COMPARING COMPARISONS IN CATEGORY LEARNING

BY

ERIN JONES HIGGINS

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology in the Graduate College of the University of Illinois at Urbana-Champaign, 2012

Champaign, Illinois

Doctoral Committee:

Professor Brian H. Ross, Chair
Associate Professor Aaron S. Benjamin
Assistant Professor Andrei Cimpian
Associate Professor Kara D. Federmeier
Professor John E. Hummel
ABSTRACT

Comparisons are central to category learning; yet very little research has been done to understand how comparisons affect what people learn. Prior work has established that different ways of learning affect what information learners acquire, suggesting that different types of comparisons may also affect learning in different ways. An important comparison-type distinction in category learning is between-category versus within-category comparisons. This paper draws on research from other domains that have looked at the role of comparisons in cognitive processes; however, the results of these studies are mixed, so it remains unclear how each type of comparison affects learning. Here, the highlighter theory of comparison learning is proposed based on the idea that studies showing a benefit for one type of comparison or the other are similar and different in systematic ways. Specifically, between-category comparisons highlight distinguishing information between categories while within-category comparisons highlight commonalities and the relational structure of items. In five experiments, one type of comparison or the other is shown to lead to higher classification performance at test and the effects of each depend on the type of information that needs to be emphasized during learning.
ACKNOWLEDGEMENTS

There are many people I would like to acknowledge who have helped me along the way. First, Brian Ross, who has been an amazing advisor and most recently has been instrumental in helping me determine what I should do with my life, which as he knows was no small feat. One would be lucky to have an advisor who cares even half as much as Brian does about his students. I want to thank my undergraduate advisor, Beth Crawford, for saying “yes” to me as an excited yet clueless freshman in college asking to work in her lab, for showing me how exciting research is, and for introducing me to the wonderful world of if statements and pivot tables in Excel. I want to thank Aaron Benjamin, Andrei Cimpian, John Hummel, and Kara Federmeier for their insightful comments and support for this project and more importantly for their guidance and support throughout the last five years.

I want to thank the members of the Ross and Hummel lab groups for their helpful comments in the early stages of this project. I also want to acknowledge Robert Molitor, Audrey Merz, Brandon Mitchell, and Kaitlin Costello, who all worked tirelessly to help me collect and code the data presented in this thesis.

Finally, I’d like to acknowledge my family and friends for their support, especially my husband, Steve Higgins, my parents, Nat and Chris Jones, and my sister, Samantha Jones.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>CHAPTER 1: INTRODUCTION</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.1 Evidence that comparison type matters</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.2 Studies suggesting between-category-comparison benefits</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.2.1 Studies showing a spacing effect in category learning</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1.2.2 Caveats to the spacing-is-better findings</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>1.3 Studies suggesting within-category-comparison benefits</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.3.1 Structure-mapping theory and analogical reasoning</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>1.3.2 Structural alignment and relational categories</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>1.3.3 What causes within-category comparisons to highlight relational</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>structure?</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>1.3.4 Surface features vs. relational structure</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>1.3.5 Structural alignment and alignable differences</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>1.4 The highlighter theory of comparison learning</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>1.5 The current experiments</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td><strong>CHAPTER 2: EXPERIMENTS 1 AND 2</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.1 Experiment 1</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>2.2 Experiment 2</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td><strong>CHAPTER 3: EXPERIMENTS 3 AND 4</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.1 Experiment 3</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>3.2 Experiment 4</td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td><strong>CHAPTER 4: INTERIM DISCUSSION OF EXPERIMENTS 1-4</strong></td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td><strong>CHAPTER 5: EXPERIMENT 5</strong></td>
<td>62</td>
</tr>
<tr>
<td>6</td>
<td><strong>CHAPTER 6: GENERAL DISCUSSION</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.1 Summary of experiments</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>6.2 Alternative explanations</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>6.2.1 Similarity explanation</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>6.2.2 Prior-knowledge explanation</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>6.2.3 Difficulty explanation</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>6.2.4 Summary of alternative explanations</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>6.3 Implications for theories of category learning</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>6.4 Comparisons in educational settings</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>6.5 Future directions</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>6.5.1 Considering the learner’s approach to each type of comparison</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>6.5.2 Order effects of between-category and within-category comparisons</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>6.5.3 Other types of comparisons</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>6.6 Conclusions</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td><strong>TABLES</strong></td>
<td>101</td>
</tr>
<tr>
<td></td>
<td><strong>FIGURES</strong></td>
<td>115</td>
</tr>
<tr>
<td></td>
<td><strong>REFERENCES</strong></td>
<td>119</td>
</tr>
<tr>
<td></td>
<td><strong>APPENDIX A</strong></td>
<td>125</td>
</tr>
<tr>
<td></td>
<td><strong>APPENDIX B</strong></td>
<td>126</td>
</tr>
<tr>
<td></td>
<td><strong>APPENDIX C</strong></td>
<td>127</td>
</tr>
<tr>
<td></td>
<td><strong>APPENDIX D</strong></td>
<td>130</td>
</tr>
<tr>
<td></td>
<td><strong>APPENDIX E</strong></td>
<td>132</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION

Effective reasoning and behavior necessitates the use of prior knowledge and past experiences. Much of this knowledge is in the form of concepts, which are mental representations of classes of items\(^1\) (i.e. categories) in the world. Concepts facilitate our understanding of current experiences by enabling us to determine what something is (e.g. is this a flower or a weed?) and how to interact with it given what it is (e.g. knowing that it is a weed leads to the decision to pull it from the ground). Knowing what category an item belongs to is powerful because it enables the use of a large body of knowledge about that class of items, which can be drawn upon for a variety of tasks such as making decisions, solving problems, making predictions, and constructing explanations. For example, diagnosing the illness a patient has makes it easier to decide on an appropriate treatment plan, knowing what type of data a researcher is working with allows her to determine the appropriate statistical test to use, and identifying that someone is a new driver helps explain why he or she put on the left-turn signal in order to turn right. Given that concepts underlie the majority of cognitive tasks, it is critical to understand how they are learned.

Learning about a category often occurs as a function of using it, so in developing an understanding of how categories are learned, it is important to consider how different kinds of active processing lead to differences in what is learned. One type of active processing that is thought to be central to category learning and use is comparisons (e.g. Spalding & Ross, 1994). In fact, as Goldstone (2010) points out, most models of categorization propose that an item is categorized on the basis of its similarity to a category prototype (i.e. a summary representation of the category; e.g. Rosch & Mervis, 1975) or to items one has already encountered and stored in long-term memory (e.g. Medin & Schaffer, 1978; Hintzman, 1986), suggesting the centrality of

\(^1\) The term *item* refers to any single instance (i.e. example) of a category that a person encounters.
Comparisons in categorization and category learning (given that learning is affected by how categories are used).

Comparisons occur all the time and can serve a variety of purposes beyond determining category membership. Imagine the following three scenarios. One, you are about to purchase a house and are deciding between two options. The first house has a large kitchen while the second house has a great backyard. Two, you encounter a car that has doors which open up instead of out and you describe that car to your friend as “out-of-the-ordinary.” Three, you are learning about different types of diseases in medical school and you learn that while a fever can be indicative of many different illnesses, a fever combined with a sore throat narrows the list of possible illnesses. In deciding which house to purchase, whether a car is out-of-the-ordinary or normal, or which symptoms are most helpful in diagnosing illnesses and which are not, you are making comparisons (between the two houses, between the car in front of you and other cars you have seen, between different illnesses). In fact, while simultaneously considering the three scenarios I just described, you may have been comparing them.

Across the three scenarios, comparisons were made for different purposes, but in each case they biased attention to certain details of the items and not others. How does making comparisons affect your understanding of a situation and the items involved? How does it affect the concept that represents the category those items are members of? As psychologists have been arguing for years, it is not the case that experiences are objectively encoded and represented in the mind, but rather, they are perceived and subsequently stored in a subjective way based on our prior knowledge and the current context of an experience. This thesis will examine how comparisons affect category learning, and specifically, the degree to which different types of comparisons lead to differences in what is learned.
While it has been suggested that comparisons play a central role in category learning (e.g. Spalding & Ross, 1994), there has been relatively little research examining their effect on learning. The small set of studies that have examined comparisons in category learning have done so with simple materials that are not representative of complex, real-world categories. In addition, these studies have yielded mixed and inconclusive results (e.g. Andrews, Livingston, & Kurtz, 2011; Hammer, Brechmann, Ohl, Weinshall, & Hochstein, 2010). A better understanding of how categories are learned requires a better understanding of comparison effects.

The majority of research examining comparisons focuses on how considering multiple items as opposed to one item at a time impacts problem solving and reasoning. The consistent finding across these studies is that comparing across two items, as opposed to independently studying each item, promotes more effective analogical reasoning (i.e. the process of identifying how aspects of one item correspond with aspects of the other) and leads to higher performance when solving novel problems (e.g. Gentner, Loewenstein, & Thompson, 2003; Gick & Holyoak, 1983; Rittle-Johnson & Star, 2007).

When considering the role of comparison in category learning, an additional level of complexity is involved. That is, there are different types of information to be learned, such as the features\(^2\) that are important for distinguishing one category from another and the feature values that are common to most members of a category. There are many different types of comparisons learners can engage in, such as comparing items from different categories, items from the same category, and within each of those comparison types, comparing items that share many feature values and items that share few feature values with each other. It is not simply a

---
\(^2\) The term feature refers to a particular part of an item. For example, a feature of a dog would be its ears. A feature value corresponds to the type of feature an item has. For example, a dog may have floppy ears as opposed to pointy ones. The floppiness of the ears would be the ear feature’s value.
question of whether comparing across items is an effective way to learn, but rather, when, how, and why particular types of comparisons are effective.

One reason for distinguishing different types of comparisons is that each is a type of active processing that may encourage learners to focus on different aspects of items. Some recent work has demonstrated that learning goals affect what category information is eventually acquired (Anderson, et al., 2002; Chin-Parker & Ross, 2002, 2004; Jones & Ross, 2011; Yamauchi & Markman, 1998). In all of these studies, classification learning (i.e. learning by determining the label for an item) was compared to inference learning (i.e. learning by predicting the value of a missing feature for an item) in an attempt to show that even a small difference in task goals (classifying a label versus predicting a feature) can affect what is learned. In many ways these two tasks are equivalent, but it has been argued that the learning goal of each is drastically different (see Markman & Ross, 2003 for a review). For classification learners, the goal is to learn the best way to distinguish between the categories being learned. For inference learners, the goal is to learn what each category is like. Learners develop category representations that reflect the learning task (Anderson, et al., 2002; Chin-Parker & Ross, 2002, 2004; Jones & Ross, 2011; Yamauchi & Markman, 1998).

These studies suggest that when categories need to be learned, either classification or inference learning could be interpreted as helpful depending on the information that is most critical for learners to acquire. More broadly they suggest that different learning tasks and conditions could be more or less beneficial depending on a learner’s goal. Therefore, if different types of comparisons lead learners to acquire different types of information, it is possible that the benefits of different types of comparisons may change depending on what information learners must acquire.
When considering the types of comparisons that may occur during category learning, one of the most critical distinctions is between-category comparison and within-category comparison. For example, when learning about birds, comparing a finch to a sparrow is likely to highlight information that distinguishes them (e.g. their different body shapes, coloring) while comparing one finch to another finch may highlight commonalities between the two birds (e.g. their beak shape). In this dissertation, I will argue that between-category and within-category comparisons provide the learner with different information about the category or categories being learned, and one or the other is beneficial (i.e. will lead to better performance on a test) depending on what information the learner needs to take away from the comparison.

In the remainder of this chapter I will draw on a variety of findings from other domains where comparisons have been studied. The pattern of findings seems to be inconsistent in terms of how between-category and within-category comparisons may affect category learning. However, I will argue that together these findings provide a framework that can be used to motivate the theory I am proposing here, which I will call the highlighter theory of comparison learning. This theory proposes that between-category and within-category comparisons highlight different information about the categories being learned. It makes the unique prediction that different types of categories modulate the effects of between-category and within-category comparisons because each type of category requires learners to focus on different information.

More specifically, the highlighter theory of comparison learning is the following: between-category comparisons highlight information that distinguishes between categories (i.e. information about what features differ across categories as well as the values for each of those features that are typical for each category) while within-category comparisons highlight
commonalities and the relational structure that items from a category share. Engaging in one type of comparison or the other will lead to differences in what information is learned.

I will argue that this theory makes the prediction that when learning to classify items as members of a category, different types of categories will benefit from different types of comparisons, as each type of category requires learners to focus on certain aspects of the items. The different types of categories I am referring to will be characterized on the basis of whether learners can rely on individual feature values of items or must rely on more complicated information, such as the relations between features, when determining category membership. While I acknowledge that the complexity of the category information learners need to consider is a continuum, here I consider two cases near the extremes to test the hypothesis: either learners can rely on feature values alone or they need to consider relations between features to determine category membership. Why this characterization was chosen instead of others will become clear after I describe the consistencies across the sets of studies showing benefits for one type of comparison or another.

The highlighter theory of comparison learning predicts that if learners can rely on feature values alone to determine category membership, then they will benefit more from between-category comparisons, as this type of comparison highlights information that distinguishes between categories (i.e. which of the items’ features tend to differ across categories, and for those features, what values tend to be typical for each category). On the other hand, if learners need to establish an understanding of the categories’ relational structures, then they will benefit from within-category comparisons, as this type of comparison highlights the common relational structure that items share.
I will return to this theory and the predictions it makes after reviewing a variety of findings that seem to provide mixed evidence for the role of each type of comparison in category learning, but when analyzed more finely will motivate the theoretical framework I am proposing here. First, I will briefly discuss a small set of studies that have focused on types of comparisons in order to provide preliminary evidence that considering how different types of comparisons affect learning, as opposed to generally considering the role of comparisons, is important.

Second, I will present studies that show that the order of items influences learning to motivate how between-category comparisons may affect category learning. Third, I will discuss research from the analogical-reasoning and problem-solving domains to establish a framework for understanding how within-category comparisons may affect category learning.

1.1 Evidence that comparison type matters

There is some evidence that different types of comparisons affect what and how the features of an item are represented. Medin, Goldstone, and Gentner (1993) showed that people perceive an ambiguous item in a way that is consistent with the comparison they engage in. For example, if shown an item with three prongs and a shorter prong that could be interpreted either as a fourth prong or as another part of the item, participants were more likely to say the item had four prongs if they had compared it to another item with four distinct prongs and three prongs if they had compared it with another item with only three distinct prongs. These results demonstrate how influential the type of comparison can be in determining how features of an item are interpreted and encoded.

The type of comparison can also affect what information is learned. Rittle-Johnson and Star (2009) showed that certain types of comparisons lead to better conceptual understanding of mathematics problems than other types. Learners studied algebra problems by either comparing
similar problems using the same solution method, different problem types using the same solution method, or the same problem solved with two solution methods. Learners who compared different solution methods for the same problem showed greater conceptual knowledge (knowledge of various algebra concepts) and procedural flexibility (the ability to generate multiple solutions if asked, as well as the ability to choose appropriate, accurate, and efficient solutions at test) than learners who compared different problem types solved with the same solution.

While these two studies used different types of comparisons than the ones of interest here, their results demonstrate that learners are sensitive to the items they are comparing. Different types of comparisons lead to differences in both how features are processed and what information is learned and ignored.

### 1.2 Studies suggesting between-category-comparison benefits

This section will focus on a set of studies that suggest engaging in between-category comparisons leads to better classification performance. In these studies, the order of items is manipulated and performance differences are interpreted as effects of different types of comparisons, under the assumption that certain item orders provide learners with opportunities to make either within-category or between-category comparisons (Kornell & Bjork, 2008). In this section, I will review several studies that demonstrate these order effects and will end with a brief discussion of the generalizability of these studies’ findings.

One of the most robust findings in the memory literature is the spacing effect. The spacing effect is a phenomenon first documented by Ebbinghaus (1885), where repetitions spaced out over a period of time lead to better recall than massed repetitions of items in a list.

---

3 The terms *spacing* and *massing* have been used to refer to two different aspects of item presentation: the amount of time between one presentation and the next and the sequence of items. Here spacing and massing will refer to the sequence of items (e.g. *abcabcab* vs. *aaabbbecc* respectively).
(Crowder, 1976). The spacing effect has been shown across a number of domains beyond memory. For instance, baseball players’ hitting performance will be higher if they practice different types of pitches in a random (i.e. spaced) sequence rather than a massed sequence (Hall, Domingues, & Cavazos, 1994).

When learning about categories, the learner’s goal is to generalize across items rather than remember specific items. When generalization is the goal, spacing different items from the same category between items from other categories can be thought of as an opportunity to compare across different types of items while massing items from the same category together can be thought of as an opportunity to compare across items of the same type. It is important to point out, however, that this is an inference and whether or not people are actually engaging in comparisons of different types depending on the order in which items are presented has not been established.

1.2.1 Studies showing a spacing effect in category learning

Kornell and Bjork (2008) found that spacing items from one category between those of other categories enhances category learning. They had participants study paintings from multiple artists and showed that when items from one category (i.e. paintings from one artist) were spaced between items from other categories, learners were better able to classify novel items from those categories at test than if they had studied the items from each category massed together. Kang and Pashler (2011) replicated this finding and validated that this effect was in fact due to the sequence of items rather than their temporal spacing. The explanation proposed by Kornell and Bjork (2008) is that spacing allows for more effective discrimination learning. By having items from one category interspersed between items of other categories, learners
could more easily compare items from different categories to see what features were diagnostic of category membership.

Kornell and Bjork’s (2008) two-part hypothesis is that spacing is effective because learners can engage in between-category comparisons, and between-category comparisons are beneficial because they teach learners how to discriminate between categories. If the first part of this hypothesis is accurate, then this study provides evidence that making between-category comparisons leads to higher classification performance.

There is evidence that spacing enhances mathematical problem solving as well. Learners who practiced problems of one type spaced between problems of other types were better able to apply appropriate solution procedures at test (Rohrer & Taylor, 2007; Taylor & Rohrer, 2010). Like Kornell and Bjork (2008), Taylor and Rohrer (2010) argue that spacing improves discriminability. Their explanation is that spacing forces learners to think more deeply about the kind of problem they have been presented with so that they can choose the appropriate solution procedure. When learning through spacing, learners need to attend to distinguishing features that cue them in to which solution to use. When learning through massing, however, learners do not have to think about distinguishing features as deeply because they already know before the problem is even presented which solution they need to use.

1.2.2 Caveats to the spacing-is-better findings

There are several caveats to the argument that spacing leads to better discrimination. First, it is important to point out that in the studies by Rohrer and Taylor (2007; Taylor & Rohrer, 2010), the benefits of spacing may have been due to the particulars of the categories learned and the way participants were tested. In the study by Rohrer and Taylor (2007), participants had to learn how to calculate the volumes of four different types of solids and in the
study by Taylor and Rohrer (2010), participants learned how to solve four categories of problem types related to a prism (find the total number of faces, corners, edges, or angles given the total number of base sides of the prism). In both cases, it was easy to determine the type of problem because it was always evident in the problem’s description (e.g. participants were prompted to find the volume of this sphere; solve for the number of faces). Participants in both studies were learning to apply equations to problems that had essentially already been categorized for them. Taylor and Rohrer (2010) acknowledge:

…In other scenarios, identifying (or classifying) the problem or task is the greater challenge. For example, different kinds of inferential statistics problems are usually stated in a superficially similar fashion…Thus learners must recognize subtle features of the problem (e.g. parametric vs. non-parametric, repeated measures vs. independent measures, etc.) that allow them to identify which kind of problem it is. (p. 845).

Taylor and Rohrer (2010) make an interesting point, but I want to argue for an even broader distinction. In the studies reporting spacing benefits, learners could rely on feature values to determine category membership. For example, in the study by Kornell and Bjork (2008), while I acknowledge that learners could have compared the more subtle aspects of the artists’ paintings, it seems more likely from the examples they provided that learners only needed a very rudimentary understanding of the feature values associated with particular artists, such as the type of brush stroke and coloring. On the other hand, there are other categories that are more difficult to distinguish, either because the items’ feature values are extremely similar or overlap across categories (making items from different categories seem very similar on the surface and therefore difficult to distinguish) as Taylor and Rohrer (2010) suggest or because they require learners to consider more complicated analyses of the items, such as considering
relations between features. For these types of categories, spacing (i.e. between-category comparison) may be less helpful or not helpful at all, and as I will argue in the subsequent section, within-category comparisons may be more helpful because they will emphasize the common relational structure that items share.

