critical role when the conceptual system is determining which type of representation to use. Vigo proposes that when an ideotype has mostly SK values of 0 or 1, the resulting representation is rule-based. Alternatively, the resulting representation is a prototype when an ideotype has mostly fractional SK values. In this way, ideotypes may be critical in the development of a unified theory of conceptual representation.

Figure 4 shows the Cartesian coordinate representation of ideotypes corresponding to the $3_2[4]$ category structures graphed in ideotype space. The ideotype points per type in Figure 4 are based on the objective learning difficulty per type as predicted by GISTM-NP.
Figure 4. Cartesian coordinates representation of ideotypes corresponding to the $3 \times [4]$ category structures graphed in ideotype space based on the objective learning difficulty per type as predicted by GISTM-NP.

Chapter Summary

In summary, it has been shown that the nature of the representation proposed in GIST is altogether different than prototype and exemplar-based theories. Whereas points in psychological space consist of actual category members (or a summary representation) in traditional theories, points in ideotype space reflect the structural manifold that is based on the detection of invariants. Further, we discussed model parameters that will become particularly important in the context of the current study. For instance, GISTM can incorporate a scaling parameter $k$ which serves to reflect individual differences.
regarding the ability to extract invariance patterns. That is, $k$ may serves as a summary index which reveals how discrimination at the category level contributes to classification performance. The model can also include sensitivity weights which reflect relative variability in the number of invariants detected by an observer. Preliminary results indicate that this model accurately identifies processing limitations that lead to individual differences in concept acquisition. It should also be noted that the framework presented above has since been used as the basis for a deterministic mathematical framework for the measurement of information (Vigo, 2011b; 2012).

Next, we review empirical work that has explored changes in conceptual representation and processing fluency that occur as a result of experience-based learning. Additionally, we discuss empirical phenomena common to perceptual expertise as it relates to the process of learning over time. Indeed, the intended contribution of the empirical and modeling sections of this paper is to interpret changes in representation that occur during learning in the context of the GISTM.
Chapter 4: Perceptual Fluency and Expertise during Learning

We first review what is known about how expert knowledge is represented. Following this, I discuss successful research approaches to expertise and perceptual fluency in areas of perceptual research. This research is meant to frame the current approach to characterizing the underlying processes of perceptual expertise during concept learning. Further, we discuss empirical and theoretical work on reaction times (RTs) during learning. Specifically, we discuss how RTs can be particularly revealing with respect to the underlying processes during category learning.

Expert Knowledge

The study of expertise as it applies to concept learning has largely been confined to how domain-specific knowledge is organized within conceptual hierarchies. For instance, knowledge in terms of conceptual representations has been shown to be organized in a particular hierarchical structure (Collins & Quillian, 1969; Quillian, 1969) and in addition to having more concepts (represented by nodes in the hierarchy) experts have more connections between these concepts than do novices (Bordage & Zacks, 1984). In this way, superior breadth and complexity in the organization and structure of conceptual representations allows the individual to obtain increasing levels of expertise over time and with increasing ease (Hambrick, 2003). Indeed, the way knowledge is organized determines ease of retrieval. Furthermore, the greater the amount of information to be searched through (as in highly complex expert domains), the more important it is that this information be organized in an efficient manner. In short, experts
not only have a large knowledge-base but the structure of such knowledge is also highly organized.

Accordingly, this more robust and complex knowledge base allows for efficient performance within the domain because of an increased ability to recognize patterns. Consequences of this increase in pattern recognition are the quick perceptions (perceptual fluency) and rapid responses (response times) that are the hallmark of expert performance. As such, many researchers view pattern recognition as the primary mechanism of expert performance (e.g. Chase & Simon, 1973; Gobert & Simon, 1996). Additionally, creating better representations leads to understanding the task domain at a deeper level (Glaser & Chi, 1988).

The core idea of the hierarchy is that categories are represented as nodes in a tree-like structure which contains superordinate and subordinate categories. As one moves up the hierarchy there are progressively larger superordinate categories that subsume (via a set inclusion relation) the smaller subordinate categories below. As such, there is greater category inclusiveness and a higher level of abstraction at higher levels within the taxonomy, where categories are related by means of class inclusion. Nodes are connected by links in the hierarchy which designate relations between concepts. Links include “is–a” and “has–part” relations between nodes (i.e. a bird “is–a” animal; a bird “has–part” wings). Further, the links are directed in the way that they either move between levels of the hierarchy (vertically) or stay within a given level (horizontally).

The different levels of abstraction in the hierarchy can be described more formally in terms of cue validity (Rosch, Mervis, Gray, & Johnson, 1976). The general idea is that
the cue validity of cue $x$ and category $y$ increases as a result of (i) increased associations between a cue $x$ within and category $y$ and (ii) increased differentiation from other contrasting categories. Basic–level category members maximize cue validity (Rosch, Mervis, Gray, & Johnson, 1976) over and above subordinate and superordinate category members. That is, basic level category members have many within category associations in that many features are shared and are highly differentiable across categories in that they share few features with members of contrasting categories. Conversely, superordinate categories share few features with one other (there are fewer within category associations) whereas subordinate categories share many features with contrasting categories (there is less across category differentiation).

Numerous studies show evidence for the preferential processing of basic–level categories. For instance, basic – level category labels are given nearly exclusively in free recall tasks (Rosch, Mervis, Gray, & Johnson, 1976) and classification RTs are faster for members of basic-level category labels (Rosch, Simpson, & Miller, 1976).

However, research has shown the influence of training and practice on category knowledge. For instance, studies comparing novice and expert knowledge show differential processing within the domain of expertise (Jolicoeur, Gluck, & Kosslyn, 1984; Tanaka & Taylor, 1991. For novices, classification RTs at the basic – level are faster than classification RTs at superordinate and subordinate levels. Accordingly, the “entry-level” in most knowledge domains is marked by faster classification RTs at the basic-level. For experts, classification RTs are equally fast at basic and subordinate levels (Johnson & Mervis, 1997; Tanaka & Taylor, 1991). This “entry level shift” occurs over
the course of learning and seems to reflect gradual processing efficiency within a domain (Joyce & Cottrell, 2004; Mack, Wong, Gauthier, Tanaka, & Palmeri, 2007; Tong, Joyce, & Cottrell, 2008).

Consider an example of how experts and novices differ with respect to their use of superordinate, basic, and subordinate categories. Tanaka and Taylor (1991) compiled a group of expert birdwatchers and a group of dog experts. Both groups were tested on both bird and dog categories so that novice performance was assumed from birdwatchers’ categorization in the dog domain (and dog experts in bird domain) and expert performance from birdwatchers in the bird domain (and dog experts in the dog domain). When asked to list features of categories in Experiment 1, novices listed more features at the basic level than at subordinate levels. However, participants in their domain of expertise did the opposite by listing features predominantly at subordinate levels. Experiment 2 used a free-naming task where participants were asked to provide a name to randomly presented pictures. Subjects provided basic-level names in their novice domain and were more likely to use subordinate-level names in their domain of expertise. In Experiment 3, a timed categorization task presented subjects with category names followed by a picture and asked them to indicate (as fast as possible) if the names matched the picture. Whereas average RTs were faster for the basic level than the subordinate level in novice domains, expert domains showed no difference between average RTs of basic and subordinate levels. Taken together, these three experiments by Tanaka & Taylor (1991) show an ease of processing effects as a result of domain-specific expertise.
Presumably, experience in the form of repeated exposure to category members creates a shift in processing resulting in efficient recognition and fluency at the subordinate level.

In all, empirical research in this area has shown that experts’ subordinate-level categories in the domain of expertise are represented much like basic-level categories in regular knowledge domains.

**Expertise and Perceptual Fluency**

The role of experience and expertise has been explored extensively in areas such as visual object processing, perceptual learning, and face processing (Gauthier, Tarr, & Bub, 2010). The common thread of these research areas is that they are concerned with the underlying processes of perceptual expertise (Palmeri & Cottrell, 2009). That is, researchers in these areas view empirical phenomena as domains of perceptual expertise and approach these phenomena accordingly by examining the effects of training and experience. Perceptual expertise in these areas is concerned with the adaptive and experience–induced processing changes in the extraction of information (Kellman & Garrigan, 2008). Accordingly, humans have shown a remarkable ability to change the way we extract information so that we may optimize performance in a domain. Experts seem to do so by developing specialized perceptual tools for analyzing stimuli in their domain of expertise. Experience through training involves developing both explicit strategies and implicit perceptual processes to more efficiently represent the structure of a given domain.
Accordingly, a distinction has been made between discovery and fluency in the course of perceptual learning (Kellman, 2002). Perceptual learning is said to involve discovery -- in the way that new task-relevant information is detected while irrelevant information is suppressed -- and fluency in the way that the detection of task-relevant information becomes automatic. In short, practice allows discovered information to become fluent and less demanding. Experience and training therefore involves discovery in the sense that new task-relevant information is learned and fluency in that the detection of task-relevant information becomes automatic. Generally, discovery effects are measured in terms of changes in response accuracies over time whereas shorter response times indicate fluency. As I mention in Chapter 1, the acquisition of expertise will be understood as the typical outcome of the learning process. In this context, observed decreases in classification errors and classification response times can be interpreted as discovery and fluency effects, respectively.

**Response Times and Practice**

In general, the time needed to perform a task decreases with practice. That is, the development of expertise is marked by increases in processing fluency in terms of response times (RTs). Specifically, studies have shown that additional practice results in RT reductions that are best characterized by a power function (Newell & Rosenbloom, 1981). The mean RT on block $N$ is characterized by the general form of the power law shown below:

\[ RT(N) = a + bN^{-c} \]  

(2)
where \( a \) is the asymptote or the limit on speed of performance for a given task, \( b \) is the difference between the initial RT upon first presentation of the task (i.e. trial/block one) and the asymptotic RT (i.e. \( a \)). \( N \) is a value for the amount of practice and is most often given by the corresponding block number in an experimental session. The learning rate parameter is given by \( c \), which captures the rate with which RTs diminish over trials/blocks.

Early research showed that power functions provide strong fits to RT data from numerous tasks where a learning effect had been shown to take place (Newell & Rosenbloom, 1981). As such, the general relationship between RTs and training became known as the power law of practice. However, numerous studies have since reported instances where decreases in RTs over time are best accounted for by exponential function curves (e.g. Delaney, Reder, Staszewski, & Ritter, 1998; Palmeri, 1999; Heathcote, Brown, & Mewhort, 2000). The exponential function characterizes the average response time (RT) on block \( N \) as follows:

\[
RT(N) = a + b \cdot e^{-c \cdot N}
\]  

(3)

where, as with the power function, \( a \) represents the asymptote, \( b \) the difference between the RT on trial one and the asymptotic RT, and \( N \) is the corresponding experiment block. Again, the learning rate parameter \( c \) is meant to capture the rate with which RTs diminish over trials/blocks of an experiment.
An example will help to show the difference between these two functions as well as illustrate their differential predictions with respect to RTs. Let’s say we have data for an experiment where participants completed a task over 25 trials. Suppose the initial trial yields a mean RT of 100 seconds and the power function asymptotes at 5 seconds (thus $a = 5, b = .95$). Further, we’ll say that our estimated learning rate parameter falls between 0 and 1, since this is typically the case (see Newell & Rosenbloom, 1981). Both power and exponential functions for these parameters are plotted in Figure 5. Functions are plotted using two values of $c$ (.25 and .75) to demonstrate the impact of this learning rate parameter. The power function predicts more substantial reductions in RTs early in learning with diminishing marginal reductions resulting from additional practice. Alternatively, the exponential function predicts that RTs decrease at a constant rate in relation to what is yet to be learned. In both cases, higher values of the learning rate parameter $c$ result in steeper RT reductions. Both functions predict diminishing marginal reductions over the course of extended practice trials.

