Metacognitive Monitoring during Category Learning:

How Success affects Future Behaviour

By © Mario E. Doyle

A Thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of

Masters of Experimental Psychology Psychology/Science

Memorial University of Newfoundland

October 2016

St. John's Newfoundland and Labrador

Abstract

The purpose of the present study was to examine how people's perceptions of their own learning, during a category learning task, matched their performance. In two experiments, participants were asked to learn natural categories, of both high and low variability, and make category learning judgments (CLJs). Variability was manipulated by varying the number of exemplars and the number of times each exemplar was presented within each category. Experiment 1 showed that participants were generally overconfident in their knowledge of low variability families, suggesting that they considered repetition to be more useful for learning than it actually was. CLJs had the largest increase when a trial was correct following an incorrect trial and the largest decrease when an incorrect trial followed a correct trial. Experiment 2 replicated these results, but also demonstrated that global CLJ ratings showed the same bias toward repetition.

Keywords: metacognition, category learning

Acknowledgments

I want to thank my supervisor Dr. Hourihan for her tremendous support throughout this whole project. It has felt like a long journey at times, but it was well worth it in the end. I would also like to thank NSERC, which provided the funding I needed to carry out this project.

Portions of Experiment 1 were previously reported in an honours thesis submitted to Memorial University in 2014.

The results from Experiment 1 and 2 included in the present thesis report were previously published in the following manuscript:

Doyle, M. E., & Hourihan, K. L. (2015). Metacognitive monitoring during category learning: How success affects future behaviour. *Memory*, 1-11. doi: 10.1080/09658211.2015.1086805.

Co-authorship Statement:

Initially, I identified one of the main ideas for the research proposal and then discussed it with my supervisor who gave me suggestions on how I could expand on this idea. From there I designed the experiments using E-prime software (a computer program) and tested participants to collect the data. Since the study used human participants and was carried out on a computer, there were not many aspects that I could not handle myself. I then analysed the data and carried out any statistical analyses using Microsoft Excel or SPSS. With respect to the published manuscript, I wrote most of it myself however my supervisor did provide extensive revisions as she was a co-author.

Table of Contents

List of Tables

List of Figures

Metacognitive Monitoring during Category Learning:

How Success affects Future Behaviour

A quick internet search reveals thousands of websites containing innumerable quotations from notable individuals that converge on the idea that we learn more from our mistakes than from our successes (Medlock, 2015). But is this commonly-held belief supported by empirical data on learning and our perceptions of our own learning behavior? The goal of the present thesis was to investigate how people perceive their progress when they are learning a *natural* category – categories that do not have strict rules that define membership inclusion criteria (Homa, Sterling, & Trepel, 1981; Kellog, Bourne, & Ekstrand, 1978), focusing on how they view their learning following a correct classification compared to an incorrect one. Metacognitive judgments – participants' assessments of the relative quality of their learning – were used to indicate how well participants thought they were learning different categories as they were performing a natural category learning task. Although the literature examining metacognition and category learning is small, there are a few studies that have examined this relationship (Higgins & Ross, 2011; Meuwese, van Loon, Lamme, & Fahrenfort, 2014; Tauber, & Dunlosky, 2015; Tauber, Dunlosky, Rawson, Wahlheim, & Jacoby, 2013; Thomas, Finn, & Jacoby, 2015; Wahlheim, Dunlosky, & Jacoby, 2011; Wahlheim & DeSoto, 2016; Wahlheim, Finn, & Jacoby, 2012). To my knowledge, the only other study on category learning that has examined metacognitive judgments on a trial-by-trial basis is by Wahlheim and DeSoto (2016). Therefore, an exploration of the pattern of metacognitive judgments throughout learning should provide valuable information on how people view their own learning progress during a category learning task.

1.1 Category learning

Any given category is made up of a number of items, called exemplars. Some categories can be grouped together to form a cluster of categories, which are organized at the superordinate level; this results in a categorical hierarchy (Bower, Clark, Lesgold, & Winzenz, 1969). For instance, oil, pastel, and watercolor all belong to the category 'PAINTINGS', which in turn belongs to the broader category 'ART'. People develop categories and superordinate categories to make it easier to organize large amounts of information. Categories are an integral part of cognitive functions and play an important role in communication (Brown, 1958), memory (Mandler, Pearlstone, & Koopmans, 1969), and learning (Ashby & Maddox, 2005). Understanding how we learn to categorize is ultimately the goal of research in category learning.

One potential issue for category learning theories is distinguishing the difference between types of categories and whether they are learned differently. Learning how to identify members of some categories, like what constitutes an oil painting versus a watercolor painting, is very straightforward because these categories have clear rules that differentiate exemplars among different categories. However, identifying an impressionistic painting versus an abstract painting is more difficult, because to a certain extent it relies on the interpretation of the viewer rather than on a strict rule. The distinction between types of categories that have clear rules versus those that do not has been referred to as ill- versus well-defined categories or rule-based versus natural categories (for a review see: Close, Hahn, Hodgetts, & Pothos, 2010). Recent studies have focused on natural (ill-defined) categories as they better resemble the types of categories people commonly use.

Understanding how people learn natural categories is complicated by the fact that natural categories are difficult to define in terms of inclusive or exclusive criteria for a given exemplar (Homa et al., 1981). For example, considering sub-categories of the category 'BIRD', members of one particular bird family differ from another bird family based on a collection of features like colour, wing span, leg size etc., but these features often overlap across families. Even expert bird watchers with a lot of experience have trouble classifying birds, because of how difficult it is to identify which group of features belongs exclusively to a particular family (Wang, Schiner, & Yao, 2008). However, even without explicit rules for organizing categories, people still manage to differentiate among different natural categories in a category learning task (e.g., Kellog et al., 1978).

The primary aim of a category learning task is to learn what constitutes a member of a particular category (Chin-Parker & Ross, 2004). One task commonly used to study category learning requires a participant to initially study a set of exemplars belonging to a number of different categories. During this study phase the participants must decide which category an item belongs to, and are then given feedback on whether they were correct. Because natural categories do not have simple rules, participants typically do not achieve 100% accuracy during the study phase; instead, they are given a certain amount of time to study, or allowed to continue until they reach an accuracy level above chance. Participants' category knowledge is subsequently tested by having to categorize novel exemplars without feedback.

In the category learning literature, there are two classes of theories that are the most well-developed. In simplified terms, exemplar-based models (e.g., Medin and Schaffer's "Context theory", 1978; Nosofsky's "categorization model", 1986; and

Minerva 2, Hintzman, 1986) suggest that people categorize novel items based on their stored memory for previously seen exemplars. The assumption that people make categorization decision based on a similarity to previous exemplars is also based on the assumption that during learning people attend to particular features that aid in diagnosing category membership (Nosofsky, 1986). One of the first exemplar models, Medin and Schaffer's context theory of categorization, was based on earlier studies by Rosch, Mervis and their colleagues who looked at category learning of ill-defined categories (e.g. Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Rosch and Mervis (1975) suggested that stimulus classification of ill-defined categories was more likely to be based on a "family resemblance" (i.e., familiarity) to previous exemplars than on underlying criterion rules. In contrast, they considered well-defined categories to have definitive rules that could be learned and applied when categorizing novel exemplars.

