LEARNING A COMPLEX CATEGORY STRUCTURE BY
CLASSIFICATION AND FEATURE INFERENCE

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Arts

by

Glen Howell

Department of Psychology
Carleton University

September 2009

©2009 Glen Howell
NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.
Abstract

Yamauchi and Markman (1998) and Rehder, Colner, and Hoffman (2009) have shown that inference learners are as successful as classification learners in classifying prototypical items at transfer, despite the novelty of the classification task. Moreover, inference learners are more likely than classification learners to select a prototypical feature in an inference transfer task, even if classification learning occurs after inference learning. I argue that these results may be specific to the simple category structure that was used. Thus, in the present experiment, participants learned a more complex category structure. During the learning phase, participants performed inference, classification, or both tasks for a total of 768 or 1920 trials. During the transfer phase, classification learners were more accurate and faster to classify prototypical items than inference learners. Furthermore, under certain conditions, they selected a prototypical feature in an inference transfer task more often than inference learners. Lastly, participants who received both learning tasks had strikingly similar transfer performance. Thus, transfer-appropriate processing (Morris, Bransford, & Franks, 1977) may be a more comprehensive explanation of the results than researchers had originally suspected.
Acknowledgments

First and foremost I would like to thank my supervisor, Dr. Guy Lacroix, for his mentorship, guidance, and encouragement through all the various stages in the production of this thesis. His continuous support, positive attitude, and passion for scientific research made the entire process an enjoyable and amazing learning experience. I would also like to thank the members of my thesis examination committee, Dr. Andrew Brook, Dr. Craig Leth-Steensen, and Dr. John Logan, for their thoughtful feedback on my thesis. Lastly, I would like to thank my girlfriend, Robyn Olivier, my parents, David and Janet Howell, as well as my brothers, Graham and John Howell, and sister, Jen Howell, for their unwavering support and encouragement throughout my academic endeavors.
# Table of Contents

Abstract.............................................................. ii
Acknowledgments................................................... iii
Table of Contents.................................................... iv
List of Tables........................................................ vi
List of Figures......................................................... vii
Introduction......................................................... 1
Category-Learning Tasks in the Laboratory........................ 2
Memory Research.................................................... 5
Comparing Classification and Inference Learning.................... 7
Critical Evaluation of the Reviewed Studies......................... 23
Goals and Hypotheses............................................... 31
Method............................................................... 33
  Participants....................................................... 33
  Stimuli............................................................ 33
  Procedure......................................................... 36
Results............................................................. 42
  Learning Phase: Accuracy and Response time Analyses.......... 43
  General Level Transfer: Accuracy and Response Time Analyses... 50
    Old Classification........................................... 53
    Old Inference................................................ 54
    New Prototype Classification............................... 56
List of Tables

Table 1 *The category structure used in Yamauchi and Markman (1998) and Rehder, et al. (2009)* ................................................................. 9

Table 2 *The category structure Johansen and Kruschke (2005) used* ................. 15

Table 3 *A summary of results from studies comparing classification and inference learning* ................................................................. 25

Table 4 *The category structure and labels used in the present experiment* .......... 26

Table 5 *Mean number of blocks to criterion as a function of Learning Order and Learning Task* ................................................................. 48

Table 6 *Mean response time at the last block of the first task and the first block of the second task* ................................................................. 50
List of Figures

Figure 1. From left to right is “Sot Maki” exemplars, SM1, SM2, and SM3 with red (darker) or yellow (lighter) head color, square or hourglass body shape, and bumpy horizontal line or vertical line chest pattern. Note that the hourglass body shape, the yellow head color, and the bumpy horizontal line chest pattern are not diagnostic of category membership.......................... 34

Figure 2. From left to right is a classification learning item and the two corresponding inference learning items for the SM1 exemplar.............. 35

Figure 3. The timeline and display for inference learning trials...................... 39

Figure 4. Mean percent correct for short (top) and extensive (bottom) learning as a function of Learning Type and Block. Block 9 (top) and Block 21 (bottom) was the first block on the second day of learning. Error bars represent standard error of the mean................................................................. 45

Figure 5. Mean response time for short (top) and extensive (bottom) learning as a function of Learning Type and Block. Block 9 (top) and Block 21 (bottom) was the first block on the second day of learning. Error bars represent standard error of the mean................................................................. 46

Figure 6. Mean percent correct for general level classification, inference, and prototype items presented as a function of Learning Type and Learning Length. Error bars are standard error of the mean.................................................. 52

Figure 7. Mean response time for general level classification, inference, and prototype items presented as a function of Learning Type and Learning Length. Error bars are standard error of the mean.................................................. 53
Figure 8. Response frequency for the four different prototype-inconsistent inference response choice conditions presented with their corresponding response choices as a function of Learning Type. 0 = prototype-inconsistent feature. 1 = a prototypical feature for the exemplar’s specific category. 2 = a prototypical feature for the exemplar’s general, but not specific category. X = a categorically invalid feature.

Figure 9. Response frequency for each specific level transfer item type presented as a function of Response Choice and Learning Type. Correct = the correct specific category label. Other = the correct general but incorrect specific category label. 1 = a prototypical feature for the exemplar’s specific category. 2 = a prototypical feature for the exemplar’s general, but not specific category. X = a categorically invalid label or feature.
LEARNING A COMPLEX CATEGORY STRUCTURE BY CLASSIFICATION AND FEATURE INFERENCE

Introduction

We live in a world that can be understood as a collection of categories. Throughout our lives, we come to label, understand, and use these categories to serve higher-order cognitive functions, such as communication, classification, inference, and problem solving. We therefore form mental representations of categories (i.e., concepts) and rely on them to execute even the simplest goal-oriented behaviour. A general goal of categorization research is to develop a theory concerning the processes involved in concept acquisition and the nature of our concepts. The classical view of the nature of categories holds that they are defined by a set of necessary and jointly sufficient features (Murphy, 2002; Medin & Smith, 1981). That is, all members of a category have one or more necessary or defining features, which only when considered together, differentiate them from all other categories. According to this view, we would come to form concepts that consist of the necessary and jointly sufficient features that define categories.

Empirical research and philosophical accounts of the nature of categories, however, have shown that real-world and artificial categories are better described as having a family resemblance structure, not a set of necessary and jointly sufficient features (Rosch & Mervis, 1975; Wittgenstein, 1953). This account proposes that category members tend to have features in common within their category (i.e., prototypical features) and they tend to have some features that are not common to other categories (i.e., diagnostic features; see, Chin-Parker & Ross, 2002, 2004). These two distinct aspects of family resemblance afford different, but very important kinds of category information. Knowledge of prototypical features serves as a good predictor for
what feature a known category member is likely to have, whereas knowledge of
diagnostic features serves as a good predictor of an object’s class. These aspects of
family resemblance are central to the use of categories and, thus, to cognitive processes
which depend on them such as inference and classification (Markman & Ross, 2003).

Category-Learning Tasks in the Laboratory

The quintessential category-learning paradigm is the feedback-driven
classification task (for classic examples, see Shepard, Hovland, & Jenkins, 1961, Posner
& Keele, 1968). In this paradigm, participants learn to classify a set of category members
(i.e., exemplars) that belong to one of two or several mutually exclusive categories in a
trial and error fashion. The exemplars consist of a number of features that may be integral
or separable (Garner, 1974). Learning is stopped once participants have reached a
criterion (e.g., 90% correct) or after a preset number of trials. Next, participants’
knowledge of the abstract category structure is probed using various transfer phase tasks
during which feedback is not provided. For example, they may be asked to classify old or
new exemplars, prototypical exemplars, single features, pairs of correlated or
uncorrelated features, or they may be required to predict the appearance of a labelled
exemplar’s missing feature (feature inference). Classification learning has been the
subject of much research within the category-learning research community (see Murphy,
2002 for a review). Recently, however, researchers have started examining other
ecologically valid ways that categories can be learned.

Feature inference is a common and important means of category learning and use
in everyday life (Markman & Ross, 2003). The paradigm is based on the idea that people
can learn about categories by making a prediction regarding an already categorized
exemplar's missing feature and then receiving feedback from the environment regarding the prediction. During feature-inference learning in the laboratory, participants are presented a set of labelled exemplars that are each missing a feature. Their task is to learn to accurately predict the appearance of the missing feature using the exemplar's category label, its remaining features, or any combination of both. Once participants learn to accurately predict missing features from the set of exemplars to a certain criterion, the learning phase is considered complete. A transfer phase is then conducted in which participants' knowledge of the categories can be probed, much like in the classification paradigm.

The research presented in this thesis is a laboratory examination of the two category-learning paradigms discussed above: Classification and inference learning. An experiment was conducted to better understand the link between category-learning tasks and concept formation. As a starting point, I adopted the classic framework from memory research, transfer-appropriate processing (Morris, Bransford, & Franks, 1977; see, Roediger, Gallo, & Geraci, 2002, for a review), to draw hypotheses regarding how memory acquired during the learning phase may affect performance in various transfer phase tasks. This framework suggested that transfer task performance measures, such as accuracy and response time, would be higher and faster, respectively, when the task performed during the learning phase matched the task performed during the transfer phase (e.g., classification learners classifying exemplars at transfer) relative to when there was a mismatch (e.g., inference learners classifying exemplars at transfer). Categorization researchers who have compared classification and inference learning are advocates of this framework (Markman & Ross, 2003; Yamauchi & Markman, 1998).
Nevertheless, some studies that have produced results that are not consistent with the predictions derived from transfer-appropriate processing. Specifically, both Yamauchi and Markman (1998) and Rehder, Colner, and Hoffman (2009) found that inference learners were as accurate as classification learners when classifying prototypical exemplars during a transfer phase. Rehder et al. also found that classification learners were as accurate as inference learners when inferring a missing feature from incomplete exemplars during a transfer phase. The present research addressed the possibility that these results, which clearly violate transfer-appropriate processing, were brought about because the category structure that participants had to learn was too simple. More specifically, it is possible that the structure could not differentiate classification and inference transfer performance well, because it did not encompass enough categories or feature values to allow for a sufficient number of alternative responses during transfer trials. Hence, the structure’s simplicity did not yield a sufficiently large range of performances to observe transfer-appropriate processing effects. Therefore, in the present research, Yamauchi and Markman’s experiments were replicated using a more complex category structure. I expected that the accuracy and response time for the learning and transfer phase results would show a clear transfer-appropriate processing pattern.

Before further describing the present experiment in detail, I will first present a brief review of memory studies related to transfer-appropriate processing, and then link this research to category-learning experiments. Next, I will present an in-depth review of the pertinent studies that compared learning of perceptual categories via classification and inference. These studies will then be critically evaluated. Finally, the goal and specific hypotheses for the present experiment will be stated.
In a classic memory study, Hyde and Jenkins (1969) asked participants to perform one of two orienting tasks on a list of words that were presented. While half the participants were instructed to rate the pleasantness of the words (thus invoking semantic processing), the other half were instructed to determine if the letter “E” occurred in the words (thus invoking a more superficial processing of the words). In a subsequent test, participants were asked to recall the previously presented words. The results from three experiments showed that participants who rated the pleasantness of the words had significantly better recall than those who determined if the letter “E” occurred in the words. This effect held when participants were told that there would be a memory test, when presentation time was doubled, and when the list was presented twice. On the basis of results such as these, Craik and Lockheart (1972) proposed a theory of memory, which suggests that different levels-of-processing lead to a gradation of strength in memory traces. For instance, the pleasantness group in Hyde and Jenkins’s experiments had better recall because they were engaged in a deeper level of processing (semantic processing) during orientation than the group who determined if the letter “E” occurred in the words.

Unfortunately, empirical research has shown the levels-of-processing framework to be irremediably flawed. Morris et al. (1977) tested the hypothesis that a semantic orienting task that invokes deep processing does not always lead to better recall, as the levels-of-processing theory suggests. They suspected that, in previous research, the semantic group had had an advantage relative to other groups who performed a task that involved shallower processing, because of a mismatch between the mode of processing at orientation and test. Therefore, Morris et al. asked their participants to perform a
semantic or phonetic orienting task and then a semantic and phonetic memory test. The authors found that participants in the semantic group had better recall than the phonetic group on the semantic memory test, thus replicating previous findings such as Hyde and Jenkins (1969) and supporting the levels-of-processing theory. More importantly, however, the authors found that participants in the phonetic group had better recall than the semantic group on the phonetic memory test. This result could not be accounted for within the levels-of-processing framework. The authors therefore suggested a new framework called transfer-appropriate processing, which emphasized the importance of examining memory test performance in relation to the orienting task goals. The framework suggests that memory performance is superior when there is a match between the task performed during orientation and test, relative to when there is a mismatch.