When data of individual participants are considered (rather than aggregated means across subjects) exponential functions appear to provide superior fits (Heathcote, Brown, & Mewhort, 2000). Additionally, consistent with component power laws Rickard (1999) found that mean RTs did not decrease as a pure power law of practice for numerosity judgment data. The current work aims to add to this line of research by fitting these functions on classification RTs during a concept learning task.
Figure 5. Power and exponential curves for two values of the learning rate parameter. In order to show the impact of the learning rate parameter \((c)\), power and exponential function curves are plotted using two values of \(c\) (.25 and .75).

Classification Response Times and Practice

Few studies have explored the relationship between classification response times (RTs) during a category learning task and the role of experience. The work reviewed in this section tested classification learning over time by testing participants on multiple experimental sessions. Taken together, this research suggests that with experience people develop perceptual expertise and classification responses become fast and automatic.

Nosofsky and Palmeri (1997) report speedups in classification RTs that occur as a function of practice on a classification task and show that a power function accurately fits RTs across sessions. Participants completed sessions of a speeded classification task in which participants are instructed to respond as quickly and as accurately as possible.
Sessions of the experiment were conducted on five consecutive days. Subjects completed a serioinformative classification task with corrective feedback where categories were composed of integral dimension stimuli (see Chapter 2 for a review). Specifically, the stimuli consisted of twelve color patches that all had a consistent red hue but varied in their degree of brightness and saturation. Within each block, six of the patches belonged to category A and six belonged to category B.

The reported data from a single participant are shown in Figure 6. Each point represents the mean RT for 60 consecutive trials. These grouped blocks are plotted as a function of block order and the resulting curve is well described by the power-law function \( r = .97 \). Nosofsky and Palmeri reported an initial mean RT of 708 milliseconds (ms), a difference of 784ms between first RT and asymptote (thus \( a = 708, b = 784 \)), and an estimated learning rate of .79. Subsequent work by Nosofsky obtained additional data using similar experimental designs (Nosofsky & Stanton, 2005; Nosofsky, Little, Donkin, & Fific, 2011).

Nosofsky and Palmeri explain the reported decreases in classification RTs using exemplar theory as well as a theory of automaticity. In short, exposure to additional exemplars over the course training results in increasing amounts of exemplar information stored in memory. This leads to better exemplar retrieval (thus better performance accuracy) and quicker exemplar retrieval over the course of training (thus faster RTs).
Figure 6. Power function fit to observed response times reported by Nosofsky and Palmeri (1997). Partial reproduction of Figure 5 from Nosofsky & Palmeri (1997). Mean response time (ms) for each grouped block of 60 trials for a single participant of Experiment 1 in Nosofsky & Palmeri (1997). Also shown is the best-fitting power function with the observed data.
Palmeri (1997) instructed participants to use an explicit rule in a dot-counting categorization task. On each trial, there were between six and eleven dots in each pattern and participants were asked to categorize the patterns based on numerosity. Subjects did so by pressing the number key corresponding to the number of dots in the pattern. The task was speeded, meaning that participants were instructed to respond as quickly as possible while minimizing errors. Performance was recorded across thirteen experimental sessions taking place within a single day. Results suggest that participants reached perceptual expertise on the task. For instance, there was a positive linear relationship between the number of dots and classification RTs. This relationship disappeared over the course of trials such that mean classification RTs did not vary by stimulus numerosity in later trials. Classification RTs in early trials seem to be reflection of explicit dot counting. The experience gained across sessions resulted in familiarity with the dot patterns. Repeated exposure to these patterns made counting unnecessary in later trials and numerosity was identified on the basis of perceiving a familiar pattern.

A subsequent transfer test presented participants with new test patterns along with old patterns. New patterns showed classification RTs that were similar to those of Session 1, where RTs were a linear function of numerosity. Consistent with previous trials, RTs on old training patterns were not influenced by numerosity.

The current work aims to add to this line of research. However, the current experimental design differs significantly in a number of ways. First, the current experiments use categorical stimuli with separable dimensions (rather than integral) in a parainformative (rather than seriinformative) concept learning task. These differences
provide considerable advantages over previous methods. The use of separable dimensions allow for generalizability in the form of straightforward comparisons with real world stimuli which may lead to applications in naturalistic environments. Direct applications are limited with integral dimensions given that it is hard to construct an integral dimension other than color.

**Chapter Summary**

In summary, this chapter reviewed work on expertise and how the concept learning literature might benefit from looking at response times in the same as way as more perceptual areas of research. Specifically, we discussed the role of discovery and fluency during learning and how power and exponential functions account for reductions in response times throughout learning. The current work contributes to this area of research by testing exponential and power functions on classification RTs during a category learning task.

Fitting our data on such curves is beneficial in that it reveals how underlying cognitive processes unfold over different time-scales during concept acquisition. Additionally, such fits may serve to compare competing learning theories that predict certain rates of RT decay. Further, such information may have practical implications for the implementation of job training programs in areas reliant on expert classification (e.g. medical diagnosis, radiology, air traffic control). As such, we fit both functions to mean RTs over the course of the three sessions.
Chapter 5: Overview of the Current Empirical Work

As reviewed above, robust learning difficulty orderings for category structures are of particular importance in concept learning research. Knowing the relative difficulty between certain category structures tells us which relational structures humans are most sensitive. However, it is rare that the stability of such orderings is tested over a period of extended learning. Further, this research rarely explores dependent variables beyond classification accuracy which may also indicate relative learning difficulty.

Like conventional work in category learning (Chapter 2), the current work uses a category with a robust learning difficulty ordering but tests these category structures over multiple experimental sessions in a way that is similar to work on perceptual expertise (Chapter 4). To the best of the author’s knowledge, the work discussed here is the first to test categorical stimuli over multiple experimental sessions.

Study 1 and Study 2 both test $3^2[4]$ category structures so that the learning difficulty ordering can be assessed across sessions. Further, both studies involve category learning tasks in which classification errors and response times (RTs) are obtained over three experimental sessions taking place on three consecutive days. This procedure is motivated by the design of previous research exploring the influence of experience/training on category learning. Precedence is shown for testing participants on consecutive days (Nosofsky & Palmeri, 1997; Nosofsky and Stanton, 2005; Nosofsky, Little, Donkin, & Fific, 2011). For instance, Nosofsky and Palmeri (1997) and Nosofsky et al. (2011) both tested participants over five consecutive days. Such studies offer benefits over experiments that use multiple sessions on a single day (Palmeri, 1997) in
that they better resemble real-world experiential learning which typically requires repeated exposure to domain specific stimuli (Myles-Worsley, Johnston, & Simons, 1988). However, a three day span also imposes inferential limitations in that the total learning time is still relatively short compared to real-world, experience-based learning. For instance, professionals who aim to master real-world tasks typically develop domain expertise across a much longer time span, often taking years (Ericsson, Krampe, & Tesch-Roemer, 1993). Previous studies in this area vary with respect to the spacing of experimental sessions, which range from a single day (Palmeri, 1997) to over twenty days (Nosofsky, Little, Donkin, & Fific, 2011). The use of many experimental sessions in previous work is largely motivated by a need for data across many blocks, i.e. enough to adequately test model predictions (see Nosofsky & Palmeri, 2015), rather than any specific learning criteria. Accordingly, three sessions tested over three days seems to represent an adequate time span for the purpose of the current work. However, systematic research on the influence of such learning session variables on concept learning is lacking in the literature and should be the goal of future research. For instance, studies could vary the total number of sessions, time intervals between sessions, blocks per session, and/or the total amount of time spent learning in terms of days/blocks.

In Study 1, each participant is exposed to a single structure type of the 3\_2[4] category family. The obtained results reveal an interesting relationship between accuracy and fluency throughout training. In terms of classification accuracy, the difficulty ordering is revealed in the first session and disappears in later sessions. In terms of RTs, the difficulty ordering is not originally present in the first session but is revealed in later
sessions. This suggests that relative concept learning difficulty is best revealed by RTs after the observer has significant experience classifying category members. As far as I know, this is the first empirical work that has explored the relationship between learning difficulty orderings of categorical stimuli and the corresponding classification RTs over multiple training sessions.

In Study 2, each participant is exposed to all six structure types of the $3_2[4]$ category family. In this case, the learning difficulty ordering remains stable over multiple experimental sessions, both in terms of proportion of errors and RTs. Testing multiple category structures per person allows us to estimate parameters at an individual level. Further, the obtained results are compared to those of Study 1 to see if the relationship between errors, RTs, and experience varies depending on the number of category structures the participant must learn. There are several possible implications that might emerge as a result of this comparison. These implications and applications are discussed in Chapter 9.

The results of both studies are interpreted in the context of GIST (Vigo, 2013; 2014) which has not yet been used to account for classification learning over multiple sessions. The model is fit to observed data across sessions in Study 1 and Study 2. Further, the results of both studies are interpreted using the parameterized model variant (Equation 1).

GIST makes several qualitative predictions. One of the basic principles underlying the theory, known as the invariance-parsimony principle, states that observers place disproportionate emphasis on fully diagnostic and fully redundant dimensions, i.e.,
dimensions whose corresponding structural kernel values are 0 and 1. The bias is captured in the model by squaring the degree of invariance, which serves to amplify the contribution of these kernels. Accordingly, the model predicts that estimated alpha values will be highest for 0-valued and 1-valued kernels. Support for this prediction was reported by Vigo (2013; 2014). Table 2 contains the average error rates per structure as reported in Experiment 1 of Vigo (2013). The structural manifold (SM) for each type is shown alongside the estimated $\alpha_d$ values per dimension. Estimated $\alpha_d$ values are highest for 0-valued SKs, yielding 1.00 in each instance. Relatively high $\alpha_d$ values were obtained for 1-valued SKs, yielding 0.90, 0.85, and 0.89 in Types 1 and 2.

Table 2

<table>
<thead>
<tr>
<th>Error</th>
<th>SM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>0.02</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>0.07</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>0.19</td>
<td>(.5, .5, .5)</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>0.18</td>
<td>(.5, .5, .5)</td>
<td>0.46</td>
<td>0.47</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>0.17</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>0.83</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>0.32</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Estimated alpha values per dimension ($\alpha_1$, $\alpha_2$, $\alpha_3$) for each of the six 3_2[4] structure types using data from Experiment 1 of Vigo (2013). Mean proportion of errors (Error) and structural manifolds (SM) are presented for each type. SKs with values of 0 and 1 along with the corresponding alphas are presented in red. This table is adapted from Table 7.5 in Vigo (2014).

These alpha parameter values suggest that observers are sensitive to the high occurrence of redundant patterns in category structures (1-valued SKs) as well as the complete absence of such patterns (0-valued SKs; $\alpha_d = 1.00$).
The invariance-parsimony principle is tested further in the current work. Studies 1 and 2 allow for a similar comparison between SK values and estimated alpha values. However, the current work provides a more demanding test the principle because it is possible to generate structural manifolds and alphas across all three sessions of each experiment. Alpha ($\alpha_d$) values are also estimated using obtained classification RT data per structure type.