Rule-based (well-defined) theories of categorization formed the basis for an alternative approach to category learning, the prototype models. These models suggest that when participants classify a novel stimulus they compare it to how well it matches the most typical exemplar in that category, called a 'prototype' (e.g., Homa et al., 1981; Jacoby, Wahlheim, & Coane, 2010). These models assume that when people are learning a new category they form a prototypical exemplar of that category, which is composed of the features they think are most typical of all the exemplars (Bransford & Franks, 1971). Support for this assumption came from Smith, Shoben, and Rips (1974) who found that people are faster at classifying more typical exemplars of a category than less typical exemplars. Additionally, prototypical information is less likely to be forgotten over large retention intervals than less typical information (e.g., Goldman & Homa, 1977).

Recent work by Chin-Parker and Ross (2004) suggests that there should be an important distinction made between diagnostic information and prototypical information. Diagnostic information (e.g., that birds fly) is more important for determining which category an item belongs to, while prototypical information is more useful in distinguishing among highly shared features within the members of a category. In a learning task where a participant must classify a member of a category of which they have no prior knowledge, individuals tend to focus on determining what cues or features best determine category membership. More attention is then paid to the features that lead to correct classification. This strategy is opposed to a prototypical model that would suggest a strategy of focusing on groups of features that are found in the majority of exemplars in a category. At the moment, there is no consensus on which model accounts for the extant of the data.

There has been substantial research on how we learn categories and the variables that affect category learning (for a review, see: Ashby & Maddox, 2005). A consistent finding in the literature is that increasing variability during learning allows people to gain a better understanding of a category, which improves classification of novel exemplars (Dukes & Bevan, 1967; Homa, 1978; Perry, Samuelson, Malloy & Schiffer, 2010; Posner & Keele, 1968; Smith, 2003; Wahlheim & DeSoto, 2016; Wahlheim et al., 2012). This is known as the variability effect. Variability refers to both the number of unique exemplars in a category and the number of times each exemplar is presented. It is difficult to specify what constitutes a “low” or “high” variability category, as past research has varied significantly in the number of exemplars presented in these conditions. For example, as their high variability condition Dukes and Bevans (1967) used a set of five stimuli

belonging to the same category and presented them four times each (each stimulus within a set had slight differences). On the other hand, Wahlheim, Finn and Jacoby's (2012) high variability condition used six different stimuli belonging to the same category, presented two times each. In general, research has shown that increasing experience with exemplars of a category increases accuracy on a categorization task (Medin & Schaffer, 1978).

Dukes and Bevan (1967) were among the first to manipulate variability during a categorization task, in which participants studied a set of photographs of men and women. There were 20 individuals in total, and the researchers created a set of five photos for each individual by varying their clothes and pose. Participants had to try to learn the name of each individual. The high variability condition had four sets of photos presented once and the low variability condition had one set presented four times. Participants were then tested on the pictures they had studied plus new pictures. Dukes and Bevan found that the number of correct responses on new items was greater for the high variability condition compared to the low variability condition. They concluded that this effect was a result of greater generalizability for the high variability condition.

A more recent study by Perry et al. (2010) looked at the effect of variability on category learning in children. They had young children (mean age of 18 months) learn words for certain objects (e.g. a bucket, a hammer, and a toothbrush). They taught the words by showing the children three pictures of the same object and saying its name. They varied the similarity between the three pictures, with the low variability condition having very similar pictures and the high variability condition having highly dissimilar pictures (the degree of similarity was previously determined by a multidimensional scaling algorithm using different features of the objects). They found that children in the

high variability condition performed better on novel items compared to children in the low variability condition. The authors suggest that this effect could be a result of children in the low variability condition attending to the shape of the objects, while children in the high variability condition did not, resulting in the former relying on a less-diagnostic feature to classify novel exemplars. An alternative explanation is that a high variability condition allowed the children to see the features of the stimuli that were not relevant to the category, which might have allowed them to disregard these features when trying to classify a novel exemplar.

There is a natural trade-off in category learning when variability is increased, as this necessarily decreases repetition of each particular exemplar when the same number of learning trials are used. Most of the studies previously cited on variability effects also demonstrate a counter effect of repetition. Conditions that have low variability, where the same exemplars are presented repeatedly, result in better test performance on the items used in the study phase. For instance, Perry et al., (2010), while showing that high variability resulted in better performance on new items, also found that low variability resulted in better performance for old items. Similarly, recent findings from Wahlheim et al. (2012) support both the variability effect and the repetition effect. In sum, increasing repetition (presentation of the same items) improves categorization of previously seen exemplars, while increasing variability (presentation of many different exemplary members of a category) improves categorization of novel items.

The effect of variability may be explained within the context of the exemplar-based models of category learning, which suggest that increasing the number of exemplars similar to a test item facilitates classification learning (Medin & Schaffer,

1978; Medin & Smith, 1981). Numerous studies have found that learning a category by studying many exemplars, or exemplars that are highly diverse in defining stimuli, increases the generalizability of that category (Homa, 1978; Homa et al., 1981; Homa & Vosburgh, 1976). However, the variability effect can equally easily be explained by the prototype model, since increasing variability may add to the stored prototypical information of a category. In order to determine which theory provides the best explanation, recent studies have used metacognitive measures.

1.2 Metacognition in category learning

Metacognition is the monitoring and knowledge of our own cognition (Flavell, 1979). Metacognitive research has mostly focused on the relation between metacognition and memory (see Dunlosky and Metcalfe, 2009, for a review). For example, one measure, judgment of learning (JOL), is used to compare people's judgments of their own learning to their actual performance (Arbuckle & Cuddy, 1969). There are also measures that ask for predicted future performance or confidence in the accuracy of past performance. The accuracy of the judgment is then measured by comparing the estimate to actual performance. One goal of metacognitive research is to identify inaccurate beliefs about cognition and develop methods to counteract these beliefs. The basis of metacognitive measures, such as JOLs, can be readily applied to other cognitive areas such as category learning.