A second caveat is that Kurtz and Hovland (1956) actually showed a benefit for massing in a category learning study. In their experiment, participants learned about geometric patterns that varied on four features (size, shape, color, and position). Each drawing could be classified into one of four categories, determined based on a rule. Within each category, two features always had the same values and the other two features’ values were varied randomly across items. With this category scheme, items from one category could frequently share the same feature values as items from another category. In order for participants to learn the categories, they had to figure out that it was the combination of two features’ values that determined category membership. Participants saw items one at a time in either a massed or spaced sequence, and those who learned through the massed sequence performed better at test (both on a classification test as well as when asked to provide a verbal description of each category). The authors suggest that in the spaced condition, the intervening items interfered with participants’ memory for prior items from the same category, so participants were unable to make effective within-category comparisons in order to determine the rule that defined each category.

Kurtz and Hovland (1956) used categories defined by a clear rule that required participants to learn which features were important for each category as well as which values for each of those features, when paired together, determined category membership. Learning these categories required learners to go beyond focusing on individual feature values and to focus

---

4The rules that defined membership in each category were as follows: all items from the *kems* category had the feature values up (position) and circle (shape), all *fovs* had the values up (position) and square (shape), all *hajs* had the values down (position) and small (size), and all *yugs* had the feature values down (position) and large (size).
instead on feature conjunctions. Massing may have been more beneficial here because the category scheme was more complex than one that only requires learners to focus on individual feature values. Comparing items from the same category may have led learners to notice the more complex category structure while comparing items from different categories hindered their ability to attend to and understand this more complicated structure.

In summary, what I have suggested here is that in the studies showing spacing benefits, learners could rely on feature values alone, while in the one study showing massing benefits, learners had to focus on more complex category information (i.e. conjunctions between features). While it is impossible to validate these observations from the results of these studies, this distinction helped motivate the theory I am proposing here for when and why each type of comparison is useful.

1.3 Studies suggesting within-category-comparison benefits

In the previously discussed studies that showed the benefits of spacing, it was argued that spacing is effective because of comparison (e.g. Kornell & Bjork, 2008), even though items were presented one at a time and participants were never actually asked to compare. Most of the research looking directly at the role of comparison has been conducted in the analogical-reasoning domain. Analogical reasoning refers to the process of identifying how aspects of one item correspond with aspects of another item. One can think of making an analogy as similar to making a within-category comparison, as the goal in each is to determine how one item is like another item (in other words, determining what can be mapped across two items is similar to determining why two items share the same label).

This section will focus on studies that demonstrate how making an analogy between two items leads learners to acquire the common relational structure that those items share. I will also
discuss how the degree to which items’ surface features are similar and different across items of the same type plays a role in noticing that those items share a common relational structure. Finally, I will briefly discuss how comparing across items of the same type may actually lead to noticing important differences between categories as well.

1.3.1 Structure-mapping theory and analogical reasoning

Making an analogy between two items makes it easier to determine what features they share. An effective analogy is meant to convey the relationship between the components of one item (the base) to those of another item (the target) through comparison. The goal is to determine what two items have in common, although differences between them are also highlighted to some degree. For example, it is popular when explaining how atoms are structured to make an analogy between an atom and the solar system (Gentner, 1983). Students already understand that planets revolve around the sun because the sun attracts them, so they are able to extend that idea to atoms. By comparing an atom to the solar system, they learn that electrons move around the nucleus of an atom because the nucleus attracts the electrons to it.

When two seemingly unrelated items, such as an atom and a solar system, are grouped together, people will consider the subtle, structural aspects of the items, such as the role each feature takes on.

When making an analogy, a mapping process is used to determine the one-to-one correspondences between the components of one item and the components of the other item. According to structure-mapping theory, proposed by Gentner (1983), this mapping process boosts attention to the relational structure of each item and takes attention away from the item’s specific feature values. This process, which Gentner refers to as structural alignment, highlights the role each feature plays and downplays the actual value for each feature. For example, when
children are learning to do division problems, they may use a word problem involving ten apples that need to be divided between two people as an analogy to help them solve a new problem involving ninety dollars that needs to be divided among three cash registers. Even though the values of the features of the problems are completely different, the role each takes on is the same across both problems. Ten apples corresponds to ninety dollars, as both take on the role of the dividend while two people corresponds to three cash registers as both take on the role of the divisor. Structure-mapping theory predicts that when relational structure needs to be learned, analogy can be a powerful tool for highlighting it even if the feature values of the various items encountered during learning are very different.

When encountering new items, using a general knowledge structure, or schema, stored in memory will lead to higher problem-solving performance than simply making an analogy using another item (Gick & Holyoak, 1983). If the schemas to be learned require learners to notice relations between features, then structure-mapping theory would predict that making analogies between items would highlight the relevant relations, leading to the acquisition of the schema. One way to ensure that learners are making analogies across items is to encourage explicit comparison between them. If learners compare across items, they should successfully acquire a schema and should be more likely to effectively solve novel problems later. Across multiple studies, this idea has been supported (Gentner, et al., 2003; Gick & Holyoak, 1983).

Some of the earliest evidence for the benefits of comparison comes from Gick and Holyoak (1983). Across a series of studies, they had participants study one or two examples of problems that required a solution using converging forces. In one story, an army general sends small groups of soldiers to attack a fortress from different locations, and in a second story, a fire fighter has multiple people spray hoses simultaneously from different locations to put out a fire.
After being exposed to one or multiple problems where the solution was to use converging forces, participants attempted to solve Duncker’s (1945) radiation problem, where the solution is to have rays converge on a tumor from multiple directions in order to shrink it. Participants who actively compared across the army general and fire fighter problems were more likely to solve the radiation problem than those who did not compare. Gick and Holyoak (1983) argue that participants who compared across problems were more likely to solve the radiation problem because comparison facilitated the development of a problem schema that represented the structural characteristics of problems requiring a converging-forces solution. For example, these participants may have learned that all of the problems present a situation that requires a large amount of force to solve; however, due to the characteristics of the barrier between the force and the object the force needs to be applied to, only a small amount of force can be used at any given spot on the barrier. Participants who did not make comparisons were not able to abstract this schema, as they focused primarily on the surface features of the problem.

Gentner, et al. (2003) demonstrated similar comparison benefits. In their study, participants either compared two examples of a negotiation strategy or studied each one independently. Those who compared the examples developed a better schema and consequently demonstrated better performance when asked to solve a novel problem requiring the same negotiation strategy. The explanation for their results is that comparisons benefit learning because they allow learners to see the underlying structure shared by both examples and ignore the surface features. On the other hand, when comparison processes are not engaged, learners focus on surface features of the examples and ignore the underlying structure.
1.3.2 Structural alignment and relational categories

Doumas and Hummel (2004) used the structure-mapping framework to motivate their study looking at the role of comparisons in relational-category learning. Relational categories are defined by the role each feature plays rather than each feature’s value. An example that illustrates this idea nicely is the category bridge. A bridge is anything that connects two locations. It can manifest itself as an impressively large structure with support beams and overarching gates or as a piece of wood or a fallen tree connecting two riverbanks. It can even refer to the idea of connecting, or transitioning between, two arguments (Gentner & Kurtz, 2005). Gentner and Kurtz (2005) point out that relational categories are at least as common as feature-based categories (categories defined by feature values rather than the relations between features) in the world.

Doumas and Hummel (2004) predicted that when learning to distinguish between categories requires learners to discover and predicate relations, analogical mapping (i.e. comparison across items from the same category) should facilitate learning. To address this prediction, they tested whether or not making a comparison across items from the same relational category affected classification performance. Participants learned about two categories of cells. Each cell had five features: its location, shape, membrane thickness, nucleus roundness, and number of organelles. A higher-order relation between the cells’ membrane thickness and the roundness of the nucleus defined category membership while the other features’ values were random across items from both categories. All participants classified items with feedback for one block of trials. Next, one group of participants was given a mapping task, where their job was to map the elements of one item to the elements of another from the same category. The other group of participants was simply told to study the two items. After the mapping task,
participants performed another block of classification trials. The learners who participated in the mapping task showed higher classification performance in the second block of trials than the learners who were only given an opportunity to study the items, demonstrating the benefits of comparing items from the same category.

Kotovsky and Gentner (1996) also demonstrated the importance of comparisons in discovering relations. In their first experiment, children of varying ages were shown a pattern of shapes and were asked to identify which of two other patterns was most like it. This pattern-matching task required children to notice relational commonalities between the presented pattern and the correct response in order to successfully perform this task. The surface similarity between the presented pattern and the responses was manipulated across trials. For half of the trials, the presented pattern and responses varied along the same dimension (e.g. the presented pattern was small circle, large circle, small circle and the correct match was the pattern small square, large square, small square). For the other half of the trials, referred to here as cross-dimension trials, the presented pattern and responses varied along different dimensions (e.g. the presented pattern was small circle, large circle, small circle and the correct match was the pattern white square, black square, white square). Within the same-dimension and cross-dimension trials, half were polarity-matched trials (as the previous two examples demonstrated) while the other half were polarity-mismatched trials (e.g. the presented pattern was white circle, black circle, white circle and the correct match was black square, white square, black square). Older children were able to successfully recognize the relational choice, regardless of the trial type. Four-year-olds had a difficult time with this task, and performed at chance on all trial types with the exception of the same dimension, polarity-matched trials.
In subsequent experiments, Kotovsky and Gentner (1996) attempted to increase four-year-olds’ performance on the pattern-matching task in various ways. In these studies, there were only two trial types: same dimension, polarity-matched trials and cross-dimension, polarity-matched trials. They found that progressive alignment (i.e. ordering the trials in such a way that learners viewed the easier same-dimension trials at the beginning of learning and the cross-dimension trials at the end of learning) and labeling (i.e. teaching children labels for the relational and non-relational answer choices before having them do the pattern-matching task)—both of which most likely encouraged children to make comparisons—improved four-year-olds’ pattern-matching performance, even for the cross-dimension trials.

Kotovsky and Gentner (1996) argue that progressive alignment was effective in the pattern-matching task because it prompted children to make comparisons across trials (due to the high similarity between trials at the beginning of the experiment). When making these comparisons, children most likely aligned their representations of the trials they were comparing on the basis of surface similarity, but when they made the comparisons, the common relational structure the trials shared was highlighted. Similarly, the label-learning task most likely encouraged children to compare across answer choices that shared labels, leading them to develop a more abstract representation than they would have had they not completed the label-learning task. Kotovsky and Gentner (1996) use these results as evidence that comparisons enabled four-year-olds to develop an abstract representation of relations.

1.3.3 What causes within-category comparisons to highlight relational structure?

One question that arises is why within-category comparisons highlight relational structure. Doumas, Hummel, and Sandhofer (2008) developed a computer model called DORA (Discovery of Relations by Analogy) in order to provide a cohesive theory about how relational
learning occurs. A key principle of the model is that comparisons are central for discovering novel relations. DORA successfully accounts for a number of results in the relational learning literature, including the progressive alignment finding by Kotovsky and Gentner (1996). While Doumas, et al.’s (2008) explanation of how comparison (and analogy) works is largely consistent with that of Gentner’s (1983) structure-mapping theory, the authors offer a more specific, algorithmic explanation for why comparisons highlight relational structure.

A key principle of the model is that comparisons play a central role in the discovery and refinement of novel relations. When DORA makes comparisons, feature values that items share have close to twice as much activation as a feature value that is unique to an item, as the shared feature value occurs across both items’ representations. Consequently, commonalities are preferred to differences. At first, one’s representation of a new category or relation may include both relevant and irrelevant commonalities. For instance, when learning the relation *round* by comparing a small red ball to a red apple, other feature values that both items share, such as the fact that both items are red and small, will also be included in the representation. As learning progresses and more comparisons are made, relevant commonalities are likely to occur more often than irrelevant ones, leading learners to eventually learn the abstract, common relational structure of items. For instance, in a subsequent comparison, if one compares a melon to a basketball, *round* is still shared by both items, but other feature values, like red, which happened to occur in both items in earlier comparisons, are not represented in these items. As more comparisons are made, irrelevant feature values (like the color red in the example) are eventually eliminated from one’s representation of a relation or category.
1.3.4 Surface features vs. relational structure

In all of the studies reviewed so far in this section (perhaps with the exception of Kotovsky & Gentner, 1996), participants have been explicitly told to make comparisons across items. In most cases, the items used in these studies shared relational structure but not surface features. An important question is whether people can make the same kinds of comparisons on their own without some sort of prompt to do so. In other words, can people notice structural commonalities across items on their own when the items do not share many surface features?

The findings by Gentner, et al. (2003) and Doumas and Hummel (2004) suggest that people fail to make comparisons across items of the same type, whose surface features vary, unless they are told to do so. In both studies, participants saw all of the same items, but the group of participants explicitly told to make comparisons outperformed the other group. However, there is some work that shows that people can sometimes spontaneously make comparisons. These spontaneous comparisons typically occur when items of the same type share surface features (e.g. one sees two of the same type of math problem which are both about car mechanics working on cars; Ross, 1984, 1987) although successful use of the comparison is determined by the relational correspondences (Ross, 1987). These findings show that when learning about items that are not similar on the surface, guided comparison is necessary in order for people to notice structural commonalities between those items. When learning about items whose surface features are similar, people are more likely to make effective comparisons on their own.

The idea that people fail to notice similarities between items that do not share many surface features is consistent with a large body of expertise research, which shows that novices tend to sort items and think about items differently than experts. For example, Chi, Feltovich,
and Glasser (1981) showed that physics novices sort physics problems based on their surface features (e.g. “these problems deal with blocks on an inclined plane”) while experts sort them by their underlying structural features (e.g. “these are conservation of energy problems”). In domains like physics where novices will be misled by the surface features of the problem, encouraging within-category comparisons may facilitate conceptual understanding.

Children have the same tendency to focus on surface features and ignore the underlying relational structure of items (e.g. Imai, Gentner, & Uchida, 1994). Imai, et al. (1994) used a task where children were shown a familiar item (e.g. an apple) and told a new label for it (e.g. “This is a dax”). Next they were shown three additional items: a perceptually similar, out-of-category item (e.g., a balloon), an item that shared the same taxonomic category but was perceptually dissimilar (e.g. a banana), and a thematic match that was also perceptually dissimilar (e.g. a knife). Their task was to choose the item that most likely shares the first item’s label. Children were more likely to select the perceptually similar, out-of-category item than the other two items.

Across a number of studies, Gentner and Namy (1999; also see Namy & Gentner, 2002) showed that when prompted to make comparisons between two items that share a taxonomic category (e.g. an apple and a pear), children are able to overcome their perceptual bias and choose the more conceptual match. What is so striking about this finding is that if either item from the comparison were presented alone, it would have resulted in the perceptual bias. This point in particular reiterates the idea that comparison is an active process, where the interpretation and encoding of each item is different because it was compared. Either item alone would have biased children to choose the perceptual, out-of-category item, but when given the opportunity to compare, children overcame their bias to focus on perceptual (i.e. surface) features and successfully recognized structural commonalities between items. In a recent follow
up study, Namy and Clepper (2010) showed that making contrasts (i.e. showing children perceptually similar items from other taxonomic categories and telling them they were not part of the target category) was not sufficient for overcoming the perceptual bias. Together these results suggest that within-category comparisons in particular highlight the common relational structure that items from a category share (e.g. the functional commonality that both apples and bananas are edible) and downplay learners’ reliance on perceptual features.

In summary, when novices in a domain encounter items whose surface features do not clearly predict category membership, they perseverate on them anyway. When the surface features are similar across items from the same category but not items from other categories, this bias to focus on surface features can lead people to acquire the relational structure (because they are more likely to spontaneously make comparisons between the items as in Ross, 1987). When the surface features are not similar across items from the same category, as is the case with most complex domains, people will still focus on surface features and will fail to notice the relational commonalities across items.

In order for novices to become experts in domains where the relational structure of items is not obvious and surface features are not similar across items, novices need to do something that will emphasize the common relational structure across the items. Structure-mapping theory explains that comparisons highlight common relational structure, leading novices to eventually acquire a schema that reflects structural rather than surface features. Therefore, in cases where novices fail to identify the relational structure of the items on their own, guided within-category comparisons will provide a means of bootstrapping their understanding of the relational structure.
1.3.5 Structural alignment and alignable differences

While the emphasis in this section has been on uncovering the common relational structure across items using structural alignment, this same process can also highlight important differences between items (Gentner and Markman 1994; Markman & Gentner, 2000). Determining the important differences that exist across items is not a trivial matter, as there are an infinite number of differences between items that one could list. For instance, a banana is different from a cat because it does not have fur or ears, it does not meow, and it is not an animal. How does one decide which differences matter and which differences do not?

Gentner and Markman (1994; Markman & Gentner, 2000) distinguish between two types of differences: alignable and non-alignable differences. An alignable difference is a feature-value difference between two items that occurs within a feature shared across both items. For instance, cars have four wheels and bicycles have two wheels. The difference between the two items is in the value of the wheel feature, which both items share. A non-alignable difference is a difference that does not have a corresponding feature across the two items. For example, a car can have a moonroof or a bicycle can have a carbon fork. There is no bicycle feature value that corresponds to a car’s moonroof (and vice versa). Alignable differences are favored over non-alignable differences. People are more likely to list alignable differences between items, are more likely to use them in making similarity judgments, and are more likely to attend to them when assessing similarities and differences between items (Markman & Gentner, 2000).

In category learning, especially when the learning goal is to distinguish between categories, noticing alignable differences is critical, as these are typically the important differences that learners should attend to. Without an understanding of the relational structure of items within each category, it is impossible to know which between-category differences are
alignable or non-alignable. Under circumstances where the relational structure is not yet understood, structure-mapping theory predicts that within-category comparisons would be helpful when one needs to identify alignable differences across items. This may seem counter-intuitive, but the idea is that one needs an understanding of relational structure before he or she could even begin to determine which differences are alignable. Once the learner has an understanding of the relational structure of items from a category, he or she now has a framework for interpreting differences between categories. In other words, while within-category comparisons may not be optimal for noticing critical between-category differences, this type of comparison may be critical for establishing the framework necessary to evaluate whether a difference is alignable or not. For example, knowing that for both cars and bicycles, a person must use some device to steer highlights an important alignable difference between cars and bicycles: for a car, a driver uses a steering wheel to steer and for a bicycle, a cyclist uses handlebars to steer. Without understanding that a person must use a device to steer a bicycle, it would be difficult for a person to determine that handlebars versus steering wheels is an important alignable difference to attend to.

1.4 The highlighter theory of comparison learning

The evidence for the benefits of within-category and between-category comparisons is mixed. On the one hand, there are studies that have demonstrated that spacing leads to better learning, and it has been suggested that this is because learners have an easier time comparing across items from different categories, leading to better discrimination ability. On the other hand, a number of studies have consistently shown that comparing two items of the same type leads to enhanced performance on a variety of tests (problem solving, classification), and in some cases, making within-category comparisons leads people to form completely different
categories than they would on their own (such as categories based on relational structure rather than surface features, as in Namy & Gentner, 2002). Based on these mixed results, it is safe to say that comparisons affect learning, but it is still not clear when and why each type of comparison is most useful. In this section, I will integrate these two very different sets of findings and use them to motivate the highlighter theory of comparison learning.

I am proposing the idea that the major difference between the sets of studies just reviewed is what information participants had to learn. As already mentioned, the studies showing a spacing benefit tended to use items whose feature values provided enough information for learners to determine category membership. On the other hand, the materials used in many of the analogical-reasoning and relational-categories studies were quite different. For these materials, participants had to consider the relational structure of the items in order to effectively interact with them.