Additionally, a principle of invariance detection underlying GIST is particularly relevant in context of the current studies. The structural equilibrium (SE) principle states that categorical stimuli with increasing degrees of SE results in easier identification of dimensions needed for rule formation. The degree of structural equilibrium $\lambda$ associated with the ideotype of a categorical stimulus $X$ is defined in terms of the proportion of 0-valued SK values in the structural manifold (plus one to avoid division by zero). Accordingly, $\lambda(X) \approx 1.33$ for structure Type V because its structural manifold is $(0, .5, .5)$ which results in a proportion of $1/3$ since one out of three SKs are 0-valued. The degree of structural equilibrium for each of the $3^2[4]$ structure types are shown in Table 3.
Table 3

<table>
<thead>
<tr>
<th>SM</th>
<th>Prop.</th>
<th>λ</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>1/3</td>
<td>1.33</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>2/3</td>
<td>1.66</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>0/3</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>0/3</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>1/3</td>
<td>1.33</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>3/3</td>
<td>2.00</td>
</tr>
</tbody>
</table>

Note: The structural manifold (SM), proportion (Prop.) of 0-valued kernels, and degree of structural equilibrium (λ) are given for each of the 3[4] structure types.
Chapter 6: Repeated Exposure to a Single Structure Type

Method

Stimuli and materials. Stimuli consisted of flasks that varied on the binary features of size (large or small), shape (circular or triangular), and color (black or white) and in accordance with the structure types associated with the $3_2[4]$ category family. Participants were tested on instances associated with one of the six category structures of the $3_2[4]$ family. An HP XW4600 workstation with a Dell 1708FP 15 in. flat panel LCD monitor (5 ms response time) was used to display the stimuli.

Participants. A total of 180 participants were tested on one of the six category structures ($n = 30$ per structure) described above. Participants completed three sessions on three consecutive days, resulting in a total of 540 experimental sessions. All participants were undergraduate students enrolled in an introductory psychology course at Ohio University.

Procedure. Participants engaged in a classification task over the course of three experimental sessions on three consecutive days. The classification task was unsupervised (without feedback) and parainformative in nature (Vigo, 2013; 2014), meaning that category members were presented in a learning phase prior to classification. The learning phase consisted of two sets of bottles spatially separated by a line in the middle of the screen. The four category members were presented above the line while the four non-members were presented below the line and the locations of members and non-members were randomized across trials. Before the experiment began, participants were told that each display represented the preferences of a bottle collector and that their task.
was to learn which bottles the collector likes. Participants were told that the four “liked” bottles were presented above a line on the screen and the four “not liked” bottles were presented below the line on the screen. After 20 s of studying the category members and non-members in the learning phase, individual flasks were serially presented for classification. Subjects were told that their task was to classify each individual bottle as either “liked” or “not liked” using the mouse: a left mouse click (labelled “Y”) meant the bottle was “liked” and a right mouse click (labelled “N”) meant the bottle was “not liked” by the bottle collector. After classification responses had been given for the eight bottles, a fixation cross appeared in the center of the screen and the next category block was shown. Participants completed three sessions on three consecutive days. Twenty-four blocks per session resulting in an experimental session lasting approximately 20 minutes each. Three twenty-four block sessions resulted in a total of seventy-two total blocks per subject. Structure instances, as well as the eight objects presented for categorization, were presented to participants randomly according to a counterbalanced design throughout the experimental session.

Classification accuracy and response times were recorded on each experimental trial. Response times were excluded from the analysis if \( RT < 0.025 \) seconds. Average response times per block were excluded if more than half of the trial RTs were missing. That is, at least four of the eight RTs per block were needed to be included in the analysis.
Results

Classification errors. Mean proportion errors per session for each of the six structure types are shown in Figure 7A. A repeated measures ANOVA on the obtained proportion of errors using session as a within-subjects variable and structure type as a between subjects variable revealed a significant main effect for both session, $F(2, 348) = 183.83, p < .001, \eta^2_p = .514$, and structure type, $F(5, 174) = 21.48, p < .001, \eta^2_p = .382$.

Additionally, there was a significant interaction between session and type, $F(10, 348) = 10.22, p < .001, \eta^2_p = .227$. In Session 1, pairwise $t$-tests show that the mean proportion errors per type conform to the classic difficulty ordering, $I < II < [III, IV, V] < VI$. Pairwise $t$-tests on the data from Sessions 2 and 3 reveal an alternative difficulty ordering in terms of classification errors, $I < II < [III, IV, VI, V]$.

Paired sample $t$-tests reveal significant reductions in mean errors between Session 1 and Session 2 for all structure types. Between Session 2 and Session 3, paired sample $t$-tests reveal significant reductions in mean errors for all types accept Type I, $t(29) = 0.626, p = 0.536$, and Type II, $t(29) = 1.18, p = 0.249$. Difference scores for errors between Sessions 1 and 2 as well as between Sessions 2 and 3 are plotted in Figure 8A for each type. The gray bars show the mean difference between proportion errors on Session 1 and Session 2 (S1-S2) whereas the white bars show the mean difference between errors on Session 2 and Session 3 (S2-S3). Paired sample $t$-tests reveal significant reductions between the two pairs of difference scores for Types I, II, IV, and VI, marginal significance for Type V, $t(29) = 1.88, p = 0.07$, and a non-significant reduction for Type III, $t(29) = 0.73, p = 0.44$. 
Further, learning curves are plotted as a function of Epoch in Figure 9A and the associated mean error values per epoch are reported in Table 4. Epochs represent the average of six successive blocks. As such, there are four Epochs per session and twelve total Epochs across all three experimental sessions.

Classification response times. Mean response times (RTs) per session for each of the six structure types are shown in Figure 7B. A repeated measures ANOVA on the obtained RTs using session as a within-subjects variable and structure type as a between subjects variable revealed a significant main effect for both session, $F(2, 348) = 104.66, p < .001, \eta_p^2 = .376$, and structure type, $F(5, 174) = 62.4, p < .001, \eta_p^2 = .642$. Additionally, there was a significant interaction between session and type, $F(10, 348) = 2.56, p = .009, \eta_p^2 = .069$. Further, RTs are plotted as a function of Epoch in Figure 9B and the associated RT values per epoch are reported in Table 4.

Paired sample $t$-tests reveal significant reductions in mean RTs between Session 1 and Session 2 for all structure types. Between Session 2 and Session 3, paired sample $t$-tests reveal significant reductions in mean RTs for all types accept type I, $t(29) = 1.70, p = 0.543$, and type III, $t(29) = 1.70, p = 0.646$. Difference scores for RTs between Sessions 1 and 2 as well as between Sessions 2 and 3 are plotted in Figure 8B for each type. The gray bars show the mean difference between RTs on Session 1 and Session 2 (S1-S2) whereas the white bars show the mean difference between RTs on Session 2 and Session 3 (S2-S3). Paired sample $t$-tests reveal significant reductions between the two pairs of difference scores for Types I, II, IV, and VI, marginal significance for Type V, $t(29) = 1.88, p = 0.07$, and a non-significant reduction for Type III, $t(29) = 0.73, p = 0.44$.  

Figure 7. (A) Mean proportion of classification errors for each of the six structure types within the 3_2[4] category family across all three sessions of Study 1. Mean classification error values and standard error bars are presented for each column. (B) Mean response times (RTs) for each of the six structure types within the 3_2[4] category family across all three sessions of the current experiment. Mean RT values (in milliseconds) and standard error bars are presented for each column. The roman numerals indicate the learning difficulty ordering according to pairwise significance tests.
Figure 8. Mean difference scores for proportion errors (A) and classification response times (B) between Session 1 and Session 2 (grey bars, S1-S2) and Session 2 and Session 3 (white bars, S2-S3) of Study 1.
<table>
<thead>
<tr>
<th>Epoch</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.06</td>
<td>.18</td>
<td>.21</td>
<td>.27</td>
<td>.25</td>
<td>.39</td>
</tr>
<tr>
<td>2</td>
<td>.02</td>
<td>.08</td>
<td>.18</td>
<td>.17</td>
<td>.21</td>
<td>.31</td>
</tr>
<tr>
<td>3</td>
<td>.01</td>
<td>.05</td>
<td>.17</td>
<td>.19</td>
<td>.19</td>
<td>.25</td>
</tr>
<tr>
<td>4</td>
<td>.01</td>
<td>.04</td>
<td>.14</td>
<td>.16</td>
<td>.19</td>
<td>.18</td>
</tr>
<tr>
<td>5</td>
<td>.01</td>
<td>.04</td>
<td>.11</td>
<td>.14</td>
<td>.14</td>
<td>.15</td>
</tr>
<tr>
<td>6</td>
<td>.00</td>
<td>.03</td>
<td>.15</td>
<td>.10</td>
<td>.15</td>
<td>.13</td>
</tr>
<tr>
<td>7</td>
<td>.00</td>
<td>.03</td>
<td>.12</td>
<td>.14</td>
<td>.13</td>
<td>.13</td>
</tr>
<tr>
<td>8</td>
<td>.00</td>
<td>.03</td>
<td>.11</td>
<td>.13</td>
<td>.12</td>
<td>.13</td>
</tr>
<tr>
<td>9</td>
<td>.01</td>
<td>.02</td>
<td>.07</td>
<td>.08</td>
<td>.09</td>
<td>.11</td>
</tr>
<tr>
<td>10</td>
<td>.00</td>
<td>.03</td>
<td>.08</td>
<td>.10</td>
<td>.09</td>
<td>.10</td>
</tr>
<tr>
<td>11</td>
<td>.00</td>
<td>.03</td>
<td>.08</td>
<td>.10</td>
<td>.10</td>
<td>.09</td>
</tr>
<tr>
<td>12</td>
<td>.00</td>
<td>.02</td>
<td>.09</td>
<td>.10</td>
<td>.11</td>
<td>.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Epoch</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>497</td>
<td>948</td>
<td>943</td>
<td>956</td>
<td>968</td>
<td>1130</td>
</tr>
<tr>
<td>2</td>
<td>317</td>
<td>791</td>
<td>797</td>
<td>841</td>
<td>855</td>
<td>1074</td>
</tr>
<tr>
<td>3</td>
<td>279</td>
<td>776</td>
<td>759</td>
<td>814</td>
<td>848</td>
<td>1088</td>
</tr>
<tr>
<td>4</td>
<td>253</td>
<td>700</td>
<td>777</td>
<td>808</td>
<td>828</td>
<td>1068</td>
</tr>
<tr>
<td>5</td>
<td>259</td>
<td>675</td>
<td>725</td>
<td>782</td>
<td>833</td>
<td>1034</td>
</tr>
<tr>
<td>6</td>
<td>192</td>
<td>655</td>
<td>716</td>
<td>805</td>
<td>830</td>
<td>1000</td>
</tr>
<tr>
<td>7</td>
<td>197</td>
<td>633</td>
<td>704</td>
<td>798</td>
<td>802</td>
<td>950</td>
</tr>
<tr>
<td>8</td>
<td>225</td>
<td>615</td>
<td>678</td>
<td>780</td>
<td>826</td>
<td>892</td>
</tr>
<tr>
<td>9</td>
<td>225</td>
<td>602</td>
<td>658</td>
<td>727</td>
<td>720</td>
<td>927</td>
</tr>
<tr>
<td>10</td>
<td>204</td>
<td>598</td>
<td>727</td>
<td>724</td>
<td>786</td>
<td>882</td>
</tr>
<tr>
<td>11</td>
<td>190</td>
<td>583</td>
<td>702</td>
<td>730</td>
<td>772</td>
<td>861</td>
</tr>
<tr>
<td>12</td>
<td>250</td>
<td>554</td>
<td>708</td>
<td>722</td>
<td>775</td>
<td>838</td>
</tr>
</tbody>
</table>