The recent use of metacognition with category learning has focused heavily on how people choose to learn. Thomas et al. (2015) demonstrated that participants' metacognitive judgments of learning were more accurate after being given an initial test with feedback and that they were sensitive to item difficulty. It has also been found that

spacing study of different categories (i.e., by studying various categories intermixed) is better for learning than blocked study (i.e., studying many exemplars from one category in a row before studying a different category; e.g., Wahlheim et al., 2011), but participants demonstrate a preference for blocking their study when given a choice (Tauber et al., 2013). A handful of studies have specifically looked at CLJs, in which participants are asked to judge how well they can classify a novel item of a particular category (Higgins & Ross, 2011; Tauber & Dunlosky, 2015; Tauber et al., 2013; Thomas et al., 2015; Wahlheim & DeSoto, 2016; Wahlheim, et al., 2011; Wahlheim et al., 2012). This is an important measure as it directly addresses the level of confidence a participant has about their ability to categorize members of a given category, which as previously mentioned is the primary aim of category learning.

One of the earliest studies to use metacognitive measures in a category learning procedure was by Homa et al. (1981) who examined the effects of variability on confidence judgments. In their experiment participants studied three different kinds of categories, with varying numbers of instances per category (i.e., variability). They tested participants by having them classify items as either one of three categories or as a 'junk' item (never before seen) and had them give confidence judgments for each of their responses. As expected by previous studies (e.g., Homa, 1978), an increase in category size was positively correlated with an increase in correct classification of new instances. Interestingly, an increase in category size resulted in a decrease in the effect of old-new similarity; that is to say, the tendency to identify a new item as a studied item based on its similarity to an old category. This means that when more instances of a category are given, it becomes more distinguishable from other categories. Confidence judgments

were found to reflect performance, suggesting that participants were able to recognize that an increase in category size improved their ability to discriminate between new and old members of a category.

A similar approach was taken by Wahlheim et al. (2012) who examined how variability and repetition affected metacognitive judgments of performance on a natural category learning task. The focus of their study was on the effect of variability and repetition on CLJs. In their study, Wahlheim et al. had participants study a set of bird species from 12 different families, in either high repetition or high variability conditions. Following the study phase, participants were given a classification test on the items they had studied plus new items. By manipulating the study items as described, classification of test items showed their respective effects on previously studied items and new items: High variability families resulted in better performance for new items, whereas high repetition families produced better performance for studied items.

More importantly, at the end of the study phase, Wahlheim et al. (2012) asked participants to make CLJs for each of the bird families. In their first experiment, participants' overall confidence in their predictions of classifying novel bird species compared to their actual performance showed that they were overconfident in their judgments (i.e., they predicted better performance than they actually obtained). In particular, the high repetition condition showed a greater discrepancy between CLJs and performance than the high variability condition, which was not due to higher CLJs but rather better classification accuracy for the high variability condition; this is called *variability neglect*. These results demonstrated that participants failed to consider variability as a benefit to category learning; this finding was replicated in their subsequent

experiments. One goal of the present thesis was to replicate Wahlheim et al.'s findings of variability neglect in CLJs while learning natural categories; specifically, I expected that participants' classifications of novel items would be more accurate in the high variability condition compared to the low variability condition, but that their CLJs would be the same for both conditions. Critically, the present experiments elicited CLJs throughout the entire learning phase, in order to determine whether changes in CLJs corresponded to changes in learning performance.

1.3 Trial-by-trial learning

Histed, Pasupathy, and Miller (2009) conducted an association learning task with monkeys in order to better understand the neurological basis of learning. The task involved monkeys making an eye movement response to the right or left depending on which particular stimulus they were shown. The monkeys were rewarded following each correct response. During the task, Histed et al. examined the firing rate of neurons in the prefrontal cortex (PFC) and the caudate nucleus (Cd) of the basal ganglia, and found evidence to support a sustained firing model of learning. What their model showed was that the information from a single trial altered the firing rate of neurons, which influenced behaviour on the next trial (Ganguli, Huh, & Sompolinsky, 2008). The researchers interpreted the results as showing that the outcome of a single trial affected the selective direction of the following trial, meaning that if the monkeys made a correct eye movement, then on the next trial they were more likely to make an eye movement in the same direction (Pasupathy & Miller, 2005).

Previously, the influence of outcome reward had been shown to affect activity during a learning task (Seo, Barraclough, & Lee, 2007; Seo & Lee, 2009). Histed et al.

(2009) showed that a given trial's outcome had a direct effect on a trial-to-trial basis and that behaviour was altered solely on the basis of the previous trial's outcome (e.g., Seo, Barraclough, & Lee, 2007; Seo & Lee, 2009). In addition, the likelihood of a correct response was found to increase following a correct trial, which was suggested to occur because of a direction selectivity effect. That is, if a response led to a correct trial then the animals were likely to make the same response on the next trial which would lead to successful learning, whereas a response leading to an incorrect trial was equally likely to be followed by the same response or a different one. In terms of category learning I predict that a given trial, for a particular category, is more likely to be correct if the previous trial is correct, and that participants' CLJs will increase accordingly.

1.4 Current thesis

The primary goal of the present thesis was to examine more closely whether CLJs reflect actual learning using a detailed analysis of the response patterns during learning. That is, are there specific patterns of behaviour during category learning that are utilized as cues to make inferences about how well one is learning a new category? The present study set to test out the metacognitive knowledge of participants during a category learning task. Specifically, the purpose was to observe whether participants judge learning to improve more following a correct trial versus an incorrect trial and whether they consider variability to be a benefit to learning. Two experiments were conducted in which participants completed a category learning task where they were shown a series of pictures belonging to natural categories (birds in Experiment 1; paintings in Experiment 2) and were asked to classify them; they were given feedback after each trial. Following feedback, participants made CLJs on the just-classified category. Categories were

manipulated for variability, meaning that while holding the number of learning trials constant, some categories only had a few exemplars presented repeatedly, while other categories had numerous unique exemplars but shown a fewer number of times. Learning was tested by presenting items from the learning task intermixed with new items and participants once again had to classify exemplars by category.

While learning was expected to result in an increase in correct classification of exemplars as the task went on, a specific trial-by-trial pattern was also expected to emerge. Similar to the results found by Histed et al. (2009), I expected to show that following a correct classification of an exemplar in a given category, the next instance from that category is more likely to be correct, as opposed to when an incorrect classification is made. It was hypothesized that correct classification trials of a particular category would increase the likelihood that the following trial on the same category would also be correct. The reason behind this prediction is based on Pasupathy and Miller's (2005) direction selectivity effect. The likelihood of making a certain response will increase if it leads to a correct trial. Since the direction selectivity effect is based on associative learning, I predict this result will apply to trials within the same category as participants are making the association between exemplar and category membership.

The effect of correct classification on subsequent trials was predicted to affect CLJs. A second hypothesis was that CLJs would be higher following correct trials than incorrect trials. This hypothesis was derived from the predictive learning models (see Luque, López, Marco-Pallares, Càmara, Rodríguez-Fornells, 2012), whereby correct trials decrease the chance of a future incorrect response and increase the perceived understanding of a concept. Also, considering CLJs were found to be generally

overconfident by Wahlheim et al. (2012), correct trials were expected to boost an individual's confidence in the understanding of a category. Lastly, metacognitive judgments were predicted to ignore the effect of variability, based on the results from Wahlheim et al. (2012).