Noticing this difference in materials led to the consideration that these studies yielded mixed results because different types of categories require an emphasis on different information. When learners can rely on feature values to determine category membership, it is more useful to highlight information that distinguishes between categories (i.e. which features are diagnostic of category membership and which values for those features are typical for each category). When learners must focus on more complicated category information, specifically the relations between features, then structural commonalities between items need to be highlighted because learners will not pick up on the items’ common relational structure on their own.

The highlighter theory proposes that each type of comparison highlights different category information. Between-category comparisons highlight the differences between one category and another. Within-category comparisons highlight commonalities across items as
well as the common relational structure that items from the same category share. In highlighting the common relational structure of items, within-category comparisons can also help learners develop a framework for determining which between-category differences are alignable. One type of comparison or the other may be beneficial for learning depending on what information needs to be emphasized.

1.5 The current experiments

Based on the integration of the empirical findings reviewed here, I proposed the highlighter theory of comparison learning, which claims that between-category comparisons highlight information that distinguishes between categories while within-category comparisons highlight commonalities and the relational structure of categories. This theory predicts that the benefits of each type of comparison will depend on the type of information one needs to acquire (either due to the type of category one is learning or due to the goals of the learner).

A straightforward way to test this theory is to find cases where learners can rely on feature values and cases requiring learners to focus on relations between features to determine category membership. The predictions are that (1) between-category comparison will be more beneficial when learners can rely on feature values to determine category membership and (2) within-category comparison will be more beneficial when learners need to acquire the relational structure of categories to determine category membership. Across five experiments, these predictions were tested by having participants learn categories through either between-category or within-category comparisons. Participants’ category knowledge was assessed with a classification test (and in some experiments, additional tests were used in conjunction with classification to determine what category knowledge participants had acquired).
I acknowledge that there may be other ways to characterize the different types of categories (e.g. general difficulty of the categories). It is important to note, however, that the highlighter theory makes specific predictions about how each type of comparison affects learning and the experiments reported below were designed to test those predictions. In other words, this was not a post-hoc explanation for comparison effects, but rather, a predictive framework for motivating the set of studies reported below. I will return to this point and elaborate on possible alternative characterizations in the General Discussion.

Experiment 1 was designed to evaluate the prediction that between-category comparisons will lead to higher classification performance when learners can rely on feature values to determine category membership. This experiment was also an attempt to replicate Kornell and Bjork’s (2008) finding using explicit comparisons rather than inferring the role of comparisons from performance differences based on item order. Experiment 2 tested the prediction that when learners need to focus on more complicated information, such as the relational structure of categories, they will benefit more from within-category comparisons than between-category comparisons. In Experiments 3 and 4, these predictions were addressed again using very different, real-world materials, slightly different learning tasks, and in Experiment 3, a within-subjects design, for the purposes of replication and generalization to a variety of different tasks and materials.

Finding in the first four experiments that the benefits of each type of comparison are dependent on the type of information one needs to learn, I test a stronger prediction made by the highlighter theory in Experiment 5. Specifically, this experiment tests the prediction that the order of items learners are presented with interacts with whether or not learners are given opportunities to explicitly make comparisons, leading to differences in performance.
To summarize, the set of studies reported below will look at the effects of between-category and within-category comparisons on learning in order to evaluate the highlighter theory of comparison learning. Different types of categories are used throughout these experiments, chosen based on whether learners can rely on individual feature values or must rely on the relational structure of items to determine category membership. This category-type distinction is used to test the claim that each type of comparison highlights different category information, leading to benefits of one type of comparison or the other depending on the information learners need to acquire.
CHAPTER 2: EXPERIMENTS 1 AND 2

The goal of Experiments 1 and 2 was to demonstrate for the first time that within-category and between-category comparisons affect category learning differently and that the benefit of each type is dependent on the kind of information that needs to be attended to during learning.

2.1 Experiment 1

Experiment 1 was designed to replicate the finding that spacing enhances category learning (Kornell & Bjork, 2008) by explicitly asking participants to make comparisons between items. The categories used in this experiment were chosen because participants could rely on the feature values alone to determine category membership, a property that I argue Kornell and Bjork’s (2008) materials also had.

The categories used in this experiment were two artificial categories of aliens, called Deegers and Koozles, which were similar to the artificial categories used in the majority of category-learning studies. The categories were determined by family resemblance (Rosch & Mervis, 1975), meaning the category members are like members of one’s family in that they share some but not all features with each other. In the most common family-resemblance-category scheme, there are two categories, and each category is represented abstractly by assigning a 1 or a 0 to each of the features and assigning an “A” or a “B” to the category labels (see Table 1). Each item presented during learning shares features with its prototype, with the exception of one feature, known as the exception feature. The exception feature is typically the opposite category’s prototypical value. For example, the prototypes of two categories, which contain four diagnostic features and two non-diagnostic features (non-diagnostic because they

---

5The prototype is essentially the best category member, as it contains more of the category’s typical features than other category members. Here, the prototype has the category’s typical value for each feature.
are not predictive of either category but are prototypical for both categories), are A11111 and B000011 (the diagnostic features are italicized and the non-diagnostic features are bolded). During learning, participants are presented with items such as A110111, where the “0” would be considered the exception feature. At test, a larger spectrum of items is shown that deviate from the prototype either in terms of the number of diagnostic features (e.g. A110110 would have typical values for three out of four diagnostic features) or the number of non-diagnostic (but still typical) features (e.g. A110110 would have typical values for four out of six features—three diagnostic features and one non-diagnostic feature).

Participants learned about two categories of aliens by either comparing two items from the same category or one item from each category across a number of trials. After learning, participants performed a classification test that included items they had seen during learning as well as new items. It was predicted that participants who learned through between-category comparison would have higher classification performance than those who learned through within-category comparison for both old and new items.

Method

Design

Participants were randomly assigned to one of the two between-subjects conditions: between-category-comparison learning or within-category-comparison learning. Two different combinations of features were used and were counterbalanced across participants.

Participants

Participants were 43 undergraduates from the University of Illinois who participated for course credit. Three participants did not follow instructions, so only 40 participants were considered in the analyses below.
**Materials**

Stimuli for both the learning and test phases were two types of fictitious aliens, called Deegers and Koozles, which varied along six binary features: arms, tails, antennae, legs, eyes, and mouth. For each feature, there was a prototypical Deeger value and a prototypical Koozle value. The categories were defined by family resemblance.

The items shown during learning always had five out of six features in common with the category prototype and one feature in common with the other category’s prototype to reflect variation among category members. Four of the features were diagnostic of category membership, meaning one value appeared 75% of the time in items in one category and only 25% of the time in items in the other category, while the two remaining features were consistently the same value across both categories’ items. This resulted in four items per category (see Table 1). The diagnosticity of the features was counterbalanced across participants, as noted above. For half of the participants, the two features that did not vary across items during learning were the mouth and tails. For the other half of the participants, the two features that did not vary were the legs and antennae.

The two features that did not vary during learning were varied at test to generate additional items. The prototypes of each category were also shown during the classification test. This resulted in 32 unique items (16 per category), which were all shown during the classification test (see Table 2). Of the 32 items, 8 were old (meaning they were shown during the learning phase) and 24 were new.
Procedure

After giving informed consent, participants were told that they would be learning about two different categories of aliens, Deegers and Koozles, and would later be tested on what they had learned.

On each trial during the learning phase, participants were shown two items along with their category labels (see Figure 1 for an example screenshot). The side of the screen each item was placed on (either right or left) was randomly determined. At the top of the screen, the between-category-comparison learners saw the prompt: “List how this Deeger [Koozle] and this Koozle [Deeger] are the same and different.” The order of the category labels always corresponded with the side of the screen each type of item was on (e.g. if a Deeger was on the left then the Deeger label was mentioned first). The within-category-comparison learners saw the prompt: “List how these Deegers [Koozles] are the same and different.” To make the category labels salient, the items and their labels were presented using different colors. Koozles were always presented in red and Deegers were always presented in blue.

Participants typed out the similarities and differences between the two items in the response box provided and then clicked on a button to submit the response and move to the next trial. There were no constraints on how much or how little participants had to write for each response and participants could take as long as they needed. There were 24 comparison trials, randomly ordered for each participant. For within-category-comparison learners, there were 12 unique comparisons that could be made, and each was presented twice. For between-category-comparison learners, there were 16 unique comparisons, so for each participant, half of the possible comparisons were randomly selected a second time (with the constraint that each item was presented an equal number of times).
After the learning phase, participants performed a classification test consisting of both old and new items. On each trial, they saw one alien and two category labels. Unlike in the learning phase, all of the information on the screen for this task was presented in black (the item as well as the labels). The participant’s job was to choose the correct category label for the alien by clicking on the appropriate box using the mouse. Participants could take as long as they needed to respond. After making a selection, the next trial appeared after a 500 ms delay. There were 32 trials and items were presented in a random order for each participant. Finally, participants were debriefed and thanked for their time. On average, the experiment took between 25-40 minutes.

Results

*Learning Phase*

Comparisons were coded based on the number of similarities mentioned in a comparison, the number of differences mentioned, the number of features mentioned overall, and the proportion of trials where category labels were used.\(^6\) In their responses, participants sometimes mentioned features that were not part of the intended category scheme and did not differ across items (e.g. body shape, head shape). For coding purposes, these were still counted as part of the number of features mentioned overall, number of similarities, and number of differences.

The properties of the comparison statements by condition are presented in Table 3. Only one property of the comparison statements is worth noting. Between-category-comparison learners were more likely to list differences between items \((M = 2.46 \text{ differences}, SD = 0.62)\) than within-category-comparison learners \((M = 2.03 \text{ differences}, SD = 0.13)\), \(t(38) = 3.09, p < 0.05\).

---

\(^6\) In subsequent analyses of comparisons (in Experiment 2 and Experiment 4), the proportion of trials where general statements were made about the items was also coded. This property was not coded in this study because there were not any real general statements made in participants’ comparisons. Comparisons were typically just lists of features that were similar and different.
0.05. This is most likely because there were more differences on average between items across trials in the between-category-comparison condition.

Test Phase

The central question is whether between-category comparisons highlight information that distinguishes between categories more effectively than within-category comparisons, as measured by higher classification performance. As predicted, between-category-comparison learners had higher overall classification accuracy ($M = 0.70$, $SD = 0.16$) than within-category-comparison learners ($M = 0.60$, $SD = 0.16$), $t$ (38) = 2.16, $p < 0.05$. Both groups performed above chance (0.5), $t$ (19) = 5.77, $p < 0.05$ and $t$ (19) = 2.65, $p < 0.05$ respectively.

When classification performance was broken down by item type, the same benefit for between-category comparisons was observed. Between-category-comparison learners had higher classification accuracy for only old items ($M = 0.69$, $SD = 0.18$) than within-category-comparison learners ($M = 0.58$, $SD = 0.18$), $t$ (38) = 2.05, $p < 0.05$. Additionally, there was a marginal effect of comparison type for only new items, with between-category comparison-learners displaying higher accuracy ($M = 0.71$, $SD = 0.17$) than within-category-comparison learners ($M = 0.60$, $SD = 0.17$), $t$ (38) = 1.94, $p = 0.06$.

Discussion

Experiment 1 demonstrated that different types of comparisons, specifically within-category and between-category comparisons, affect what people learn about categories. Between-category-comparison learners showed higher classification performance than within-category-comparison learners. These results are consistent with Kornell and Bjork’s (2008) finding that categories whose members are spaced between other category’s members during learning are learned better than categories whose members are massed together.
There are multiple explanations for the results of Experiment 1, and here I acknowledge the three most plausible ones. First, between-category comparisons are always better than within-category comparisons. Second, any time the learning goal is to acquire information that distinguishes categories, between-category comparison will be the most effective way to learn that information. Third, between-category comparisons are effective here because of the types of categories that were learned, and with other types of categories, within-category comparisons could be more beneficial for learning. Experiment 2 was designed to test these possibilities.

2.2 Experiment 2

Experiment 1 showed a benefit for between-category comparisons. However, there is also preliminary evidence that within-category comparisons could be helpful for learning. Kurtz and Hovland (1956) showed that participants learned more when items from the same category were massed together. In addition, there is evidence from the problem-solving and analogical-reasoning literatures showing that comparing two items that share the same relational structure can be helpful (e.g. Gentner, et al., 2003; Namy & Gentner, 2002).

Experiment 2 followed up on the predictions made by the structure-mapping framework (Gentner, 1983). Specifically, it tested the prediction that learners will benefit from structural alignment between two items when the relational structure of the categories is difficult to notice and learners need to use that information to determine category membership. According to structure-mapping theory, comparing two items of the same type highlights the relational structure they share (Gentner, 1983).

This experiment used the same materials and tests from Rehder and Ross (2001) and Erickson, et al. (2005). The materials were two categories of machines, called Morkels and Krenshaws, defined by whether or not the feature values of an item made sense together (i.e.
form a coherent machine) or not. The same feature values could appear in either category, making it difficult to abstract the structure of the categories by solely focusing on the individual values (see Appendix A).

Rehder and Ross (2001) and Erickson, et al. (2005) both used a novel classification test to see if learners had acquired the categories. In this test, participants were shown items that had completely new feature values and had to classify them. What makes this set of materials so powerful is that if participants do not learn what determines category membership (whether or not the feature values make sense together), then they will be at chance when trying to classify these new items. If they figure out the abstract coherence-based rule that determines category membership, it should be relatively easy for them to classify the new items even though the feature values are unfamiliar.

If within-category comparisons focus learners on what each category is like, then this is similar to the benefit that inference learning provides. Erickson, Chin-Parker, and Ross (2005) showed that for these abstract, coherence-based categories, participants who learned through inference learning were better able to classify novel items at test than participants who learned through classification. This was an atypical finding, because for many studies that compared classification and inference, classification learners tended to perform better on full-feature classification tests (e.g. Anderson, et al., 2002; Yamauchi & Markman, 1998). Erickson, et al. (2005) argue that the reason for the inference benefit is that inference learning focuses the learner on what each category is like and how the features of the items fit together.

In this experiment, participants learned the two coherence-based categories through either within-category or between-category comparisons. On each learning trial, two items, each missing one feature, were presented one at a time along with their category label. Participants
inferred the missing feature and received feedback for each item independently, and then compared the two items they had just seen\textsuperscript{7}. The two items were either from the same category or different categories depending on the learning condition the participant was assigned to. After learning, participants performed two classification tests that assessed participants’ relational knowledge. In the first test, participants were presented with novel items that they had to classify. The feature values used to construct these items were completely new and unfamiliar, so the only way to successfully classify these items was to use the abstract coherence-based relation. In the second test, participants were provided with two out of the three features of an item they had studied during the learning phase and had to classify the item based on only those features. This test assessed participants’ knowledge of relations as well; however, to be successful in this task, participants could have either used specific relational knowledge (i.e. their knowledge of feature correlations within items they encountered during the learning phase) or the abstract coherence-based relation. The prediction was that participants who learned through within-category comparisons would perform better on both types of classification tests than participants who learned through between-category comparisons.

Method

Design

Participants were randomly assigned to one of the two between-subjects conditions: between-category-comparison learning or within-category-comparison learning.

Participants

Participants were 28 undergraduates from the University of Illinois who participated for course credit.

\textsuperscript{7} Based on pilot data, the decision was made to add these two inference trials, as these categories were difficult for participants to learn by just doing the comparison part of the trial. Adding in this additional opportunity for actively processing the items boosted performance, allowing us to see important comparison differences.
**Materials**

The materials were the same as those used in studies by Rehder and Ross (2001) and Erickson, et al. (2005). They were two categories of machines, called Morkels and Krenshaws. There were different stimuli used during the learning and test phases of the experiment. During learning, there were three items per category. Each item had three features and each feature had three potential values: where it operated (land, water, the stratosphere), what it picked up (harmful solids, spilled oil, dangerous gaseous ions), and what tool it used to do its task (a shovel, a sponge, an electrostatic filter). For each item, the feature values were always presented in the same order (where it operated, what it picked up, what it used to do its task). Each value for each feature could appear across items from either category.

What determined each item’s category was whether or not the three features made sense together, as determined by Rehder and Ross (2001). If the three features made sense together (e.g. operates on land, works to gather harmful solids, has a shovel), it was a Morkel. If the three features did not make sense together (e.g. operates on land, works to clean spilled oil, has an electrostatic filter), it was a Krenshaw (see Appendix A the complete list of category items).

For the novel classification test, a completely new set of Morkels and Krenshaws were used (also from Rehder & Ross, 2001, and Erickson, et al., 2005). Each item had the same three features, but the values were ones that were never seen during learning (e.g. operates on the beach, works to remove broken glass, has a sifter). There were six new items per category (See Appendix B for a complete list of the items).

For the feature-pair classification test, feature values from the learning phase were used to form 18 unique feature pairs (9 per category).
Procedure

After giving informed consent, participants were told that they would be learning about two different categories of machines, Morkels and Krenshaws and would later be tested on what they had learned. They were also told that the same features of the machines could appear across both types of machines and that two machines of the same type may have different features.

During the learning phase, participants compared items 18 times, and each comparison consisted of multiple parts: two inference trials (where an item was missing a feature and participants had to figure out the best value for the missing feature) and a trial where participants listed similarities and differences between the two items they had just seen during the inference trials. There were three items per category, each with three features, creating a total of nine unique inference trials per category. Each of the 18 unique inference trials was presented twice during learning in a random order, and each time it was randomly paired with another inference trial. For the between-category-comparison condition, the other inference trial had to be from the other category. For the within-category-comparison condition, the other inference trial had to be from the same category. The order of the two inference trials was randomly determined. To emphasize the category labels, the items and the category labels were presented in matching colors across both the inference trials and the comparison (Morkels were always blue, Krenshaws were always magenta).

For each inference trial, participants were presented with an item that had two out of three of its features and its label. Below the item there were three possible values for the missing feature. Participants had to click on what they thought was the missing feature of the presented item. They could take as long as they wanted to respond. After making a choice, they were presented with feedback in the form of “Correct!” (in green) or “Incorrect” (in red) along with
the complete item (and with the previously missing feature highlighted in either green or red depending on whether they were correct or incorrect). They could study the feedback as long as they wanted and had to press the spacebar to continue. There was a 500 ms delay before the next inference trial started.

After completing two inference trials, participants were presented with the two items they had just seen, side-by-side (the comparison part of the trial was set up in the same way as it was in Experiment 1). The side of the screen an item was placed on during the comparison (either right or left) was randomly determined. In the between-category-comparison condition, participants were prompted to “List how this Morkel [Krenshaw] and this Krenshaw [Morkel] are similar and different.” The order of the category labels always corresponded with the side of the screen each type of item was on (e.g. if a Morkel was on the left then the Morkel label was mentioned first). In the within-category-comparison condition, participants were prompted to “List how these Morkels [Krenshaws] are similar and different.” Participants typed out the similarities and differences between the two items in the response box provided and then clicked on a box to submit the response and move to the next trial. There were no constraints on how much or how little participants had to write for each response and participants could take as long as they needed to respond. Once participants responded, there was a 500 ms delay before the next comparison started.

After learning, participants completed two different tests, which were always in the same order: novel classification followed by feature-pair classification. The order remained the same across participants because the novel classification test was less likely to interfere with feature-pair classification performance.
For the novel classification test, participants were presented with items one at a time in the center of the screen (items were in black) along with the two category labels (which remained in the same colors they were presented in during learning). Participants clicked on the label that they thought went with the presented item. They did not receive any feedback on their choices. There were 12 trials presented in a random order for each participant.

For the feature-pair classification test, participants were presented with only two features at a time and based on just those features they had to determine if the item was a Morkel or a Krenshaw. Like the novel classification test, participants were presented with an item in the center of the screen (items were in black) and they had to click on the label that they thought went with the pair of features (labels were in the same colors they were presented in during learning). The feature values that were used were the same as those used in the learning phase. There were 18 trials, presented in a random order for each participant. Finally, participants were debriefed and thanked for their time. On average, the experiment took between 25-40 minutes.