Note: Mean proportion of classification errors and response times (in milliseconds) given per Epoch for each of the six structure types of the $3_2[4]$ category family in Study 1. Epochs represent the average of six successive blocks such that four Epochs correspond to an experimental session of the current study. Dotted lines delineate the four Epochs corresponding to each session.
Figure 9. Mean proportion of classification errors (A) and response times (B) in Study 1 plotted as a function of Epoch (average of six successive blocks).
**GISTM**

**Model fits.** Without parameters, the GISTM accounts for 86% of the variance in classification performance in Session 1 ($R^2 = .86, p = .007$). The following fits using the parameterized model were done using the data within each session. Using the variant of GISTM presented in Equation 1 above, the model accounts for 99.5% of the variance ($R^2 = .995, p < .001$) in Session 1. Estimated parameter values for Session 1 data are as follows: $k = 0.23, a_1 = 0.41, a_2 = 0.37, a_3 = 1.00$. For Session 2, the parameterized GISTM accounts for 96% of the variance ($R^2 = .957, p < .001$). Estimated parameter values for Session 2 data are as follows: $k = 0.01, a_1 = 0.01, a_2 = 0.08, a_3 = 0.75$.

For Session 3, the parameterized GISTM accounts for 93% of the variance ($R^2 = .933, p = .002$). Estimated parameter values for Session 3 data are as follows: $k = 0.01, a_1 = 0.01, a_2 = 0.22, a_3 = 0.73$.

**Scaling parameter $k$.** In order to obtain overall parameter estimates, $k$ and $\alpha_d$ values must be estimated simultaneously on the data across all three sessions. When doing so on the mean proportion of errors, the following $k$ values are obtained: Session 1 ($k = .14$); Session 2 ($k = .24$); Session 3 ($k = .25$). These $k$ values indicate better overall discrimination by participants across sessions and reflect the overall improvement in performance reported above. The $k$ values additionally capture the large improvement in accuracy between Sessions 1 and 2 as well as the relatively small gains between Sessions 2 and 3. These values show evidence that large improvements in overall discrimination occur early in learning with diminishing marginal improvements in discrimination occurring with additional practice.
Overall parameter estimates for $k$ and $\alpha_d$ parameters were also obtained using the classification RT data. When doing so, the following $k$ values are obtained: Session 1 ($k = .10$); Session 2 ($k = .13$); Session 3 ($k = .13$).

**Alpha ($\alpha_d$) sensitivity parameters.** Estimated $\alpha_d$ are shown in Table 5. These values reflect the mediation of ideotype formation (Vigo, 2013, 2014). Recall that these $\alpha_d$ values provide a test of the invariance-parsimony principle which underlies the GISTM. Unlike previous studies, the obtained data allows us to generate alphas across multiple experimental sessions. The structural manifolds (SM) in Table 5 reflect the predictions of the GISTM-NP and, for the sake of comparison, are held constant across sessions (unique invariance signatures for each session are obtained when plotting the ideotypes in Figure 10). The data from Session 1 seems to corroborate previous research consistent with the invariance-parsimony principle. That is, estimated $\alpha_d$ values are highest for 0-valued SKs (yielding 1.00 in each instance) and 1-valued SKs yield very high values in Types 1 and 2. 0-valued SKs consistently yield alpha values of 1.00 across sessions. However, the alpha values of 1-valued SKs fall after Session 1. Also, alphas for middle-valued SKs (0.5) consistently rise across all three sessions.
Table 5

*Study 1 alpha values for errors.*

<table>
<thead>
<tr>
<th>Session</th>
<th>Error</th>
<th>SM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>0.03</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>0.09</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>0.18</td>
<td>(.5, .5, .5)</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>0.20</td>
<td>(.5, .5, .5)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>0.21</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>0.28</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-1</td>
<td>0.01</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>0.03</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>0.12</td>
<td>(.5, .5, .5)</td>
<td>0.22</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>0.13</td>
<td>(.5, .5, .5)</td>
<td>0.45</td>
<td>0.44</td>
<td>0.42</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>0.14</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>0.14</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Estimated alpha values using the obtained mean errors (Error) across all three sessions of Study 1.

The reduction of alpha for 1-valued SKs makes sense in terms of optimal classification. Type 1 and Type 2 learners are increasingly emphasizing diagnostic 0-valued SKs while de-emphasizing completely redundant dimensions.

Although alphas for middle-valued SKs consistently rise across sessions, there is no change in terms of emphasis towards particular SKs – a result that is consistent with model assumptions. Most interesting are the values obtained for $\alpha_2$ and $\alpha_3$ within Type 5. These alphas begin at 0.01 in Session 1, climb to about 0.5 in Session 2, and finish at 1.00 in Session 3. This seems to suggest that Type 5 learners seem to take into account
the partial redundancy of dimensions with middle-valued SKs of 0.5. Alpha ($\alpha_d$) values were also obtained using the obtained classification RT data. The obtained alpha values estimated using the mean RT data are shown in Table 6 and largely reflect the obtained alpha values using error data.

Table 6

*Study 1* alpha values for response times

<table>
<thead>
<tr>
<th>Session 1</th>
<th>RT</th>
<th>SM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>366</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>804</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>0.53</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>819</td>
<td>(.5, .5, .5)</td>
<td>0.73</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>854</td>
<td>(.5, .5, .5)</td>
<td>0.43</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>875</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>0.65</td>
<td>0.59</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>1,090</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
</tr>
<tr>
<td>3[4]-2</td>
</tr>
<tr>
<td>3[4]-3</td>
</tr>
<tr>
<td>3[4]-4</td>
</tr>
<tr>
<td>3[4]-5</td>
</tr>
<tr>
<td>3[4]-6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
</tr>
<tr>
<td>3[4]-2</td>
</tr>
<tr>
<td>3[4]-3</td>
</tr>
<tr>
<td>3[4]-4</td>
</tr>
<tr>
<td>3[4]-5</td>
</tr>
<tr>
<td>3[4]-6</td>
</tr>
</tbody>
</table>

Note: Estimated alpha values using the obtained mean classification response times (RT) across all three sessions of Study 1.

**Ideotype space.** Ideotypes per structure type are plotted as points in ideotype psychological space in Figure 10. These ideotypes are obtained based on the error data.
from Study 1. Invariance signatures are plotted for each session based on the structural manifold. Notice the stability of the points for Types II and VI across sessions. In sessions 1 and 2, points for Types III, IV and V cluster together. In session 3, however, Type V diverges from this clustering and is represented in a location near that of Type I.

The behavior of these points across sessions supports the structural equilibrium (SE) principle underlying GIST. Types with the highest degrees of SE (II & VI) show the most stability across sessions. Types with no SE, i.e. Types with no 0-valued SKs, (III & IV) show less stability and shift their location in the space across sessions. Further, these points cluster together in each session. Types with SE $\lambda(X) \approx 1.33$ (I and V) also change location across sessions, moving closer to one another in the space so that they nearly overlap by Session 3. Taken together, the behavior of these points seems to reveal a trend in which ideotypes for structures with the same degree of SE cluster together over the course of the three sessions. In sum, the obtained ideotypes show supporting evidence for the underlying principle of structural equilibrium in GIST. Specifically, category structures with comparable degrees of SE are similarly represented. Further, these representations grow increasingly similar over the course of learning.
Summary of Study 1 Results

To review, we found that a robust learning difficulty ordering does not remain stable in terms of proportions of classification errors across stages of learning. Specifically, the $3_2[4]$ ordering is obtained in Session 1 but disappears in Sessions 2 and
3. That is, the classic learning difficulty ordering deviates significantly upon repeated practice and exposure to category instances when learning takes place over consecutive days. Learning difficulty in terms of reaction times (RTs) shows the opposite relationship over time. Specifically, the $3_2[4]$ ordering is not obtained with mean RTs per type in Session 1 and 2 but is obtained in Session 3. That is, whereas classification errors may nearly cease as a result of practice, the relative degree of learning difficulty is revealed in the RT data. We view this as a significant because it shows that even when individuals reach a relative degree of expertise in terms of accuracy, the objective (or original) degree of difficulty continues to influence processing and can be revealed by behavioral markers other than accuracy. Early in learning, relative difficulty between concepts may be reliably revealed using measures of classification accuracy. Late in learning, relative difficulty between concepts might be best revealed using classification RTs. The obtained errors and RTs inform our understanding of category learning difficulty orderings and of how structural or relational information influences concept learning. Indeed, increased sensitivity over time to category structures might change the nature of classification performance, but not the nature of concept acquisition.

These results have implications for those interested in expertise development if one views expertise as the normative outcome of learning (see Palmeri & Cottrell, 2009). In this way, the development of expertise (in terms of its behavioral markers) can be described by accurately and precisely describing changes in knowledge representation that results from learning. Viewing expertise as the endpoint of learning affords the use of formal concept learning models as well as their associated representational theories as
a theoretical base for the study of expertise development. The way in which classification performance unfolds over time due to experience and training is fundamental to research on both the acquisition of perceptual skills (i.e. perceptual expertise) and expert classification. For instance, the difference between novice and expert classification decisions have been examined in jobs that require years of studying highly specialized image sets, including radiology (Myles-Worsley, Johnston, & Simons, 1988; Evans, Cohen, Tambouret, Horowitz, Kreindel, & Wolfe, 2011; Drew, Vo, Olwal, Jacobson, Seltzer, & Wolfe, 2013; Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe, 2013), cytology (Evans, Wolfe, Tambouret, & Wilbur, 2010; Tambouret, Evans, Wolfe, & Wilbur, 2010), dermatology (Norman, Rosenthal, Brooks, Allen, & Muzzin, 1989). The reported relationship between classification accuracy and the time course of classification decisions throughout the learning process may have implications for training in these fields.
Chapter 7: Repeated Exposure to Six Structure Types

Study 2 tests each participant on all six of the $3^2[4]$ category structures across three experimental sessions. Testing participants on multiple category structures allows us to estimate parameters at an individual level. Parameter estimates per subject were not possible from the resulting data of Study 1 since parameters are estimated by maximizing the coefficient of determination between the mean errors and the corresponding GISTM predictions per type. Data was collected for only a single type in Study 1, so no coefficient of determination is possible.

Testing individuals on all six structure types is further motivated by the resulting comparison with the results of Study 1 that is made possible. For instance, relationship between errors, RTs, and experience found in Study 1 may vary depending on the number of category structures the participant must learn. In this way, our manipulation implicitly tests the extent to which domain specificity is necessary for expert classification. Different results suggest that category learning in one domain (within a certain structure type) does not generalize to other domains. On the other hand, similar results would suggest that repeated classification over time increases category learning performance in general and across domains.