2.1 Experiment 1

2.2 Method

2.2.1 Participants. Participants included 35 Memorial University students; one participant did not complete the procedure. Twenty participants received course credit in exchange for participation and 15 were compensated $10 for participating.

2.2.2 Materials. The experiment was conducted on a computer, using E-prime 2.0 (Psychology Software Tools, Pittsburgh, PA). The participants were shown a series of pictures of birds centered on a white background accompanied by a list of bird family names labelled 1-6 directly underneath. A total of 96 pictures were used, which included 16 different exemplars for each of the following six families: finch, jay, oriole, sparrow, flycatcher, and warbler. Nearly all exemplars depicted different species of birds within each family. The photos were selected from a set of photos used by Tullis, Benjamin, and Ross (2011) (originally obtained from various internet sources). For each participant, the program randomly selected three of the families to be assigned to the Diverse (high variability) condition; the remaining three families were assigned to the Repeated (low variability) condition. Diverse families contained 12 different birds shown once each during each block of the learning phase and Repeated families contained four different birds shown three times each during each block of the learning phase. Four studied bird

pictures (for the Diverse families, a random four of the studied exemplars were selected) and four new bird pictures from studied families were presented on the test.

2.2.3 Design. Category type (Diverse vs. Repeated) and learning block (the learning phase was composed of two blocks, with 72 trials per block) were manipulated within-subjects. A mean accuracy measure was obtained from an average score across all trials, with separate accuracy scores for the two blocks of the learning phase and the test phase. Similarly, a mean CLJ measure was obtained by averaging all responses across trials for each block of the learning phase. Additional analyses of individual trial response contingencies were conducted as described below.

2.2.4 Procedure. Participants were tested individually in a small cubicle. The experiment began with an introduction explaining that the task was to classify bird species based on the family they belonged to, as well as to learn the families in general. Participants were informed that they would be given feedback on each response, and they would then be asked to provide a CLJ. Participants were told to use the keyboard to make their responses, and to take as much time as they needed to complete each trial. The experiment began once the participant pressed a key to start.

Learning trials were self-paced. On each trial a slide presented a picture of a bird at the center of the screen that took up approximately half of the screen, with the six family names listed underneath and labelled 1-6. Participants selected a response by pressing the number key on the keyboard corresponding to the family name shown on screen. A response input display appeared on screen for participants to view and correct if a typing mistake was made. Following their choice a feedback slide told them if they were correct or incorrect and displayed the correct answer, including the picture of the

just-classified bird, for 3000 ms. The next slide asked them to make a category learning judgment (CLJ) with the following prompt: “How confident are you that you know the (FamilyName), that if shown a different bird in this family you could identify it?”, where (FamilyName) was the bird family that had been presented prior to the CLJ. Participants made their choices on a rating scale of 0-100, 0 being “not confident at all” and 100 being “completely confident”. They were encouraged to make use of the whole range of the scale. The learning phase was composed of two blocks of 72 trials. Exemplars from six bird families, including three Diverse category types (twelve birds presented once) and three Repeated category types (four birds presented three times each) were presented in random order in Block 1. Block 2 presented the same bird exemplars as in Block 1, but in a new random order. Participants were not explicitly informed about the different types of categories.

The test phase began directly after the last learning trial ended. Before starting the first test trial, participants were informed that the test phase was about to begin and that they would no longer be given feedback on their classification accuracy. They were instructed to try to correctly classify as many bird species as possible. Test trials began once the participants indicated they were ready. The appearance for each test trial was identical to the study trials, except that feedback and CLJs were not included. Following a response, the experiment immediately moved on to the next trial. The test phase consisted of 48 trials: for each Repeated category type, each studied item was shown plus four new items; for each Diverse category type, a random four of the possible twelve studied items were presented plus four new items.

2.3 Results

Statistical reliability was measured at $p < 0.05$ in all of the analyses to be discussed. I first considered test performance in relation to CLJs, and then explored the learning trials in more detail. Data from four participants were eliminated from all analyses to be discussed, due to failure to complete the experiment or failure to follow instructions. Some analyses report different degrees of freedom, which occurred whenever a measure contained missing data from a participant (e.g. when a participant did not have any incorrect trials for a category).

2.3.1 Test Phase

2.3.1.1 ***Mean Performance***. Test accuracy for old and new items (see Table 2.1) was analyzed using a 2 (item status: new vs. old) x 2 (category: Diverse vs. Repeated) repeated measures ANOVA. There was a main effect of item status (F (1, 30) = 101.29, MSE = 0.022, $p < 0.001$, $\eta_p^2 = 0.77$), but no main effect of category type ($p = 0.50$). Old items were categorized better than new items, in general, but there was also a significant interaction (F (1, 30) = 31.16, MSE = 0.012, $p < 0.001$, $\eta_p^2 = .51$). Old items from Repeated categories were categorized better than old items from Diverse categories, but there was no significant difference between category types in performance on new items.

2.3.2 Learning Phase. Classification accuracy and CLJs across the learning phase are displayed in Figure 2.1. This figure primarily demonstrates the increase in accuracy for both Repeated and Diverse categories as the learning phase progressed, while also showing that CLJs increased accordingly. Participants' CLJs were much closer to their actual performance on Diverse categories than Repeated categories, even as they were performing better on the Repeated category items.

Table 2.1 *Mean Performance and Category Learning Judgments (CLJs) in Learning and Test Phases*

	Experiment 1		Experiment 2	
	Category Type			
	Diverse	**Repeated**	**Diverse**	**Repeated**
Learning Phase Accuracy	0.49	0.63	0.63	0.71
	(0.20)	(0.02)	(0.026)	(0.019)
Learning Phase CLJs				
Correct	57.29	58.97	61.19	61.87
	(3.91)	(4.30)	(4.54)	(3.74)
Incorrect	38.59	39.56	43.03	41.71
	(3.67)	(3.74)	(4.18)	(4.18)
End of learning CLJs				
Old Items	-	-	77.21	76.00
			(3.03)	(3.04)
New Items	60.88	68.70	64.80	62.93
	(5.14)	(4.96)	(3.29)	(3.05)
Test Phase Accuracy				
Old Items	0.67	0.81	0.79	0.91
	(0.20)	(0.15)	(0.028)	(0.021)
New Items	0.51	0.43	0.68	0.60
	(0.19)	(0.20)	(0.028)	(0.036)

Note: For Experiment 1, end of learning CLJs were calculated using the mean of the last 3 CLJs per category. For Experiment 2, end of learning CLJs were the mean of the global CLJs per category type. Numbers in parentheses are standard error of the mean.