Results

Learning Phase

Inference Performance. Inference performance was analyzed separately for the inferences made during the first 9 comparison trials (Block 1) and the inferences made during the second 9 comparison trials (Block 2) to look at learning over time. If within-category comparisons are more beneficial for learning categories requiring learners to notice their relational structures, then within-category-comparison learners should acquire the categories more quickly than between-category-comparison learners. Even in Block 1, within-category-comparison learners performed better on inference trials ($M = 0.65$, $SD = 0.18$) than between-category-comparison learners ($M = 0.49$, $SD = 0.16$), $t (26) = 2.65$, $p < 0.05$. Both groups performed above chance ($t$
within-category-comparison learners continued to perform better on the inference trials ($M = 0.82, SD = 0.29$) than between-category-comparison learners, who barely improved from their Block 1 performance ($M = 0.54, SD = 0.29$), $t(26) = 3.04, p < 0.05$. Once again, both groups performed above chance ($t(13) = 9.66, p < 0.05$ and $t(13) = 2.52, p < 0.05$ respectively).

**Comparison Statements.** Comparisons were coded based on the number of similarities mentioned in a comparison, the number of differences mentioned, the number of features mentioned overall, the proportion of trials where category labels were referred to, and whether or not any general statements were made about one or both items.

General statements were defined as any statement where the items were discussed as a whole or where features were discussed in a more abstract way than simply their surface values. For example, one participant wrote, “…Both are well-equipped to do the task” when comparing two Morkels and wrote, “…None of the properties within one Krenshaw fit together” when comparing two Krenshaws. Another participant wrote, “…Both Morkels are well suited for the environment they operate in.” These were all considered general statements. Statements that were not considered general were most commonly just lists of surface feature similarities and differences between the items (e.g. “One operates on land and the other operates in water”).

The properties of the comparison statements by condition are presented in Table 4. Only two properties of the comparisons statements are worth noting. One, between-category-comparison learners mentioned labels on a higher proportion of trials ($M = 0.60, SD = 0.48$) than within-category-comparison learners ($M = 0.24, SD = 0.34$). Two, within-category-comparison learners made general statements about the items on a marginally higher proportion of trials ($M$...
= 0.46, SD = 0.31) than between-category-comparison learners (M = 0.21, SD = 0.38), t (26) = 1.87, p = 0.07.

The finding that between-category-comparison learners used labels in more of their responses than within-category-comparison learners is likely because within-category-comparison learners did not need to distinguish between the categories in their responses. The major difference between between-category-comparison and within-category-comparison conditions was the finding that within-category-comparison learners used general statements about the items more often. This finding is consistent with the idea that within-category comparisons help learners see beyond the individual surface features to determine the common underlying structure of the items.

Test Phase

Novel Classification Test Performance. The central question was whether learning through within-category comparisons led to higher novel classification performance than learning through between-category comparisons. It was predicted that within-category comparisons would be more helpful because they highlight the common relational structure shared across items. As predicted, within-category-comparison learners were significantly more accurate in classifying novel items (M = 0.70, SD = 0.24) than between-category-comparison learners (M = 0.52, SD = 0.17), t (26) = 2.16, p < 0.05. Additionally, within-category-comparison learners performed above chance (0.5), t (13) = 3.02, p < 0.05, while between-category-comparison learners did not, t (13) = 0.62, p > 0.05.

Feature-Pair Classification Test Performance. As predicted and consistent with the results of the novel classification test, within-category-comparison learners were significantly more accurate in classifying pairs of features they had seen during learning (M = 0.81, SD = 0.23)
than between-category-comparison learners \( (M = 0.53, SD = 0.22) \), \( t (26) = 3.30, p < 0.05 \).

Additionally, within-category-comparison learners performed above chance \( (0.5) \), \( t (13) = 4.94, p < 0.05 \), while between-category-comparison learners did not, \( t (13) = 0.48, p > 0.05 \).

**Discussion**

Across two different classification tests, within-category-comparison learners performed better than between-category-comparison learners. In order to acquire these coherence-based categories, learners had to look beyond the individual surface features of the items to determine their relational structures. These findings are consistent with what structure-mapping theory would predict, which is that within-category comparisons highlight structural commonalities across items. In addition, it is consistent with the theoretical framework proposed here, which predicts that when learners need to focus on the relational structures of the categories, they will benefit more from within-category comparisons because within-category comparisons highlight the common relational structure across items.

The results of this experiment, in combination with the results of Experiment 1, demonstrate that the type of category one is learning about determines the type of comparison that is most helpful for learning. These two very different effects were predicted by the highlighter theory of comparison learning. The major distinction between the categories used in Experiment 1 and the categories used in Experiment 2 was whether or not learners could rely on feature values or more complicated category information (relations between features) to determine category membership.
CHAPTER 3: EXPERIMENTS 3 AND 4

Experiments 1 and 2 demonstrated that the type of comparison affects what people learn. In addition, it seems that the type of comparison that is helpful is dependent on the type of category one is learning about. The theoretical framework proposed here argues that these results are due to between-category comparisons highlighting distinguishing feature-value information between categories and within-category comparisons highlighting the relational structure of items.

Experiments 3 and 4 were designed to provide more evidence for this theory. If the same effects can be demonstrated when using slightly different tasks and, more importantly, when using complex real-world categories, then this will further increase the plausibility of this theoretical framework.

3.1 Experiment 3

To make the claim that between-category comparisons are more beneficial when learners can rely on feature values alone, the benefit of between-category comparisons needs to be demonstrated across a wide range of stimuli that share this characteristic. It was also important to demonstrate the effect with a complex, real-world set of materials, so in Experiment 3, participants learned about real-world categories of birds. While I acknowledge that families of birds have relational structures that one could learn, this information is not necessary to determine category membership for the stimuli used here (and focusing on feature values is sufficient).

In addition, in Experiments 1 and 2, participants were explicitly told to compare across items. When learning in the real world, however, a learner is not always prompted to compare items. One question that arises is whether comparison, and more specifically the type of
comparison, matters when it is incidental to the task and no explicit analyses of similarities and differences are required. Experiment 3 addresses this question. Participants learned about real-world categories of birds by performing a modified same/different judgment task. Participants were presented with a bird and its label and then asked to choose the label for another bird. The two label choices included one that was the same as the first bird’s label and one that was different. Rather than responding *same* or *different*, learners chose either the same label or a different label, making comparison helpful, but not necessary for the task.

Finally, if it is the case that spacing out items is beneficial because it enables comparison (e.g. Kornell & Bjork, 2008), then it might also be the case that presenting the two items to be compared on the same screen should enhance this benefit, compared with presenting the two items one after the other as in the spacing effect studies. To test this idea, items appeared either simultaneously or sequentially across participants.

Participants learned about six different families (i.e. categories) of birds: three in a way that encouraged between-category comparisons and three in a way that encouraged within-category comparisons. They either learned about the bird families by viewing two items simultaneously or sequentially. To determine whether one type of comparison was more beneficial, as well as whether comparison benefits are enhanced when items are side-by-side, all participants performed a novel classification test after learning. Novel classification is a typical means of assessing category knowledge, as it addresses the degree to which participants are able to generalize what they have learned. It was predicted that participants would perform better on a novel classification test for bird families learned in a context encouraging between-category comparisons than those learned in a context encouraging within-category comparisons.
Method

Design

Unlike in the previous two experiments, in this experiment there was a within-subject manipulation of comparison type across all participants: out of six bird families participants had to learn, three were learned in a way that encouraged between-category comparisons and three were learned in a way that encouraged within-category comparisons.

Participants were randomly assigned to one of the two between-subjects conditions: either all items were presented on the screen in pairs (pairs condition) or they were all presented one at a time (singles condition). In addition, the six bird families (i.e. categories) were divided into two groups of three and assignment of bird family to comparison condition (within-category or between-category) was counterbalanced across participants, resulting in 2 counterbalancing conditions. In addition, there were two different list orders counterbalanced across participants. Finally, assignment of bird family to list position was also counterbalanced using a Latin-squares design, resulting in three counterbalancing conditions. In total there were 12 counterbalancing conditions.

Participants

Participants were 48 undergraduates from the University of Illinois who participated for course credit.

Materials

Stimuli for both the learning and test phases were color images of birds from six families in the Passeriformes order, compiled by Jacoby, Wahlheim, and Coane (2010) from www.whatbird.com. The six bird families were finches, flycatchers, swallows, thrushes, vireos, and warblers (See Figure 2). For each family, six unique items were used in the study phase and
Six unique items were used for the novel classification test (assignment of item to either the study or test phase was the same across all participants).

As mentioned above, the assignment of bird family to learning condition was counterbalanced across participants. For half of the participants, finches, vireos, and thrushes were learned through within-category comparisons while swallows, flycatchers, and warblers were learned through between-category comparisons (and vice-versa for the other half).

There were two different learning sequences generated with various constraints to ensure that the bird families learned through between-category and within-category comparisons were adequately spaced between other bird families’ items. This was to ensure that the only difference between the two conditions was what participants had compared within a trial rather than how spaced or massed the items were across trials. The order was constrained such that the number of other families spaced between presentations of items from one family was similar across the between-category and within-category conditions. An additional constraint was that the sequence of within-category and between-category trials had to be unpredictable, so that the classification responses participants were making were not predictable by the previous trial’s correct response. Finally, for bird families in the between-category-comparison condition, an item from each family had to appear equally often with one from each of the other bird families in that condition. Assignment of bird family to comparison condition and assignment of bird family to list position were counterbalanced across participants, resulting in 12 unique lists.

While the ordering of bird families was pre-determined, the actual items that were presented were randomly drawn from the list of all six items from that family. Across both conditions, each item appeared two times.
Procedure

After giving informed consent, participants were told that they would be learning about six different families of birds and would later be tested on what they had learned. Before beginning the learning phase of the experiment, participants were asked to rate their knowledge (on a 1-7 scale, where a one means very little to no knowledge and a seven is a significant amount of knowledge) of each bird family. This was to make sure that no participants were included who were already familiar with the bird families used in the experiment. All participants gave low ratings for all bird families, so this aspect of the experiment is not discussed further.

On each trial during the learning phase, participants in the pairs condition were presented with two birds. Below the bird on the left side of the screen was its label (e.g. “This is a Finch”) and below the bird on the right side was a prompt (e.g. “Is this a Finch or a Warbler?”). The participant’s job was to determine which of two labels went with the bird picture on the right side of the screen and click on that label. One option was always the same label as the left-side bird’s label. Essentially participants were making a same/different judgment, only they chose either the same or a different category label instead (see Figure 3 for an example screenshot). Participants were never told that they should use the other bird or compare across items.

For bird families in the within-category-comparison condition, the correct answer to the prompt was always the same label as the label for the bird on the left. For bird families in the between-category-comparison condition, the correct answer was always the label that did not match the label for the bird on the left. After making a response, the participant received feedback in the form of “Correct!” or “Incorrect”. Regardless of the correctness of the
participant’s response, the two bird pictures along with their correct labels and the corrective feedback remained on the screen for 12 seconds.

The only major difference for participants in the singles condition was that the bird pictures were presented one at a time. First, participants viewed a bird and its label (e.g. “This is a Finch”) for eight seconds. Next, participants viewed a different bird and a prompt (e.g. “Is this a Finch or a Warbler?”). The participant had to determine which of two labels went with the bird, and one option was always the same label as the preceding bird’s label. Feedback was similar to feedback in the pairs condition, except that only one bird was shown with its label and the feedback remained on the screen for six seconds. As in the pairs condition, participants were never told that they should use the other bird picture or make comparisons.

There were 36 learning trials (where a learning trial in the singles condition was the combination of the first item where the label was given and the second item where the label was asked about).

After learning, participants made family level category-learning judgments (CLJs), where they predicted how they would perform on a novel classification test for each bird family. Participants used a number pad and predicted their performance on a scale of 16% (chance) to 100%. There were six trials, one for each bird family.

Next, participants performed a novel classification test. On each trial, they were given an item and had to classify it by clicking on one of the six labels presented on the screen. After choosing a label, participants indicated how confident they were in their selection on a scale between 16% (chance) and 100%. There were 36 trials (six items were presented from each bird family) and they were presented in a different random order for each participant. Finally,
participants were debriefed and thanked for their time. On average, this experiment took 30-40 minutes.

Results

Learning Performance

The main learning finding is that learners who studied items in pairs had marginally higher learning accuracy than learners who studied items one at a time; however, comparison type did not affect learning accuracy. A 2x2 mixed factorial ANOVA showed that overall learning accuracy was marginally higher for pairs learners (M = 0.75, SD = 0.10) than singles learners (M = 0.70, SD = 0.08), F (1, 46) = 3.51, p = 0.07. There was no main effect of comparison type on learning accuracy (between-category items: M = 0.73, SD = 0.13 and within-category items: M = 0.71, SD = 0.12), F (1, 46) = 0.758, p > 0.05.

There was also an interaction between presentation type and comparison type, F (1, 46) = 4.13, p < 0.05, so follow up paired t-tests were run. For singles learners, between-category-comparison trials were marginally more accurate (M = 0.73, SD = 0.14) than within-category-comparison trials (M = 0.67, SD = 0.10), t (23) = 1.73, p = 0.10. For pairs learners, there was no difference between comparison types (between-category-comparison trials: M = 0.74, SD = 0.11; within-category-comparison trials: M = 0.76, SD = 0.12), t (23) = 1.00, p > 0.05.

Novel Classification Test Performance

The main goal of this study was to see if between-category comparison was more beneficial for category learning even when it was incidental to the task. A secondary goal was to see if the method of presentation—pairs or singles—affected the degree to which participants effectively compared across items. A 2x2 mixed factorial ANOVA showed a main effect of comparison type. Classification accuracy for novel items in categories that were learned through
between-category comparisons ($M = 0.42, SD = 0.18$) was higher than it was for items in
categories learned through within-category comparisons ($M = 0.36, SD = 0.19$), $F (1, 46) = 4.88,$
$p < 0.05$. There was a marginal main effect of presentation type. Pairs learners were marginally
more accurate ($M = 0.43, SD = 0.16$) than singles learners ($M = 0.35, SD = 0.16$), $F (1, 46) = 2.83,$
$p = 0.10$. There was no interaction between presentation type and comparison type, $F (1, 46) = 0.70,$
$p > 0.05$, suggesting that learners benefited from between-category comparisons
regardless of whether they viewed the items simultaneously or sequentially (means and standard
deviations by condition are presented in Table 5). All metacognition results are reported in
Appendix C because any differences seem to be driven by the observed accuracy differences, so
the ratings did not provide any additional, meaningful results in relation to the questions of
interest here.

Discussion
In a task where comparisons were incidental to the learning task and specific analyses of
similarities and differences across items were not required, the type of comparison was still
critical for learning. After learning, birds from families that were learned in a situation that
encouraged between-category comparisons were classified more accurately than birds from
families learned in a situation that encouraged within-category comparisons. This was true
regardless of whether items were presented one at a time or in pairs.
There was a marginal benefit to studying items in pairs rather than one at a time. While
learners could have noticed that items were paired in the singles condition (because trials
alternated between learners being provided a label for the item and learners classifying the item),
it would have been more difficult for them to make effective comparisons than in the pairs
condition. While the finding that between-category comparisons were more beneficial than
within-category comparisons for both the singles and pairs conditions suggests that the order of items matters, the marginal benefit of the pairs condition over the singles condition demonstrates that having the ability to compare across items, regardless of item order, is beneficial for learning about items whose category membership can be determined by individual feature values.

The main result, that novel birds from families learned through between-category comparisons were classified better than novel birds from families learned through within-category comparisons, provides more evidence for the explanation that between-category comparisons are beneficial when learners can rely on feature values to determine category membership.

3.2 Experiment 4

Experiment 4 evaluated the explanation that within-category comparisons are helpful for learning when learners need to focus on items’ common relational structure. Learners were taught about two different mathematical concepts, permutations and combinations. Participants either compared within-category (two permutations problems) or between-category (a permutations problem and a combinations problem) during learning. After learning, participants classified novel problems by determining which formula was appropriate. For another set of novel problems, learners were given the correct formula and had to appropriately apply it to solve the problem. The prediction was that within-category comparisons would be more beneficial than between-category comparisons, as measured by higher classification and problem-solving performance.
Method

Design

Participants were randomly assigned to one of the two between-subjects conditions: within-category-comparison learning or between-category-comparison learning. Within each learning condition, the order of the problems being compared was counterbalanced across participants, resulting in four different counterbalancing conditions.

Participants

Participants were 12 undergraduates from the University of Illinois who participated for course credit.

Materials

Four problems, two permutations and two combinations problems, were written for the learning phase (see Appendix D). Within each problem type, one problem referred to animate objects (e.g. students) while the other referred to inanimate objects (e.g. clothing). Regardless of the between-subjects manipulation, an animate-object problem was always paired with an inanimate-object problem for a comparison to ensure that the values of the surface features could not serve as a relevant cue for problem type for either comparison condition.

Four problems, two of each type, were written for the classification test. Another four, two of each type, were written for the problem-solving test (see Appendix D). An equal number of animate-object and inanimate-object problems were written for each problem type and each test.

Procedure

After giving informed consent, participants were told that they would be learning about two different categories of math problems and would later be tested on what they had learned.
Each participant was given a packet that included seven pages and a cover sheet. On the first page after the cover sheet was a sheet that detailed how to solve a particular class of problems (either permutations or combinations; see Appendix E). On the second page was a problem of that type. Participants had to identify the variables and apply the formula, but did not have to do the arithmetic. Once satisfied, they could turn the page and see the solution. This process repeated either for a problem of the same type, if the participant was in the within-category-comparison condition, or for a problem of a different type if the participant was in the between-category-comparison condition. After completing the second problem, the last page had the following prompt: “List the similarities and differences between the problems and their solutions.” Participants could write as much or as little as they wanted and take as long as they needed to get through the entire packet. When finished, the participant turned in the packet and received a second packet, which was in the same format. Across the two packets there were a total of four problems, which resulted in two comparisons (one per packet).

After learning, participants were given another packet that included the classification test followed by the problem-solving test (presented in the same order for each participant, as the classification test was of primary interest). For the classification test, participants were shown four novel problems, one per page, and had to circle which of the two formulas (permutations or combinations) was the correct one for each problem. The formulas were appropriately marked with a $P$ or a $C$, so even if the participant had not memorized the formulas, he or she should have been able to recognize which was which. The second test was the problem-solving test. Participants were shown four novel problems, one per page, along with the appropriate formula. Their task was to solve the problem up to the arithmetic. They could take as much time as they needed. For each test, the problems were randomly ordered for each participant. Finally,
participants were debriefed and thanked for their time. On average, this experiment took 30-40 minutes.

Results

*Comparison Statements*

Comparisons were coded based on the number of similarities mentioned in a comparison, the number of differences mentioned, the number of features mentioned overall, the proportion of trials where category labels were referred to, and whether or not any general statements were made about one or both items.

General statements were defined as any statement where the items as a whole were referred to or where abstract statements where made about the features of the problems. It is important to point out that this way of coding ended up capturing two types of statements, and in the future, these should be coded more finely (but due to the small number of participants, it did not make sense to do this here). One type of general statement that occurred most frequently was simply a general comment about the problems that did not provide any explicit details about the problems’ structures but suggested that they may have been considering structure. For example, one participant commented, “Both require the same set of steps to get to a solution.” Another participant noted, “There are different types of variables to keep in mind for each type of problem.” The other type of general statement was one that provided more specific details about the problems’ structures. Only one participant used this type of statement throughout the comparisons. For example, this participant commented, “In problem 1 it was necessary to know the number of objects being chosen, but in problem 2, it was necessary to know the number of choices you are interested in.” The current coding system did not distinguish between these two
types of general statements because there was such a small number of the second type (as they really only occurred in one participant’s data).

The properties of the comparison statements by condition are presented in Table 6. Across all aspects of the comparison statements, there were no statistically significant differences between conditions. However, while there was no statistically significant difference in the proportion of trials where general statements were made about the items (between-category comparison; \( M = 0.25, SD = 0.45 \) and within-category comparison; \( M = 0.50, SD = 0.27 \)), the trend in the data is in the direction consistent with the idea that within-category comparisons shifts learners’ focus away from the specific features of each problem and toward the structure of the problem. It is likely that the lack of a significant effect is due to the small number of observations in each condition as well as the large amount of variability across participants.