Method

Stimuli and materials. As in Study 1, the stimuli consist of flasks that varied on the binary features of size (large or small), shape (circular or triangular), and color (black or white) and in accordance with the structure types associated with the $3^2[4]$ category family. Participants were tested on instances associated with all six category structures of
the $3_2[4]$ family. An HP XW4600 workstation with a Dell 1708FP 15 in. flat panel LCD monitor (5 ms response time) was used to display the stimuli.

**Participants.** 40 participants were tested on the six category structures. Precedence for the current sample size can be found in previous research using well-defined category structures, which typically sample from between 30 to 40 participants (see Nosofsky, Gluck, Palmeri, McKinley, and Glauthier, 1994; Feldman, 2000; Vigo, 2013; Vigo, Zeigler, and Halsey, 2013; Vigo, Evans, and Owen, 2014; Vigo & Zeigler, 2016).

High effect sizes make relatively small samples common in this area of research. For instance, a power analysis (where $\alpha = .05$, Cohen’s $d = .08$) shows that a sample of 30 participants is sufficient for the current study.

Participants completed three sessions on three consecutive days. All participants were undergraduate students enrolled in an introductory psychology course at Ohio University.

**Procedure.** Participants engaged in a classification task over the course of three experimental sessions on three consecutive days. The classification task is unsupervised (without feedback) and parainformative in nature (Vigo, 2013; 2014), meaning that category members are presented in a learning phase prior to classification. The learning phase consists of two sets of bottles spatially separated by a line in the middle of the screen. The four category members are presented above the line while the four non-members are presented below the line and the locations of members and non-members are randomized across trials. Before the experiment began, participants are told that each
display represents the preferences of a bottle collector and that their task was to learn
which bottles the collector likes. Participants are told that the four “liked” bottles were
presented above a line on the screen and the four “not liked” bottles were presented
below the line on the screen. After 20s of studying the category members and non-
members in the learning phase, individual flasks are serially presented for classification.
Subjects are told that their task is to classify each individual bottle as either “liked” or
“not liked” using the mouse: a left mouse click (labelled “Y”) meant the bottle was
“liked” and a right mouse click (labelled “N”) meant the bottle was “not liked” by the
bottle collector. After classification responses are given for the eight bottles, a fixation
cross appears in the center of the screen and the next category block was shown.
Participants completed three sessions on three consecutive days. Six instances of each
structure type were presented randomly within each session, resulting in a thirty-six block
session lasting approximately 30 minutes each. Structure instances, as well as the eight
objects presented for categorization, are presented to participants randomly according to a
counterbalanced design throughout the experimental session.

Classification accuracy and response times were recorded on each experimental
trial. Response times were excluded from the analysis if RT < 0.025 seconds. Average
response times per block were excluded if more than half of the trial RTs were missing.
That is, at least four of the eight RTs per block were needed to be included in the
analysis.
Empirical Predictions

**Classification errors.** Consistent with the results of Study 1, we expect that the classic concept learning difficulty ordering for the six $3_2[4]$ structures will be observed in Session 1 when the proportion errors are averaged per structure type: $I < II < [III, IV, V] < VI$. However, we predict that *classification* errors will be reduced over the course of the sessions such that the data will yield an alternative ordering consistent with Study 1: $I < II < [III, IV, VI, V]$. Further, I predict that nearly all types will show significant reductions in mean errors between Sessions 1 and 2 as well as between Sessions 2 and 3. It is possible that disproportionately large learning effects will occur in one of the two gaps between sessions (from S1 to S2 or from S2 and S3). If we had to predict which is more likely to see disproportionately high error reduction, previous research would suggest it would occur early in learning between S1 and S2. A large change in learning difficulty such as this between S1 & S2 could mean that near perfect performance has been achieved by S2 making significant differences in performance between S2 & S3 unlikely.

**Classification response times.** We predict that the obtained mean response times (RTs) per type for each session of Study 2 will replicate the RT results on Study 1. Specifically, we predict that alternative difficulty orderings will be obtained for Session 1 and Session 2. By Session 3, however, mean RTs per type will show the classic difficulty ordering: $I < II < [III, IV, V] < VI$. Further, I predict that nearly all types will show significant reductions in mean RTs between Sessions 1 and 2 as well as between Sessions 2 and 3.
**Statistical Analyses**

In order to test the predictions presented above, the following data analyses are conducted on the obtained data from Study 2. This section simply presents a sketch of the statistical analyses performed --skip to the “Results” section below to view the results of these analyses on the obtained data.

A two-way repeated measures analysis of variance (ANOVA) is performed on the obtained mean proportion of errors per type across all three sessions. In Study 1, session was entered as a within-subjects variable and structure type was a between-subjects variable. In Study 2, however, both session and type is entered as within-subject variables since participants are all exposed to the same six structure types. The two-way repeated measures ANOVA is used to reveal any main effects for the independent variables (session and type) on the dependent variable (proportion error) and uncover possible interactions between the variables (session x type).

Additionally, pairwise $t$-tests are used to show whether the mean proportion of errors per type conform to the classic difficulty ordering ($I < II < [III, IV, V] < VI$). That is, a total of fifteen pairwise $t$-tests are performed per session in order to establish learning difficulty orderings for sessions 1, 2, and 3. The use of pairwise $t$-tests to establish learning difficulty orderings is consistent with the previous empirical work on these structures (e.g., Shepard, Hovland, & Jenkins, 1961; Nosofsky, Gluck, Palmeri, McKinley, & Glaughier, 1994; Vigo, 2013; Vigo, Zeigler, & Halsey, 2013; Vigo, Evans, & Owens, 2014).
Further, pairwise t-tests are used to reveal significant reductions in mean errors between Session 1 and Session 2 as well as between Session 2 and Session 3. These tests are performed across sessions for all six structure types, resulting in a total of twelve paired sample tests (six between each session).

I also test whether there are disproportionate reductions in mean errors between sessions. To do so, obtain difference scores are obtained between mean errors per type on Sessions 1 and 2 as well as between Sessions 2 and 3 for each subject. This results in two groups of difference scores per structure, where each score corresponds to the difference in individual performance across sessions. Pairwise t-tests between these groups reveal whether there are significantly different reductions between the first two and second two sessions.

Finally, all of the abovementioned analyses are conducted using mean classification response times (RTs) as the dependent variable. To summarize, these include the following analyses on mean RTs: (1) a two-way repeated measures ANOVA on the mean classification RTs per type across all three sessions; (2) pairwise t-tests within each session to reveal the relative difficulty orderings in terms of classification RTs; (3) pairwise t-tests used to reveal significant reductions in classification RTs between sessions; and (4) pairwise t-tests on the obtained RT difference scores.

**GISTM Predictions**

**Scaling parameter $k$.** Recall that the scaling parameter $k$ serves to reflect individual differences regarding the ability to extract invariance patterns. By indicating the overall degree of discrimination between differing category structures, $k$ serves as a
summary index which reveals how discrimination at the category level contributes to
classification performance. By estimating $k$ values per subject we can determine the
proportion of subjects that discriminate above or below certain levels. There are several
ways to perform such an analysis with the obtained data. Each of the procedures
discussed below are performed for both the obtained mean proportion errors and mean
classification response times (RTs). Again, this section simply presents a sketch of the
parameter estimation process--skip to the “GISTM” section below to view the obtained
model predictions.

First, $k$ is estimated at the family level across sessions. That is, I estimate
parameter $k$ on a per subject basis for all six structures. That is, a single $k$ value is
estimated for all six structures per subject across sessions. Recall that in previous work
participants show better discrimination over sessions, reflecting better overall
improvement on the task. Since higher $k$ values indicate better overall discrimination by
the observer, larger predicted $k$ parameter values are obtained across sessions. $k$ values
also can be used to capture relative improvement in accuracy between sessions, such as
the large gains from Sessions 1 to Session 2 as well as the relatively small gains between
Sessions 2 and Session 3 in Study 1. That is, large improvements in overall
discrimination occur early in learning along with diminishing marginal improvements in
discrimination as a result of additional practice.

Second, $k$ is estimated at the family level per session. Recall the work reviewed
above (Vigo, Evans, & Owens, 2014) which demonstrated the ability of the scaling
parameter $k$ to reflect group differences. Participants in this study were tested on all six
structure types. Estimated $k$ values for each population in the study were reported as follows: adults ($k = .76$); non-ADHD adolescents ($k = .54$); adolescents with ADHD ($k = .29$). Analogous use of parameter $k$ is performed in Study 2. Instead of estimates of $k$ per population, $k$ is estimated per person based on the classification accuracy for each session of the three part study. That is, I explore the discrimination levels per person across the three sessions. Here, three $k$ values are estimated at the structure family level per person, one for each session. Structure level discrimination values rise if better discrimination is achieved by participants over the course of learning.

**Alpha parameters and ideotype space.** Recall that GISTM can also include sensitivity weights which reflect relative variability – due to noise, limitation of resources, and other factors – in the number of invariants detected by an observer. Similar to Study 1, $\alpha_d$ values are estimated using the data from each session of Study 2. The obtained alpha values show additional support for the invariance-parsimony principle. Specifically, observers place disproportionate emphasis on fully diagnostic and fully redundant dimensions, i.e., dimensions with SKs of 0 or 1. The model predicts that estimated alpha values will be highest for 0-valued and 1-valued SKs.

Further, ideotypes are plotted as representations in ideotype psychological space. Further support is found for the structural equilibrium (SE) principle in term of the location of ideotype points across session. Ideotypes for structures with the same degree of SE cluster together over the course of the three sessions. Consistent with Study 1, types with the highest degrees of SE (II & VI) show the most stability in ideotype location across sessions. Recall that these results support the structural equilibrium
principle by showing that category structures with comparable degrees of SE are similarly represented by category learners.

**Results**

**Classification errors.** Mean proportion errors per session for each of the six structure types are shown in Figure 11A. A two-way repeated measures analysis of variance ANOVA on the obtained proportion of errors using session and structure type as within-subjects variables revealed a significant main effect for both session, $F(2, 78) = 37.64, p < .001, \eta^2_p = .491$, and structure type, $F(5, 195) = 58.40, p < .001, \eta^2_p = .600$. Additionally, there was a significant interaction between session and type, $F(10, 390) = 4.50, p < .001, \eta^2_p = .103$. In all sessions, pairwise $t$-tests show that the mean proportion errors per type conform to the classic difficulty ordering, I < II < [III, IV, V] < VI.

Paired sample $t$-tests reveal significant reductions in mean errors between Session 1 and Session 2 for all structure types, except Type I which showed marginal significance, $t(39) = -1.706, p = 0.096$. Conversely, paired sample $t$-tests revealed that mean errors between Session 2 and Session 3 were non-significant for all types. Difference scores for errors between Sessions 1 and 2 as well as between Sessions 2 and 3 are plotted in Figure 12A for each type. The gray bars show the mean difference between proportion errors on Session 1 and Session 2 (S1-S2) whereas the white bars show the mean difference between errors on Session 2 and Session 3 (S2-S3). Paired sample $t$-tests reveal significant reductions between the two pairs of difference scores for Types II and III (II: $t(39) = 2.26, p = 0.03$; III: $t(39) = 2.95, p = 0.005$) and a non-
significant reduction for Types I, IV, V, and VI (I: \( t(39) = -1.385, p = 0.174 \); IV: \( t(39) = 0.709, p = 0.492 \); V: \( t(39) = 1.144, p = 0.26 \); VI: \( t(39) = 1.324, p = 0.193 \)).