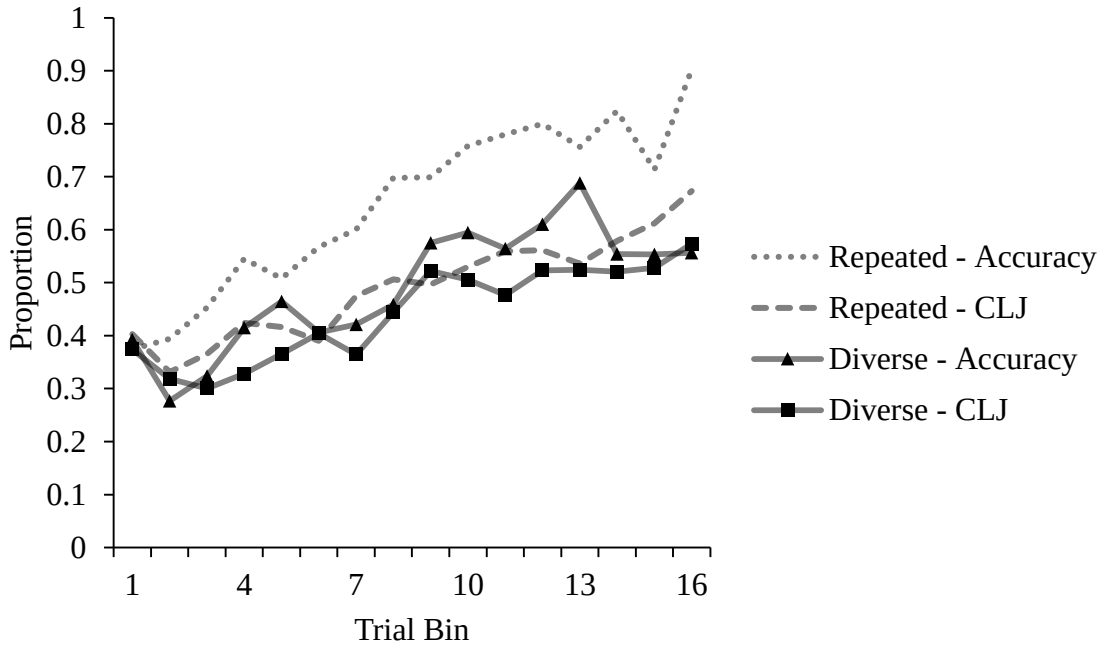

Figure 2.1 Mean classification accuracy and CLJs for Diverse and Repeated categories across both blocks of learning trials in Experiment 1. Each trial bin includes nine trials.

2.3.2.1 ***Mean Performance***. Mean CLJs following correct vs. incorrect responses (displayed in Table 2.1) were analyzed using a 2 (Block: 1 vs. 2) x 2 (category type: Diverse vs. Repeated) x 2 (response: correct vs. incorrect) repeated measures ANOVA. There were main effects of Block (F (1, 29) = 27.85, MSE = 407.25, $p < 0.001$, η_p^2 = 0.49) and response (F (1, 29) = 55.26, MSE = 393.96, $p < 0.001$, $\eta_p^2 = 0.66$), but no main effect of category type ($p = 0.38$), and an interaction between block and response (F (1, 29) = 7.53, MSE = 43.10, $p = 0.010$, $\eta_p^2 = 0.21$). As predicted, CLJs were higher for correct trials than for incorrect trials. The magnitude of the difference between CLJs following correct responses versus incorrect responses was smaller in Block 2 (M = 33.45, SD = 30.58) than in Block 1 (M = 42.75, SD = 28.51), and was unaffected by

category type. This likely indicates a perceived decrease in the overall rate of learning across blocks.

2.3.2.2 ***Trial-by-trial Accuracy***. For each participant, accuracy data from each trial (n) were sorted into four bins on the basis of whether the current trial under consideration and the previous trial within the same family were correct or incorrect: correct (n) following correct (n-1), where *n* and *n-1* were trials corresponding to the same family; correct-incorrect; incorrect-correct; and incorrect-incorrect. These four bins were tabulated separately for both Diverse and Repeated categories (see Table 2.2). Each participant was then given a score based on the proportion of trials that fell into each bin. A nonparametric sign test was used to determine whether categorization responses were more likely to be correct following a correct trial than following an incorrect trial. Wilcoxon's rank sign test was chosen because it assesses whether there is a difference between the mean ranks of two related samples that come from the same population. For Repeated categories, correct-correct differed from correct-incorrect ($Z = -4.76$, $p < 0.001$), while incorrect-correct did not differ from incorrect-incorrect ($Z = -0.361$, $p = 0.72$). When the same analysis was performed for Diverse categories, correct-correct once again differed from correct-incorrect ($Z = -2.15$, $p = 0.031$), however incorrect-incorrect also differed from incorrect-correct ($Z = -2.59$, $p = 0.010$).

2.3.2.3 ***Trial-by-trial CLJs***. The numeric change in CLJs from one trial to the next within a given family was sorted into four bins on the basis of response accuracy contingencies, as described above for accuracy data, but considering Block 1 and Block 2

separately.[1] The values are displayed in Table 2.2. The mean change in CLJ for each of the eight bins were analyzed using a 2 (Block: 1 & 2) x 4 (bin: correct-correct, correct-incorrect, incorrect-correct, incorrect-incorrect) repeated measures ANOVA. There was no main effect of Block (p = .53), but the main effect of bin was significant (F (3, 90) = 27.88, MSE = 244.41, p < .001, η_p^2 = 0.48) as was the Block x Bin interaction (F (3, 90) = 10.78, MSE = 28.42, p < .001, η_p^2 = 0.26).

Because the Block x Bin interaction was significant, follow-up analyses on differences among bins were conducted separately for Block 1 and Block 2. In Block 1, an initial one-way ANOVA was significant (F (3, 120) = 47.01, MSE = 104.652, p < .001, η_p^2 = 0.54). Follow-up pairwise comparisons (using the Bonferroni correction for multiple comparisons) showed that all but two of the bins were significantly different from one another (all ps < .001), while bin 1 (correct-correct) did not differ from bin 4 (incorrect-incorrect). Participants increased their CLJs the most on correct trials when the previous trial had been incorrect, whereas correct responses following correct responses resulted in a significantly smaller increase in CLJs (t (30) = 6.57, p < .001, t (30) = 2.99, p = .005, respectively) When participants were incorrect and had been incorrect on the previous trial, CLJs did not increase or decrease significantly (p = .53), and when an incorrect response followed a correct response, CLJs decreased significantly (t (30) = -6.13, p < .001). That is, CLJs tended to change the most when response accuracy changed, and changed only slightly (if at all) when response accuracy was the same as on trial n – 1.

[1]I initially considered Diverse and Repeated families separately in this analysis, but found no significant differences based on category type. I therefore collapsed over this variable in the reported analysis.