**Test Phase**

**Classification Test.** If within-category comparisons highlight the common relational structure of the items being compared, then this type of comparison should be more beneficial than between-category comparisons for learning about probability problems whose surface features vary and whose structure is difficult for learners to acquire. As predicted, within-category-comparison learners had higher classification accuracy (\( M = 0.83, SD = 0.13 \)) than between-category-comparison learners (\( M = 0.54, SD = 0.25 \)), \( t (10) = 2.57, p < 0.05 \). In addition, within-category-comparison learners performed above chance (0.5), \( t (5) = 6.32, p < 0.05 \), while between-category-comparison learners did not, \( t (5) = 0.42, p > 0.05 \). This finding suggests that when categories have items whose structure is difficult for learners to determine, within-category comparisons are more beneficial for acquiring conceptual knowledge.
**Problem-Solving Test.** Within-category-comparison learners showed better (though not reliably better) problem solving ability ($M = 0.91$, $SD = 0.13$) than between-category-comparison learners ($M = 0.75$, $SD = 0.27$), $t(10) = 1.35$, $p > 0.05$. For this test, participants were given the formula and asked to plug in the appropriate values for each variable. Given the small number of participants and the ceiling effects for the within-category-comparison learners, it is not surprising that we did not find a statistically significant difference here, though the effect was in the predicted direction.

**Discussion**

In Experiment 4, participants who compared problems of the same type were better able to classify novel problems during the test phase than participants who compared different types of problems. This finding is consistent with the results of Experiment 2, and provides more evidence for the idea that within-category comparisons are helpful under circumstances where learning the relational structures of the categories is necessary.

The lack of effect in the problem-solving test was likely due to ceiling effects in the within-category-comparison condition. The problem-solving test was relatively easy compared to the novel classification test. Participants were given a problem and the equation they should use and their task was to simply write out the solution based on the values given in the problem. Despite these ceiling effects, the trend in the data is consistent with the difference found in the novel classification test between within-category and between-category-comparison learners.
CHAPTER 4: INTERIM DISCUSSION OF EXPERIMENTS 1-4

Previous work looking at between-category and within-category comparisons produced mixed results, making it difficult to determine the effect of each type of comparison on learning. Across Experiments 1-4, which used a diverse set of tasks and materials, there were clear consistencies between the types of comparisons that were beneficial for learning and the types of categories participants had to acquire. These results provide a clearer picture of how each type of comparison affects learning.

In Experiments 1 and 3, learners could rely on feature values to determine category membership for aliens and birds respectively. In Experiment 1, learners either wrote down similarities and differences between two items from the same category or two items from different categories. In Experiment 3, learners engaged in a modified same/different task that encouraged, but did not require, learners to make comparisons. Across both experiments, between-category-comparison learners displayed higher classification performance than within-category-comparison learners.

In Experiments 2 and 4, learners had to rely on more complicated information, and specifically they had to rely on the relational structures of the items, to determine category membership for coherence-based machines and probability problems respectively. In both experiments, learners wrote down similarities and differences between two items from the same category or two items from different categories after interacting with those items one at a time (either by predicting a missing feature or solving a problem). Across both experiments, within-category-comparison learners demonstrated higher classification performance (and higher feature-pairs classification performance in Experiment 2) than between-category-comparison learners.
The results of Experiments 1-4 are consistent with the highlighter theory of comparison learning, which is that between-category comparisons highlight feature value differences across categories while within-category comparisons highlight commonalities and the relational structure of categories (and one or the other will be beneficial for learning depending on the type of information one needs to learn). It is important to point out that the highlighter theory proposed in this thesis as well as the results of these four experiments were not put together in a post-hoc fashion. Structure-mapping theory predicted within-category-comparison benefits (Gentner, 1983) and prior work on the spacing effect suggested that between-category comparisons are helpful for category learning (Kornell & Bjork, 2008).
CHAPTER 5: EXPERIMENT 5

To understand how comparisons affect learning, it is important to look at effects of different types of comparisons, but it is also critical to understand how each type of comparison affects learning relative to circumstances where no comparisons are made. Additionally, it is not clear from the results of Experiments 1-4 whether particular orders of items are enough to generate comparison effects or whether learners must be explicitly asked to make comparisons.

In Experiment 3, where category membership could be determined by attention to feature value differences, these issues were addressed by including a singles condition, where learners were shown items one at a time, and a pairs condition, where learners were shown two items side-by-side. Even though both the singles and pairs learners benefitted more from the between-category-comparison item order than the within-category-comparison item order, seeing the items side-by-side was marginally more helpful than seeing items one at a time, suggesting the importance of making comparisons (rather than simply having a particular order of items) when learners can rely on feature values to determine category membership. Across Experiments 1-4, there was no corresponding experiment that addressed how each comparison type affects learning relative to a no-comparison condition with items whose category membership requires attention to relational structure. Experiment 5 was designed to address this question.

Experiment 5 more finely examined the effects of each type of comparison on learning by considering between-category-comparison learning and within-category-comparison learning relative to learning without comparisons (where items were presented one at a time in an order yoked to either the between-category-comparison or within-category-comparison condition) under circumstances where learners need to acquire the categories’ relational structures.
When learners do not know the categories’ relational structures at the beginning of learning, making between-category comparisons may hinder their ability to acquire the items’ relational structure relative to doing no comparisons at all, as between-category comparisons may lead learners to focus on the wrong information (the feature values). Additionally, while it is clear from Experiments 2 and 4 that within-category comparisons are more helpful than between-category comparisons when learning categories that require knowledge of relational structure, it is not clear from those experiments whether prompting learners to make within-category comparisons is more helpful than not prompting learners to make comparisons but providing them with the same item order. There is evidence suggesting the importance of prompting people to make comparisons when the relational structure of the items is difficult to learn (e.g. Doumas & Hummel, 2004; Gentner, et al., 2003; Ross, 1984, 1987). On the other hand, order effects studies (e.g. Kurtz & Hovland, 1956; Kornell & Bjork, 2008) suggest that learners may make comparisons on their own, so it is possible that each no-comparison condition will mimic the comparison condition its item order is based on.

In this experiment, the same coherence-based categories from Experiment 2 were used and participants were randomly assigned to one of four learning conditions: between-category-comparison learning (BC learning), no-comparison learning with the item order yoked to the between-category-comparison condition (BN learning), within-category-comparison learning (WC learning), or no-comparison learning with the item order yoked to the within-category-comparison condition (WN learning).

After learning, relational knowledge was assessed through two tests: novel classification and feature-pairs classification. Both the novel classification test and the feature-pairs classification test assess relational knowledge, but each assesses it at a different level of
specificity. For both tests, focusing on individual feature values will not be helpful because, during learning, each feature value appears equally often across both categories. It is the relations between the features that matter for category membership. Therefore, in order to be successful on either or both of these tests, participants must focus on how feature values go together, rather than on individual feature values. In the novel classification test, participants are presented with items made up of completely new and unfamiliar feature values. The only way to successfully classify these items is to understand that the feature values of items from one category make sense together while the feature values of items from the other category do not make sense together. On the other hand, in order for participants to be successful on the feature-pairs classification test, participants could either use the abstract, coherence-based relation or more specific relational knowledge (i.e. knowledge of feature correlations that occur within items from the learning phase; for example, knowing that when the feature values operates on land and has a shovel appear together, the item is in the Morkel category). If participants display high feature-pairs classification performance but lower novel classification performance, it suggests that participants are relying to some degree on their knowledge of specific feature correlations, rather than the abstract coherence-based relation, to classify the feature-pairs.

The prediction was that if within-category comparisons highlight items’ common relational structure in a way that learners would not be able to do on their own, then WC learning should lead to higher feature-pairs and novel classification performance than WN learning. In addition, if between-category comparisons actually draw learners’ attention away from the items’ relational structure, then BC learning should lead to lower levels of feature-pairs classification and novel classification performance than BN learning. Finally, if the order of items encourages learners to make comparisons to some degree, then of the learners who are not given
opportunities to explicitly compare, WN learning should lead to higher performance on the feature-pairs and novel classification tests than BN learning.

Method

Design

The design of the experiment was 2 (within-category or between-category item order) x 2 (comparison or no-comparison), resulting in four between-subjects conditions: between-category-comparison learning (BC learning), no-comparison learning with the item order yoked to the between-category-comparison condition (BN learning), within-category-comparison learning (WC learning), or no-comparison learning with the item order yoked to the within-category-comparison condition (WN learning). Participants were randomly assigned to one of the four conditions.

Participants

Participants were 60 undergraduates recruited from the University of Illinois, who participated for course credit.

Materials

The materials were the same as those used in Experiment 2.

Procedure

The procedure closely followed the procedure from Experiment 2. The major difference was that participants performed 36 comparisons (rather than 18 as in Experiment 2). This number was chosen in an attempt to decrease the likelihood that floor effects would decrease the chances of observing differences between conditions.

Participants in the BC and WC conditions compared items 36 times, and as in Experiment 2, each comparison consisted of multiple parts: two inference trials (where an item
was missing a feature and participants had to figure out the best value for the missing feature) and a trial where participants listed similarities and differences between the two items they had just seen. The inference trials were set up in the same way as they were in Experiment 2. After making a response for each inference trial, the feedback screen (which included feedback in the form of “Correct!” in green or “Incorrect” in red along with the complete item, with the previously missing feature highlighted in either green or red depending on whether the response was correct or incorrect) was displayed for 2010 ms (the average amount of time participants spent studying this feedback screen in a pilot experiment). As in Experiment 2, after two inference trials, participants were shown the two items they had just seen and were told to list similarities and differences between them (between-category-comparison learners saw: “List how this Morkel [Krenshaw] and this Krenshaw [Morkel] are similar and different” and within-category-comparison learners saw: “List how these Morkels [Krenshaws] are similar and different”). Participants could take as long as they needed to write out their comparisons.

Participants in the BN condition performed 72 inference trials in an order that was yoked to the BC condition (for every two items, they saw one from each category). Participants in the WN condition performed 72 inference trials in an order that was yoked to the WC condition (every two items were from the same category). The inference trials were exactly the same as they were for each of the two comparison conditions; however, to equate for the amount of time the BC and WC learners were exposed to the items while writing their comparisons, the no-comparison learners viewed each inference trial’s feedback screen for 17956 ms, which was 2010 ms plus half of the total time that it took for comparison learners to write their comparisons (as determined by a pilot experiment).
As in Experiment 2, after the learning phase, participants completed a novel classification test followed by a feature-pair classification test. For the novel classification test, participants were presented with items one at a time in the center of the screen (items were in black) along with the two category labels (which remained in the same colors they were presented in during learning). Participants clicked the label that they thought went with the presented item and did not receive any feedback on their choices. There were 12 randomly ordered trials. For the feature-pairs classification test, participants were presented with two features they had seen during learning and had to determine based on only those features if the item was a Morkel or a Krenshaw. Like the novel classification test, participants were presented with an item in the center of the screen in black font and had to choose the label they thought went with the pair of features (labels were in the same colors they were presented in during learning). There were 18 randomly ordered trials. Finally, participants were debriefed and thanked for their time. On average, the experiment took between 40-50 minutes.

Results

*Learning Phase*

*Inference Performance.* Inference performance was analyzed separately for the inferences made during the first 18 comparison trials (Block 1) and the inferences made during the second 18 comparison trials (Block 2) to look at learning over time. If within-category comparisons are more beneficial for learning categories requiring learners to notice their relational structures, then WC learners should show higher inference trial accuracy than WN learners. If between-category comparisons lead learners to focus on the wrong information, then BC learners should show lower inference trial accuracy than BN learners. If the order of items encourages learners
to make comparisons even when they are not prompted to do so, then WN learners should show higher inference trial accuracy than BN learners.

Table 7 displays Block 1 accuracy by condition. Table 8 displays Block 2 accuracy by condition. There was no difference between the WC and WN conditions or the BC and BN conditions; however, there was a main effect of item order, where learners in both within-category item order conditions showed higher accuracy on the inference trials across both blocks of learning than learners in both between-category item order conditions.

A 2 (within-category or between-category item order) x 2 (comparison or no comparison) ANOVA was run on Block 1 inference accuracy and showed a main effect of item order, $F(1, 56) = 4.80, p < 0.05$. Learners who saw items in the within-category order, regardless of whether they were given an opportunity to compare or not, performed more accurately on inference trials ($M = 0.67, SD = 0.16$) than learners who saw items in the between-category order ($M = 0.58, SD = 0.17$). There was no main effect of comparison, $F(1, 56) = 0.36, p > 0.05$, and no interaction between item order and comparison, $F(1, 56) = 0.00, p > 0.05$.

Planned independent-sample t-tests showed that there were no significant differences in performance between WC and WN learners, $t(28) = 0.14, p > 0.05$, or BC and BN learners, $t(28) = 0.13, p > 0.05$. Additionally, there was no significant difference in performance between WN learners ($M = 0.67, SD = 0.19$) and BN learners ($M = 0.57, SD = 0.18$), $t(28) = 1.42, p > 0.05$.

A 2 x 2 ANOVA was run on Block 2 inference accuracy and showed a main effect of item order, $F(1, 56) = 7.98, p < 0.05$. Learners who saw items in the within-category order, regardless of whether they were given an opportunity to compare or not, performed more accurately on inference trials ($M = 0.87, SD = 0.18$) than learners who saw items in the between-
category order ($M = 0.71, SD = 0.24$). There was no main effect of comparison, $F (1, 56) = 0.50, p > 0.05$, and no interaction between item order and comparison, $F (1, 56) = 1.57, p > 0.05$.

Planned independent-sample t-tests showed that there were no significant differences in performance between WC and WN learners, $t (28) = 0.45, p > 0.05$, or BC and BN learners, $t (28) = 1.23, p > 0.05$. Additionally, there was no significant difference in performance between WN learners ($M = 0.86, SD = 0.22$) and BN learners ($M = 0.77, SD = 0.22$), $t (28) = 1.09, p > 0.05$.

**Comparison Statements.** Comparisons were coded based on the number of similarities mentioned, the number of differences mentioned, the number of features mentioned overall, the proportion of trials where category labels were referred to, whether or not any abstract statements were made about one or both items, and the proportion of trials where learners made general rather than specific statements about features.

Abstract statements were defined as any statement where the items were discussed in an abstract way (e.g. “…Both [Morkels] are well-equipped to do the task.” and “…Both Morkels are well suited for the environment they operate in.”). General statements were defined as any statement made about a particular feature in a general way (e.g. “This Morkel uses a different tool from the other Morkel” rather than something more specific like “This Morkel uses a shovel while the other Morkel uses a sponge”). If learners made a general statement about one or more features during a comparison, that trial was coded as general. If both general and specific statements were made about the same feature on a trial, the general statement did not count (as it was clear that the learner did focus on that feature’s specific feature values for that particular trial).
The properties of the comparison statements by condition are presented in Table 9. There are two properties of the comparison statements that are worth noting. One, WC learners used general statements on a marginally higher proportion of trials \((M = 0.51, \, SD = 0.42)\) than BC learners \((M = 0.24, \, SD = 0.37)\), \(t (28) = 1.81, \, p = 0.09\). This finding is consistent with the highlighter theory’s claim that between-category comparisons focus learners on feature value differences (leading these learners to make more specific rather than general comparison statements than the within-category learners). Two, BC learners used category labels on a higher proportion of trials \((M = 0.56, \, SD = 0.41)\) than WC learners \((M = 0.14, \, SD = 0.16)\), \(t (28) = 3.65, \, p < 0.05\). This is likely because BC learners had to differentiate between two categories in their responses while WC learners did not.

Test Phase

The two most important results for this experiment are novel classification and feature-pairs classification performance. Here I present the feature-pairs classification test first, as it is more useful to think of these tests in terms of the level of specificity they are testing rather than the order in which they were performed during the experiment. Across both tests, the central questions are (1) whether within-category-comparison learners show more relational knowledge than yoked, no-comparison learners and (2) whether between-category-comparison learners show less relational knowledge than yoked, no-comparison learners. It was predicted that within-category comparisons would be more helpful for learning relational structure because they highlight the common structure shared across items while between-category comparisons would actually be detrimental to learning relational structure because they focus learners on distinguishing feature values (and focusing on feature values here would not help with learning the categories). Additionally, it was predicted that if the order of items encourages learners to
make comparisons even when they are not prompted to do so, and if within-category comparisons more effectively highlight the relational structure of categories, then WN learners should show higher inference trial accuracy than BN learners.

*Feature-Pair Classification Test Performance.* The feature-pairs classification test assessed participants’ knowledge of specific feature correlations from the items they had studied during the learning phase. Table 10 displays feature-pairs accuracy by condition. The predictions were supported: WC learners outperformed WN learners, and BN learners outperformed BC learners. Additionally, there was an effect of item order, where learners who were given the within-category item order outperformed learners who were given the between-category item order.

A 2 x 2 ANOVA was run and showed a main effect of item order. Learners who saw items in the within-category order, regardless of whether they were given an opportunity to compare or not, were better at classifying feature pairs \( (M = 0.90, SD = 0.18) \) than learners who saw items in the between-category order \( (M = 0.70, SD = 0.20) \), \( F (1, 56) = 18.63, p < 0.05 \).

There was no main effect of comparison, \( F (1, 56) = 0.05, p > 0.05 \). Importantly, there was an interaction between item order and comparison, \( F (1, 56) = 8.04, p > 0.05 \).

Follow-up independent t-tests were run and showed that BC learners classified feature pairs worse \( (M = 0.63, SD = 0.19) \) than BN learners \( (M = 0.76, SD = 0.21) \), \( t (28) = 2.01, p < 0.05 \), while WC learners classified feature pairs marginally more accurately \( (M = 0.96, SD = 0.12) \) than WN learners \( (M = 0.84, SD = 0.22) \), \( t (28) = 1.93, p = 0.06 \). Additionally, WN learners performed 8% higher than BN learners, though this difference was not statistically significant, \( t (28) = 0.92, p > 0.05 \). All conditions performed above chance (BC: \( t (14) = 2.75, p < 0.05 \); BN: \( t (14) = 4.87, p < 0.05 \); WC: \( t (14) = 15.42, p < 0.05 \); WN: \( t (14) = 6.11, p < 0.05 \)).
**Novel Classification Test Performance.** The novel classification test assessed participants’ knowledge of the abstract, coherence-based relation that defined the two categories. Table 11 displays novel classification accuracy by condition. Learners who were given an item order that encouraged within-category comparisons performed significantly better on the novel classification test than learners who were given an item order that encouraged between-category comparisons. The predicted interaction, that WC learners would outperform WN learners and BC learners would perform worse than BN learners, was not found, likely due to floor effects in the BC condition and due to no-comparison learners making comparisons to some degree across trials. These issues are discussed in further detail below.

A 2 x 2 ANOVA showed a main effect of item order, $F(1, 56) = 11.01, p < 0.05$. Learners who saw items in the within-category order, regardless of whether they were given an opportunity to compare or not, were better at classifying new items ($M = 0.77, SD = 0.21$) than learners who saw items in the between-category order ($M = 0.58, SD = 0.23$). There was no main effect of comparison, $F(1, 56) = 0.38, p > 0.05$, and no interaction between item order and comparison, $F(1, 56) = 0.41, p > 0.05$.

Planned independent t-tests were run. There was no statistically significant difference between BC learners ($M = 0.54, SD = 0.25$) and BN learners ($M = 0.62, SD = 0.22$), $t(28) = 0.84, p > 0.05$, even though BN learners performed 8% better than BC learners on the test. It is likely that this comparison was heavily impacted by floor effects, as BC learners’ performance was not statistically different from chance, $t(14) = 0.68, p > 0.05$ (and 73% of participants in the BC condition had a proportion of correct responses of less than 0.58). Without the floor effects, it is possible, given the 8% difference in accuracy, that BN learners would have reliably outperformed BC learners on this task.
The real mystery of the novel classification performance data was the lack of difference between WC learners \((M = 0.77, SD = 0.23)\) and WN learners \((M = 0.77, SD = 0.19)\), \(t(28) = 0.00, p > 0.05\). Prior research has demonstrated the benefits of making within-category comparisons relative to not making comparisons (e.g. Gentner, et al. 2003; Doumas & Hummel, 2004). On the other hand, work by Kurtz and Hovland (1956) suggests that learners can make use of item order to make effective within-category comparisons. More broadly, the growing body of work on order effects and category learning (e.g. Kornell & Bjork, 2008; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010) suggests that item order may be a critical component for encouraging learners to make comparisons. Therefore, the most plausible explanation for why there was no difference between WC and WN learners on the novel classification test is that WN learners were making comparisons across trials. Evidence that WN learners may have been taking advantage of the order of items comes from the finding that WN learners classified novel items significantly better than BN learners, \(t(28) = 2.10, p = 0.05\). The only explanation for the difference between these two no-comparison conditions was the order of the items.