**Classification response times.** Mean response times (RTs) per session for each of the six structure types are shown in Figure 11B. A two-way repeated measures analysis of variance ANOVA on the obtained RTs using session and structure type as within-subjects variables revealed a significant main effect for both session, \( F(2, 78) = 33.94, p < .001, \eta_p^2 = 0.465 \), and structure type, \( F(5, 195) = 203.59, p < .001, \eta_p^2 = 0.839 \). Additionally, there was a significant interaction between session and type, \( F(10, 390) = 3.13, p < .001, \eta_p^2 = 0.074 \). In Session 1, paired sample \( t \)-tests show that the mean RTs per type conform to the classic difficulty ordering, I < II < [III, IV, V] < VI. Paired sample \( t \)-tests showed significant differences between Type IV and Type V structures in both Session 2, \( t(39) = 2.646, p = 0.012 \), and Session 3, \( t(39) = 2.082, p = 0.044 \). Other than these differences, all other paired sample \( t \)-tests conformed to the classic difficulty ordering in Sessions 2 and 3.

Between Session 1 and Session 2, paired sample \( t \)-tests reveal significant reductions in mean RTs across types, with the exception of marginal significance shown by Type IV, \( t(39) = 1.81, p = 0.079 \), and Type VI, \( t(39) = 1.85, p = 0.073 \). Between Session 2 and Session 3, paired sample \( t \)-tests reveal significant reductions in mean RTs for all types except Type VI, \( t(39) = 1.73, p = 0.092 \). Difference scores for RTs between Sessions 1 and 2 as well as between Sessions 2 and 3 are plotted in Figure 7.2B for each type. The gray bars show the mean difference between RTs on Session 1 and Session 2 (S1-S2) whereas the white bars show the mean difference between RTs on Session 2 and
Session 3 (S2-S3). Paired sample $t$-tests showed no significant differences between the two pairs of difference scores for any types.
Figure 11. (A) Mean proportion of classification errors for each of the six structure types within the $3_2[4]$ category family across all three sessions of Study 2. Mean classification error values and standard error bars are presented for each column. (B) Mean response times (RTs) for each of the six structure types within the $3_2[4]$ category family across all three sessions of Study 2. Mean RT values (in milliseconds) and standard error bars are presented for each column. The roman numerals indicate the learning difficulty ordering according to pairwise significance tests.
Figure 12. Mean difference scores for proportion errors (A) and classification response times (B) between Session 1 and Session 2 (grey bars, S1-S2) as well as Session 2 and Session 3 (white bars, S2-S3) of Study 2.
GISTM

Model fits. Without parameters, the GISTM accounts for 91% of the variance in classification performance in Session 1 ($R^2 = .909, p = .003$). The following fits using the parameterized model were done using the data within each session. Using the variant of GISTM presented in Equation 1 above, the model accounts for 99.9% of the variance ($R^2 = .998, p < .001$) in Session 1. Estimated parameter values for Session 1 data are as follows: $k = 0.63$, $a_1 = 0.11$, $a_2 = 0.91$, $a_3 = 0.99$. For Session 2, the parameterized GISTM accounts for 98% of the variance ($R^2 = .981, p < .001$). Estimated parameter values for Session 2 data are as follows: $k = 0.97$, $a_1 = 0.36$, $a_2 = 0.53$, $a_3 = 0.99$. For Session 3, the parameterized GISTM accounts for 99% of the variance ($R^2 = .986, p < .001$). Estimated parameter values for Session 3 data are as follows: $k = 0.77$, $a_1 = 0.32$, $a_2 = 0.62$, $a_3 = 0.99$.

Scaling parameter $k$. In order to obtain overall parameter estimates, $k$ and $\alpha_d$ values must be estimated simultaneously on the data across all three sessions. When doing so on the mean proportion of errors, the following $k$ values are obtained: Session 1 ($k = .52$); Session 2 ($k = .73$); Session 3 ($k = .81$). These $k$ values indicate better overall discrimination by participants across sessions and reflect the overall improvement in performance reported above. The $k$ values additionally capture the large improvement in accuracy between Sessions 1 and 2 as well as the relatively small gains between Sessions 2 and 3. These values show evidence that large improvements in overall discrimination occur early in learning with diminishing marginal improvements in discrimination occurring with additional practice.
Overall parameter estimates for $k$ and $\alpha_d$ parameters were also obtained using the classification RT data. When doing so, the following $k$ values are obtained: Session 1($k = .12$); Session 2 ($k = .15$); Session 3 ($k = .19$).

**Alpha ($\alpha_d$) sensitivity parameters.** Estimated $\alpha_d$ using the obtained mean errors from Study 2 are shown in Table 7. Recall that the obtained $\alpha_d$ values in Study 1 supported the invariance-parsimony principle underlying the GISTM. The estimated $\alpha_d$ values in Study 2 are similar to those obtained in Study 1. For instance, estimated $\alpha_d$ values are highest for 0-valued SKs (yielding 1.00 in each instance) and 1-valued SKs yield very high values in Types 1 and 2. 0-valued SKs consistently yield alpha values of 1.00 across sessions. Once again, the alpha values of 1-valued SKs fall after Session 1. The reduction of alpha for 1-valued SKs makes sense in terms of optimal classification. As in Study 1, Type 1 and Type 2 learners are increasingly emphasizing diagnostic 0-valued SKs while de-emphasizing completely redundant dimensions. Also consistent with Study 1, alphas for middle-valued SKs (0.5) rise across all three sessions. Again, this suggests that when classifying Type 5, participants are increasing taking into account the partial redundancy of dimensions with middle-valued SKs of 0.5 across sessions.

Alpha ($\alpha_d$) values were also obtained using the obtained classification RT data. The obtained alpha values estimated using the mean RT data are shown in Table 8 and largely reflect the obtained alpha values using error data.
<table>
<thead>
<tr>
<th>Session</th>
<th>Error</th>
<th>SM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>0.01</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>0.10</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>0.88</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>0.17</td>
<td>(.5, .5, .5)</td>
<td>0.43</td>
<td>0.43</td>
<td>0.42</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>0.18</td>
<td>(.5, .5, .5)</td>
<td>0.34</td>
<td>0.33</td>
<td>0.34</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>0.18</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>0.48</td>
<td>0.52</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>0.28</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session</th>
<th>Error</th>
<th>SM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>0.01</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>0.05</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>0.10</td>
<td>(.5, .5, .5)</td>
<td>0.83</td>
<td>0.82</td>
<td>0.84</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>0.13</td>
<td>(.5, .5, .5)</td>
<td>0.50</td>
<td>0.62</td>
<td>0.63</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>0.13</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>0.85</td>
<td>0.91</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>0.21</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session</th>
<th>Error</th>
<th>SM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>0.01</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>0.05</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>0.09</td>
<td>(.5, .5, .5)</td>
<td>0.83</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>0.11</td>
<td>(.5, .5, .5)</td>
<td>0.66</td>
<td>0.66</td>
<td>0.69</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>0.11</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>0.17</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Estimated $\alpha$ values using the obtained mean errors (Error) across all three sessions of Study 2.
Table 8

*Study 2 alpha values for response times.*

<table>
<thead>
<tr>
<th>Session1</th>
<th>RT</th>
<th>SM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>348</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>827</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>0.47</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>932</td>
<td>(.5, .5, .5)</td>
<td>0.32</td>
<td>0.29</td>
<td>0.08</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>904</td>
<td>(.5, .5, .5)</td>
<td>0.34</td>
<td>0.33</td>
<td>0.34</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>900</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>0.55</td>
<td>0.50</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>1,066</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session2</th>
<th>RT</th>
<th>SM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>311</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>710</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>0.68</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>807</td>
<td>(.5, .5, .5)</td>
<td>0.55</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>840</td>
<td>(.5, .5, .5)</td>
<td>0.48</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>783</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>986</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session3</th>
<th>RT</th>
<th>SM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3[4]-1</td>
<td>278</td>
<td>(0, 1, 1)</td>
<td>1.00</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>604</td>
<td>(0, 0, 1)</td>
<td>1.00</td>
<td>1.00</td>
<td>0.73</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>725</td>
<td>(.5, .5, .5)</td>
<td>0.63</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>763</td>
<td>(.5, .5, .5)</td>
<td>0.54</td>
<td>0.55</td>
<td>0.59</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>722</td>
<td>(0, .5, .5)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>931</td>
<td>(0, 0, 0)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Estimated alpha values using the obtained mean classification response times (RT) across all three sessions of Study 2.

**Ideotype space.** Ideotypes per structure type are plotted as points in ideotype psychological space in Figure 13. These ideotypes are obtained based on the error data from Study 2. Invariance signatures are plotted for each session based on the structural manifold. Recall that the behavior of the ideotypes in Study 1 supported the structural equilibrium (SE) principle underlying GIST. Specifically, ideotype points for structures with the same degree of SE clustered together over the course of the three sessions. Ideotype points for the six structures show the same clustering behavior in Study 2.
Types with no 0-valued SKs where $\lambda(X) \cong 1.00$ (III & IV) have ideotype points that cluster together in each session. Types with SE $\lambda(X) \cong 1.33$ (I and V) move closer to one another across sessions. Types with the highest degrees of SE (II & VI) show relative stability across sessions. The obtained ideotypes supports the idea that category structures with comparable degrees of SE are similarly represented and show supporting evidence for the underlying principle of structural equilibrium in GIST.
**Figure 13.** Cartesian coordinates representation of ideotypes corresponding to the $3_2[4]$ category structures graphed in ideotype space. Plot (A) shows ideotypes based on the objective learning difficulty per type as predicted by GISTM-NP. Plots (B), (C), and (D) show ideotype points for Sessions 1, Session 2, and Session 3, respectively. Plots A through D are based on structural manifolds reflecting perceived invariance detection based on the error data from Study 2.

**Individual discrimination $k$.** Testing each participant on all six structure types allows us to conduct individual parameter estimates. Accordingly, $k$ was estimated on the error and RT data across sessions for each subject. This results in three $k$ values per
participant, one for each of the three sessions. By estimating $k$ values at this level, we can determine the percentage of our participants that discriminate at or below a certain level.

Figure 14 shows the cumulative discrimination performance across the three sessions of Study 2 when estimated on errors. On the error data in Figure 14, about 85% of participants discriminate at 1.0 or below during Session 1, whereas only about 45% of participants discriminate at this level during Session 3. Notable is the large range of discrimination performance across all sessions.

Figure 15 shows the cumulative discrimination performance across the three sessions of Study 2 when estimated on mean RTs per subject. Note the much smaller range of discrimination values within sessions. In addition, there are much smaller differences in $k$ values between sessions when estimated on the RT data.