Table 2.2 *Trial-by-Trial Accuracy Contingencies during Learning Trials and Corresponding Changes in Category Learning Judgments (CLJs) in Experiment 1*

	Trial Type			
	Correct-Correct	**Correct-Incorrect**	**Incorrect-Correct**	**Incorrect-Incorrect**
Proportion of Learning Trials				
Diverse	0.28	0.21	0.21	0.29
	(0.025)	(0.007)	(0.006)	(0.022)
Repeated	0.45	0.16	0.19	0.20
	(0.029)	(0.0078)	(0.0083)	(0.021)
Change in CLJs				
Block 1				
Diverse	+3.36	-13.40	+14.77	-0.98
	(1.81)	(2.30)	(2.01)	(0.84)
Repeated	+3.02	-17.38	+15.59	-0.12
	(1.39)	(3.08)	(3.05)	(1.25)
Block 2				
Diverse	+2.72	-8.15	+10.76	-1.16
	(0.83)	(2.46)	(2.31)	(0.61)
Repeated	+0.86	-9.17	+12.64	-1.13
	(0.60)	(3.02)	(2.98)	(1.19)

Note: Numbers in parentheses are standard error of the mean.

A similar analysis of Block 2 also showed that all but two of the bins were significantly different from one another (all *p*s < .019), and bin 1 and 4 did not differ. Changes in CLJs across bins followed the same pattern as in Block 1, but the magnitude of the increase or decrease in CLJs was smaller for Incorrect-Correct bins (t (30) = 4.82,

$p < .001$), Correct-Incorrect (t (30) = -3.26, p = .003) and Correct-Correct (t (30) = 3.51, p = .001), relative to Block 1.

2.3.2.4 *Calibration of CLJs*. Calibration scores were obtained by taking the average of the last three CLJs per category (as a measure of performance at the end of the learning phase) and subtracting a corresponding accuracy score (e.g., CLJs for diverse – accuracy for old, diverse items). Calibration scores for diverse and repeated categories can be seen in Figure 2.2. Values above zero indicate overconfidence, those under zero indicate underconfidence and zero indicates perfectly accurate scores.

Calibration scores for diverse and repeated categories were analyzed using a 2 (Item status: Old vs. New) x 2 (Category type: diverse vs. repeated) repeated measures ANOVA. There was a main effect of item status (F (1, 30) = 101.288, MSE = 0.022, η_p^2 = 0.771, $p < 0.001$), but no main effect of category type (p = 0.103). There was also a significant interaction (F (1, 30) = 31.163, MSE = 0.012, η_p^2 = 0.51, $p < 0.001$). One-sample t-tests indicated that calibration scores for old (t (30) = -2.966, p = 0.006) and new items (t (30) = 4.896, $p < 0.001$) in repeated categories were different from zero, indicating significant underconfidence and overconfidence, respectively. This was not true for the Diverse condition (p = 0.095 for new items and p = 0.229 for old items).

2.4 Discussion

The purpose of Experiment 1 was to determine whether accuracy on a single learning trial affects subsequent performance on the learning task and whether participants' metacognitive judgments were in line with their performance. The results of the test phase indicated a trend that new items for diverse categories were better classified

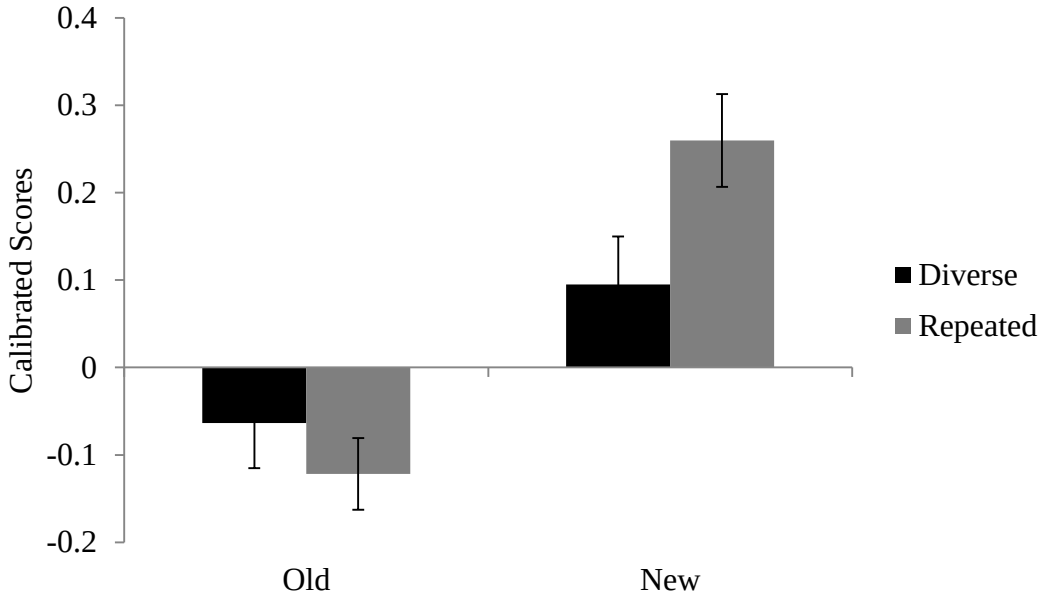

Figure 2.2 Calibration scores (average of last 3 CLJs minus test accuracy) for Diverse and Repeated categories in Experiment 1. *Error bars* represent standard errors of the means.

than new items for repeated categories; however, this finding failed to reach significance. Repetition had the expected benefit for classification of old Repeated items compared to old Diverse items. Participants' CLJs were greater following correct responses versus incorrect responses; they increased as the learning phase progressed, and they did not differ based on category type. Thus, participants did not consider variability when giving their trial-by-trial CLJs. Contrary to Wahlheim et al.'s (2012) finding and our prediction, new items from Diverse categories were not classified better than new items from Repeated categories (p = .087). However, as previously noted there is a numeric trend in the expected direction, so it is possible that the null result is due to the relatively small number of items tested and potential variation in item difficulty. In their third and fourth

experiment Wahlheim and DeSoto (2016) also failed to find a variability effect using a very similar procedure to the Wahlheim et al. (2012) study. Another finding to consider in the current experiment is that participants were not well calibrated on repeated categories, in general, meaning that the predictions for how well they thought they would classify old and new items were numerically quite different from the observed accuracies. This indicates overconfidence in repetition, rather than a neglect of variability (especially considering that diverse categories were well calibrated). This is in line with Wahlheim et al.'s finding that participants were greatly overconfident for new items in their repeated condition.

In regard to trial-by-trial accuracy, a correct trial was more likely to be followed by another correct trial than by an incorrect trial, while incorrect trials were equally likely to be followed by correct or incorrect trials. An interpretation of the data would suggest that a trial was more likely to be correct if the previous trial (from the same category) was also correct. This supports my prediction and the direction selectivity effect (Histed et al., 2009); successful behavior leads to better learning than making mistakes. The implications of this finding are discussed below.