The parallel in category-learning research has recently been drawn from studies that have compared classification and inference learning. These two learning tasks are thought of as requiring different goals, strategies, and modes of processing. For example, studies have shown that participants who are engaged in classification learning focus their attention on as many diagnostic features as needed to solve the categorization problem (Shepard, Hovland, & Jenkins, 1961). Participants who are engaged in inference learning, however, focus their attention primarily on category labels, in order to learn prototypical features within each category in an attempt to predict missing features more accurately (Rehder et al., 2009). Lastly, classification learners have been shown to become sensitive to diagnostic features only, whereas inference learners have been shown to become sensitive to both diagnostic and prototypical features (Chin-Parker & Ross, 2002, 2004).
The goal of studies comparing classification and inference learning has generally been to determine whether or not the learning task affects concept formation. That is, does classification learning yield a different concept than inference learning? If so, specifically how are the concepts different? To determine the nature of the concept acquired through the classification and inference learning process, researchers usually employ a transfer phase that consists of both classification and inference trials (see Chin-Parker & Ross, 2002, 2004, for a different methodology). According to transfer-appropriate processing, this methodology confounds the results because participants’ memory for items is superior when the link between the learning task and the transfer task is conserved (e.g., classification learners classifying exemplars at transfer), whereas their memory for the same items is inferior when the link is broken (e.g., inference learners classifying exemplars at transfer). Therefore, basing a conclusion on concept acquisition on transfer data such as these is problematic (Markman & Ross, 2003; Yamauchi & Markman, 1998). I will now review the pertinent category-learning studies that have compared classification and inference learning.

Comparing Classification and Inference Learning

Yamauchi and Markman (1998) were the first to rigorously study the link between the type of category learning task and concept acquisition. In their first experiment, they assigned participants to one of three conditions: inference learning, classification learning, or a condition that involved a mix of both inference and classification learning trials. They reasoned that inference and classification learners would form qualitatively different concepts because the strategy required to learn to accurately classify is different from the strategy required to learn to accurately predict
missing features. They suspected that inference learners would pay attention to how features tend to appear within each category, due to the presence of the category label during learning trials, whereas classification learners would pay attention to a few diagnostic features or entire exemplars, in an attempt to learn what category each exemplar belongs to. Due to this difference in strategy, the authors hypothesized that inference learners would form a representation that is consistent with the category prototypes (i.e., the set of modal values for each feature within a category) and classification learners would form dimensional rules with some exceptions (Nosofsky, Palmeri, & Mckinely, 1994) or an exemplar representation (Medin & Schaffer, 1978; Nosofsky, 1986).

Table 1 displays an abstract description of the category structure Yamauchi and Markman (1998) implemented in their experiments (also used in Ahn & Medin, 1992; Reed, 1972; Medin, Wattenmaker, & Hampson, 1987; among many other studies). The structure consists of two categories with four exemplars each. Each exemplar has four features that can take on one of two values (i.e., 0 or 1). The prototype for Category A, displayed at the bottom of Table 1 (item A0), has the value 1 for all four features, while the prototype for Category B has the value 0 for all four features.

In the learning phase, classification learners were instructed to learn to accurately classify the eight exemplars (prototypes were not presented in the learning phase). They were free to use the appearance of any combination of the four binary valued features to meet the task demands. Inference learners, on the other hand, were instructed to learn to accurately predict the appearance of each exemplar’s diagnostic features. For item A1 in Table 1, the form, size, and color, were diagnostic, each with the value 1, whereas
Inference vs. Classification Learning

position, with value 0, was diagnostic of the opposing category and was not used as a learning phase item. With three items for each of the eight exemplars, there was a total of 24 inference learning items. Participants were free to use the category label, the appearance of the three remaining features, or any combination of the two to meet the task demands. Lastly, mixed learners were given intermixed classification and inference blocks of 8 trials each. After participants selected a response on any given learning trial, they were provided feedback and allowed 3 seconds to study the correct response.

Participants in all conditions were allowed up to 30 blocks of 8 trials (240 trials) to achieve 90% correct in 3 consecutive blocks. Once this criterion was met, the learning phase ended.

Table 1

*The category structure used in Yamauchi and Markman (1998) and Rehder, et al. (2009)*

<table>
<thead>
<tr>
<th>Feature Dimensions</th>
<th>Feature Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category A</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>A1</td>
<td>1 1 1 0</td>
</tr>
<tr>
<td>A2</td>
<td>1 1 0 1</td>
</tr>
<tr>
<td>A3</td>
<td>1 0 1 1</td>
</tr>
<tr>
<td>A4</td>
<td>0 1 1 1</td>
</tr>
<tr>
<td>A0</td>
<td>1 1 1 1</td>
</tr>
</tbody>
</table>

*Note.* Feature dimensions 1 through 4 were form (0 = triangle and 1 = square), size (0 = small and 1 = large), color (0 = green and 1 = red), and position (0 = right and 1 = left), respectively. The value 0 in Category A and the value 1 in Category B are prototype-inconsistent values. Exemplars A0 and B0 are the category prototypes.
In the transfer phase, all participants classified a random mixture of the 8 old items (learning phase items) and the two category prototypes. Then, they made inference decisions for a random mixture of the 24 old inference items as well as 8 new inference items. For the new inference items, the eight exemplars were presented with a missing feature value that is inconsistent with its category prototype (e.g., for item A1 in Table 1, the position of the stimulus, which has a value 0, was queried). These items are hereafter referred to as prototype-inconsistent inference items. Yamauchi and Markman (1998) cleverly used these items to detect a difference in category representation for classification and inference learning. The authors hypothesized that if inference learning yields a prototype representation and classification learning yields rules and exceptions or an exemplar representation, then inference learners would incorrectly select a prototypical feature value more often than classification learners when presented with a prototype-inconsistent inference item (e.g., for item A1 in Table 1, inference learners would select the value 1).

The learning phase results showed that inference learners required on average approximately half the number of blocks than the classification learners to reach the criterion ($M = 6.5$ vs. $M = 12.3$). The transfer results revealed that classification learners were more accurate than inference learners at classifying the old items ($M = 92\%$ vs. $M = 77\%$) and inference learners were more accurate than classification learners at making inference decisions for the old items ($M = 94\%$ vs. $M = 81\%$). These results are exactly what would be expected from transfer-appropriate processing (Morris et al., 1977), because accuracy was higher when the learning type matched the transfer item type. The results, however, also revealed that classification and inference learners were equally
successful in classifying the prototype exemplars (both $M = 96\%$), a result that is not consistent with transfer-appropriate processing. Finally, when presented with prototype-inconsistent inference items, inference learners were more likely to incorrectly select a prototypical feature value than classification learners ($M = 86\%$ vs. $M = 64\%$). This result was particularly supportive of the authors’ hypothesis that inference learners form a prototype representation and classification learners form rules and exceptions or an exemplar representation. Nevertheless, they noted that these results might not reflect a real difference in category representation because of the impact of transfer-appropriate processing on transfer performance.

In order to eliminate a transfer-appropriate processing explanation for the difference in prototype-inconsistent inference transfer performance in the first experiment, Yamauchi and Markman (1998) conducted a second experiment in which they assigned participants to one of two conditions that consisted of both learning tasks presented one after the other, with either classification first or inference first. Hence, by the end of the learning phase, participants in both groups had been given a chance to become proficient at both tasks and to process the stimuli both ways. The authors hypothesized that if inference learning yields a prototype representation, then inference-first learners would complete the subsequent classification learning task in fewer blocks than the classification-first learners, because a prototype representation is sufficient for highly accurate classification. On the other hand, if classification learning yields rules and exceptions or an exemplar representation, then the classification-first learners will not complete the subsequent inference learning task in fewer blocks than the inference-first learners, because rules and exceptions or an exemplar representation are insufficient
for highly accurate inference. The results were consistent with this hypothesis. Inference-
first learners met the classification learning criterion in fewer blocks than classification-
first learners ($M = 7.9$ and $M = 12.5$, respectively). Also, inference-first learners met the
inference learning criterion in marginally fewer blocks than classification-first learners
($M = 7.8$ vs. $M = 9.2$, respectively). This outcome suggests that the representation formed
during inference learning facilitated the subsequent classification learning task. It also
suggests that the representation formed during classification learning did not facilitate,
and even slightly hindered, the subsequent inference learning task performance,
presumably because the category representation had to be altered to meet the inference
learning criterion.

The transfer phase results showed that classification-first and inference-first
learners were equally accurate when making classification decisions for the old items ($M
= 95\%$ vs. $M = 91\%$) and the prototype items ($M = 88\%$ vs. $M = 81\%$). The results also
showed, however, that inference-first learners were more accurate than classification-first
learners in making inference decisions for the old items (95\% vs. 88\%). Finally, the key
finding from the first experiment was replicated. That is, when presented with prototype-
inconsistent inference items, inference-first learners were more likely to incorrectly select
a prototypical feature value than classification-first learners ($M = 86\%$ vs. $M = 54\%$).
Most importantly, this result could not have been explained by transfer-appropriate
processing because both groups were proficient at the classification and inference
learning tasks. Hence, the authors suggested that it was due to the hypothesized effect of
the different strategies for classification and inference learning on concept formation.
To test the hypothesis further, Yamauchi and Markman (1998) fit Nosofsky’s (1986) Generalized Context Model (GCM), to the classification and inference transfer data acquired in both experiments. The GCM assumes that participants store in memory all learning phase exemplars as sets of features. When making a subsequent categorization decision, the model computes the similarity of the features of the to-be-categorized item to all previously stored exemplars. The probability of a stimulus being categorized correctly is equal to its similarity with members of its category divided by its similarity to other memorized exemplars. The model had four free parameters that corresponded to the amount of attention paid to each feature. To accommodate for inference in the model, the authors treated the category labels as equivalent to features. The model fitting results showed that the GCM accounted well for classification transfer data after classification and inference learning in Experiment 1, but not after classification-first and inference-first learning in Experiment 2. The attention weights for these model fits were relatively high for three out of the four features. The GCM accounted very poorly for the inference transfer data from all learning conditions. The attention weights in this case were high for the category label and nearly zero for the three remaining features. Overall, the authors suggested that the model fitting results, although preliminary, were consistent with the idea that different strategies were adopted and representations used for inference and classification.

Based on the observed differences in prototype-inconsistent inference transfer accuracy in Experiments 1 and 2, the authors concluded that inference learning yielded a prototype representation and classification learning yielded dimensional rules and exceptions or an exemplar representation. As the authors explained, these results also
suggest that the representation formed during the inference portion of the inference-first learners training regimen and the classification portion of the classification-first learners training regimen were robust enough to undergo the subsequent learning task without significant modification.

In subsequent modeling of Yamauchi and Markman's (1998) data, Kruschke, Johansen, and Blair (1999) showed that a slightly different GCM could account for the transfer data in all conditions. Contrary to expectation, however, a similarly parameterized prototype model fit the inference transfer data for inference learners slightly worse than the exemplar model (.040 vs. .056 root mean squared deviations). This suggests that inference learners may have acquired an exemplar representation, not a prototype representation, as Yamauchi and Markman suggested. In further research, Johansen and Kruschke (2005) reasoned that it is possible that the simple category structure used in Yamauchi and Markman’s experiments could not differentiate exemplar from prototype representation. Therefore, in two experiments, they replicated Yamauchi and Markman’s first experiment using a category structure that has been shown to distinguish exemplar and prototype representation (Medin & Schaffer, 1978; Smith & Minda, 2000).

Table 2 presents an abstract description of this structure. It is known as the 5-4 structure (Smith & Minda, 2000), because one category has 5 exemplars and the other has 4. The categories are very difficult to learn via classification, because either all exemplars need to be memorized or a disjunctive rule involving three specific dimensions needs to be formed. For example, in Table 2, the categories can be accurately divided if dimension 2 is ignored and a best two out of three rule is applied. Specifically, if two out of three
values on dimensions 1, 3, and 4, are 1, the exemplar belongs to Category A, otherwise it belongs to Category B. Notably, the structure is difficult to learn via inference as well because the cue validity of the category label for predicting feature values is low, especially compared with a more simple structure like that used in Yamauchi and Markman's experiments.

Table 2

*The category structure Johansen and Kruschke (2005) used*

<table>
<thead>
<tr>
<th>Feature Dimensions</th>
<th>Feature Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category A 1 2 3 4</td>
<td>Category B 1 2 3 4</td>
</tr>
<tr>
<td>A1 1 1 1 0</td>
<td>B1 1 1 0 0</td>
</tr>
<tr>
<td>A2 1 0 1 0</td>
<td>B2 0 1 1 0</td>
</tr>
<tr>
<td>A3 1 0 1 1</td>
<td>B3 0 0 0 1</td>
</tr>
<tr>
<td>A4 1 1 0 1</td>
<td>B4 0 0 0 0</td>
</tr>
<tr>
<td>A5 0 1 1 1</td>
<td>B0 0 0 0 0</td>
</tr>
<tr>
<td>A0 1 1 1 1</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* The physical implementation of the abstract feature dimensions was randomly assigned using four of the five following features: head shape (round or square), nose direction (pointing up or down), tail length (short or long), antenna shape (straight or curved), and number of legs (four or eight). The value 0 in Category A and the value 1 in Category B are prototype-inconsistent values. Exemplars A0 and B0 are the category prototypes.

In the first experiment, Johansen and Kruschke (2005) instructed participants to learn about categories via classification or inference. For every learning trial, participants
were allowed 20s to make their decision and then 30s to study the item when presented with corrective feedback. Participants completed a fixed number of learning trials (25 blocks or 225 trials). The learning criterion was initially 90% correct in the last block of learning, but so few participants met this criterion in the classification condition that a less conservative one had to be used (75% correct in the last block). 26 out of 59 participants in the classification condition and 41 out of 58 participants in the inference condition met the criterion. The results replicated Yamauchi and Markman’s (1998) critical finding. Namely, when presented a prototype-inconsistent inference item, inference learners were more likely than classification learners to select a prototypical value ($M = 98\%$ vs. $M = 78\%$).