All learners with the exception of BC learners (as noted above) performed above chance (BC: \(t(14) = 0.68, p > 0.05\); BN: \(t(14) = 2.10, p = 0.05\); WC: \(t(14) = 4.49, p < 0.05\); WN: \(t(14) = 5.54, p < 0.05\).

*Follow-up Analysis of WC Learners.* It was striking to see that the predictions were supported in the feature-pair classification test, but not in the novel classification test. These results suggest two possibilities, both based on the idea that the level at which items are analyzed determines the level of abstractness a learner is able to develop. One possibility is that WC learners were focusing on the specifics of the items they had learned about due to the nature of the comparison prompt (which asked learners to list similarities and differences between the items). This focus
on the specifics of the items may have caused them to learn the feature correlations within the items better than the other learning groups, but also to develop less abstract category representations than they would have otherwise. A second possibility is that WN learners were focusing more generally on the items than WC learners because WN learners had to rely on their imperfect memory of the items to make comparisons (making it difficult for them to focus on specific feature values, leading to worse performance on the feature-pairs classification test). One or both of these could have occurred.

The first possibility can be addressed with these results; however, the second one cannot be tested with the data from this experiment, as it would require some measure of WN learners’ memory across the learning trials. To test the first possibility, I looked at WC learners’ novel classification performance as a function of the degree to which they made general versus specific statements about the items they were learning about. If focusing on more general information about the items leads to better abstraction, then learners who made more general statements about the items in their comparisons will show higher novel classification performance.

WC learners were sorted into groups based on the proportion of trials where they made general (rather than specific) statements about features (where 1.0 indicates that a participant always made general statements and 0.0 indicates that a participant never made general statements). Figure 4 displays the distribution of participants. As Figure 4 shows, there were three different types of WC learners: those who almost always made general statements, those who sometimes made general statements, and those who rarely made general statements (and made specific statements about feature values instead). The four learners who sometimes made general statements were not included in this analysis. The prediction was that learners who focused on general similarities and differences between items would show higher novel
classification performance (as evidence that they had abstracted the coherence relation) than learners who focused on specific similarities and differences between the items.

The five learners who almost always made general statements were compared to the six learners who almost always made specific statements. Participants in the general-comparison group made significantly more general statements ($M = 0.99$, $SD = 0.01$) than the participants in the specific-comparison group ($M = 0.04$, $SD = 0.03$), $t (9) = 69.78$, $p < 0.05$. Due to ceiling effects across both groups, there was no significant difference in feature-pairs classification accuracy between the general-comparison group ($M = 1.00$, $SD = 0.00$) and the specific-comparison group ($M = 0.92$, $SD = 0.17$), $t (9) = 1.15$, $p > 0.05$. Importantly, participants in the general-comparison group performed significantly better on the novel classification test ($M = 0.98$, $SD = 0.04$) than participants in the specific-comparison group ($M = 0.68$, $SD = 0.23$), $t (9) = 2.83$, $p < 0.05$. The general-comparison WC learners performed significantly better than the WN learners, $t (18) = 2.41$, $p < 0.05$, but there was no reliable difference between the specific-comparison WC learners and the WN learners, $t (19) = 0.92$, $p > 0.05$.

It is possible that the observed difference in novel classification performance was simply because the general-comparison WC learners had learned the categories immediately and their comparison statements reflected that knowledge. In order for this explanation to be plausible, it would mean that general-comparison learners had picked up the abstract, coherence-based rule at the beginning of learning (because they all made general statements on almost 100% of the trials). Therefore, if learners’ comparison statements were a reflection of learners’ knowledge of the abstract coherence-based rule rather than a reflection of the way the learners approached the comparison task, then the general-comparison WC learners should show higher accuracy in the

---

8 A similar effect is shown even if the 4 participants are included in the analysis and the two comparison groups are determined by a median split: participants in the general-comparison group performed marginally better on the novel classification test ($M = 0.88$, $SD = 0.20$) than participants in the specific-comparison group ($M = 0.65$, $SD = 0.22$), $t (13) = 2.00$, $p = 0.07$. 

75
first block of the learning task than the specific-comparison learners. On the other hand, if the learner’s approach to the comparison task affected their ability to learn the abstract, coherence-based rule, then there should be no difference in accuracy between general-comparison and specific-comparison learners in the first block of the learning task. General-comparison learners’ Block 1 accuracy ($M = 0.63, SD = 0.13$) was similar to specific-comparison learners’ Block 1 accuracy ($M = 0.63, SD = 0.16$), $t(9) = 0.04$, $p > 0.05$, consistent with the idea that the way the learner approaches the comparison affects his or her ability to learn the categories.

This analysis provides some evidence that the benefits of within-category comparisons may depend on the way the learner approaches the comparison. It does not rule out the possibility that WN learners were making more general comparisons due to imperfect memory for the specifics of the items, but it does suggest that part of the reason there was no difference between WC and WN learners on the novel classification test was due to the way that some of the WC learners approached the comparison task.

Discussion

In Experiment 5, between-category and within-category comparisons were considered relative to two no-comparison conditions whose item orders were yoked to each of the comparison conditions’ item orders. Most importantly, the predicted interaction between item order and whether or not learners were prompted to compare was observed in the feature-pairs classification test. This test assessed learners’ knowledge of specific relational information (i.e. the feature correlations they had observed in the items they had learned about). WC learners outperformed WN learners, demonstrating that there are performance benefits for learners who make within-category comparisons when relational structure must be highlighted. BC learners performed significantly worse than BN learners. It is typically assumed that making
comparisons should help learning; however, here I demonstrate a case where making comparisons was actually detrimental to learning. This is likely because between-category comparisons led learners to focus on the wrong information (feature-value differences across items, which were not as helpful here without an understanding of relational structure). To the best of my knowledge, this is the first time that learning through comparisons was shown to be detrimental to performance relative to learning without comparisons.

Additionally, across the learning phase, novel classification test, and feature-pairs classification test, learners who saw items in a within-category item order throughout learning outperformed learners who saw items in a between-category item order throughout learning, regardless of whether learners were prompted to make comparisons or not. These results are consistent with the results of Experiments 2 and 4 in that they demonstrate the benefits of within-category comparisons relative to between-category comparisons for categories in which the relational structure must be used to determine category membership.

These findings provide strong evidence in favor of the highlighter theory of comparison learning. Between-category and within-category comparisons highlight different information about categories. When relational structure needs to be learned, within-category comparisons effectively highlight the relevant information while between-category comparisons lead learners to focus on the wrong information.

The assumption, especially in the analogical-reasoning literature, has been that two is better than one—that comparing is better than not comparing at all. This assumption held up when only within-category comparisons were considered relative to no comparisons (e.g. Doumas & Hummel, 2004; Gentner & Namy, 1999; Gentner, et al., 2003) and when there was no manipulation of the degree to which learners could compare (e.g. Kornell & Bjork, 2008).
However, the results of Experiment 5 show that comparison effects are more complicated than they initially seemed. There are circumstances where making comparisons will lead learners to focus on the wrong information and leave them worse off than they were had they not made any comparisons in the first place.

On the other hand, the novel classification test showed a main effect of item order, but no interaction between item order and whether or not learners made comparisons. Floor effects most likely made it difficult to see a reliable difference in novel classification performance between BC learners and BN learners (the proportion of correct responses on this test was 0.58 or lower for 73% of BC learners and 60% of BN learners). Additionally, BN learners performed 8% better than BC learners on the task, suggesting that the predicted difference may have shown up if floor effects had not been an issue. In contrast, the feature-pairs classification test was easier, so floor effects were less likely to occur (the proportion of correct responses on this test was 0.56 or lower for 40% of BC learners at 33% of BN learners), and the predicted difference was observed between the BC and BN conditions.

The surprising result was the lack of a difference between WC learners and WN learners on the novel classification test. This result is not consistent with prior work showing within-category comparison benefits relative to learning without comparisons (e.g. Doumas & Hummel, 2004; Gentner, et al. 2003; Namy & Gentner, 1999). The most plausible explanation is that WN learners took advantage of the order of items to make within-category comparisons. Items were displayed for a long amount of time (17956 ms) to ensure that time was not confounded across the comparison and no-comparison conditions. Therefore, it would have been easy for WN learners to compare two back-to-back items, as their memory for previously viewed items would have been strong. Evidence that these learners may have been comparing comes from
considering the novel classification performance difference between the two no-comparison conditions, where the only difference between those conditions was the order of items (the order was either yoked to the within-category-comparison item order—the WN condition—or the between-category-comparison item order—the BN condition). WN learners performed 15% better than BN learners on the novel classification test, consistent with the idea that WN learners were taking advantage of the order of items to make effective comparisons.

However, it must be more complicated than this everybody-made-comparisons explanation, as this explanation would predict that there would be no difference between the WC and WN conditions in the feature-pair classification test. First, it is important to point out that these two tests assess different types of relational knowledge. The feature-pairs classification test assesses item-specific relational knowledge (though participants could have used abstract knowledge to perform this task) while the novel classification test assesses abstract relational knowledge. As I briefly explained in the Experiment 5 results section, there are two plausible explanations for seeing different patterns of results across the two tests. Both explanations assume that the way the learner approaches the comparison is what determines the level of abstractness he or she achieves. The first possibility is that the prompt used to elicit comparisons in the WC condition (“List the similarities and differences between these Morkels [Krenshaws]”) was focusing learners on the specifics of the items. This focus may have led the WC learners to learn the feature correlations that occurred within the items more easily than learners in the other conditions, but also to focus less on the abstract relational structure of the categories than they would have without the prompt. The second possibility is that the WN learners were making comparisons across items by relying on their memory for the item they had just seen and comparing it to the item they were currently viewing. By making comparisons in this way, these
learners may have been thinking more generally about the items than the WC learners, as they most likely had imperfect memory for the specifics of each item.

One or both of these explanations is possible, though it is impossible to address the second one with the data from this experiment. The first explanation was addressed by focusing on individual differences across WC learners. WC learners who made general statements on a higher proportion of comparison trials during learning were more likely to abstract the coherence relation and were better able to classify novel items. This finding suggests the importance of considering how the learner approaches a comparison and suggests that instructing learners to think more broadly about items when making within-category comparisons may lead them to abstract the relational structure more easily (and may lead to differences between the WC and WN conditions in the novel classification test).

In summary, this experiment demonstrated both the benefits and hindrances of making different types of comparisons during learning. The finding that learning through between-category comparisons leads to worse performance relative to learning without making comparisons (but seeing items in the same order) when relational structure must be learned is novel and provides strong evidence in favor of the highlighter theory. The finding that learning through within-category comparisons leads to better performance relative to learning without making comparisons (but seeing items in the same order) suggests the importance of providing learners with the opportunity to explicitly compare across items (at least when it comes to learning specific relational information). In addition, the effect of individual differences observed within the WC condition suggests an important direction for future work, which I will discuss along with other future directions in more detail below.
CHAPTER 6: GENERAL DISCUSSION

Comparisons have been suggested as central to category learning (e.g. Spalding & Ross, 1994), yet very few studies have actually addressed the role of comparisons in learning categories. Recent work demonstrates that the type of active processing one engages in during learning affects what is learned (e.g. Anderson, et al., 2002; Chin-Parker & Ross, 2002, 2004; Jones & Ross, 2011; Yamauchi & Markman, 1998), suggesting the importance of considering how different types of comparisons influence what information learners focus on and acquire.

6.1 Summary of experiments

I proposed the highlighter theory of comparison learning, which was used as a framework for generating the predictions made for the set of experiments reported here. The major claim of this theory is that each type of comparison highlights different category information. Between-category comparisons highlight feature-value differences between categories. Within-category comparisons highlight commonalities as well as the relational structure shared across items from the same category. Evidence consistent with the highlighter theory was demonstrated in Experiments 1-5, using a wide range of materials and tasks.

In Experiments 1 and 3, participants learned about items whose category membership could be determined by feature values alone. Experiment 1 used artificial categories of aliens while Experiment 3 used real-world bird categories. Across both experiments, between-category comparisons were more beneficial than within-category comparisons, as demonstrated by classification performance. In Experiments 2 and 4, participants learned about items whose category membership required attention to the relational structure of the categories. Experiment 2 used artificial, coherence-based machine categories while Experiment 4 used two types of real-world probability problems. Across both experiments, classification performance was higher for
participants who learned through within-category comparisons than for those who learned through between-category comparisons. For both types of categories, comparison effects were demonstrated with artificial categories, where there was a significant amount of control over the characteristics of the items and the category schemes, and then with real-world categories that I argue had the same properties as their artificial-category counterparts (either feature values or relational structure had to be learned in order to determine category membership).

Experiments 3 and 5 more finely considered the benefits of between-category and within-category comparisons by comparing each type to a condition where comparisons were less likely to occur. In Experiment 3, learners either saw one or two items at a time during learning. Presenting two items at a time resulted in marginally higher performance on the novel classification test than presenting one item at a time (regardless of item order), suggesting that when learners can rely on individual feature values to determine category membership, both within-category and between-category comparisons are beneficial to some degree.

In Experiment 5, learners were either given a chance to make comparisons or not when learning about coherence-based machines. Participants who were not explicitly asked to make comparisons were shown items one at a time in an order yoked to either the between-category-comparison or within-category-comparison condition. For the feature-pairs classification test, a measure of learners’ specific-item relational knowledge, learners who made within-category comparisons outperformed learners who were not asked to make comparisons (but saw the same item order), and learners who made between-category comparisons performed worse than learners who were not asked to make comparisons (but saw the same item order). When determining category membership requires attention to the relational structure of items, making between-category comparisons actually interferes with learners’ ability to focus on relational
structure, and making within-category comparisons enhances learners’ ability to focus on relational information. To my knowledge this is the first time that comparisons have been shown to negatively affect learning. The assumption has always been that comparisons are helpful and prior work has consistently supported this view. However, these studies either always considered within-category comparisons relative to no comparisons (e.g. Doumas & Hummel, 2004; Gentner & Namy, 1999; Gentner, et al., 2003) or never manipulated the degree to which learners could compare (e.g. Kornell & Bjork, 2008).

Even though the results from the novel classification test did not show the same interaction, there are two reasons to believe that this was due to the no-comparison learners making comparisons. One, there was a difference in performance between the two no-comparison conditions on the novel classification test (where learners who saw items in the same order as the within-category-comparison learners performed better than learners who saw items in the same order as the between-category-comparison learners), and two, there was a main effect of item order across both the novel classification and feature-pairs classification tests (learners who saw a within-category item order outperformed learners who saw a between-category item order, regardless of whether or not they were prompted to compare). Both of these effects are best explained by assuming that the no-comparison learners used the item orders they were provided with to make either between-category or within-category comparisons.

Lastly, in Experiment 5 a post-hoc analysis of individual differences across within-category-comparison learners’ comparison statements revealed an interesting pattern of results. Learners who made more general comparison statements were more likely to acquire the abstract, coherence-based relation that determined category membership than learners who made more specific comparison statements, as evidenced by novel classification accuracy. This is
preliminary evidence that the success of a comparison not only depends on what is being compared, but also on how the learner approaches the comparison. This idea is discussed further in the future directions section of this chapter.

6.2 Alternative explanations

Despite the predictiveness of the highlighter theory, there are a number of alternative accounts of these data. These alternative accounts do not challenge the idea that different types of comparisons lead to differences in what is learned, but instead challenge the way that the different types of categories used in these experiments were distinguished. Below I will address three of the most plausible alternative accounts. The first account distinguishes the types of categories in terms of within-category and between-category similarity, the second distinguishes them in terms of the degree to which learners already have relevant prior knowledge, and the third distinguishes them in terms of difficulty.

6.2.1 Similarity explanation

One alternative explanation for the results reported here is the similarity explanation. This explanation is motivated by a study by Carvalho and Goldstone (presented at the Annual Meeting of the Psychonomic Society, 2011), which looked at the effects of different item orders—interleaving and blocking—on learning, under the assumption that interleaving encourages between-category comparisons while blocking encourages within-category comparisons. Participants learned categories of blobs that had a single defining feature. The blobs either shared many other features with other blobs, both within and across categories (high similarity), or few other features, both within and across categories (low similarity). They demonstrated that when within-category and between-category similarity were both low (i.e.

---

9 Interleaving is their terminology for *spacing* (i.e. presenting an item from one category, followed by an item from a different category) while blocking is their terminology for *massing* (i.e. presenting an item from one category followed by an item from the same category).
blobs within a category and across categories were perceptually very different from each other), blocking items from a category together led to higher categorization performance than interleaving items from one category between those of other categories. When within-category and between-category similarity were both high (i.e. blobs within a category and across categories were perceptually similar to each other), interleaving led to better categorization performance than blocking.

Carvalho and Goldstone’s (2011) explanation for these results is that within-category comparisons focus learners on commonalities, so they are more helpful when there are relatively few common feature values across items from a category. Between-category comparisons focus learners on differences, so they are more helpful when very few feature values differ across categories.

Another way of phrasing Carvalho and Goldstone’s (2011) explanation is to say that when there is a critical feature that is difficult to see, the most helpful comparison is determined by whether it becomes easier to notice that feature when focusing on within-category commonalities or between-category differences. In their study, there was always just one critical feature that was difficult for learners to notice. The degree to which between-category or within-category comparisons were effective in helping learners notice that feature depended on whether or not there was high or low similarity among the other features across the items being learned. A major challenge with attempting to extend this explanation to the studies reported here is that it is unclear whether it extends beyond a simple, defining feature category scheme. In the experiments reported here, the categories were more complex. There were always multiple features that mattered, so learning the categories was not simply about whether or not learners could notice one feature in particular.
Another issue with Carvalho and Goldstone’s (2011) results and their proposed explanation is that it is unclear how between-category and within-category similarity interact with each other. In their study, between-category and within-category similarity were always confounded (if between was high, so was within and if between was low, so was within). The similarity explanation offers no framework for making predictions under circumstances where between-category similarity is high and within-category similarity is low, or vice versa.

Finally, it is not clear that the similarity explanation would have predicted between-category comparison benefits in Experiment 3. In Experiment 3, participants learned about six categories of birds, and while I do not have a systematic analysis of the similarity space of the bird categories, there did seem to be a large amount of variation within the categories. For instance, warblers drastically varied in color, markings, and size (e.g. one was yellow and black and plump while another was light brown and relatively skinny). Despite significant variation between birds from the same family, between-category comparisons were more helpful for learning these categories.

In future work, I intend to look more closely at cases where the similarity explanation and the highlighter theory make different predictions. There are two ways I intend to do this. First, I will pit the relations versus feature values manipulation against the similarity manipulation. For example, I could systematically vary the within-category and between-category similarity of the bird families and look at this manipulation’s effect on the benefits of each type of comparison. The highlighter theory would predict between-category-comparison benefits across all manipulations of similarity while the similarity explanation would predict within-category-comparison benefits when within-category similarity is low and between-category-comparison benefits when between-category similarity is high.
Second, I will vary the within-category similarity of the probability problems from Experiment 4 (and keep the between-category similarity between the two problem types high across the similarity manipulation). With this design between-category and within-category similarity are unconfounded, but more importantly, the predictions made for the highlighter theory and similarity explanation are different when between-category and within-category similarity are both high. The highlighter theory would predict within-category comparison benefits regardless of the similarity manipulation while the similarity explanation would predict between-category comparison benefits when within-category similarity is high.