The changes in the magnitude of CLJs showed that participants' judgments of their own learning are sensitive to the feedback they receive. When a correct response is made following an incorrect response, this is perceived to be the result of a large gain in overall learning of the category, leading to a large increase in CLJ magnitudes. Conversely, when an incorrect response is made following a correct response, this results in a reduction in CLJ magnitudes, perhaps in an attempt to correct a perceived overconfidence from the prior correct trial. For repeated responses (correct-correct and

incorrect-incorrect), CLJ magnitudes change very little from trial to trial, reflecting a belief that learning is not changing substantially. This finding was surprising and novel, but has wide implications on the nature of how people perceive learning. Therefore, one of the goals of a second experiment was to replicate this finding.

3.1 Experiment 2

The purpose of Experiment 2 was primarily to replicate the findings of Experiment 1 and extend them to different stimuli. The stimuli that I chose were previously used by Kornell and Bjork (2008). The set of stimuli were 10 pictures of landscape or skyscape paintings from six different artists, chosen because of their relatively low popularity, therefore making them less likely to be known by participants. Kornell and Bjork used these stimuli to observe the effects of spaced (interleaved) versus massed (blocked) study. They predicted that massed study would help learning more than spaced, because it would allow participants to compare the similarities among exemplars. Contrary to their prediction they found that spaced study was better than massed study, although participants judged massed to be more beneficial. As previously mentioned, this "massing illusion" has since been supported by follow up studies (e.g., Tauber et al., 2013).

Another addition to Experiment 2 was that global CLJs (i.e., single judgments about each artist category) were elicited after participants completed the learning phase and which asked for separate confidence ratings for both old and new items for each category. This procedure differs from Experiment 1, which only asked for participants to give a CLJ rating after every trial (i.e., item-level CLJs). Since item-level CLJs specifically state to give a confidence judgment based on future performance on novel

items, it is difficult to interpret the relation between item CLJs and performance on old items at test. By including global CLJs for both new and old items it would allow for a better determination of whether participants' CLJs are accurate predictors of their test performance on the two types of items. It was expected that global CLJs would show a bias toward Repeated categories over Diverse categories, similar to the item-level CLJs in Experiment 1.

3.2 Method

3.2.1 Participants. Participants included 34 Memorial University students. All participants received course credit in exchange for participation.

3.2.2 Materials. Instead of pictures of birds, Experiment 2 used pictures of paintings, selected from materials originally used by Kornell and Bjork (2008). I used paintings from the following six artists: Pessani, Stratulat, Wexler, Braque, Seurat, and Cross. Besides changing pictures, the number of stimuli and repetitions in each condition was identical to Experiment 1.

3.2.3 Design. The design was identical to Experiment 1, with the addition of global CLJs (for both Old and New items) elicited at the end of the learning phase.

3.2.4 Procedure. The procedure for Experiment 2 was almost identical to the procedure for Experiment 1, besides the change in stimuli and the addition of global CLJ ratings. Before starting the learning phase participants gave a CLJ rating for each of the artists, simply to indicate whether any participants had knowledge of the artists prior to the experiment. Upon completion of the learning phase participants were asked to give two CLJ ratings for each category, called global CLJs. They were first given the following prompt: "Please give an overall confidence judgment on how likely you would

be to correctly identify the paintings you just studied by [Artist]". This was followed by a very similar prompt, except it was for asking for a confidence rating for new items from that artist. The order of the artists was random, however participants always gave a rating for old items followed by a rating for new items for the same artist.

3.3 Results

Statistical reliability was measured at $p < 0.05$ in all of the analyses to be discussed. As I did in Experiment 1, I first consider test performance in relation to CLJs, and then explore the learning trials in more detail. Data from three participants were eliminated from all the analyses to be discussed because of failure to follow instructions. As in Experiment 1, the degrees of freedom for the following analyses change depending on whether there are missing values.

3.3.1 Test Phase.

3.3.1.1 ***Mean Performance.*** Test accuracy for old and new items (see Table 3.1) was analyzed using a 2 (item status: new vs. old) x 2 (category: Diverse vs. Repeated) repeated measures ANOVA. There was a main effect of item status (F (1, 30) = 186.68, MSE = 0.007, $p < 0.001$, $\eta_p^2 = 0.86$), but no main effect of category type ($p = 0.66$). Old items were categorized better than new items, in general, but there was also a significant interaction (F (1, 30) = 16.6, MSE = 0.019, $p < 0.001$, $\eta_p^2 = 0.36$). Old items from Repeated categories were categorized better than old items from Diverse categories, but there was no difference between category types in performance on new items. However, as we found in Experiment 1, there is a numeric trend in the predicted direction (new items from Diverse categories being categorized better than Repeated categories).

Table 3.1 *Trial-by-Trial Accuracy Contingencies during Learning Trials and Corresponding Changes in Category Learning Judgments (CLJs) in Experiment 2*

	Trial Type			
	Correct-Correct	**Correct-Incorrect**	**Incorrect-Correct**	**Incorrect-Incorrect**
Proportion of Learning Trials				
Diverse	0.45	0.17	0.19	0.19
	(0.033)	(0.0091)	(0.0092)	(0.022)
Repeated	0.56	0.14	0.17	0.13
	(0.029)	(0.0098)	(0.01)	(0.013)
Change in CLJs				
Block 1				
Diverse	+4.01	-11.75	+15.22	+0.12
	(0.66)	(2.06)	(1.92)	(0.66)
Repeated	+4.94	-11.95	+13.51	-0.29
	(1.002)	(1.22)	(1.47)	(0.83)
Block 2				
Diverse	+2.15	-7.16	+7.93	-0.54
	(0.58)	(1.54)	(1.41)	(1.08)
Repeated	+1.88	-5.40	+6.50	-1.17
	(0.38)	(1.00)	(1.39)	(1.30)

Note: Numbers in parentheses are standard deviations.

3.3.2 Learning Phase. Classification accuracy and CLJs across the learning phase are displayed in Figure3.1. Similar to Figure 2.1, participants' CLJs were again much closer to their actual performance on Diverse categories than Repeated categories, even as they were performing much better on the Repeated category items.