Recall that the scaling parameter $k$ serves to account for individual differences regarding the ability to extract invariance patterns and serves as a summary index that reveals how discrimination at the category level contributes to classification performance. Larger $k$ values indicate better discrimination and relatively higher discriminability between distances means relatively higher sensitivity to invariance patterns.
Figure 14. Cumulative percentage of discrimination for all three sessions of Study 2. A single $k$ value was estimated for mean errors within each session, per subject.
Figure 15. Cumulative percentage of discrimination for all three sessions of Study 2. A single $k$ value was estimated for mean RTs within each session, per subject.
Summary of Study 2 Results

We found that when participants are given all six of the 3\_2[4] structures the learning difficulty ordering remains stable over multiple experimental sessions. Recall that in Study 1, when subjects were given a single type, the learning difficulty ordering was shown in Session 1 but disappeared in Sessions 2 and 3. Such differences between Study 1 and Study 2 emphasizes the fundamental role of context during classification. When a single type is tested in isolation over the course of multiple sessions, differences in classification performance disappear between types. However, when such types are continually presented in the context of the other types, the relative difficulty remains over the course of learning. It is particularly interesting that the ordering was obtained in terms of significance tests despite reductions in proportion errors per type across sessions (with the exception of Types I and II in Session 2 and 3).

Further, the classic learning difficulty ordering was achieved or approximated in terms of reaction times (RTs) in all three sessions. The only exception was significant differences between Types IV and V in Sessions 2 and 3. In Study 1, the ordering was not obtained for RTs until Session 3. In other words, the relative processing difficulty is present over the course of learning when category structures are presented in the context of the other structures, but not present when structures are given in isolation. We view the RT stability in Study 2 as further evidence in support of the role of context during learning.

Finally, support for the GISTM is found in Study 2 in much the same way as in Study 1. Specifically, $k$ and $\alpha_d$ values using the obtained data from study 2 closely
resemble the estimates from Study 1. These obtained parameter values further support the underlying principles of invariance-parsimony and structural equilibrium.

Correspondingly, the behavior of ideotype points across sessions closely resemble the behavior of points in Study 1. This supports the idea that ideotype points are being fine-tuned properly regardless of whether the points are obtained between participants (as in Study 1) or within participants (as in Study 2).
Chapter 8: Additional Response Time Analyses for Study 1 and Study 2

Additional analyses are conducted on the response time (RT) data from Study 1 and Study 2. These analyses are useful in terms of ongoing debates in the field. For instance, the relationship between classification RTs and practice is explored by estimating the best fitting power and exponential functions across the obtained data (Figures 16 and 17). Further, response time distributions for each structure type are obtained across all three sessions (Figure 18).

Recall that there is a debate regarding whether decreases in RTs that result from practice are best accounted for by power function or exponential function curves (e.g. Delaney, Reder, Staszewski, & Ritter, 1998; Palmeri, 1999; Heathcote, Brown, & Mewhort, 2000). The power function predicts more substantial reductions in RTs early in learning with diminishing marginal reductions resulting from additional practice. Alternatively, the exponential function predicts that RTs decrease at a constant rate in relationship to what is yet to be learned. In both cases, higher values of the learning rate parameter $c$ result in steeper RT reductions.

Fitting our data on such curves is beneficial in that it reveals how underlying cognitive processes unfold over different time-scales during concept acquisition. Additionally, such fits serve to compare competing learning theories that predict certain rates of RT decay. Further, such information may have practical implications for the implementation of job training programs in areas reliant on expert classification (e.g. medical diagnosis, radiology, air traffic control). As such, we fit both functions to mean RTs over the course of the three sessions.
As such, both functions were fit to mean RTs over the course of the three sessions of Study 1. A power function best fit the mean RTs per epoch when combined across all types (Figure 16). Further, power functions provided the best fit to the majority of 3_2[4] category types. See Table 9 for a comparison of learning rate parameters and the coefficients of determination for each function across types. Figure 17 shows mean classification response time per type plotted as a function of epoch as well as power and exponential fits. The fits of both functions in terms of the coefficients of determination were very similar, making it difficult to say if one function does indeed better account for classification RT data over time.

*Figure 16. Mean classification response time (ms) per epoch when averaged across all types in Study 1. Epochs are the average of six consecutive experimental blocks, or 48 consecutive trials (6 blocks * 8 trials each = 48 total trials). As such, each session is represented by four epochs each (sessions are delineated by the gray lines in the figure).*
Table 9

*Study 1 learning rates and fits per type*

<table>
<thead>
<tr>
<th></th>
<th>Power</th>
<th></th>
<th>Exponential</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>c</td>
<td>R^2</td>
<td>c</td>
<td>R^2</td>
</tr>
<tr>
<td>3[4]-1</td>
<td>1.29</td>
<td>0.94</td>
<td>0.83</td>
<td>0.94</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>0.23</td>
<td>0.99</td>
<td>0.31</td>
<td>0.98</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>0.96</td>
<td>0.92</td>
<td>0.62</td>
<td>0.91</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>0.32</td>
<td>0.91</td>
<td>0.29</td>
<td>0.86</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>0.47</td>
<td>0.81</td>
<td>0.35</td>
<td>0.78</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>0.01</td>
<td>0.83</td>
<td>0.01</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Note: Learning rate parameters (c) and coefficient of determination (R^2) for power and exponential functions when mean RTs are taken for epochs within each type in Study 1.
Figure 17. Power and exponential functions when mean RTs are taken for epochs within each type in Study 1.
Figure 18. Response time distributions for each structure type across all three sessions for the experiment in Study 1. Dotted lines depict the cumulative percentage.
Similar analyses were performed using the obtained data from Study 2. Figure 19 shows the best fitting power and exponential functions when estimated on mean RTs per epoch across all types. Recall that there are more blocks total per session in Study 2, so epochs are the average of nine consecutive experimental blocks (instead of six, as in Study 1), or 72 consecutive trials. As such, each session is represented by four epochs each. Again, the fits of both functions in terms of the coefficients of determination were similar, making it difficult to say if one function does indeed better account for classification RT data over time. Its seems that in this case, the power function best accounts for the steep decline in mean RTs early in learning while the exponential function better accounts for the mean RTs later in learning.

Best fitting functions were also obtained for each type. Figure 20 shows mean classification response time per type in Study 2 plotted as a function of epoch, along with the power and exponential fits. Note that although there were more overall blocks per session in Study 2 (36 total blocks), there were fewer blocks per type (6 blocks per type) than in Study 1 since all six types were tested. Six blocks per type results in eighteen total blocks over the course of three sessions. Therefore, Figure 20 consists of nine total epochs where each epoch is the average of two consecutive same-type blocks. Table 10 summarizes the fits by providing a comparison of learning rate parameters and the coefficients of determination for each function across types. The maximized fits per type produce similar r-squared values between the two functions, making it again difficult to say which function accounts best for classification RTs. Notable, however, are the relatively low learning rate parameters (c) obtained in Study 1 as compared to Study 2.
These smaller learning rate values capture the smaller rate with which RTs are diminishing over blocks of Study 2. It seems that RTs diminish at a quicker rate when individuals are tasked with categorizing objects from the same structure type (as in Study 1) rather than objects from various structure type of the same category (as in Study 2).

Figure 19. Mean classification response time (ms) per epoch when averaged across all types in Study 2. Epochs are the average of nine consecutive experimental blocks, or 72 consecutive trials (9 blocks * 8 trials each = 72 total trials). As such, each session is represented by four epochs each (sessions are delineated by the gray lines in the figure).
Figure 20. Power and exponential functions when mean RTs are taken for epochs within each type in Study 2.
Table 10

*Study 2 learning rates and fits per type*

<table>
<thead>
<tr>
<th></th>
<th>Power</th>
<th></th>
<th>Exponential</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c$</td>
<td>$R^2$</td>
<td>$c$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>3[4]-1</td>
<td>0.10</td>
<td>0.83</td>
<td>0.15</td>
<td>0.86</td>
</tr>
<tr>
<td>3[4]-2</td>
<td>0.10</td>
<td>0.95</td>
<td>0.38</td>
<td>0.95</td>
</tr>
<tr>
<td>3[4]-3</td>
<td>0.10</td>
<td>0.91</td>
<td>0.10</td>
<td>0.84</td>
</tr>
<tr>
<td>3[4]-4</td>
<td>0.16</td>
<td>0.81</td>
<td>0.32</td>
<td>0.79</td>
</tr>
<tr>
<td>3[4]-5</td>
<td>0.10</td>
<td>0.93</td>
<td>0.26</td>
<td>0.91</td>
</tr>
<tr>
<td>3[4]-6</td>
<td>0.10</td>
<td>0.90</td>
<td>0.39</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Note: Learning rate parameters ($c$) and coefficient of determination ($R^2$) for power and exponential functions when mean RTs are taken for epochs within each type in Study 2.
Chapter 9: Discussion

In summary, the reported data from the experiments above provide empirical evidence for each of the following:

1. Classification RTs reveal learning difficulty.
2. Stable difficulty orderings in terms of accuracy depend on sampling procedures.
3. Evidence that conceptual expertise is domain specific.
4. Evidence that conceptual expertise is accompanied by a speedup in processing.

Further, the theoretical contributions of the current work concern GIST. Theoretical takeaways include each of the following:

1. Observers place disproportionate emphasis on fully diagnostic and fully redundant dimensions
2. Structures with comparable degrees of structural equilibrium are similarly represented.
3. Additional experience with these structures results in progressively similar representations.

General Discussion

In this section, we present a general discussion of the work presented above. After a summary of the reported results from Study 2, we review the empirical contribution of this work with respect to the concept learning and perceptual expertise literatures. After our empirical discussion, we review the theoretical contributions in terms of both concept learning theory and perceptual expertise.
The reported data from Study 2 shows that if participants are randomly presented instances of all six structure types, the difficulty ordering remains stable across all three sessions. This stability is present in terms of mean accuracy (errors) and nearly shown in terms of mean response times (only marginal p-values between Types IV and V create deviations from the ordering).

Recall that learning difficulty orderings for categorical stimuli have long provided an empirical foundation for concept learning and categorization research (e.g. Shepard, Hovland, & Jenkins, 1961; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Feldman, 2000; Vigo, 2013; Vigo, Zeigler, & Halsey, 2013; Vigo, Evans, & Owens, 2014). The current experiments were designed to extend this empirical work in two critical areas. First, the stability of the $3_2[4]$ difficulty ordering was tested by obtaining classification data over multiple sessions. Second, classification RTs were treated as a measure of concept learning difficulty by considering the relative RT orderings. The conventional approach in this area of research has sought to establish difficulty orderings in terms of mean classification accuracy (proportion of correct/incorrect classification responses) from a single learning session. I am unaware of a study testing the stability of such orderings over a period of extended learning. The observed results of Study 2 seem to be the first to show that this learning difficulty ordering remains stable over time and across multiple learning sessions. This is significant because it is possible that the $3_2[4]$ learning difficulty might only be obtained when categorical stimuli are novel. It has remained an empirical question whether the ordering remains after the categorical stimuli are no longer novel. The stability across sessions of Study 2 provides further support for
the robust learning difficulty ordering associated with the $3^2[4]$ structures. Indeed, given the number of replications of the ordering, one might consider it an empirical law of human classification learning.