6.2.2 Prior-knowledge explanation

Another alternative explanation for the interaction between comparison type and category type is the prior-knowledge explanation\(^\text{10}\). The idea is that the amount of prior knowledge one has about the categories to be learned determines the benefits of each type of comparison. Prior knowledge may lead learners to develop expectations about the relevance of particular features or relations and may help them constrain their attention to information that is important. Once learners know what information is important, between-category comparisons will be effective because learners will already know what alignable differences they should be paying attention to. Consequently, this explanation predicts that between-category comparisons will be helpful when learners already have a significant amount of prior knowledge while within-category comparisons will be helpful when learners have very little prior knowledge of the categories being learned (or when their prior knowledge leads them to establish the wrong expectations).

The explanation regarding how within-category and between-category comparisons are beneficial is still based on the idea that between-category comparisons highlight differences across categories while within-category comparisons highlight commonalities. By highlighting

---

\(^\text{10}\) I thank Andrei Cimpian for pointing out this possibility.
commonalities, within-category comparisons help learners determine which features are important for category membership. Therefore, when a learner has very little prior knowledge (or if a learner’s prior knowledge causes him or her to have the wrong expectations about the categories to be learned), within-category comparisons will be more helpful. On the other hand, between-category comparisons are only helpful once learners have determined which features are important. Determining the important features can either occur through activation of relevant prior knowledge or through engaging in within-category comparisons. Once learners figure out the important features, between-category comparisons will be more beneficial for learning to effectively distinguish between categories, as they highlight feature-value differences between categories.

This prior-knowledge explanation could explain the Experiment 1-4 results. Between-category comparisons were more helpful for the aliens and bird categories because participants had substantial prior knowledge about both sets of categories (at least in the way the aliens were designed here—they looked very similar to people), and consequently, knew what features to pay attention to. Within-category comparisons were more helpful for the machine and probability problems, as both of these sets of categories were less familiar to learners (and while participants may have had prior knowledge about machines, they most likely did not have experience with machines that are classified based on whether or not the features make sense together).

This explanation offers an interesting framework, but it needs further specification, as it is unclear in some circumstances what predictions it would make. There are four cases I will address here where the prior-knowledge view as I understand it does not make clear predictions. Each of these cases raises important questions that this view needs to further specify.
The first case is Experiment 4’s results (where within-category comparisons were more beneficial for learning about two categories of permutations problems). The participants involved in Experiment 4 had a significant amount of prior knowledge about the construction of math problems as well as a reference sheet that cued them in to the relevant details of the problems, yet within-category comparisons were more beneficial than between-category comparisons. Appendix E shows the general instructions that learners were given about permutations and combinations problems. These instructions provided learners with a framework for determining the relevant pieces of information for a problem as well as how to solve the problem. In addition, the undergraduate students who participated in the experiment most likely had expectations about math problems that were consistent with the problems they were presented with (because they have been doing math problems like these for years in school); specifically, they most likely knew that the surface features of a problem are not a good predictor of category membership. Given that the participants had a significant amount of prior knowledge available to them throughout the learning phase of this experiment, the prior-knowledge view would have most likely predicted that between-category comparisons would be more beneficial, yet the opposite result as found. One possibility is that the learners did not have enough prior knowledge to benefit from between-category comparisons. If that is the case, then how much prior knowledge is enough?

The second case is Experiment 5’s results. Specifically, it is difficult for this view to explain the finding that between-category-comparison learners actually showed worse performance than the yoked, no-comparison learners. It is not clear that this explanation provides any reason for comparisons to lead to worse performance. Rather, this explanation argues that one type may be helpful while another is not (and therefore learners engaging in the
less helpful comparisons should show similar performance to a group that did not compare at all). A more complete version of this explanation will need to specify exactly how between-category comparisons would hurt learning relative to learning without comparisons when learners do not have prior knowledge about the categories they are learning about.

A third case is the set of studies by Namy and colleagues (Gentner & Namy, 1999; Namy & Gentner, 2002; Namy & Clepper, 2010). They showed in numerous studies that four-year-olds were able to categorize based on taxonomic category membership if they had compared two items from the same taxonomic category first. Children who did not make comparisons (or made comparisons across items from different taxonomic categories) were more likely to categorize based on perceptual similarity. In all of these studies, children had to give the “real” English names for each of the items they had seen during the category-matching task. In order to be included in the study, they had to provide the correct name for most of the items. Given that all of the children included in the analyses easily met this inclusion criterion, they must have had some prior knowledge of the items presented during the task.

The prior knowledge explanation would predict that between-category comparisons should be more beneficial for these children because they entered the experiment with prior knowledge about the items they were interacting with. In the study by Namy and Clepper (2010), children either compared two items from the same taxonomic category or two items from different taxonomic categories (that were perceptually similar). Children who compared items from the same taxonomic category were more likely to categorize based on taxonomic category membership than children who compared items from different taxonomic categories. Even though the different-taxonomic-categories condition used in this study is slightly different from the between-category-comparison condition from the studies reported here (as the items were
from many different taxonomic categories, making this a “Category A versus not A” kind of comparison rather than a “Category A versus Category B” comparison), the prior-knowledge explanation would not have predicted the within-category-comparison benefit relative to this condition. As was the case with the Experiment 4 results, an alternative possibility is that the children did not have enough prior-knowledge to benefit from between-category comparisons. How much prior knowledge is sufficient to enable children to benefit from between-category comparisons rather than within-category comparisons on this task?

A fourth case is the study by Carvalho and Goldstone (2011) described above. In their study, participants learned about unfamiliar blobs that either shared many features with each other or few features with each other. The prior-knowledge explanation would predict that within-category comparisons should be more beneficial regardless of item similarity, as the prior knowledge and expectations one has about the blobs and their features is most likely low and does not change across the similarity manipulation. The results, however, show an interaction between item similarity and comparison type. These results raise the question: are there particular circumstances where learners do not need prior knowledge to benefit from between-category comparisons?

The prior-knowledge explanation provides an interesting framework for exploring comparison effects. It is possible that amount of prior knowledge is the right characterization of these results, or at the very least, prior knowledge may be part of a network of factors that influence how comparisons affect learning. To help understand the role of prior knowledge, I plan to do a number of studies that manipulate the amount of prior knowledge participants start with when learning different types of categories. For example, I could manipulate the amount of relevant prior knowledge learners have available to them when learning about categories that
require attention to relational structure. The prior-knowledge explanation would predict that between-category comparisons would be more beneficial when learners have prior knowledge about the categories to be learned, even if the categories to be learned require attention to relational structure.

6.2.3 Difficulty explanation

A third alternative explanation for the results of the studies reported here is the difficulty explanation. The different types of categories could have been characterized by their degree of difficulty, and the results of Experiments 1-4 could be explained by arguing that the aliens and birds were easier to learn than the machines and probability problems. Between-category comparisons were more helpful when the categories were easier to learn while within-category comparisons were more helpful when the categories were more difficult to learn.

One challenge to this view is that the difficulty of the bird families from Experiment 3 relative to the coherence-based categories from Experiments 2 and 5 does not map on to this explanation. The difficulty view would have predicted that the coherence-based categories from Experiments 2 and 5 were harder to learn than the bird families from Experiment 3. Learning performance from participants in the no-comparison condition from Experiment 3 (the singles condition) was compared to learning performance from participants in the no-comparison conditions from Experiment 5 (regardless of item order). In both experiments, the number of learning trials was the same; however, in Experiment 3, participants classified birds on half of the trials and studied a bird and its label on the other half of the trials, whereas in Experiment 5, participants inferred the missing feature of an item on every trial. If performance during only the first block in Experiment 5 is considered (to equate the number of active processing trials), Experiment 5 no-comparison learners performed on average 29% better than chance (which was
33%) while Experiment 3 no-comparison learners performed on average 20% better than chance (which was 50%). If performance from both blocks of trials in Experiment 5 is considered (to equate for the number of exposures to items), Experiment 5 no-comparison learners performed on average 40% better than chance.

Taking into account the difference in chance across Experiments 3 and 5 (33% and 50% respectively), there was no real difference in learning performance when equating for the number of active processing trials, and if anything, the difference goes in the opposite direction when all trials from Experiment 5 are considered (that is, learners are showing higher levels of performance for what would be considered the more difficult categories). The difficulty explanation is not sufficient for explaining these data. I acknowledge that comparing learning performance across two experiments that used different tasks and materials is not an ideal measure of difficulty. Despite the differences in learning tasks and materials, it does seem that the difficulty explanation would have predicted that the coherence-based categories would be harder to learn (and therefore learners should have shown worse learning performance).

On the other hand, is possible that difficulty does determine the degree to which within-category and between-category comparisons are helpful, and here I simply examined one way that difficulty could be realized (by distinguishing categories based on whether learners could just rely on individual feature values or had to rely instead on the relational structure of the items). Despite this observation, it is not clear that the difficulty account would have predicted that between-category comparisons would be detrimental to learning as shown in the feature-pairs classification test.

While I intend to address this explanation in future work, there are two reasons why it will be challenging to do so. First, it is unclear what an appropriate measure of difficulty would
be. Second, and related to the first reason, difficulty can be realized in a number of ways. For instance, difficulty could mean varying the degree to which learners must focus on relational structure versus salient feature values, varying within-category similarity, or increasing the variation in feature values for a particular feature. Without a more specific account of difficulty, it is impossible to quantify how difficult a set of categories is relative to another set of categories. The best way to determine whether difficulty is the best characterization for these categories is to attempt to establish that a number of factors that arguably give rise to difficulty (e.g. requiring learners to focus on relational structure, using categories that are difficult to distinguish) affect the benefits of each type of comparison in similar ways.

6.2.4 Summary of alternative explanations

In summary, there are plausible alternative explanations for the results of the experiments reported here and I intend to address each more directly in future work. While it is possible that one of these explanations better captures the reason for differential comparison-type benefits, another possibility is that each of these explanations has complementary effects on the benefits of each type of comparison. For example, perhaps for categories where relational structure must be learned, having prior knowledge may reduce the benefits for within-category comparisons while having no prior knowledge may lead to larger benefits for within-category comparisons. Additionally, it is possible that within-category and between-category similarity interact with the predictions of the highlighter theory in interesting ways.

Regardless of the best way to distinguish the types of categories used in the experiments reported here, the results of Experiments 1-5 very clearly show that within-category and between-category comparisons affect what information learners acquire. Finally, it is important to note once again that the highlighter theory, which was generated based on a substantial body
of prior research, successfully predicted the Experiment 1-5 results.

6.3 Implications for theories of category learning

Categories are learned and used in a variety of ways, but of the many ways to learn categories, comparisons have been suggested as central to learning (e.g. Spalding & Ross, 1994). Comparisons are made on a daily basis and affect what information learners pay attention to and learn about. A complete understanding of category learning requires an intricate understanding of the role of comparisons. This thesis focused on two types of comparisons that are critical for category learning—between-category and within-category comparisons—and showed that each type of comparison leads learners to focus on different information.

Additionally, these results provide more evidence in favor of the claim made by Markman & Ross (2003), that different ways of learning affect what is learned about categories. The type of comparison a learner engages in dictates what information he or she acquires during learning. Thus, any theory or model that attempts to account for how categories are learned must make the assumption that different types of active processing lead to differences in what is learned.

6.4 Comparisons in educational settings

A natural extension of these findings is to consider the role of comparisons in the classroom. In fact, Rittle-Johnson and Star (2009) note that there is some evidence from case studies of teachers and analyses of cultural differences in educational practices that comparisons benefit mathematics education. Despite this observation, very little research has focused on determining the most effective ways to compare or even why comparisons are helpful. Without this level of specification, it would be easy for teachers to potentially miss an opportunity where certain types of comparisons would have improved learning or to encourage students to make the
wrong types of comparisons (e.g. telling them to make between-category comparisons under circumstances where within-category comparisons are most beneficial).

There is still much left to be understood about the role of comparisons in category learning, but as a first step, the results of the set of studies reported here suggest that when students must learn concepts that require them to learn relational structure (e.g. different types of algebra problems, different types of physics problems), teachers should encourage students to engage in within-category comparisons. When learners can focus on feature values rather than more complicated information (e.g. when learning to differentiate between different artists’ paintings in art history class), between-category comparisons will be more beneficial.

6.5 Future directions

I have already addressed some of the future directions necessary for the short-term. Specifically, future work will address the likelihood that other ways of characterizing the different types of categories could lead to the same comparison effects. In this section, I provide three examples of additional future directions.

6.5.1 Considering the learner’s approach to each type of comparison

Across Experiments 1-5, a number of different tasks were used to elicit comparisons in order to ensure generalizability. Additionally, there was no reason to believe that a particular way of doing between-category or within-category comparisons was better than another. However, after considering why the predicted interaction was not found in the novel classification test in Experiment 5, I considered the possibility that the way comparisons were being prompted might have affected how learners approached the comparison task. By focusing learners on both similarities and differences, within-category comparisons may have been less helpful than they would have been without such a specific prompt. In Experiments 1-4, this was
less of an issue because the performance differences between the between-category-comparison and within-category-comparison conditions were so large. In Experiment 5, where each type of comparison was compared to a no-comparison condition that shared the same order of items, the way the prompt was phrased may have been critical for observing differences.

Given this observation, an important question is raised: what are the most effective ways to make between-category and within-category comparisons? The essence of a within-category comparison is that it focuses learners on commonalities rather than differences. By asking learners to also focus on differences in some of the experiments reported here, within-category comparisons may have become less effective. Instructing learners more specifically to focus on commonalities may enhance the benefits of within-category comparisons. Similarly, it is possible that between-category comparisons became less helpful when the comparison prompt asked learners to focus on similarities as well as differences, rather than differences alone.

Furthermore, in Experiment 5, learners who wrote general statements about the items’ features, rather than statements about each feature’s specific value, were better able to abstract the coherence relation. This finding leads to the prediction that when learners are told to focus more generally on the items’ features, the benefits of within-category comparisons will be enhanced. An interesting extension of this prediction would be to consider the cases where between-category-comparison learning is more effective. If the benefit of between-category comparisons is that they focus learners on feature value differences, then when learners are told to think more specifically about the items when comparing, the benefits of between-category comparisons will be enhanced. Future work will consider these questions and more generally will consider how different ways of eliciting comparisons may affect the degree to which they benefit learning. The results of this work will be important for more developing a more detailed
understanding of how comparisons affect learning. Additionally, understanding the most effective ways to prompt comparisons has significant implications for instructing teachers on how to implement comparisons in the classroom.

6.5.2 Order effects of between-category and within-category comparisons

In Chapter 1, I argued that within-category comparisons could actually help learners make distinctions between alignable and non-alignable differences. I claimed that differences could only be interpreted as alignable if the relational structure of the items was understood. If this claim is true, then for categories requiring learners to focus on relational structure, within-category comparisons are a necessary first step in learning categories. However, once the relational structure is understood, learners could potentially benefit from making between-category comparisons because they have now established a framework for interpreting differences between categories as either alignable or non-alignable. The prediction would be that once learners have established the relational structure of the items, making between-category comparisons would be more effective than making additional within-category comparisons because within-category comparisons are not effective for noticing important between-category differences.

In future work, this prediction could be addressed by having participants learn the coherence-based categories from Experiments 2 and 5 through within-category then between-category comparison, between-category then within-category comparison, between-category comparison only, or within-category comparison only. One challenge for this experiment is determining when learners are ready to move from one type of comparison to the other. One possibility is to use inference performance during learning as an indicator. When learners hit a particular learning criterion, they move to the other comparison type. The prediction is that
within-category followed by between-category comparison is the most effective learning condition.

6.5.3 Other types of comparisons

Only between-category and within-category comparisons were considered in the studies reported here. As was pointed out in Chapter 1, there are many ways to differentiate types of comparisons. As an example, within a category one could compare items that are very similar to or very different from each other. When and why might it be beneficial to make each of these types of comparisons? The benefits of progressive alignment (e.g. Kotovsky & Gentner, 1996), and some new work demonstrating the benefits of fading (Casale & Pashler, presented at the Annual Meeting of the Psychonomic Society, 2011), suggest that it may be more helpful to compare very similar items at first followed by less similar items later. Perhaps making comparisons across very similar items allows learners to use similarity to bootstrap their understanding of relational structure. When the less similar items are compared later on, learners will already have a better understanding of what to pay attention to and the less similar items will be effective in ensuring that learners develop an abstract representation of the relational structure. Future work will address this and other distinctions between various types of comparisons.

6.6 Conclusions

In this thesis, a large body of research from the analogical-reasoning, problem-solving, and categorization domains was used to generate the highlighter theory of comparison learning, which proved effective in predicting the effects of different types of comparisons on category learning. Consistent with the highlighter theory, Experiments 1-5 showed that the benefits of each type of comparison depended on the type of information learners needed to focus on in
order to determine category membership. Within-category comparisons were more beneficial when relational structure had to be learned while between-category comparisons were more beneficial when learners could rely on individual feature-values to determine category membership. This is a promising first step in understanding how comparisons affect what information learners acquire about categories. Given the centrality of comparisons in category learning, efforts should be made to continue examining these comparison effects in further detail.
Table 1: *Family-Resemblance Category Structure for Experiment 1*

<table>
<thead>
<tr>
<th>Item type</th>
<th>Deeger</th>
<th>Koozie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototype</td>
<td>111111</td>
<td>000011</td>
</tr>
<tr>
<td>Learning exemplars</td>
<td>111011</td>
<td>000111</td>
</tr>
<tr>
<td></td>
<td>110111</td>
<td>001011</td>
</tr>
<tr>
<td></td>
<td>101111</td>
<td>010011</td>
</tr>
<tr>
<td></td>
<td>011111</td>
<td>100011</td>
</tr>
</tbody>
</table>

*Note:* Prototypes were not shown during the learning phase. The last two features were always the “1” value during learning, reducing the set of learning items to 4 per category.
Table 2: Items presented during the classification test for Experiment 1

<table>
<thead>
<tr>
<th>Deeger</th>
<th>Koozle</th>
</tr>
</thead>
<tbody>
<tr>
<td>111111</td>
<td>000011</td>
</tr>
<tr>
<td>111110</td>
<td>000001</td>
</tr>
<tr>
<td>111101</td>
<td>000010</td>
</tr>
<tr>
<td>111011</td>
<td>000111</td>
</tr>
<tr>
<td>110111</td>
<td>001011</td>
</tr>
<tr>
<td>101111</td>
<td>010011</td>
</tr>
<tr>
<td>011111</td>
<td>100011</td>
</tr>
<tr>
<td>111100</td>
<td>000000</td>
</tr>
<tr>
<td>111010</td>
<td>000110</td>
</tr>
<tr>
<td>111001</td>
<td>000101</td>
</tr>
<tr>
<td>110110</td>
<td>001010</td>
</tr>
<tr>
<td>110101</td>
<td>001001</td>
</tr>
<tr>
<td>101110</td>
<td>010010</td>
</tr>
<tr>
<td>101101</td>
<td>010001</td>
</tr>
<tr>
<td>011110</td>
<td>100010</td>
</tr>
<tr>
<td>011101</td>
<td>100001</td>
</tr>
</tbody>
</table>
Table 3: *Experiment 1 Comparison Coding by Condition*

<table>
<thead>
<tr>
<th></th>
<th>Between-category-comparison condition</th>
<th>Within-category-comparison condition</th>
</tr>
</thead>
<tbody>
<tr>
<td># Features Mentioned</td>
<td>6.63 (2.42)</td>
<td>6.70 (2.75)</td>
</tr>
<tr>
<td># Similarities Mentioned</td>
<td>4.16 (2.46)</td>
<td>4.66 (2.76)</td>
</tr>
<tr>
<td># Differences Mentioned</td>
<td>2.46 (0.62)</td>
<td>2.03 (0.13)</td>
</tr>
<tr>
<td>Proportion of Trials where Labels were Used</td>
<td>0.55 (0.46)</td>
<td>0.32 (0.44)</td>
</tr>
</tbody>
</table>

*Note:* The average number of features mentioned exceeds the actual number of features defined in the category scheme. This is because some participants actually listed features such as body shape and head shape as part of the similarities and differences they noticed, which were the same across all items.
Table 4: *Experiment 2 Comparison Coding by Condition*