To test for the hypothesized difference between the representation formed during inference and classification, the authors applied three categorization models to the inference and classification transfer data. Specifically, they fitted the GCM (Nosofsky, 1986), a prototype model, and a rule-based model to the transfer data. The rule-based model assumed that participants mentally represented a series of bi-directional label-to-feature and feature-to-label rules. For inference learners, this meant that the modal feature value on each dimension for each category would be associated with the category label in a series of separate bi-directional rules. Thus, for a category structure with 4 dimensions and two categories, there would be 8 separate rules, one for each dimension and category label combination.

The model fitting results of the inference learners’ transfer data, measured in root mean squared deviations, showed that the prototype and rule-based models both fit the average inference (0.030 and 0.030 vs. 0.140) and average classification (0.039 and 0.039 vs. 0.039).
.135) data better than the GCM. In contrast, the results for the classification learners’
transfer data showed that the exemplar model provided a marginally better fit than the
prototype model to the average inference (.108 vs. .116) and average classification (.086
vs. 096) transfer data. These modeling results did support Yamauchi and Markman’s
(1998) hypothesis that classification learners acquire an exemplar representation. It was
not clear, however, whether inference learners acquired a prototype representation or a
rule-based representation.

Therefore, in a second experiment, the authors instructed participants to learn the
5-4 category structure through two different inference conditions: Inference of prototype-
consistent values only, which replicates Yamauchi and Markman’s (1998) methodology
(labeled here as the consistent group) or inference of prototype-inconsistent values only
(labeled the inconsistent group). This clever manipulation differentiated a prototype
representation from a rule-based representation because the prototype representation for
each condition would be exactly the same, whereas the rule-based representation for each
condition would be diametrically opposed. As a result, predictions derived from these
models for the learning and transfer results are in opposition. Specifically, if a set of bi-
directional label-based rules were mentally represented during inference learning, the two
inference groups would need a similar number of learning blocks to reach the learning
criterion and their transfer results would be diametrically opposed. Conversely, if a
prototype representation was formed during learning, the inconsistent group would not be
able to meet the learning criterion and transfer performance would be similar to that of
the consistent group.
At transfer, both groups were presented with the old classification items, new classification items as well as all possible inference items. In addition, there were inference items that no longer contained the category label. In support of Johansen and Kruschke’s (2005) rule-based model, a similar number of participants in both the consistent and inconsistent groups met the learning criterion of 90% correct for the last block of training (29 out of 50 and 26 out of 47, respectively). The inconsistent group classified 40% of the transfer items correctly, while the consistent group classified 72% of those items correctly. Moreover, inference transfer performance was essentially diametric for the two groups. For the prototype-consistent items, the consistent group inferred 88% of the feature values correctly, whereas the inconsistent group inferred 22% of the values correctly. For the prototype-inconsistent items, the consistent group inferred only 13% of the feature values correctly, whereas the inconsistent group inferred 88% of the values correctly.

The model fitting results showed that the rule-based model accounted better for the inconsistent group’s inference transfer performance than the prototype model. Specifically, for label-present inference transfer items, the rule-based model provided a better fit to the data than the prototype model (.122 vs. .242, measured in root mean squared deviations). Likewise, for label-absent inference transfer items, the rule-based model provided a better fit than the prototype model (.075 vs. .091). As was found in the first experiment, the rule-based and prototype models fit the consistent group’s inference and classification transfer data approximately the same. In summary, the authors concluded that the prototype model was mimicking a set of bi-directional label-based rules in the first experiment and that only the rule-based model could account for the
transfer performance of both inference learning groups in the second experiment. They therefore suggested that inference learning does not promote the development of a prototype category representation, as a prototype model assumes and as Yamauchi and Markman (1998) proposed, but it does promote bi-directional category-to-feature and feature-to-category rules. Johansen and Kruschke’s (2005) use of a category structure that was difficult for both classification and inference learners allowed them to uncover a hypothesized difference in representation between classification and inference that could not have been uncovered using a simple category structure (see Table 1).

Recently, Rehder et al. (2009) examined classification and feature inference learners’ eye movements throughout the course of learning. They obtained eyetracking data to evaluate two previously proposed accounts of inference learning. The first is Johansen and Kruschke’s (2005) set-of-rules model described above and the second is Markman and Ross’s (2003) hypothesis that inference learning motivates participants to learn the internal structure of categories (i.e., prototypes). The authors reasoned that if the set-of-rules account is true, inference learners should fixate on the category label only during learning trials. On the other hand, if inference learning motivates participants to learn the internal structure of categories, they should fixate not only on the category label but also on non-queried features during learning trials.

In the first experiment, the authors replicated Yamauchi and Markman (Experiment 1, 1998) using the structure displayed in Table 1 (with a different physical manifestation of the structure) and recorded participants’ eye movements throughout the learning phase. The learning phase results showed that inference learners reached the learning criterion in fewer blocks than classification learners ($M = 7.9$ vs. $M = 13.2$). The
transfer accuracy results showed that classification learners were more accurate than inference learners when classifying the old items ($M = 98\%$ vs. $M = 77\%$), as transfer-appropriate processing would predict. Inference learners, however, were slightly more accurate than classification learners when classifying the category prototypes ($M = 100\%$ vs. $M = 97\%$), although it appears both groups were at ceiling, a point that will be discussed later. Moreover, classification learners made inference decisions for the old items as accurately as inference learners ($M = 90\%$ vs. $M = 93\%$). The previous two results are inconsistent with transfer-appropriate processing. Lastly, when presented with a prototype-inconsistent inference item, inference learners were more likely than classification learners to select a prototypical feature ($M = 93\%$ vs. $M = 67\%$) just as Yamauchi and Markman (1998) and Johansen and Kruschke (2005) found.

The authors presented eye movement data for the inference learners in terms of probability of fixation on the three non-queried features or the category label as well as duration of fixation. They found that inference learners were likely to fixate (84\%), and spent much of their time fixating (78\%) the non-queried dimensions during the first block of the learning phase. In contrast, they were very likely to fixate (83\%), and spent a chance amount of time fixating (22\%) the category label during the first block of the learning phase. By the end of the learning phase, their tendency to fixate on the non-queried dimensions was still relatively high (>40\%) and time spent fixating was still significantly above zero. In contrast, throughout the learning phase, they were very likely to fixate (>90\%) and gradually spent more of their time fixating the category label (60\% at end). In summary, these results are consistent with previous findings that inference learners rely on the category label to make feature inferences (Yamauchi & Markman,
However, these results also show that non-queried dimensions are attended to during the learning phase and thus may be incorporated into the inference learners' concept. The authors hypothesized that non-queried dimensions may be attended to in order to increase accuracy in future feature prediction trials (labeled an anticipatory learning strategy hypothesis).

In order to further examine why inference learners attend non-queried dimensions, Rehder et al. (2009) conducted another experiment. This time participants were instructed to learn to predict the appearance of only two of the four features. This allowed the authors to determine whether non-queried features were being attended to for the purpose of learning the internal structure of the categories, Markman and Ross's (2003) hypothesis, or for the purpose of increasing accuracy on future feature prediction trials, the anticipatory learning strategy hypothesis. They reasoned that if inference learners were trying to learn the internal structure of the categories, the two never-queried dimensions would be fixated often during trials and for a fair amount of time. In contrast, if participants were engaged in an anticipatory learning strategy, they would attend to the two queried dimensions, but not the two never-queried dimensions, because those dimensions are not useful for making more accurate feature predictions on the two queried dimensions.

The transfer results showed that classification of the old items was slightly, but not significantly lower than in the first experiment ($M = 73\%$). This suggests that inference learners may only use one or two features to make classification decisions. Inference transfer performance for the queried versus never-queried dimensions was compared across old and prototype-inconsistent inference trials. Participants selected
prototypical values more often for the queried than the never-queried dimensions ($M = 92\%$ vs. $M = 63\%$) as was predicted by the anticipatory learning strategy hypothesis. However, prototypical values were selected for never-queried dimensions significantly above the 50 percent chance rate. The eyetracking data suggested that at first participants were attending to never-queried features, but that as the learning phase progressed, they stopped attending to those features. The data also showed that attention was mostly on the category label and the queried dimensions. Thus, again, the data did not support the hypothesis that inference learning motivates participants to learn the internal structure of the categories, nor did it fully support the assumption of the set-of-rules model. Participants simply seemed to behave in a way that allowed them to cope with the task demands.

The authors suspected that it is possible that participants attended to and learned about the never-queried features during feedback, which would account for the above chance tendency to select prototypical features for never-queried dimensions. Thus, the authors presented eyetracking data that was collected during the 3 seconds of post-trial feedback. The data showed that never-queried features were attended to during feedback with a moderate probability earlier in the learning phase (~40\%) and with a decreasing probability later on in the learning phase (~20\%). The authors also thought that never-queried feature learning might have occurred incidentally at the start of the learning phase, while participants were figuring out which features would be queried. Hence, in a final experiment, the authors allowed participants to press a button to terminate feedback and they told participants which dimensions would be queried. The results showed substantially less fixation on never-queried dimensions and also chance performance on
Inference vs. Classification Learning

the never-queried inference transfer items. Furthermore, a one-feature classification test that followed learning revealed chance performance on never-queried features and ceiling performance on queried features.

In summary, Rehder et al. (2009) have shown that inference learning does not invoke in participants a motivation to learn about the categories internal structure per say (i.e., prototypes), nor does it promote exclusive feature-to-category and category-to-feature bi-directional rules, but it seems to promote learning of multiple feature-to-category relationships through an anticipatory learning strategy. However, for the present thesis, it is vital to note that predictions from transfer-appropriate processing were violated in this study. That is, in the first experiment, classification and inference learners were equally accurate when making feature-inferences ($M = 93\%$ vs. $M = 90\%$) and they were equally accurate when classifying the category prototypes ($M = 100\%$ vs. $M = 97\%$).

Critical Evaluation of the Reviewed Studies

In this thesis, whether differences in strategies and representation for classification and inference learning exist is not a matter for debate. Yamauchi and Markman (1998), Johansen and Kruschke (2005), and Rehder et al.'s (2009) research, among other studies (Chin-Parker & Ross, 2002, 2004), have made it clear that different strategies and category representations are acquired through the classification and inference learning process. Rather the foci of the thesis are the violations of transfer-appropriate processing (Morris et al., 1977) that were found in the reviewed studies.

In the context of category-learning research, transfer-appropriate processing predicts that transfer task performance should be best when the task performed during the
learning phase matches the task performed during the transfer phase (e.g., classification learners classifying exemplars at transfer) relative to when there is a mismatch (e.g., inference learners classifying exemplars at transfer). In direct violation of these predictions, Yamauchi and Markman (1998) as well as Rehder et al. (2009) found that inference learners were as accurate as classification learners when classifying the category prototypes at transfer. An inspection of Table 3 shows that, in both studies, transfer task accuracy for prototype classification was at ceiling. Also, Rehder et al. found that classification learners were as accurate as inference learners when they made feature-inferences for old inference items presented at transfer. Again, an inspection of Table 3 shows that old inference accuracy in Rehder et al.'s experiment was close to ceiling.

This thesis addressed the possibility that prototype and old inference accuracy were at ceiling because the category structure was too simple, as it was comprised of two categories with eight exemplars that varied on four binary dimensions (see Table 1). The simplicity of the structure and its equivocality in distinguishing different representations was the reason why Johansen and Kruschke (2005) used a more complicated category structure in their investigation. I thus decided to build on Johansen and Kruschke's strategy and built an even more complex category structure. It included 24 exemplars that belonged to one of four general categories. Each exemplar was composed of three features that could take one of nine values (see Table 4). With four possible categories, as opposed to only two, chance performance was lower (25% vs. 50%) which created a larger performance range with which to observe differences between the classification and inference learners’ classification accuracy at transfer.
Table 3

A summary of results from studies comparing classification and inference learning

<table>
<thead>
<tr>
<th>Study</th>
<th>Learning Task</th>
<th>Number of Learning Blocks</th>
<th>Transfer Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Old Classification</td>
</tr>
<tr>
<td>Yamauchi &amp; Markman (1998)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp 1</td>
<td>Classification</td>
<td>12.3</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>Inference</td>
<td>6.5</td>
<td>.77</td>
</tr>
<tr>
<td>Exp 2</td>
<td>Classification-First</td>
<td>12.5 9.2</td>
<td>.91</td>
</tr>
<tr>
<td></td>
<td>Inference-First</td>
<td>7.9 7.8</td>
<td>.95</td>
</tr>
<tr>
<td>Johansen &amp; Kruschke (2005)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp 1</td>
<td>Classification</td>
<td>25</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>Inference</td>
<td>9</td>
<td>.76</td>
</tr>
<tr>
<td>Exp 2</td>
<td>Consistent</td>
<td>9</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>Inconsistent</td>
<td>20</td>
<td>.40</td>
</tr>
<tr>
<td>Rehder et al. (2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp 1</td>
<td>Classification</td>
<td>13.2</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td>Inference</td>
<td>7.9</td>
<td>.77</td>
</tr>
</tbody>
</table>

Note. Dashes indicate the data were not presented. Transfer task performance is presented as proportion correct. Consistent = prototype-consistent inference items. Inconsistent = prototype-inconsistent inference items.
Table 4

*The category structure and labels used in the present experiment*

<table>
<thead>
<tr>
<th>Spec</th>
<th>Exemplar</th>
<th>Color</th>
<th>Shape</th>
<th>Pattern</th>
<th>Spec</th>
<th>Exemplar</th>
<th>Color</th>
<th>Shape</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sot</td>
<td>SM1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SM2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Wok</td>
<td>WH1</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SM3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td>WH2</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>IM1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td>WH3</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Iko</td>
<td>IM2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>Sin</td>
<td>SH1</td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>IM3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td></td>
<td>SH2</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>VC1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td></td>
<td>SH3</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Vin</td>
<td>VC2</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>Hap</td>
<td>HT1</td>
<td>7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>VC3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td></td>
<td>HT2</td>
<td>7</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>KC1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td></td>
<td>HT3</td>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Kol</td>
<td>KC2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>Ana</td>
<td>AT1</td>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>KC3</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td></td>
<td>AT2</td>
<td>8</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

*Note.* The feature values 0 through 8 represent different perceptual manifestations of each feature. Feature value 0 is non-diagnostic of the general and specific categories. Category prototypes are not displayed here but, for example, the “Sot Maki” prototype is “1 1 1” and the “Maki” prototype is bimodal “1 1 1” and “2 2 2”. Spec = Specific category label.
An interesting aspect of this structure was that the prototype-inconsistent values (e.g., the value 0 in Table 4) were not diagnostic of any of the four categories. In the structures displayed in Table 1 and 2, however, the prototype-inconsistent values (e.g., value 0 in Category A and value 1 in Category B) were diagnostic of the opposing category. Therefore in the studies that were reviewed, when participants were presented with a prototype-inconsistent inference item, they had to select one of two features. There was a prototypical feature for the item’s category (e.g., value 1 in Category A) and a feature that was diagnostic of an opposing category (e.g., value 0 in Category A). It is possible that the difference in the classification and inference learners’ tendency to select this prototypical feature stemmed from the fact that the alternative response is one that is diagnostic of the opposing category, instead of a genuine difference in preference for the prototypical feature. Hence, the structure that was used in the present experiment could evaluate this possibility.