However, the current work reveals that the stability of this ordering depends on an additional factor. Specifically, the stability of the ordering in terms of errors seems to depend on the number of category structure types observers are exposed to during learning. When category structures are randomly sampled from all six types (as in Study 2) there is stability across sessions. However, when participants are given instances of a single type (as in Study 1) the ordering varies across sessions. The source of this instability likely stems from the relative degree of expertise participants are able to achieve when presented with so many instances of the same category type. The task of Study 1 learners seems relatively less demanding than that of Study 2 learners who are repeatedly exposed to category structures of varying degrees of complexity. Indeed, Study 1 learners provided more accurate classifications than Study 2 learners in terms of aggregate mean classification errors. See Figure 21 and Figure 22 for side-by-side comparison of the respective error and RT data from both studies. The mean proportions of errors are lower in Study 1 across every type (with the exception of Type I). This result highlights the importance of methodology in concept learning research. Category learning experiments using categorical stimuli vary in terms of the sampling procedures used. For instance, Shepard, Hovland, and Jenkins (1961) presented participates with all six structures whereas Nosofsky, Gluck, Palmeri, McKinley, & Glauthier (1994) tested participants on only two of the six structures. Vigo and his collaborators (Vigo, 2014;
Vigo, Zeigler, & Halsey, 2013; Vigo, Evans, & Owens, 2014; Vigo & Zeigler, 2016) have employed consistent procedures when using the $3_2[4]$ structures, sampling from all six structure types. Previous studies that did not sample from all six structures still observed the ordering. The ordering was obtained because mean errors were taken based on the averaging across a single experimental session. The current findings suggest that if participants continued learning a limited set of category structures over the course of multiple sessions, we could expect the ordering to disappear over time.
Figure 21. Study 1 and Study 2 comparison of errors. (A) shows the mean proportion of errors per type across sessions in Study 1. (B) shows the mean proportion of errors per type across sessions in Study 2.
Figure 22. Study 1 and Study 2 comparison of RTs. (A) shows the mean classification RTs per type across sessions in Study 1. (B) shows the mean classification RTs per type across sessions in Study 2.
When considering possible explanations for the different orderings between studies, it might be helpful to consider previous research in the area of expertise. Work in this area has established the domain specific nature of expertise (Voss, Greene, Post, & Penner, 1983; Bédard & Chi, 1992). That is, expertise develops in the domain of application and does not typically generalize to other areas. For example, if experts with doctoral degrees are given problems outside of their subject; their performance is the same as undergraduate students (Voss, Greene, Post, & Penner, 1983). The relative gains in classification accuracy for Study 1 learners in the current study seem to support the idea of domain specific expertise and suggest that the same may be true during categorization. The better accuracy of Study 1 learners may have resulted simply due to receiving more overall instances of a given type than Study 2 learners. The implication is that conceptual expertise may develop acutely for specific categories in which we receive an abundance of experience. Admittedly, this claim may be premature on this basis of this work alone and future research should involve testing the relative domain specificity or generality of concept learning.

An additional empirical contribution of the current study concerns the obtained RT orderings and the treatment of classification RTs as a measure of concept learning difficulty. Previous research has not typically explored dependent variables other than classification accuracy. The results of Study 1 show that RTs might be a better measure of relative difficulty when measuring performance in an area where individuals have gained a lot of experience. These results suggest that as subjects’ performance increases to perfect or near perfect levels of classification, reaction times still reveal key
differences in learning difficulty. This is a result that might be useful in many real world applications. For instance, consider a field where experts rarely make mistakes. It may be difficult to study their performance in this area using measures of accuracy given the limited number of errors. Our results suggest that RTs can be used as a measure of task difficulty, allowing researchers the ability to order the relative difficulty of these tasks.

Additionally, our reported results are consistent with hallmarks of perceptual expertise. Recall that previous research has shown that expertise development is accompanied by a speedup in processing, marked by quicker RTs over time (e.g. Newell & Rosenbloom, 1981; Delaney, Reder, Staszewski, & Ritter, 1998; Palmeri, 1999; Heathcote, Brown, & Mewhort, 2000). Previous research has explored whether these quicker RTs are best accounted for by power function or exponential function curves (e.g. Delaney et al., 1998; Palmeri, 1999; Heathcote et al., 2000). For example, classification decisions have been shown to be well fit by a power function (Nosofsky & Palmeri, 1997). Both of the current studies saw significant reductions in mean RTs over the course of learning. Despite extensive curve fitting on the obtained RT data, general conclusions about the superiority of power and exponential functions are hard to make based on these analyses. This is because the obtained fits were largely similar between the two functions. We hope that this work informs future work that aims to describe the form of observed reductions in classification RTs over the course of learning. The power function predicts more substantial reductions in RTs early in learning with diminishing marginal reductions resulting from additional practice while the exponential function predicts that RTs decrease at a constant rate in relation to what is yet to be learned.
We interpret the results of both experiments in terms of the Generalized Invariance Structure Theory (GIST; Vigo, 2013; 2014), as well as its core formal mathematical model of classification behavior (the GISTM). Our theoretical discussion was largely confined to the GIST because of the inability of other models to make straightforward predictions for the current tasks. Hopefully this work encourages more flexible modelling in this area, allowing models to make dynamic predictions based on experience and training.

With respect to the GIST, the current empirical work tested the validity of two underlying principles. Support was shown for the invariance parsimony principle, which states that observers place disproportionate emphasis on fully diagnostic and fully redundant dimensions. As predicted, estimated alpha values were highest for 0-valued and 1-valued kernels. This finding adds to a growing number of results (Vigo, 2013; 2014) supporting the idea that humans are sensitive to the low and high occurrence of redundant patterns in categorical stimuli, while patterns with partial redundancy are difficult to ascertain. Further, sensitivity to fully diagnostic dimensions was maintained across sessions while sensitivity to fully redundant dimensions was reduced. Support was also shown for the structural equilibrium (SE) principle, which states that categorical stimuli with a higher proportion of fully diagnostic 0-valued SKs (thus higher SE) results in easier identification of dimensions needed for rule formation. Ideotypes per structure type were plotted as points in ideotype psychological space across all sessions. The movement of the points in psychological space supports the SE principle. Specifically, ideotypes for structures with the same degree of SE cluster together over the course of the
three sessions. The implication is that category structures with comparable degrees of SE are similarly represented and that these representations grow increasingly similar over the course of learning. Combined, these results support GIST’s invariance–based account of categorization behavior along with its proposal that a process of invariance pattern detection underlies category learning. See Figures 23 – 25 for to compare the location of ideotype points between studies. The locations of ideotype points become more stable over the course of learning so that the locations are nearly identical between studies in Session 3 (see Figure 25).

In all, we showed how the parametric variant of the GISTM advances our understanding of concept representations over the course of learning.
Figure 23. Ideotypes points for each $3_2[4]$ category structures plotted based on Session 1 data from Study 1 (left) and Study 2 (right).
Figure 24. Ideotypes points for each 3_2[4] category structures plotted based on Session 2 data from Study 1 (left) and Study 2 (right).
Figure 25. Ideotypes points for each 3_2[4] category structures plotted based on Session 3 data from Study 1 (left) and Study 2 (right).
Limitations and Future Directions

The current work tested classification performance in a rather specific domain, using the $3_2[4]$ category structure types in a parainformative task that did not include corrective feedback. Advantages of using such a precise method include high internal validity and results that are easily comparable to many previous studies. However, it is likely that such a well-defined specific methodology also reduces the generalizability of the results. Accordingly, future research should employ several methodological extensions in order to expend the current findings. Future studies might use a serioinformative task (instead of a parainformative task) and/or provide corrective feedback to participants. Perhaps most critical, however, is that similar studies be conducted using category families that vary in their size and dimensionality by sampling from the 84 category structures studied by Vigo (2013). Additionally, this work should use stimuli that better resemble real world categories. It should be noted that recent work has already begun to do so (Vigo, Evans, & Owens, 2014; Vigo & Doan, 2015). The use of well-defined category structures is a scientifically responsible starting point since they represent the simplest cases and provide clear quantities that are comparable across experiments. However, we are assuming that the same cognitive and perceptual processes taking place during these experiments also take place in the real world when processing everyday stimuli. This is an assumption that should be tested in future work. In addition, although the choice of testing participants on three consecutive days is based on convention, it is likely an arbitrary interval. As such, future work might explore the influence of varying intervals between sessions of classification learning.
Possible future research might involve ways eye-tracking and/or mouse-tracking can be used to explore the relationship between classification and expertise. Eye-tracking technology has had a long and relatively successful history in the vision sciences (see Mele & Federici, 2012). Such success is in large part because studies have revealed that saccadic eye movements do indeed reflect cognitive processes (Kowler, 1991; Viviani, 1989; 1990; Yarbus, 1967). However, eye-tracking research requires costly equipment and necessitates time-intensive empirical work. Visual search can be alternatively measured using mouse-tracking technology which is considerably easier and less costly. Many recent studies have shown that mouse-tracking is highly correlated with eye movements (Chen, Anderson, & Sohn, 2001; Cooke, 2006; Guo & Agichtein, 2010; Huang, White, & Dumais, 2011; Rodden & Fu, 2007; Rodden, Fu, Aula, & Spiro, 2008; Chen & Lim, 2013).

For instance, future work might explore the nature of the relationship between visual search during category learning and higher order information processing. Vigo, Zeigler, and Halsey (2013) reported an inverse relation between object-based fixation time (i.e. maintenance of gaze at a single location) and the information conveyed by category members during the acquisition of a novel concept. That is, average fixation times per object during category learning were shown to be inversely proportional to the amount of information conveyed by an object about the category as a whole. A novel model of whole-object visual search (information model of fixation; IMF), based on representational information theory (RIT; Vigo, 2011b), was shown to fit the reported data. Future work could use mouse-tracking on a classification task that takes place over
several days. Such work will aim to explore the processes that underlie expert
classification by describing how visual search strategies aid in the acquisition of concepts
of varying complexity. As in the original study (Vigo, Zeigler, & Halsey, 2013), the
results would be interpreted in the context of predictions made by the classic prototype
account of category learning (Rosch, 1978), the multiplicative prototype models of
category learning (Estes, 1986; Nosofsky & Zaki, 2002), and an information-theoretic
account (Vigo, 2011b) which measures the subjective information conveyed by members
of categorical stimuli.

Another research project in this area might explore possible differences in the
time-course of classification RTs between experts and novices. Research has shown that
experts are more efficient than novices largely because they differentially allocate the
time-course of task specific responses. Specifically, experts across many domains have
been shown to take longer than novices early in processing, when first initiating a
response (Ranganathan & Carlton, 2007; Shank & Haywood, 1987; Sim & Kim, 2010;
Sanchez, Sicilia, Guerrero & Pugnaire, 2005; Kobus, Proctor & Holste, 2001; Incera &
McLennan, 2015). In the context of classification research, the question is as follows: Do
expert classifiers follow the same qualitative differences in processing when compared to
novice classifiers? In future research might explore the dynamic time course of
classification RTs by comparing experts and novices using mouse-tracking technology.
Using MouseTracker software (Freeman & Ambady, 2010), the proposed research
question can be explored by analyzing possible differences in the time elapsed between
the start of a trial and the first mouse movement (i.e. initiation time). Mouse-tracking
allows initiation times to be recorded in-route to giving classification responses on the screen.

In conclusion, the way in which classification accuracy and classification response times change over time supports underlying principles of GIST and demonstrates how the parametric variant of the GISTM advances our understanding of concept representations over the course of learning. Combined, the reported results support GIST’s invariance–based account of categorization behavior along with its proposal that a process of invariance pattern detection underlies category learning. It is likely that many lines of empirical work in experience—based learning might be unified under this theoretical framework. It is my hope that this work inspires future research on concept learning and that the robustness of these empirical and theoretical relationships are tested in new domains of application.
References


*Cognition and Categorization*. Lawrence Erlbaum Associates.