<table>
<thead>
<tr>
<th></th>
<th>Between-category-comparison condition</th>
<th>Within-category-comparison condition</th>
</tr>
</thead>
<tbody>
<tr>
<td># Features Mentioned</td>
<td>2.91 (0.16)</td>
<td>2.62 (0.94)</td>
</tr>
<tr>
<td># Similarities Mentioned</td>
<td>1.01 (0.11)</td>
<td>1.01 (0.72)</td>
</tr>
<tr>
<td># Differences Mentioned</td>
<td>1.90 (0.18)</td>
<td>1.62 (0.66)</td>
</tr>
<tr>
<td>Proportion of Trials where Labels were Used</td>
<td>0.60 (0.48)</td>
<td>0.24 (0.34)</td>
</tr>
<tr>
<td>Proportion of Trials where General Statements were Made</td>
<td>0.21 (0.38)</td>
<td>0.46 (0.31)</td>
</tr>
</tbody>
</table>
Table 5: *Experiment 3 Novel Classification Performance Broken Down by Comparison Type and Presentation Type*

<table>
<thead>
<tr>
<th></th>
<th>Between-category-comparison condition</th>
<th>Within-category-comparison condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pairs condition</td>
<td>0.46 (0.19)</td>
<td>0.39 (0.17)</td>
</tr>
<tr>
<td>Singles condition</td>
<td>0.37 (0.16)</td>
<td>0.32 (0.21)</td>
</tr>
</tbody>
</table>
Table 6: Experiment 4 Comparison Coding by Condition

<table>
<thead>
<tr>
<th></th>
<th>Between-category-comparison condition</th>
<th>Within-category-comparison condition</th>
</tr>
</thead>
<tbody>
<tr>
<td># Features Mentioned</td>
<td>2.42 (1.24)</td>
<td>2.00 (1.44)</td>
</tr>
<tr>
<td># Similarities Mentioned</td>
<td>1.08 (0.56)</td>
<td>0.84 (0.74)</td>
</tr>
<tr>
<td># Differences Mentioned</td>
<td>1.33 (0.79)</td>
<td>1.17 (0.88)</td>
</tr>
<tr>
<td>Proportion of Trials where Labels were Used</td>
<td>0.33 (0.41)</td>
<td>0.33 (0.41)</td>
</tr>
<tr>
<td>Proportion of Trials where General Statements were Made</td>
<td>0.25 (0.27)</td>
<td>0.5 (0.45)</td>
</tr>
</tbody>
</table>
Table 7: *Experiment 5 Block 1 Inference Learning Performance, Broken Down by Item Order and Comparison Type*

<table>
<thead>
<tr>
<th></th>
<th>Between-category Item Order</th>
<th>Within-category Item Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison Learners</td>
<td>0.58 (0.15)</td>
<td>0.68 (0.15)</td>
</tr>
<tr>
<td>No-Comparison Learners</td>
<td>0.57 (0.18)</td>
<td>0.67 (0.19)</td>
</tr>
</tbody>
</table>
Table 8: *Experiment 5 Block 2 Inference Learning Performance, Broken Down by Item Order and Comparison Type*

<table>
<thead>
<tr>
<th></th>
<th>Between-category Item Order</th>
<th>Within-category Item Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison Learners</td>
<td>0.66 (0.26)</td>
<td>0.89 (0.14)</td>
</tr>
<tr>
<td>No-Comparison Learners</td>
<td>0.77 (0.22)</td>
<td>0.86 (0.22)</td>
</tr>
</tbody>
</table>
Table 9: *Experiment 5 Comparison Coding by Condition*

<table>
<thead>
<tr>
<th></th>
<th>Between-category-comparison condition</th>
<th>Within-category-comparison condition</th>
</tr>
</thead>
<tbody>
<tr>
<td># Features Mentioned</td>
<td>2.68 (0.72)</td>
<td>2.86 (0.61)</td>
</tr>
<tr>
<td># Similarities Mentioned</td>
<td>0.93 (0.23)</td>
<td>0.81 (0.45)</td>
</tr>
<tr>
<td># Differences Mentioned</td>
<td>1.73 (0.10)</td>
<td>2.07 (0.57)</td>
</tr>
<tr>
<td>Proportion of Trials where Labels were Used</td>
<td>0.56 (0.41)</td>
<td>0.14 (0.16)</td>
</tr>
<tr>
<td>Proportion of Trials where Abstract Statements were Made</td>
<td>0.02 (0.04)</td>
<td>0.11 (0.19)</td>
</tr>
<tr>
<td>Proportion of Trials where General Statements were Made</td>
<td>0.24 (0.37)</td>
<td>0.51 (0.44)</td>
</tr>
</tbody>
</table>
Table 10: *Experiment 5 Feature-Pairs Classification Performance, Broken Down by Item Order and Comparison Type*

<table>
<thead>
<tr>
<th></th>
<th>Between-category Item Order</th>
<th>Within-category Item Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison Learners</td>
<td>0.63 (0.19)</td>
<td>0.96 (0.12)</td>
</tr>
<tr>
<td>No-Comparison Learners</td>
<td>0.76 (0.21)</td>
<td>0.84 (0.22)</td>
</tr>
</tbody>
</table>
Table 11: *Experiment 5 Novel Classification Performance, Broken Down by Item Order and Comparison Type*

<table>
<thead>
<tr>
<th></th>
<th>Between-category Item Order</th>
<th>Within-category Item Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison Learners</td>
<td>0.54 (0.25)</td>
<td>0.77 (0.23)</td>
</tr>
<tr>
<td>No-Comparison Learners</td>
<td>0.62 (0.22)</td>
<td>0.77 (0.19)</td>
</tr>
</tbody>
</table>
Table C1: *Experiment 3 Average CLJ Ratings by Comparison Type and Presentation Type*

<table>
<thead>
<tr>
<th>Condition</th>
<th>Between-category-comparison condition</th>
<th>Within-category-comparison condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pairs condition</td>
<td>0.52 (0.16)</td>
<td>0.49 (0.17)</td>
</tr>
<tr>
<td>Singles condition</td>
<td>0.39 (0.17)</td>
<td>0.40 (0.18)</td>
</tr>
</tbody>
</table>
Table C2: *Experiment 3 Average Confidence Ratings by Comparison Type and Presentation*

<table>
<thead>
<tr>
<th>Type</th>
<th>Between-category-comparison condition</th>
<th>Within-category-comparison condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pairs condition</td>
<td>0.49 (0.15)</td>
<td>0.47 (0.18)</td>
</tr>
<tr>
<td>Singles condition</td>
<td>0.37 (0.19)</td>
<td>0.37 (0.21)</td>
</tr>
</tbody>
</table>
Table C3: *Experiment 3 Average Within-Participant Gamma Correlations by Comparison Type and Presentation Type*

<table>
<thead>
<tr>
<th></th>
<th>Between-category-comparison condition</th>
<th>Within-category-comparison condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pairs condition</td>
<td>0.43 (0.38)</td>
<td>0.38 (0.41)</td>
</tr>
<tr>
<td>Singles condition</td>
<td>0.31 (0.54)</td>
<td>0.44 (0.60)</td>
</tr>
</tbody>
</table>
FIGURES

Figure 1: *Experiment 1 example between-category-comparison trial.*

List how this *Koozle* and this *Deeger* are the same and different.
Figure 2: Examples of bird pictures from each bird family used in Experiment 3.

Finch
Flycatcher
Sparrow

Thrush
Vireo
Warbler
Figure 3: *Experiment 3, example within-category-comparison trial from the pairs condition.*
Figure 4: Distribution of WC learners based on proportion of general comparisons.
REFERENCES


Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*, 393-408.


## APPENDIX A

Learning Materials for Experiment 2 & Proposed Experiment 5

<table>
<thead>
<tr>
<th>Morkels</th>
<th>Krenshaws</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operates on land</td>
<td>Operates on land</td>
</tr>
<tr>
<td>Works to gather harmful solids</td>
<td>Works to clean spilled oil</td>
</tr>
<tr>
<td>Has a shovel</td>
<td>Has an electrostatic filter</td>
</tr>
<tr>
<td>Operates on the surface of the water</td>
<td>Operates on the surface of the water</td>
</tr>
<tr>
<td>Works to clean spilled oil</td>
<td>Works to collect dangerous gaseous ions</td>
</tr>
<tr>
<td>Has a spongy material</td>
<td>Has a shovel</td>
</tr>
<tr>
<td>Operates in the stratosphere</td>
<td>Operates in the stratosphere</td>
</tr>
<tr>
<td>Works to collect dangerous gaseous ions</td>
<td>Works to gather harmful solids</td>
</tr>
<tr>
<td>Has an electrostatic filter</td>
<td>Has a spongy material</td>
</tr>
</tbody>
</table>
APPENDIX B

Test Materials for Experiment 2 & Proposed Experiment 5

---

**Novel Morkels**

- Operates in highway tunnels
- Works to remove carbon dioxide
- Has a large intake fan

- Operates in swamps
- Works to remove malaria-ridden mosquitoes
- Has a finely-woven net

- Operates in war zones
- Works to gather shards of metal
- Has a large magnet

- Operates in parks
- Works to gather discarded paper
- Has a metal pole with a sharpened end

- Operates on the seafloor
- Works to remove lost fishing nets
- Has a hook

- Operates on the beach
- Works to remove broken glass
- Has a sifter

---

**Novel Krenshaws**

- Operates in highway tunnels
- Works to remove lost fishing nets
- Has a sifter

- Operates in swamps
- Works to remove broken glass
- Has a metal pole with a sharpened end

- Operates in war zones
- Works to gather discarded paper
- Has a finely-woven net

- Operates in parks
- Works to gather shards of metal
- Has a hook

- Operates on the seafloor
- Works to remove malaria-ridden mosquitoes
- Has a large intake fan

- Operates on the beach
- Works to remove carbon dioxide
- Has a large magnet

---

126
APPENDIX C

Experiment 3: Metacognition Results

There are three ways to look at the metacognitive judgments of performance: average values of participants’ judgments, calibration, and resolution. Calibration is the degree to which one’s metacognitive judgment matches one’s actual performance (Nelson, 1996). This measurement can show whether participants are well calibrated or if they are over-confident or under-confident. Resolution is the degree to which one’s metacognitive judgments correlate with performance on individual items. Resolution is measured by the within-subject gamma correlation between metacognitive judgments and test performance (Nelson, 1996). For the category-learning judgments (CLJs, which were family level predictions of novel classification accuracy made before participants began the novel classification test), the absolute values of the judgments as well as the calibration between judgments and accuracy were analyzed. For confidence ratings (which were made on an item-by-item basis after participants classified each item), the absolute values of the ratings, calibration, and resolution were analyzed.

Category-learning Judgments

A 2x2 mixed factorial ANOVA was run on the CLJs, and showed that the judgments were higher for pairs learners (M = 0.50, SD = 0.15) than singles learners (M = 0.39, SD = 0.16), F (1, 46) = 6.21, p < 0.05. There was no main effect of comparison type, F (1, 46) = 0.29, p > 0.05, nor was there an interaction between presentation type and comparison type, F = 1.26, p > 0.05 (see Table C1 for a complete breakdown of the CLJ means and standard deviations).

Calibration was calculated as the difference between participants’ actual proportion of correct trials and their CLJs (which were participants’ predictions of the percentage of trials they would classify correctly). Higher numbers indicate larger differences between predicted and
actual performance (the closer to zero, the more accurate). Positive numbers indicate over-confidence while negative numbers indicate under-confidence.

A 2x2 mixed factorial ANOVA was run and showed that CLJs were marginally better calibrated if participants had learned about the bird families through between-category comparison ($M = 0.04, SD = 0.22$) than within-category comparison ($M = 0.08, SD = 0.20$), $F(1, 46) = 3.50, p = 0.07$. There was no difference in calibration between pairs learners and singles learners, $F(1, 46) = 0.37, p > 0.05$, and there was no interaction between presentation type and comparison type, $F = 0.45, p > 0.05$. Comparison type was manipulated within-subjects and as noted above, there was no difference in average rating across the two comparison conditions. The difference was in accuracy across the within-category-comparison and between-category-comparison conditions and this difference is what caused a difference to show up here.

Confidence Ratings

One person failed to make confidence ratings, so only 39 participants’ data was analyzed. A 2x2 mixed factorial ANOVA showed that confidence ratings, which were made for each item after each novel classification trial, were higher for pairs learners ($M = 0.48, SD = 0.16$) than singles learners ($M = 0.37, SD = 0.20$), $F(1, 45) = 4.62, p < 0.05$. There was no main effect of comparison type, $F(1, 45) = 0.15, p > 0.05$, nor was there an interaction between presentation type and comparison type, $F(1, 45) = 0.69, p > 0.05$ (see Table C2 for a complete breakdown of the confidence means and standard deviations).

Calibration was calculated as the difference between participants’ actual proportion of correct trials and their averaged confidence ratings. A 2x2 mixed factorial ANOVA showed that confidence ratings were better calibrated if participants had learned about the bird families through between-category comparison ($M = 0.01, SD = 0.25$) than within-category comparison
(\(M = 0.06, SD = 0.20\)), \(F(1, 45) = 113.18, p < 0.05\). There was no main effect of presentation type, \(F(1, 45) = 0.07, p > 0.05\), nor was there an interaction between presentation type and comparison type, \(F(1, 45) = 2.62, p > 0.05\). As noted above, comparison type was manipulated within-subjects, and there was no difference in average confidence rating across items in the two comparison conditions. The difference was in accuracy and this difference is what caused a difference to show up here.

Resolution was calculated as the mean within-participant gamma correlation between item-by-item confidence ratings and accuracy. A 2x2 mixed factorial ANOVA was run and showed no main effects of either comparison type (\(F(1, 45) = 0.20, p > 0.05\)) or presentation type (\(F(1, 45) = 0.093, p > 0.05\)) as well as no interaction between the two factors (\(F(1, 45) = 0.88, p > 0.05\)). Means and standard deviations are displayed in Table C3.
APPENDIX D

Experiment 4 Probability Problems

Problems used during learning

Permutations:

The FC United soccer club is having tryouts. Coach Tom wants to assess each player’s ability, so he divides up the 27 soccer players into two teams for a scrimmage. Team B has 14 players, but only 11 can be on the field at a time. How many ways can the 11 different field positions be filled by Team B?

Fresh Fields grocery store is restocking their cereal aisle. They have 20 types of cereal to stock, but there are only 15 shelf positions. Each shelf position can have one cereal type. How many different ways can the 20 cereal types be arranged on the shelves?

Combinations:

The University’s Salsa club is having a class for beginner men. 16 men signed up for the class. There are more instructors during the Tuesday section of the class than the Thursday section. The men need to be divided up so that 10 attend class on Tuesday and 6 attend class on Thursday. They are assigned to a section randomly. How many ways can the men be selected for the Tuesday section of the class?

Kelly is a homeowner who is trying to decorate her patio with flowers. She goes to the garden store and buys 12 different flowers to plant in pots. When she returns, she realizes she only has 8 flowerpots. How many different ways could Kelly pick the flowers to plant in the 8 flowerpots?

Problems used during the test phases (the first two of each type were used in the classification test and the second two were used in the problem solving test).

Permutations:

Shadyside Elementary School is having its annual candy sale, and the 28 students in Mrs. Miller’s class will be going door to door to sell the candy. 10 of the students have never participated in the candy sale, so Mrs. Miller wants these students to pair up with students who have participated before. Of the 18 students who have participated, only 13 are willing to pair up with the 10 students who have never sold candy. The students who have never participated get to choose their partners one at a time, with the choosing going by age, and with the youngest student choosing first. How many ways could the students who have never participated in the candy sale choose their more experienced partners?
Jillian has to create a code for her security system. The system has 10 possible symbols and 5 different symbols need to be used for the code. No symbols can be repeated. How many different sequences of symbols can be generated?

7 new graduate students will join Big University’s biology department in Fall 2011. An office is set up with 10 desks. To be fair, the graduate students will claim their desk based on the order that they arrive to school. How many ways can the students choose the desks?

Good Times band is performing a show in a few days. They are able to perform 11 songs during their set. The manager of the band has to choose the order of the songs from a list of 42 possible songs. He randomly chooses 11 numbers between 1-42, and these correspond to the songs that will be played. How many different ways can the 11 songs be chosen?

**Combinations:**

Children at the Playhouse Preschool are going on a field trip to the zoo. There are 8 children who are going on the trip and there are 2 parents driving them to the zoo in their cars. One car is a red mini van that can hold 5 children. How many different ways can the children be assigned to the red mini van?

Steve is packing for his trip to the beach. He is flying and wants to avoid paying the luggage fee at the airport, so he opts to pack his clothing in a small carry-on bag. He has 15 shirts in his dresser, but can only take 7 with him. How many ways can he choose shirts to pack in his bag?

At the Mega Marketing firm, the 19 office employees need to divide themselves into three committees to plan various office events. The committees are chosen randomly, by employees drawing a piece of paper with the committee assignment written on it, out of a hat. The first committee is in charge of employee birthday celebrations, and this committee requires 7 members. How many ways can the employees be picked for the birthday celebrations committee?

The Great Buys department store is assessing their inventory for spring. In the men’s clothing department, there are 16 different brands being sold; however, there are only 6 different mannequins that can be used to display clothing. The manager decides to choose the 6 brands to be displayed by randomly choosing the names from a bowl. How many different ways can the brands be chosen?
APPENDIX E

Experiment 4 Instructions

First, all participants were told about factorials.

Factorial Instructions:

A number’s factorial is the product of all the numbers less than or equal to that number.

Factorial $m$, denoted by $m!$, is:

$$m! = m \cdot (m - 1) \cdot (m - 2) \cdot \ldots \cdot (2) \cdot (1)$$

Example: the factorial of 6 is

6 factorial, or $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$.

Next, each time they encountered a problem during learning, they were given instructions for how to solve that problem type. Here are the instructions for each type.

Permutations Instructions:

PERMUTATIONS

Suppose that there is a set of objects and someone chooses some number of objects from the set. What is the probability that a specific set of the objects are chosen in a particular order? To find this probability, you must first figure out how many possible orderings of objects there might be. The number of possible orderings is called the number of permutations.

To find the probability of one particular order being chosen, you need to keep three things in mind.

First, you must figure out who is actually doing the choosing.

Second, you must figure out which set of objects is being chosen from. The number of objects in this set is represented by the variable $n$. That is, $n$ is the number of objects that are being chosen from.

Finally you must know how many objects are being chosen from the set in a particular order. This number is represented by the variable $r$. That is, $r$ is the number of choices that you are interested in.

To find the number of permutations, you need to know how many different orderings of $r$
objects can be chosen from \( n \) total objects. To get this, multiply the number \( n \) then \((n-1)\) etc., until you get \( r \) terms. If that is the number of different orders, the probability on any particular order is just 1 divided by that or,

\[
\frac{1}{(n) (n-1) (n-2) \cdots (n-r+1)}
\]

In which:
- \( n \) is the number of objects in the set (the set being chosen from).
- \( r \) is the number of objects being chosen.

Combinations Instructions:

COMBINATIONS

Suppose that you have a set of items. From this set, you remove a certain number of particular items to form a subset. What is the probability that any specific subset of items will be removed? The number of all possible subsets is called the number of combinations.

To find the number of combinations there are three sets of items that you must keep in mind.

The first is the total set of items from which the subset is being removed. The number in this set is represented by the variable \( j \).

The second set is made of the items that are being removed from the subset. This number is represented by the variable \( h \).

The third set is made up of the items that were in set \( j \) but were NOT removed. The number in this set is represented by the variable \( (j - h) \). This is the number of items that is left over after the subset has been removed.

To find out the number of different subsets possible with \( h \) objects, you take the factorial of the total set and divide it by the product of the factorials of the two subsets, to get \( j!/[h! \ (j - h)!] \). Since this is the number of different subsets, the probability of any particular subset is just 1 divided by that number, or

\[
\frac{1}{j! /[h! \ (j - h)!]}
\]

In which:
- \( j \) is the total number of items
- \( h \) is the subset that is removed from \( j \)
- \( (j - h) \) is the subset that is NOT removed from \( j \).