Furthermore, given the vital role that prototype-inconsistent inference items have played in distinguishing category representations for classification and inference, the present study examined the impact of manipulating response choice for these items. For example, for the SM1 exemplar presented in Table 4, participants were offered as response options the prototype-inconsistent value (i.e., the value 0), the prototypical feature for SM1’s specific category (i.e., the value 1), as well as two categorically invalid features (e.g., values 3 through 8). This condition replicated the response choices offered for prototype-inconsistent inference items in the reviewed studies. In addition to this response choice condition, however, there were three other conditions. A description of each of these conditions follows, again using the SM1 exemplar as an example. In one
condition, the prototype-inconsistent value was presented along with three categorically invalid features. In another condition, the prototype-inconsistent value was presented with a prototypical feature for SM1's general, but not specific category, (i.e., the value 2), and two other categorically invalid features. In a final condition, the prototype-inconsistent value was presented with both prototypical features and one categorically invalid feature. These response choice conditions were implemented to more systematically determine response preference for prototype-inconsistent inference transfer.

Another feature of the structure presented in Table 4 that differs from the structures presented in Tables 1 and 2, was the hierarchical nature of the category structure. This deviation from the structures used in previous studies was justifiable because research has shown that categories are organized in a hierarchical fashion (Rosch & Mervis, 1975; Murphy, 2002). In the present study, participants were only provided information about the general categories during learning trials. However, during feedback, information about the exemplars' specific category was presented. In the transfer phase, classification and inference items were presented with specific category labels in order to assess the extent of specific level category learning. The hierarchical nature of the structure was incorporated as a preliminary step to studying whether there are differences between classification and inference in terms of acquiring knowledge at different levels of a category hierarchy.

Out of the studies reviewed above, Yamauchi and Markman (1998) were the only researchers to assess how classification and inference learners responded to the critical prototype-inconsistent inference items, in a case where transfer-appropriate processing
clearly could not explain the results (i.e., the classification-first and inference-first conditions). For the present study, in addition to a classification and inference learning condition, there was also a classification-first and an inference-first learning condition. This allowed for a reassessment of the two learning orders, in terms of their learning and transfer phase performance and in particular an assessment of how they would respond to prototype-inconsistent inference items when response options were systematically manipulated.

Across the reviewed studies, there were considerable differences in the amount of time participants were exposed to stimuli during corrective feedback. For instance, Yamauchi and Markman (1998) and Rehder et al. (Experiments 1, 2, and 3, 2009) allowed 3 seconds to study items during feedback, Johansen and Kruschke (2005) allowed up to 30 seconds, and Rehder et al. (Experiment 4) allowed a self-paced study time. Because transfer phase analyses involved comparing learning groups, it was imperative to control important extraneous variables between the learning groups, such as exposure time to stimuli during feedback. It seemed likely that a longer interval or self-paced study time would not have been the best way to control this variable because it would have allowed greater variance in actual study time and could have therefore lead to differences in incidental learning between groups (Rehder et al., 2009). Thus, in the present research, feedback was limited to a brief, but sufficient study time interval for all learning groups.

Another related difference was the number of blocks to criterion and thus the amount of exposure to the stimuli between learning groups. As was noted earlier, there have been two criteria typically employed to decide when participants have become
sufficiently proficient at their learning task. Either the number of trials is fixed a priori and only participants who perform beyond a preset accuracy level are included in the transfer analyses, or a learning criterion is set a priori and the learning phase ends once participants met it. Yamauchi and Markman (1998) and Rehder et al. (2009) employed the latter option, whereas Johansen and Kruschke (2005) employed the former. In the present experiment, I followed Johansen and Kruschke’s lead and fixed the number of learning phase blocks and did not include participants in transfer analyses if they did not meet a certain learning criterion. This way, participants’ exposure to stimuli during the learning phase was controlled for as best as possible.

Across the reviewed studies, the learning phase consisted of a limited number of learning blocks (see Table 3). In particular, in Yamauchi and Markman’s (1998) first experiment, inference learners completed 6.5 blocks and classification learners completed 12.3 blocks on average. In their second experiment, classification-first learners completed 12.5 blocks of classification and 9.2 blocks of inference, while the inference-first learners completed 7.9 blocks of classification and 7.8 blocks of inference. In the present experiment, the number of learning blocks was manipulated, with a short session designed to be comparable to the reviewed studies, and an extensive session designed to allow for asymptotic performance at the end of the learning phase.

In a final note about the reviewed studies, accuracy was the main dependent measure in the transfer phase analyses that were reported. Response time, however, can provide additional insight regarding the learning task difficulty and transfer of knowledge acquired through the various learning tasks. In fact, transfer-appropriate processing (Morris et al., 1977) makes specific predictions for the classification, inference, and
mixed learners’ response times to make classification and inference decisions at transfer, which will be outlined below.

**Goals and hypotheses**

The goal of the experiment was to examine the effect of classification, inference, and mixed category learning of a complex category structure on participant’s subsequent ability to make classification and inference decisions, over the course of a short or an extensive learning session. For the learning analyses that involved the mixed learning conditions, two potential outcomes can be expected. If Yamauchi and Markman (1998) are correct that inference learning facilitates classification learning but that classification learning does not facilitate inference learning, then classification-first learners should require more blocks to meet the inference learning criterion than inference-first learners. However, with a complex category structure and extensive learning, I expected that both classification and inference learning should facilitate the transition into a subsequent learning task in the mixed learning conditions. Specifically, I expected that classification-first learners would require fewer blocks than inference-first learners to meet the inference learning criterion and inference-first learners would require fewer blocks than classification-first learners to meet the classification learning criterion. For the learning length manipulation, I expected that participants assigned to extensive learning would have higher accuracy and lower response time at the end of the learning phase than those assigned to short learning.

For the transfer phase, I expected that accuracy and response time to the old classification and old inference items as well as the new prototype items would yield a clear transfer-appropriate processing pattern (Morris et al., 1977). More specifically, I
expected that inference learners would be more accurate and faster than classification learners when making inference decisions for the old items. Likewise, I expected that classification learners would be more accurate and faster than inference learners when making classification decisions for the old items, as well as the new prototype items. For the mixed learning conditions, I expected that classification-first and inference-first learners would be equally accurate and fast when making classification decisions for the old items and the new prototype items as well as when making inference decisions for the old items. Moreover, it was reasonable to expect that the classification-first and inference-first learners’ response time and accuracy for all old classification and inference items as well as the prototypes would be in between the observed values for the inference and classification learners. This pattern was expected because the mixed learners spent half the number of blocks performing each learning task relative to the classification and inference learners. Lastly, I expected that participants assigned to an extensive learning session would have higher accuracy and lower response time for the old classification, inference, and new prototype items than those assigned to a short session.

For the prototype-inconsistent inference items, two potential outcomes could be expected. If Yamauchi and Markman (1998) were correct about the differences in representation between inference and classification learners, then inference and inference-first learners would be more likely than classification and classification-first learners to select a prototypical feature value when presented with a prototype-inconsistent inference item. Specifically, they should select a prototypical feature for the item’s specific category and a prototypical feature for the item’s general, but not specific
category (e.g., the values 1 and 2, respectively, for the SM1 exemplar in Table 4) more often when these choices are available as response options. Also, I expected that classification learners would select the prototype-inconsistent value more often than inference learners, because classification learning has been shown to yield a representation that encompasses the prototype-inconsistent values more than inference learning (e.g., an exemplar vs. a prototype representation). Finally, classification-first and inference-first learners were expected to share the same tendency to choose prototypical features, an outcome that is consistent with transfer-appropriate processing (Morris et al., 1977).

Method

Participants

122 undergraduate students, who were recruited through Carleton University’s Psychology Experiment Sign-Up System, volunteered to participate in the experiment for bonus percentage points that were applied to their final grade in a first or second year psychology course. Participants assigned to short learning were granted 2 bonus percentage points and those assigned to extensive learning were granted 4 bonus percentage points.

Stimuli

The learning phase stimuli were 24 fictional creatures that belonged to one of four general categories, and to one of two more specific categories. The creatures were composed of three varying features that were perceptually implemented using nine feature values: Head color (red, blue, green, yellow, brown, orange, grey, purple, or white), body shape (square, triangle, octagon, oval, diamond, hour glass, parallelogram,
half moon, or a cross), and chest pattern (circles, vertical lines, bumpy horizontal lines, exes, stars, arrows, addition signs, hatched lines, or curved diagonal lines). The creatures' other features did not vary (e.g., number of antennas and eyes were the same for all creatures). Table 4 presents the category structure along with the chosen category labels. Each creature belonged to a general category (e.g., “Maki”) and to one of two more specific categories (e.g., “Sot Maki” or “Iko Maki”). For instance, exemplar SM1 was a “Sot Maki”, whereas exemplar IM1 was an “Iko Maki”. Figure 1 shows an example of three “Sot Maki” exemplars.

![Figure 1. From left to right are “Sot Maki” exemplars, SM1, SM2, and SM3 with red (darker) or yellow (lighter) head color, square or hourglass body shape, and bumpy horizontal line or vertical line chest pattern. Note that the hourglass body shape, the yellow head color, and the bumpy horizontal line chest pattern are not diagnostic of category membership.](image)

It was likely that the salience of the nine feature values selected for each dimension differed. In order to address this concern, one option was to partially counterbalance the feature values within and across dimensions using a Latin square. However, this would have resulted in 729 partial counterbalancing conditions. Thus, it
was clear that partial counterbalancing was not a realistic option. Therefore, prior to the onset of the learning phase, each participant received a random assignment of the perceptual feature values to the abstract category structure and the structure that resulted was used throughout the experiment (e.g., Johansen & Kruschke, 2005; Smith, Redford, Gent, & Washborn, 2005). With the random assignment of the feature values to the category structure, it was very unlikely that any particular salient feature value or groups of feature values within or across dimensions and categories spuriously affected the learning and transfer phase results.

The learning phase stimuli for classification learners were the 24 fictional creatures presented without their general category label (e.g., “Maki” for exemplar SM1 in Table 4). The learning phase stimuli for inference learners were the 24 creatures presented with their general category label and without one of two diagnostic feature values. Because each exemplar consisted of two diagnostic feature values, there were 48 unique inference learning stimuli. Figure 2 displays an example of a classification item with its two corresponding inference items.

![Classification and Inference Items](image)

*Figure 2.* From left to right is a classification learning item and the two corresponding inference learning items for the SM1 exemplar.
The transfer stimuli were the old classification and old inference items, the category prototypes (e.g., for the “Maki” category, there were two prototypes: an exemplar with values 1, 1, 1 and an exemplar with values 2, 2, 2), and prototype-inconsistent inference items (e.g., the SM1 exemplar was missing its chest pattern with value 0). Specifically, there were 48 old classification items (24 with general labels and 24 with specific labels), 96 old inference items (48 with general labels and 48 with specific labels), 48 prototype items (24 with general labels and 24 with specific labels), and 96 prototype-inconsistent inference items (with general labels only). In total, there were 288 transfer phase stimuli.

Procedure

Participants were tested individually on a Dell Dimension 5150 desktop computer (with an Intel Pentium 4 processor) using E-Prime software version 1.1 (Schneider, Eschman, & Zuccolotto 2002a; 2002b). This software was used to display the experimental instructions, present the stimuli, and collect participants’ response time and accuracy. Participants were randomly assigned to a classification, inference, or to one of two mixed-learning conditions with classification-first or inference-first. They were also randomly assigned to a short or an extensive learning condition. Finally, they were also randomly assigned to a counterbalancing condition, in which one group received inference transfer first and the other received classification transfer first.

The experiment was conducted in two phases. For the learning phase, inference learners were instructed to predict the creatures’ missing feature value based on its category label and the remaining feature values. Classification learners were instructed to predict the creatures’ general category label based on its feature values. Participants were
Inference vs. Classification Learning

informed that at first they would have to guess, but that by studying the correct response during feedback, they could attain 100 percent accuracy.

Learning trials proceeded as follows. Seven horizontally arranged asterisks appeared in the centre of the computer screen for 500 ms. Then, for inference learners, a creature with a missing diagnostic feature appeared in the middle of the computer screen with its appropriate general category label directly above it. Four response choices were presented horizontally below the creature. The choices were randomly associated with the letters “A”, “S”, “K”, or “L”. The four choices included the correct feature value and a randomly chosen diagnostic feature value from each of the three remaining general categories. The display for classification learners was different in that the creature was complete, the general category label was not presented above, and the four choices below were the correct general category label and the three other general category labels. Participants were provided up to 6000 ms to select their response by pressing the corresponding “A”, “S”, “K”, or “L” key on the keyboard. Once the response was selected, feedback was provided immediately. Incorrect responses were accompanied by a tone that sounded for 500 ms. If participants failed to respond within 6000 ms, a tone sounded for 500 ms, the correct response was displayed, and a message encouraging them to make faster responses was presented below the creature. The feedback remained on the screen and participants were provided 2000 ms to study the labeled and complete creature. The creature’s specific category label (e.g., “Sot Maki”) was displayed on the screen during corrective feedback. Note also that classification, inference, and mixed-learners were allowed the same amount of time to study the exact same complete and labeled creature during feedback. A 250 ms inter-stimulus-interval occurred, and then the
trial was complete. Figure 3 displays an example of the timeline and the arrangement of the stimuli, the category label, and the four response choices on the computer screen for an inference learner.

A learning block for classification and inference learners consisted of 48 trials. Inference learners inferred two different diagnostic feature values for the 24 creatures (e.g., color and form for the SMI exemplar in Table 4), whereas classification learners classified each creature twice. Thus, within one block, the feedback for both inference and classification learners consisted of two presentations of each creature with its specific label. The order of presentation of the creatures within each block was determined randomly at the onset of each block. Participants’ average response time and accuracy for the block were displayed on the computer screen at the end of every block, before the following block began. These procedures were identical for the classification-first and inference-first learners.

A pilot study was conducted in order to determine how many learning blocks would be required for classification and inference learners to achieve 90% correct. Participants completed 30 learning blocks. The results showed that classification and inference learners required on average 8.09 ($SD = 3.51$) and 12.62 ($SD = 5.78$) blocks, respectively, to achieve 90% correct. In order to ensure that a large proportion of participants in the short learning condition in the present study would have achieved 90% correct by the end of the learning phase, they were assigned 16 blocks (768 trials). Participants in the extensive learning condition were assigned an additional 24 blocks for a total of 40 learning blocks (1920 trials). Participants took 3.25 hours on average to complete the pilot experiment. After completing the experiment, I asked participants how
many more learning blocks they would have been able to do. The modal response was “no more blocks”. Thus, I decided to split the learning phase for the present experiment into two consecutive days, with 8 blocks each for the short learning condition and 20 blocks each for the extensive learning condition. Classification-first learners started the second day with inference learning and inference-first learners started the second day with classification learning. Extensive learners were offered a short break at the end of every 10th block (480 trials).

**Figure 3.** The timeline and display for inference learning trials.

Following the learning phase on the second day, participants were presented the transfer phase instructions. They were told that the purpose of the transfer phase was to
test how quickly and accurately they could make classification and inference decisions. All participants were told that feedback would not be provided. Because the classification and inference groups solely focused on classification and inference, respectively, during the learning phase, an explanation of the task that was unknown was provided. They were shown, on the computer screen, an example of a creature missing either the label (for inference learners) or a feature (for classification learners). They were told that the creature was not a member of the categories they learned and that it was provided in order to present an example of the unfamiliar task. The task that they were not familiar with was contrasted with the one they were familiar with. For instance, a classification learner was told that an inference would require them to predict the appearance of a missing feature based on the category label and the remaining features, instead of predicting the category label based on all of the features. Classification-first and inference-first learners were already familiar with both tasks. Therefore, they were not shown an example or provided an explanation of either task.

The transfer trials were the same as the learning trials, except for the following two differences. First, the 2000 ms feedback was not provided. Secondly, for classification trials that assessed specific level knowledge (old and prototype), specific labels were presented as response options (e.g., “Sot Maki” and “Iko Maki” for exemplar SM1 in Table 4) along with two specific labels of a randomly selected alternative category (e.g., “Vin Cura” and “Kol Cura”). For old inference items that assessed specific level knowledge, the exemplar’s specific label was presented above the creature. For prototype-inconsistent inference items, there were the 24 possible items (e.g., in Table 4, each exemplar was missing the value 0) presented four times, each time with four
different response choices. In one condition, the prototype-inconsistent feature value (i.e., 0) was presented with 3 categorically invalid features (hereafter represented as 0XXX). In a second condition, the prototype-inconsistent value was presented with a prototypical feature of the exemplar’s specific category (e.g., value 1 for exemplar SM1 in Table 4) as well as two categorically invalid features (hereafter represented as 01XX). In a third condition, the prototype-inconsistent feature value was presented with a prototypical feature of the exemplar’s general, but not specific category (e.g., value 2 for exemplar SM1 in Table 4), as well as two categorically invalid features (hereafter represented as 02XX). In the fourth condition, the prototype-inconsistent feature value was presented with both prototypical features and one categorically invalid feature (hereafter represented as 012X).

The classification, inference, and prototype-inconsistent inference transfer trials were not inter-mixed. The reason the transfer tasks were separated was that the cognitive demand of task switching could have adversely impacted the response time and accuracy dependent measures (Monsell, 2003). The order of presentation of the transfer tasks was partially counterbalanced, with the prototype-inconsistent inference trials always presented as the last transfer task. Thus, half of the participants were presented classification, inference, and then prototype-inconsistent inference trials, while the other half were presented inference, classification, and then prototype-inconsistent inference trials. The order of presentation of items within each transfer task was randomized. Following the transfer phase, participants were debriefed and thanked for their participation.
Results

107 out of the 122 participants who signed up to participate in the experiment completed Sessions 1 and 2. Of those 15 participants who did not complete the experiment, 7 did not show up for Session 2 and the data for 8 participants were lost because of run-time errors. There were an unequal number of participants in the groups with 13 in short classification, 18 in short inference, and 16 in the extensive inference condition. There were 12 participants in all 5 other groups.

For all of the following response time analyses, incorrect responses were discarded. Then, responses that were made in 150 ms or less were also discarded as they were produced in less time than it takes to make a simple choice in a choice reaction time experiment (Ratcliff, 1993). Finally, using a non-recursive shifting z-score criterion set at 2.5 standard deviations above or below the mean, response time outliers within each cell in the factorial design were also discarded from analyses (Van Selst & Jolicoeur, 1994). This procedure eliminated 2.40% of the learning phase response times and an average of 2.28% for each of the transfer item types.

For all of the following analyses of variance (ANOVAs), if there was a violation of the homogeneity of variance assumption, as indicated by Levene’s test of equality of error variances ($p <= .05$), Dunnett T3 post-hoc pairwise comparisons were used to further explore significant main effects. If the homogeneity of variance assumption was not violated, Tukey HSD post-hoc pairwise comparisons were conducted instead. For all of the following chi-square tests of independence, the assumption that there are at least five observations in each cell was confirmed. Also, if the chi-square was significant, post-hoc pairwise comparisons were conducted on the independent variable within each level.
of the dependent variable using a Bonferroni correction. These comparisons were considered significant if $p < .05$.

**Learning phase: Accuracy and response time analyses**

.94% of all responses were deleted from the learning data set because participants failed to respond within the 6-second time limit. To examine the pattern of these timed-out responses across the groups, a 2 x 4 randomized-groups analysis of variance (ANOVA) with Learning Length (short vs. extensive) and Learning Type (classification, inference, classification-first, and inference-first) was performed on mean number of timed-out responses. Because the Learning Length groups differed greatly in the number of blocks they completed (16 vs. 40), the relative mean number of timed-out responses was artificially inflated for the short learning group. Thus, only the first 16 blocks were used in the analysis. The interaction and main effects were not statistically significant, $p >= .20$. Therefore mean number of timed-out responses did not differ across the groups.

The accuracy and response time measures for the final block of the learning phase were submitted to separate 4 x 2 randomized groups ANOVAs with Learning Type and Learning Length as between-subjects variables. The goal of these analyses was to determine if there were any systematic differences in accuracy and response time between the different Learning Types and Learning Lengths at the end of the learning phase. Figure 4 and 5 display mean accuracy and response time, respectively, for all learning blocks as a function of Learning Type and Learning Length.

The ANOVA on accuracy for the final learning block yielded a significant main effect of Learning Type, $F(3,99) = 4.31, p < .01, \text{partial } \eta^2 = .12, MSE = .02$, and Learning Length, $F(1,99) = 5.28, p = .024, \text{partial } \eta^2 = .051$. The interaction between
Learning Type and Learning Length, however, was not significant, $F(3,99) = 1.09, p = .358$, partial $\eta^2 = .032$. Post-hoc comparisons showed that inference learners ($M = 83.8\%$, $SD = 19.1\%$) were less accurate at the final block of the learning phase than inference-first learners ($M = 95.5\%$, $SD = 10.1\%$), $t(56) = -3.03$, $SE = 3.87\%$, $p = .022$, and marginally less accurate than classification-first ($M = 93.6\%$, $SD = 8.9\%$), $t(56) = -2.64$, $SE = 3.74\%$, $p = .063$, and classification learners ($M = 94.7\%$, $SD = 15.5\%$), $t(57) = -2.44$, $SE = 4.51\%$, $p = .10$. The significant main effect of Learning Length indicated that participants who were assigned extensive learning ($M = 94.9\%$, $SD = 10.3\%$) had higher accuracy in the final block of the learning phase than those who were assigned short learning ($M = 87.7\%$, $SD = 18.2\%$). As can be seen in Figure 4, at the end of the learning phase, accuracy was high in the short learning condition (although less so for inference learners) and at asymptote for extensive learners.

The ANOVA on response time yielded a significant main effect of Learning Type, $F(3,99) = 4.43, p < .01$, partial $\eta^2 = .12$, $MSE = 269929$. Learning Length, however, $F(1,99) = .039, p = .84$, and the interaction between Learning Type and Learning Length were not significant, $F(3,99) = .091, p = .97$. Post-hoc comparisons showed that classification learners ($M = 1422$ ms, $SD = 463$ ms) were faster to respond at the final block of the learning phase than inference ($M = 1864$ ms, $SD = 595$ ms), $t(47) = -3.22, p < .01$, and classification-first learners ($M = 1849$ ms, $SD = 527$), $t(45) = -3.24, p = .025$. The other comparisons were not significant, $p > .23$. As can be seen in Figure 5, the differences at the final block of the learning phase seem to be a result of a consistently higher response time needed to perform the inference task relative to the classification task.
Figure 4. Mean percent correct for short (top) and extensive (bottom) learning as a function of Learning Type and Block. Block 9 (top) and Block 21 (bottom) was the first block on the second day of learning. Error bars represent standard error of the mean.
Figure 5. Mean response time for short (top) and extensive (bottom) learning as a function of Learning Type and Block. Block 9 (top) and Block 21 (bottom) was the first block on the second day of learning. Error bars represent standard error of the mean.
To determine whether there were facilitation effects on learning (i.e., whether classification learning facilitated subsequent inference learning and whether inference learning facilitated subsequent classification learning), the number of blocks to reach 90% correct in three consecutive blocks, Yamauchi and Markman's (1998) criterion, to a 2 x 2 mixed ANOVA with Learning Order (classification-first vs. inference-first) as a between-subjects variable and Learning Task (classification vs. inference) as a within-subjects variable. Unfortunately a comparison involving the short and extensive learning conditions could not have been conducted because only 1 participant in the short inference-first condition and 3 in the classification-first condition met the learning criterion for the first learning task. Even using a less stringent criterion (75% correct in one block), only 4 participants in the inference-first condition would have been included. Thus, only participants who were assigned to an extensive learning session (12 classification-first and 10 inference-first) were included in the following ANOVA.

It yielded a significant interaction between Learning Order and Learning Task, $F(1,20) = 23.03, p < .001, MSE = 8.97$, partial $\eta^2 = .53$, as well as a significant main effect of Learning Task, $F(1,20) = 57.10, p < .001$, partial $\eta^2 = .74$. The main effect of Learning Order was not significant, $F(1,20) = .41, p = .53$. Table 5 displays mean number of blocks to reach 90% correct in three consecutive blocks as a function of Learning Order and Learning Task. A simple main effect analysis of Learning Order for each Learning Task showed that inference-first learners required more blocks to reach criterion than the classification-first learners for the first learning task, $t(20) = 3.40, SE = 1.59, p < .01$. After having completed the first task, however, inference-first learners required marginally fewer blocks to complete the second task than the classification-first.
learners, $t(20) = 1.56$, $SE = 2.11$, $p = .13$. To directly compare results with Yamauchi and Markman (1998), a simple main effect analysis on Learning Task for each Learning Order was computed. For the inference learning task, participants required fewer blocks to reach criterion if classification was performed prior to inference $t(20) = -3.28$, $SE = 2.41$, $p < .01$. For classification learning, participants required fewer blocks to reach criterion if inference was performed prior to classification, $t(20) = -5.32$, $SE = 1.09$, $p < .001$. Hence, when participants were provided a substantial amount of learning blocks, both learning tasks significantly facilitated performance on the subsequent learning task. Recall that Yamauchi and Markman had found that inference learning facilitated classification learning, but that classification learning did not facilitate inference learning.

Table 5

<table>
<thead>
<tr>
<th>Learning Order</th>
<th>Learning Task</th>
<th>Inference</th>
<th>Classification</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inference-First</td>
<td>Inference</td>
<td>16.4 (1.2)</td>
<td>5.20 (1.6)</td>
<td>10.8 (1.2)</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification-First</td>
<td>Inference</td>
<td>11.0 (1.1)</td>
<td>8.5 (1.4)</td>
<td>9.8 (1.1)</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>13.7 (.79)</td>
<td>6.9 (1.1)</td>
<td>9.9 (.8)</td>
</tr>
</tbody>
</table>

*Note.* Standard error of the mean is in parentheses.

In order to analyze how switching learning tasks affected response time for classification-first and inference-first learners, response times for all mixed learners were submitted to a $2 \times 2 \times (2)$ mixed ANOVA with Learning Order and Learning Length as a
between-subjects variables and Task Switch (last block of first task vs. first block of second task) as a within-subjects variable. The interaction between Learning Order and Task Switch was significant, $F(1,44) = 41.08, p < .001$, partial $\eta^2 = .48$, $MSE = 138110$. The main effect of the Task Switch was also highly significant, $F(1,44) = 58.11, p = .001$, partial $\eta^2 = .57$. All other effects failed to reach significance, all $F_s < 1.60$ and $p \geq .21$.

Table 6 displays mean response time at the last block of the first task and the first block of the second task as a function of Learning Order and Learning Task. A simple effects analysis of Learning Order on each level of Task Switch was computed. The analysis showed that inference-first learners were slower than classification-first learners during the final block of the first learning task, $t(46) = 4.59, SE = 125, p < .001$. In contrast, classification-first learners were slower than inference-first learners at the first block of the second learning task, $t(46) = 2.82, SE = 142, p < .01$. This difference was a result of the inference task requiring more time to complete than the classification task. Next, a simple main effect analysis of the Task Switch on each level of Learning Order was performed. The results showed that the Task Switch (from Block 8 to 9 and Block 20 to 21 in Figure 5 for short and extensive learners respectively) had a significant impact on response time for the classification learners, $t(46) = -9.92, SE = 107, p < .001$. However, the Task Switch had a negligible impact on response time for the inference-first learners, $t(46) = -.86, SE = 107, p = .40$.

In summary, at the end of the learning phase, inference learners were less accurate than inference-first learners and marginally less accurate than classification and classification-first learners. Also, inference and classification-first learners were slower than classification learners at the end of the learning phase, due to a difference in the
amount of time required to perform the learning tasks. Notwithstanding these results, all four learning groups' accuracy and response time were at asymptote for 10 or more learning blocks in the extensive learning condition (see Figure 4 and 5). Participants assigned to an extensive learning session were more accurate at the end of the learning phase than those assigned a short learning session. Lastly, in contrast to Yamauchi and Markman (Experiment 2, 1998), both classification-first and inference-first learners showed significant facilitation effects on learning when switching from their first learning task to their second learning task. Switching learning tasks affected response time for the classification-first learners, but not for the inference-first learners, suggesting that the classification task is easier than the inference task.

Table 6

<table>
<thead>
<tr>
<th></th>
<th>Learning Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Order</td>
<td>Inference</td>
</tr>
<tr>
<td>Inference-First</td>
<td>2061 (88)</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
</tr>
<tr>
<td>Classification-First</td>
<td>1490 (88)</td>
</tr>
<tr>
<td>Total</td>
<td>1776 (62)</td>
</tr>
</tbody>
</table>

Note. Standard error of the mean is in parentheses.

General level transfer: Accuracy and response time analyses
Participants who did not achieve 75% correct or higher for the final learning phase block were not included in any of the following analyses due to insufficient learning. Including all participants did not change the qualitative pattern of results. In total, there were 12 participants who did not meet this criterion, including 1 from each of the short classification, extensive inference, extensive classification, short inference-first, and extensive classification-first conditions, and 7 from the short inference condition. In addition, one participant was excluded from all transfer analyses because of a failure to comply with the experimental instructions. The data collected from the remaining 94 participants were used for the following transfer phase analyses. There were an unequal number of participants in the groups with 12 in each of the short classification, short classification-first, and extensive classification conditions; 11 in each of the short inference, extensive classification-first, and short inference-first conditions; and finally, 14 in the extensive inference condition.

3.06% of all responses were deleted from the transfer data set because participants failed to respond within the 6-second time limit. To examine the pattern of these timed-out responses across the groups, a 2 x 4 ANOVA was performed on the participants’ mean number of timed-out responses with Learning Length and Learning Type as between-subjects variables. The interaction and main effects were not statistically significant with $p > .58$ Thus, there was no systematic pattern of timed-out responses across the groups.

In order to test how well general level category knowledge acquired during the learning phase was used to perform the transfer tasks, participants’ accuracy and response time for each general level item type (old classification, old inference, new prototype, and
new prototype-inconsistent inference) were submitted separately to 4 x 2 x 2 randomized-groups ANOVAs. The three randomized-groups variables were Learning Type, Learning Length, and Counterbalancing Group (classification transfer first vs. inference transfer first). Because Counterbalancing Group did not yield a significant main effect or interaction in any of the following ANOVAs, it was excluded from the analyses. Figure 6 and 7 present mean accuracy and response time, respectively, for general level classification, inference, and prototype transfer items as a function of Learning Type and Learning Length.

Figure 6. Mean percent correct for general level classification, inference, and prototype items presented as a function of Learning Type and Learning Length. Error bars are standard error of the mean.
Figure 7. Mean response time for general level classification, inference, and prototype items presented as a function of Learning Type and Learning Length. Error bars are standard error of the mean.

**Old classification.** The ANOVA on general classification accuracy yielded a significant main effect of Learning Type, $F(3, 86) = 16.05, p < .001$, partial $\eta^2 = .36$, $MSE = .015$. The interaction between Learning Length and Learning Type and the main effect of Learning Length, in contrast, were only marginally significant, $F(3, 86) = 2.52, p = .064$, partial $\eta^2 = .081$, and $F(1, 86) = 3.83, p = .053$, partial $\eta^2 = .043$, respectively. Post-hoc comparisons showed that inference learners ($M = 74.5\%, SD = 21.9\%$) were significantly less accurate when making general classification decisions than, in order, classification-first ($M = 93.0\%, SD = 8.8\%$), $t(44) = -3.75, p < .01$, inference-first ($M = $
93.6%, $SD = 8.8\%$), $t(45) = -3.88, p < .01$, and classification ($M = 96.3\%, SD = 4.7\%$), $t(45) = -4.67, p < .001$, learners. All other comparisons were not significant, $p > .50$. As can be seen in Figure 6, the marginal interaction between Learning Type and Learning Length as well as the main effect of Learning Length were due to inference learners’ improvement in making classification decisions when they underwent extensive learning ($M = 82.6\%, SD = 16.2\%$) relative to short learning ($M = 65.7\%, SD = 24.5\%$).

The ANOVA on general classification response time yielded a significant main effect of Learning Type, $F(3,86) = 11.49, p < .001$, partial $\eta^2 = .29$, $MSE = 241217$, and a significant main effect of Learning Length, $F(1,86) = 5.97, p = .017$, partial $\eta^2 = .065$. The interaction between Learning Type and Learning Length was not significant, $F(3,86) = .92, p > .10$. The post-hoc comparisons showed that inference learners ($M = 2535$ ms, $SD = 596$ ms) made classification decisions more slowly than, in order, classification-first ($M = 2000$ ms, $SD = 424$ ms) $t(44) = 3.69, p < .01$, inference-first ($M = 1935$ ms, $SD = 453$ ms), $t(45) = 4.19, p < .001$, and classification ($M = 1708$ ms, $SD = 525$ ms), $t(45) = 5.77, p < .001$, learners. The main effect of Learning Length indicated that the short learners were faster ($M = 1911$ ms, $SD = 544$ ms) than the extensive learners ($M = 2163$ ms, $SD = 593$ ms). This counter-intuitive finding may have been due to a fatigue effect.

Old inference. The ANOVA on general inference accuracy revealed a significant effect of Learning Type, $F(3,86) = 15.43, p < .001$, partial $\eta^2 = .35$, $MSE = .022$. The interaction between Learning Length and Learning Type and the main effect of Learning Length, however, were not significant, $F(3,86) = 1.12, p > .10$, and $F(1,86) = .70, p > .10$, respectively. Post-hoc comparisons showed that inference learners ($M = 82.3\%, SD = 12.5\%$) were significantly more accurate when making general inference decisions than,
in order, inference-first ($M = 65.1\%, SD = 14.2\%), t(45) = 4.40, p < .001, classification-
first ($M = 64.3\%, SD = 18.0\%), t(44) = 3.93, p < .01, and classification ($M = 53.2\%, SD
= 13.6\%), t(45) = 7.64, p < .001. learners. Also, inference-first learners were significantly
more accurate than classification learners, $t(46) = 2.96, p = .028$, and, although it was
only a trend, classification-first learners were marginally more accurate than
classification learners, \( t(45) = 2.37, p = .13 \).

The ANOVA on general accuracy presented above, did capture the learners’
propensity to select the correct or incorrect feature. However, recall that two out of the
four response choices were categorically valid (e.g., in Table 4, for SM1, values 1 and 2
were possible response options). Thus, a secondary analysis was computed in order to
determine if the learning groups differed in their tendency to select the incorrect but
categorically valid response choice (e.g., value 2 for the SM1 exemplar). A chi-square
test of independence was performed with Learning Type as an independent variable and
Response Choice as a dependent variable. There were three qualitatively distinct response
choices, including the correct prototypical feature for the exemplar’s specific category
(value 1 for SM1), a prototypical feature for the exemplar’s general, but not specific
category (value 2 for SM1), and a categorically invalid feature (e.g., for SM1, any value
from 3 through 8).

The overall chi-square was significant, \( \chi^2 (6, N = 94) = 248.32, p < .001 \). Pairwise
comparisons showed that inference learners more frequently selected the correct
prototypical feature for the exemplar’s specific category than the other three learning
groups, and classification-first and inference-first learners also selected this response
more frequently than the classification group. This pattern of results was also observed in
the ANOVA on general inference accuracy presented above. However, a novel finding that was not revealed by the ANOVA was that classification learners more frequently selected the prototypical feature for the exemplar's general, but not specific category (i.e., value 2 for exemplar SM1 in Table 4), than inference and inference-first learners. Also, both mixed learning groups more frequently selected this feature than inference learners. Lastly, classification learners selected categorically invalid responses more frequently than the three other learning groups, and inference-first learners selected these responses more frequently than the inference and classification-first groups.

Finally, the ANOVA on general inference response time showed that the main effects and interaction were not significant, $F, <= 1.84$. Although the main effect of Learning Type was not significant, the pattern of means across the Learning Types was as expected. Specifically, inference learners ($M = 2211$ ms, $SD = 602$ ms) made inference decisions faster than, in order, classification-first ($M = 2373$ ms, $SD = 456$ ms), inference-first ($M = 2390$ ms, $SD = 524$ ms), and classification ($M = 2594$ ms, $SD = 666$ ms) learners.

**New prototype classification.** The ANOVA on general prototype accuracy showed a significant main effect of Learning Type, $F(3,86) = 11.57, p < .001$, partial $\eta^2 = .29$, $MSE = .016$. The interaction between Learning Type and Learning Length was marginally significant, $F(3,86) = 2.16, p = .10$, and the main effect of Learning Length was not significant, $F(1,86) = .70, p > .10$. Post-hoc comparisons revealed that inference learners ($M = 78.5\%, SD = 21.9\%$) were significantly less accurate when making general classification decisions for the prototypes than, in order, classification ($M = 95.5\%, SD = 5.1\%$), $t(45) = -3.63, p < .01$, inference-first ($M = 95.8\%, SD = 10.5\%$), $t(45) = -3.43, p$
Inference vs. Classification Learning

<= .01, and classification-first (M = 96.2%, SD = 6.3%), t(44) = -3.72., p < .01, learners. All other comparisons failed to reach significance, p > .90. The marginal interaction between Learning Type and Learning Length was due, again, to the inference learners’ improvement in making classification decisions after short learning (M = 71.3%, SD = 21.9%) relative to an extensive learning (M = 85.0%, SD = 20.7%).

The ANOVA on general prototype response time yielded a significant main effect of Learning Type, F(3,86) = 13.95, p < .001, partial \( \eta^2 = .33 \), MSE = 174747, and a marginally significant main effect of Learning Length, F(1,86) = 3.31, p = .072, partial \( \eta^2 = .037 \). The interaction between Learning Type and Learning Length was not statistically significant, F(3,86) = .44, p = .73. Post-hoc comparisons showed that inference learners (M = 2271 ms, SD = 430 ms) made classification decisions for the category prototypes more slowly than, in order, classification-first (M = 1789 ms, SD = 416 ms) t(44) = 3.91 , p <= .001, inference-first (M = 1660 ms, SD = 432ms), t(45) = 5.01, p < .001, and classification (M = 1520 ms, SD = 389 ms), t(45) = 6.16, p < .001, learners. The marginally significant main effect of Learning Length indicated that the short learners had a tendency to respond faster (M = 1723 ms, SD = 484 ms) than the extensive learners (M = 1884 ms, SD = 509 ms). This counter-intuitive result may again have been due to a fatigue effect.

*New prototype-inconsistent inference.* The prototype-inconsistent inference items differed in terms of the response choices that were presented to participants. There were four response choice conditions, represented as 0XXX, 01XX, 02XX, and 012X. Each response choice condition was analyzed separately using chi-square tests of independence with Learning Type as the independent variable and Response Choice as the dependent
variable. Learning Length was not included because it did not affect the qualitative pattern of results. Also, including it would have violated an assumption of chi-square analysis (i.e., > 5 observations per cell) in the 012X condition. Figure 8 displays response frequency for the different prototype-inconsistent inference items and their associated response choices as a function of Learning Type.

For the OXXX condition, there were two qualitatively different response choices, which included the prototype-inconsistent feature and a categorically invalid feature. The chi-square was significant, $\chi^2(3, N = 94) = 55.20, p < .001$, indicating that the response choice was dependent on the Learning Type. Post-hoc comparisons revealed that inference and classification-first learners selected the prototype-inconsistent feature more frequently than classification and inference-first learners. They also selected categorically invalid features less frequently than classification and inference-first learners.

For the 01XX condition, there were three qualitatively different response choices, which included the prototype-inconsistent feature, a prototypical feature for the exemplar's specific category, and categorically invalid features. The chi-square was significant, $\chi^2(6, N = 94) = 121.52, p < .001$. As can be seen in Figure 8, response frequency for the prototype-inconsistent feature shifts dramatically from the OXXX to the 01XX condition, as a prototypical feature is introduced as a response choice. The pairwise comparisons revealed that classification learners selected the prototype-inconsistent feature more frequently than the other three groups; classification learners selected the prototypical feature for the exemplar's specific category less than the three other groups (which replicates Johansen & Kruschke, 2005; Rehder et al., 2009; Yamauchi & Markman Experiment 1, 1998) and classification learners were more likely
to select a categorically invalid feature than the other three groups. Unlike what Yamauchi and Markman (Experiment 2, 1998) found, the difference in the tendency for classification-first and inference-first learners to select the prototypical feature was not significant.

Figure 8. Response frequency for the four different prototype-inconsistent inference response choice conditions presented with their corresponding response choices as a function of Learning Type. 0 = prototype-inconsistent feature. 1 = a prototypical feature for the exemplar’s specific category. 2 = a prototypical feature for the exemplar’s general, but not specific category. X = a categorically invalid feature.

For the 02XX condition, there were again three qualitatively different response choices, including the prototype-inconsistent feature, a prototypical feature for the
exemplar's general, but not specific category, and categorically invalid features. The chi-square was significant, $\chi^2 (6, N = 94) = 182.93, p < .001$. As can be seen in Figure 8, the data reveal a very different trend from the 01XX condition, when a different prototypical feature was introduced as a response option. The pairwise comparisons within each level of Response Choice showed that inference learners selected the prototype-inconsistent feature more frequently than the other three groups. The inference-first learners also selected this feature more than the classification and classification-first learners. Unlike the 01XX condition, classification learners selected a prototypical feature more frequently than the inference learners; the classification-first learners had the strongest tendency to select this feature; and the inference-first learners selected this feature more often than the inference learners. Lastly, the classification-first learners selected categorically invalid features less frequently than the other three groups.

For the final response condition 012X, there were four qualitatively different response choices, including the prototype-inconsistent feature, a prototypical feature for the exemplar's specific category, a prototypical feature for the exemplar's general, but not specific category, and a categorically invalid feature. The chi-square was significant, $\chi^2 (9, N = 94) = 118.49, p < .001$. Both classification and inference learners selected the prototype-inconsistent feature more frequently than classification-first and inference-first learners. Inference learners selected a prototypical feature for the exemplar's specific category more frequently than the classification and classification-first learners. Also, classification-first and inference-first learners selected this feature more frequently than classification learners. As can be seen in Figure 8, this trend reverses almost completely for the prototypical feature of the exemplar's general, but not specific category, much
like the shift that was observed going from the 01XX to the 02XX response condition. Classification and classification-first learners selected this feature more frequently than inference and inference-first learners. Also, inference-first learners selected this feature more often than inference learners. Finally, the differences in preference for categorically invalid features across the four learning groups were not significant.

Specific level transfer: Response frequency analyses

During the learning phase, participants performed their assigned task(s) at the general category level, but they received feedback that included information about the exemplars' general and specific category. At transfer, inference items included the exemplar's specific and general category label (e.g., "Sot Maki"). Furthermore, classification and prototype items were presented with specific level response options. For example, when participants had to decide what category the SM1 exemplar belonged to (see Table 4), they were presented the exemplar's specific category label ("Sot Maki" and "Iko Maki") as well as another category's two specific labels as response choices. In order to determine preference for various response choices, participants' response frequency for each item type was analyzed separately using chi-square tests of independence with Learning Type as the independent variable and Response Choice as the dependent variable. Learning Length was not included because of violations of chi-square assumptions. Including it did not change the qualitative pattern of results. Figure 9 presents response frequency for each specific level transfer item type as a function of Response Choice and Learning Type.

Old classification. There were three qualitatively different response choices for all specific classification items, including the exemplar's correct specific label (e.g., for item
SM1 in Table 1, "Sot Maki"), the exemplar’s incorrect specific but correct general label (e.g., for the previous example, "Iko Maki"), and a categorically invalid label (e.g., "Hap Timo"). The chi-square test of independence was significant, \(\chi^2(6, N = 94) = 49.88, p < .001\). Post-hoc comparisons revealed that inference learners selected the correct label more frequently than the classification-first learners. Furthermore, they selected the incorrect specific, but correct general label less frequently than the other three learning groups. Lastly, they selected more categorically invalid labels than the other three learning groups.

Old inference. Similar to the general inference items, the specific inference items were accompanied by three qualitatively different response choices, including a prototypical feature for the exemplar’s specific category, a prototypical feature for the exemplar’s general, but not specific category, and categorically invalid features. The chi-square test was significant, \(\chi^2(6, N = 94) = 217.43, p < .001\). Post-hoc comparisons revealed a pattern of results that was similar to general inference performance. Specifically, inference learners selected a prototypical feature for the exemplar’s specific category more frequently than the other three learning groups. Also, the classification-first and inference-first learners selected this feature more frequently than the classification learners. In contrast, classification and classification-first learners selected a prototypical feature for the exemplar’s general, but not specific category, more frequently than inference and inference-first learners. Also, inference-first learners selected this feature more often than the inference learners. Lastly, classification learners selected a categorically invalid feature more frequently than the other three learning groups and
inference and inference-first learners selected a categorically invalid feature more frequently than classification-first learners.

![Figure 9](image-url)

*Figure 9.* Response frequency for each specific level transfer item type presented as a function of Response Choice and Learning Type. Correct = the correct specific category label. Other = the correct general but incorrect specific category label. 1 = a prototypical feature for the exemplar’s specific category. 2 = a prototypical feature for the exemplar’s general, but not specific category. X = a categorically invalid label or feature.

*New prototype classification.* The specific prototype items had the same three qualitatively different response choices as the specific classification items. The chi-square test was significant, $\chi^2(6, N = 94) = 82.60, p < .001$. The comparisons showed that inference and inference-first learners selected the exemplar’s correct specific label
more frequently than the classification-first learners. Furthermore, they selected the exemplar's incorrect specific but correct general label less frequently than classification and classification-first learners. Finally, inference learners selected more categorically incorrect labels than all other learning groups.

Discussion

The present experiment examined the impact of classification, inference, and mixed category learning on participants' subsequent ability to classify and make feature inferences. Previous category learning studies have shown that classification and inference learners are equally accurate when classifying prototypes at transfer (Rehder et al., 2009; Yamauchi & Markman, 1998). Furthermore, Rehder et al. have shown that classification and inference learners are equally accurate when making feature inferences at transfer. These results directly violate transfer-appropriate processing (Morris et al., 1977). I have argued that the violations are specific to the simple category structure that was used. The goal of the present experiment was therefore to replicate the previous studies using a more complex category structure, while controlling for stimulus exposure between learning types and manipulating learning length.

The following predictions were made. Transfer accuracy for the old classification, old inference, and new prototype items was expected to be higher when the learning task matched the transfer task, relative to a mismatch. Furthermore, participants were expected to be faster to respond when the learning task matched the transfer task, a dependent measure that was not examined in the reviewed studies. Finally, mixed learners' accuracy and response time were expected to be in between the observed values for the classification and inference learners. In accordance with the reviewed studies,
when presented with prototype-inconsistent inference items, inference learners were expected to select prototypical features more often than classification learners. Lastly, mixed learners were expected to share the same tendency to select prototype-inconsistent and prototypical features.

**Transfer Performance**

As was predicted, the results showed that classification learners were more accurate and faster than inference learners when making general classification decisions for the old items. Furthermore, and contrary to previous findings (Rehder et al., 2009; Yamauchi & Markman, 1998), classification learners were also more accurate and faster than inference learners when making general classification decisions for the category prototypes. Thus, this result suggests that violations of the idea of transfer-appropriate processing in past research were due to a ceiling effect on accuracy, not a genuine difference between learning types. The mixed learners were more accurate and faster than the inference learners to make general classification decisions for the old and prototype items. Contrary to expectation, however, they were as accurate and fast as the classification learners, although the observed means for both accuracy and response time were in the hypothesized order. A glance at Figure 6 shows that the expected result may not have been found because the groups were near ceiling for accuracy. The category structure was therefore complex enough to distinguish classification and inference learners’ ability to classify exemplars at transfer. It appears, however, that it was not able to further distinguish classification and mixed learners’ ability to classify exemplars at transfer, due to a ceiling effect.
Performance for the general inference items was consistent with transfer-appropriate-processing as well. Specifically, and contrary to previous research (Rehder et al., 2009), inference learners were more accurate than classification learners when making general inference decisions for the old items. This again suggests that the previous violation of transfer-appropriate processing was due to a ceiling effect on inference accuracy, not a genuine difference between learning types. As was predicted, the mixed learners were less accurate than inference learners but more accurate than classification learners when making inference decisions for the old items. For response time, the pattern of observed means was the same as the observed pattern for accuracy. The mean differences, however, did not reach significance (see Figure 7). Nevertheless, the category structure was sufficiently complex to yield the predicted pattern of results for general inference accuracy across all learning groups.

The inference and classification learners' response choice frequency for prototype-inconsistent inference items were in accordance with previous findings in some conditions, but were not in others. First, as was predicted, inference learners selected the prototypical feature for the exemplar's specific category more frequently than classification learners (i.e., the value 1 for the SM1 exemplar in Table 4). This effect held when the only categorically valid alternative response choice was the prototype-inconsistent feature (the value 0 for all exemplars in Table 4; the 01XX condition in Figure 8) and when this feature was accompanied by the prototypical feature for the exemplar's general, but not specific category (the value 2 for exemplar SM1 in Table 4; the 012X condition in Figure 8). These results replicate previous findings (Johansen & Kruschke, 2005; Rehder et al., 2009; Yamauchi & Markman, Experiment 1, 1998).
Contrary to expectation, however, classification learners selected the prototypical feature for the exemplar's general, but not specific category, more frequently than inference learners. This effect held when the only categorically valid alternative response choice was the prototype-inconsistent feature (the 02XX condition) and when this feature was presented with the prototypical feature for the exemplar's specific category (012X condition). Furthermore, inference learners selected the prototype-inconsistent feature more often than classification learners, when it was presented with categorically invalid features (the 0XXX condition). These results contradict previous findings. Specifically, research has consistently shown that classification learners have a stronger tendency than inference learners to select prototype-inconsistent features (Johansen & Kruschke, 2005; Rehder et al., 2009; Yamauchi & Markman, 1998). Together these results suggest a general difference between the concepts acquired during the classification and inference learning processes. More precisely, the data suggest inference learners incorporated more information about the prototypical features of the specific categories than classification learners. If this interpretation is correct, the analyses on the specific level performance should have provided converging evidence. Specifically, inference learners should have made more correct specific level classifications and feature-inferences than classification learners.

The specific level analyses, however, revealed that classification and inference learners selected the exemplar's correct specific label equally often for both the old classification and new prototype items. Classification learners selected the incorrect specific, but correct general category label more often than inference learners. Lastly, as transfer-appropriate processing predicts, inference learners selected more categorically
invalid labels than classification learners. These results show that inference learners were less willing than classification learners to ascribe an exemplar an incorrect specific, but correct general category label. The addition of the specific label to the old inference items did not change the learners’ response choice frequency. For both the general and specific items, inference learners selected the prototypical feature for the exemplar’s specific category more frequently than classification learners. In contrast, classification learners selected the prototypical feature for the exemplar’s general, but not specific category, more frequently than inference learners. Lastly, as transfer-appropriate processing predicts, classification learners selected more categorically invalid features than inference learners.

In summary, these results suggest that inference learners were more sensitive to the specific category divisions than classification learners, in accordance with the prototype-inconsistent results. This interpretation is consistent with Rehder et al.’s (2009) anticipatory learning strategy hypothesis. The exemplar’s specific label, displayed during feedback in the learning phase, provided an additional source of information that would have increased future feature predictions. More precisely, the label would have suggested to learners that the feature values in one specific category tend to be different from those in another specific category (e.g., in Table 4, the values 1 for “Sot Maki” vs. the values 2 for “Iko Maki”). In fact, within-category correlation learning has been demonstrated in previous inference learning studies (Chin-Parker & Ross, 2002; 2004), but not in a study with a hierarchical category structure. The preceding results need to be interpreted with caution, however, because when there was match between the learning and transfer tasks
for classification learners, there was a mismatch for inference learners. Hence, a transfer-appropriate processing effect confounded the results.

In contrast, the mixed learners’ performance for the prototype-inconsistent and specific items cannot be explained by transfer-appropriate processing, because both groups had engaged in inference learning for exactly the same number of trials. As was expected, both groups selected the prototypical feature for the exemplar’s specific category equally often (i.e., value 1 for exemplar SM1 in Table 4). This held when the only categorically valid alternative response choice was the prototype-inconsistent feature (the 01XX condition) and when this feature was presented with the prototypical feature for the exemplar’s general, but not specific category (the 012X condition). As can be seen in Table 3, Yamauchi and Markman (Experiment 2, 1998) found that inference-first learners selected a prototypical feature more often than the classification-first learners. Based on this result, the authors made a distinction between the learners’ concepts, without the burden of a transfer-appropriate processing explanation. However, in the present experiment, when a more complex category structure was used, Yamauchi and Markman’s results were not replicated.

Surprisingly, there were two notable differences in the mixed learners’ prototype-inconsistent response choice frequency. Specifically, the classification-first learners selected the prototypical feature of the exemplar’s general, but not specific category, more frequently than inference-first learners (the 02XX condition), whereas the inference-first learners selected the prototype-inconsistent feature more frequently. Secondly, when all other response choices were categorically invalid (the OXXX condition), classification-first learners selected the prototype-inconsistent feature more
frequently than the inference-first learners. These results suggest a learning order effect, in which classification-first and inference-first learners acquired slightly different aspects of the category structure. The specific level performance could provide more information about exactly what information the mixed learners acquired.

The specific level analyses for the mixed learners revealed few differences between the groups. There were no differences in the groups' specific classification response choice frequency. However, for the specific inference items, the classification-first learners selected a prototypical feature for the exemplar's general, but not specific category, more frequently than inference-first learners (i.e., the value 2 for exemplar SM1 in Table 4). Also, for specific classification of the prototype items, inference-first learners selected the exemplar's correct specific category label more frequently than classification-first learners. Furthermore, classification-first learners selected the exemplar's incorrect specific, but correct general category label more frequently than inference-first learners.

In summary, as predicted, the classification and inference transfer data clearly demonstrated the performance benefits that result from a match between learning and transfer tasks. In the mixed learning conditions, when transfer-appropriate processing was controlled for, there were no differences in transfer accuracy or response time for the general classification, inference, and prototype items. Although the mixed-learners differed in certain prototype-inconsistent and specific level response choice frequency, the overall pattern of results did not provide a compelling argument for a general difference in the concepts acquired by the mixed learning procedures, unlike what has been previously reported (Yamauchi & Markman, 1998). The outcome of the present
experiment indicates that the violations of transfer-appropriate processing in the reviewed studies are specific to the simple category that was used. In addition to transfer performance, learning performance also offers a source of information for evaluating task difficulty and category representation.

*Learning Performance*

An important part of the learning data was the examination of the transition from the first task to the second task for the mixed learners. Recall that Yamauchi and Markman (1998) found that inference learning facilitates subsequent classification learning, but that classification learning slightly hinders subsequent inference learning. In contrast to their findings, the results in the present experiment showed a facilitation effect for both extensive mixed learning conditions. These results suggest that classification-first learners in the present experiment acquired a representation that is sufficient to make inferences, in contrast to the representation acquired by the classification-first learners in Yamauchi and Markman's experiment. The learning results acquired in their study were not replicated in the present study when a more complex category structure was used. The observed differences between the two studies may be a product of more controlled stimulus exposure, the relative amount of information that the structures embody, or it may be related to the nature of the category structures. An examination of learning task difficulty may provide some insight.

The learning data show that the inference task was more difficult than the classification task. A glance at Figure 4 and 5 shows that mean inference accuracy was consistently lower, and mean response time was consistently slower, than mean classification accuracy and response time across all learning blocks in the short and
extensive learning conditions. Inference learners were also less accurate at the final block of the learning phase than the other learning groups. Furthermore, inference learners required more blocks to reach 90% correct in three consecutive blocks than classification learners. Lastly, the task switch from classification to inference significantly increased response time (the spike at block 9 and 21 in Figure 5), whereas the switch from inference to classification did not increase response time. These learning results conflict with previous findings. Yamauchi and Markman (1998) as well as Rehder et al. (2009) found that the classification learning task was more difficult than the inference learning task (see Table 3).

Although differences in task difficulty for inference and classification learning of various category structures is not new to the categorization literature (see, Chin-Parker & Ross, 2004; Johansen & Kruschke, 2005; Yamauchi, Love, & Markman, 2003), it is instructive to determine why the effect of task difficulty in the present experiment is in opposition to previous findings. The most parsimonious explanation resides in the nature of the category structures that were learned. In Yamauchi and Markman's structure, the prototype-inconsistent features were diagnostic of category membership. For example, these features for exemplars A1 to A4 (i.e., the value 0 in Category A in Table 1) are diagnostic of category B. Therefore, if classification learners attended to these features during a trial, they may incorrectly classify the exemplar as a category B member. In the present structure, the prototype-inconsistent features are not diagnostic of category membership. For example, the value 0 in Table 4 is equally associated within and across all the general categories. Therefore, if classification learners attended to these non-diagnostic features, they would not classify the item incorrectly, at least based on these
values alone. Thus, the diagnosticity or lack thereof for the prototype-inconsistent features can account for the difference in task difficulty for classification learners between the two structures. In fact, a study by Markman and Maddox (2003) found that classification learning was much easier when the category structure had non-diagnostic prototype-inconsistent values than a similar category structure with diagnostic prototype-inconsistent values, when there were multiple feature manifestations on each dimension, which was the case in the present experiment.

It is also possible that the prototype-inconsistent features in the present experiment made inference learning more difficult than in Yamauchi and Markman’s (1998) experiment. The most recent account of inference learning (Rehder et al., 2009) suggests that participants attend to non-queried features in addition to the category label during learning trials in an attempt to increase accuracy for future feature predictions. In effect, they incidentally encode multiple feature-to-category associations in a partial exemplar representation. In the case of the structure Yamauchi and Markman used, when a feature was queried, there were always two diagnostic feature values and one prototype-inconsistent value. Hence, if a non-queried dimension were attended to during learning trials, 66% of the time a diagnostic feature dimension would be encoded and could aid in future feature prediction. For the remaining 33%, the prototype-inconsistent feature, which was highly associated with the opposing category, would be incidentally encoded and could contribute to making an incorrect future feature prediction. In the present structure, when a feature was queried, there was always one diagnostic feature and one prototype-inconsistent feature. Hence, if a non-queried dimension were attended to during learning, 50% of the time a diagnostic feature would be encoded and used in
future feature predictions trials, which is less than the other structure. For the remaining 50%, the prototype-inconsistent value, which was not diagnostic and equally prototypical of the opposing categories, would be incidentally encoded. This would not benefit future feature prediction, however, because this value was equally associated with the categories, hence it had no predictive power. Therefore, on the one hand, Yamauchi and Markman’s inference learners had more diagnostic features to incidentally encode, but at the same time, the prototype-inconsistent features could contribute to incorrect predictions. On the other hand, inference learners in the present experiment had fewer diagnostic features to incidentally encode, but at the same time, the prototype-inconsistent features could not have contributed to incorrect predictions. Thus, based on the anticipatory learning strategy hypothesis, it is not clear why the inference task was difficult in the present study and easy in Yamauchi and Markman’s study.

Rather than incidental encoding of prototype-inconsistent values, the difference in difficulty could have been related to the difference in the number of possible features and categories for inference trials. In the present research, there were four categories with eight queried feature values on each of the three dimensions. On any given inference trial, four of the eight feature values were selected as response choices. In Yamauchi and Markman’s (1998) experiment, there were only two categories with two queried feature values on each of four dimensions. On any given inference trial, inference learners had to make a binary decision. The amount of information that inference learners had to process and represent in the present experiment could account for the difference in difficulty between the two studies.
Finally, in the present experiment, the difference in classification and inference task difficulty may have also been related to the amount of information that learners needed to process and represent in order to meet the task demands. It is reasonable to think that inference learning was more difficult than classification learning, due to a difference in the amount of information that each learner needed to acquire to achieve equivalent accuracy. Inference learners needed to acquire prototypical knowledge on all three of the feature dimensions, with eight possible feature values each, to achieve perfect accuracy. Classification learners, on the other hand, needed to acquire diagnostic knowledge on only two of the three feature dimensions, to discover the division of the four categories and achieve perfect accuracy. Hence, inference learners had to acquire more information to achieve equivalent accuracy.

Limitations

The complex category structure that was used did not completely dissociate transfer performance for all the learning types as was predicted. Although the ceiling effects on inference and classification transfer were eliminated for the classification and inference learners, there was still a ceiling effect on classification of old and prototype items for mixed learners. It seems that creating a category structure that equates task difficulty for inference and classification transfer may be more intricate than previously assumed.

Another limitation was a counterintuitive effect of learning length on response time for old classification, in which extensive learners made slower general classifications than short learners. This disparity may have been due to a fatigue effect caused by the difference in learning blocks between the groups. Specifically, participants
assigned to short learning completed 8 blocks whereas those assigned to extensive
learning completed 20 blocks on the second day of the learning phase, just prior to the
transfer phase (384 vs. 960 trials). Also, Figure 7 shows that mean response time and
variance were high, especially in the inference condition. This was due to a
methodological constraint in assigning response choices to response keys on trials. For
classification learners, the response key to response choice assignment was randomized
every trial for the four labels. For inference learners, it was randomized every trial but
with a selection of four of the eight possible features for one of three possible queried
dimension. Thus, the learners could not make one-to-one mappings of the response
choices to the response keys, which made response time higher and more variable, but
even more variable for inference learners because of the number of different features that
they could have considered. Hence, although the predicted pattern of mean response time
was observed for the various transfer item types, some of the mean differences did not
reach significance due to high variability.

Another limitation was related to the hierarchical structure of the categories. The
general and specific categories were implemented to make the category structure more
complex and to examine differences in category learning at both category levels.
However, the hierarchical structure of the categories did not conform to all the principles
of real-world category hierarchies. For instance, according to the property inheritance
principle, all features true of the general categories must also be true of the specific
categories (Murphy, 2002). Therefore, because the Maki category in Table 4 has feature
values 0, 1, and 2, the “Sot Maki” and “Iko Maki” categories needed to also have these
values. This was a limitation of the study, but it does not discredit the observed
differences in general transfer performance, the level at which participants needed to learn to meet the task demands.

*Future research*

There is an ongoing effort in the categorization research community to branch out from the quintessential classification learning task and study new ways that categories can be learned (see, Markman & Ross, 2003 for a review). The present thesis is a contribution to this effort. Through comparing classification and inference learning and transfer using a complex category structure, the present work makes novel contributions to a growing body of research.

Firstly, there has been some work on examining how multiple feature manifestations affect classification learning of a category structure with non-diagnostic as opposed to diagnostic prototype-inconsistent features (e.g., Markman & Maddox, 2003). That research has shown that classification learning of these two structures is equally difficult with single feature manifestations, but as the number of feature manifestations increases, the diagnostic structure is much more difficult to learn than the non-diagnostic structure. Thus, for classification learning, number of feature manifestations interacted with category structure type. The structure in the present thesis is the non-diagnostic type and it has multiple feature manifestations (3 for each general category). In accordance with what Markman and Maddox found, classification learning was easy relative to the diagnostic structure that Yamauchi and Markman (1998) used. Inference learning, on the other hand, was difficult relative to the diagnostic structure that Yamauchi and Markman used. This suggests that a non-diagnostic structure with multiple feature manifestations makes inference learning more difficult than a diagnostic structure with single feature.
manifestations. Thus the interaction between number of feature manifestations and category structure type reported by Markman and Maddox may also interact with learning task.

Secondly, the use of a more complex category structure in the present work allowed for a manipulation of response choice for classification and inference items that have been critical in distinguishing category representation. There were multiple categorically valid and invalid features for both the inference and classification learning and transfer tasks, instead of binary choices. This type of complexity can provide researchers more fine-grained distinctions between what is learned during inference and classification by determining exactly what features learners’ prefer in the face of various response choices.

Finally, the reviewed studies used category structures that consisted of not much more than the minimal amount of information that is required to create artificial categories. Unfortunately, the use of a simple structure has created ceiling effects on transfer performance and violations of a robust memory theory. Conclusions regarding what is learned in classification and inference in these experiments are therefore questionable. In the present experiment, when a complex structure was used, typical transfer-appropriate processing results were obtained. Unfortunately, however, there was scant evidence to suggest differences in concept acquisition based on the mixed learners performance, the only conditions which could not have been explained by transfer-appropriate processing.
References


Markman, A. B., & Maddox, W. T. (2003). Classification of exemplars with single- and multiple-feature manifestations: the effect of relevant dimension variation and