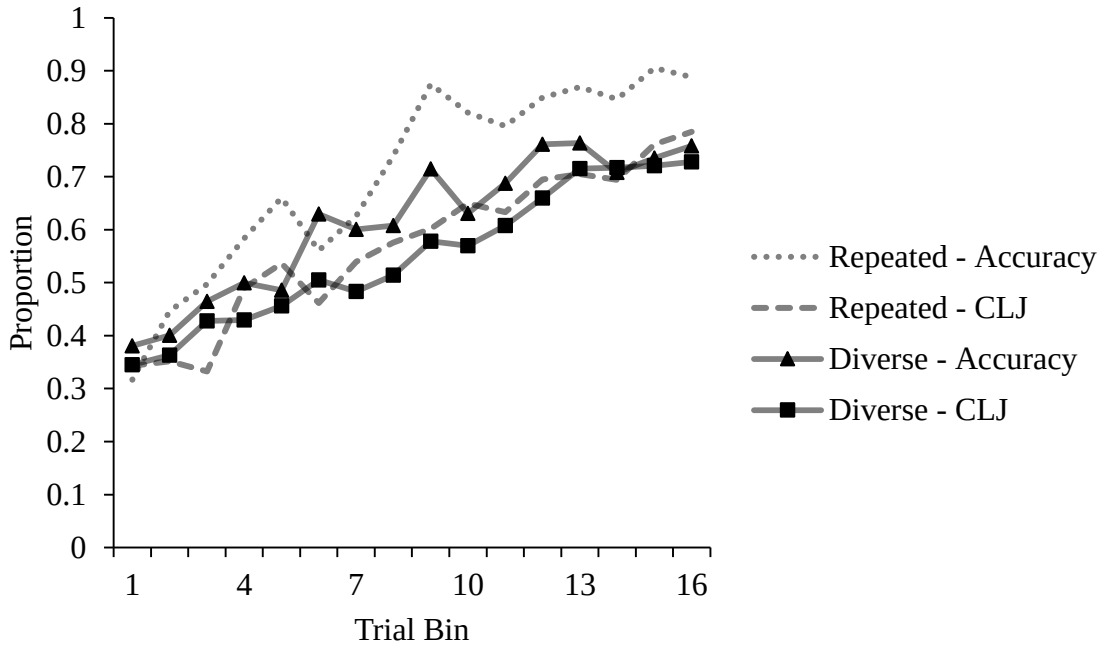

Figure 3.1 Mean classification accuracy and CLJs for Diverse and Repeated categories across both blocks of learning trials in Experiment 2. Each trial bin includes nine trials.

3.3.2.1 ***Mean Performance***. Mean CLJs following correct vs. incorrect responses (displayed in Table 2.1) were analyzed using a 2 (Block: 1 vs. 2) x 2 (category type: diverse vs. repeated) x 2 (response: correct vs. incorrect) repeated measures ANOVA. There were main effects of Block (F (1, 25) = 38.11, MSE = 506.66, $p < 0.001$, η_p^2 = 0.60) and response (F (1, 25) = 149.42, MSE = 127.75, $p < 0.001$, $\eta_p^2 = 0.86$), but no main effect of category type ($p = 0.39$), and no significant interactions (all $ps > 0.125$). This means that CLJs for both correct and incorrect responses increased from Block 1 to Block 2 and CLJs for correct responses were higher than CLJs for incorrect responses.

3.3.2.2 ***Trial-by-trial Accuracy***. Responses were binned and analyzed exactly as in Experiment 1 (see Table 3.1). For Repeated categories, correct-correct differed from

correct-incorrect (Z = -6.34, p < 0.001). Incorrect-correct also differed from incorrect-incorrect (Z - -2.63, p = 0.009). When the same analysis was performed for Diverse categories, correct-correct differed from correct-incorrect (Z = -5.98, p < 0.001), while incorrect-correct did not differ from incorrect-incorrect (Z = -1.53, p = 0.13).

3.3.2.3 ***Trial-by-trial CLJs***. The mean change in CLJ for each of the eight bins (see Table 3.1) were analyzed using a 2 (Block: 1 & 2) x 4 (bin: correct-correct, correct-incorrect, incorrect-correct, incorrect-incorrect) repeated measures ANOVA. There was a significant main effect of Block (F (1, 30) = 7.08, MSE = 12.69, p < .012, η_p^2 =0.19), Bin (F (3, 90) = 85.78, MSE = 49.23, p < .001, η_p^2 = 0.74), and a Block x Bin interaction (F (3, 90) = 11.76, MSE = 36.47, p<.001, η_p^2 = 0.28).

Because the Block x Bin interaction was significant, follow-up analyses on differences among bins were conducted separately for Block 1 and Block 2. In Block 1, an initial one-way ANOVA was significant (F (3, 120) = 86.56, MSE = 41.41, p<.001, η_p^2 = 0.69). Follow-up pairwise comparisons (using the Bonferroni correction for multiple comparisons) showed that all four bins were significantly different from one another (all ps < .036). Participants increased their CLJs the most on correct trials when the previous trial had been incorrect, whereas correct responses following correct responses resulted in a significantly smaller increase in CLJs (t (30) = 9.32, p < .001, t (30) = 6.59, p < .001, respectively) When participants were incorrect and had been incorrect on the previous trial, CLJs did not increase or decrease significantly (p = .87), and when an incorrect response followed a correct response, CLJs decreased significantly (t (30) = -7.98, p < .001). A similar analysis of Block 2 also showed that all bins, except bins 1 and 4, were significantly different from one another (all ps < .001). Changes in CLJs across bins

followed the same pattern as in Block 1, but the magnitude of the increase or decrease in CLJs was smaller for Incorrect-Correct bins (t (30) = 5.93, p < .001), Correct-Incorrect (t (30) = -6.12, p = .001) and Correct-Correct (t (30) = 4.47, p < .001), relative to Block 1.

3.3.2.4 ***Calibration of CLJs***. Calibration scores for Experiment 2 were first calculated the same way as Experiment 1 (i.e., averaging the final three CLJs per category and subtracting the corresponding accuracy score). Calibration scores for diverse and repeated categories can be seen in Figure 3.2. Calibration scores for diverse and repeated categories were analyzed using a 2 (Item status: Old vs. New) x 2 (Category type: diverse vs. repeated) repeated measures ANOVA. There was a main effect of item status (F (1, 30) = 146.198, MSE = 0.009, η_p^2 = 0.830, p < 0.001), but no main effect of category type (p = 0.555). This means that CLJs more accurately predicted performance on old items than on new items. There was also a significant interaction (F (1, 30) = 16.095, MSE = 0.019, η_p^2 = 0.349, p < 0.001). One-sample t-tests indicated that calibration scores for old (t (30) = -4.085, p < 0.001) and new items (t (30) = 2.360, p = 0.025) in Repeated categories were different from zero, indicating significant overconfidence. Once again this was not true for the Diverse condition (p = 0.214 for new items and p = 0.255 for old items).

With the inclusion of global CLJ ratings another calculation for calibration scores was obtained by simply taking the global CLJ ratings and subtracting the corresponding accuracy measure (e.g., CLJs for old items – accuracy on old items). These new calibration scores for diverse and repeated categories can be seen in Figure 3.3. Calibration scores for Diverse and Repeated categories were analyzed using a 2 (Item

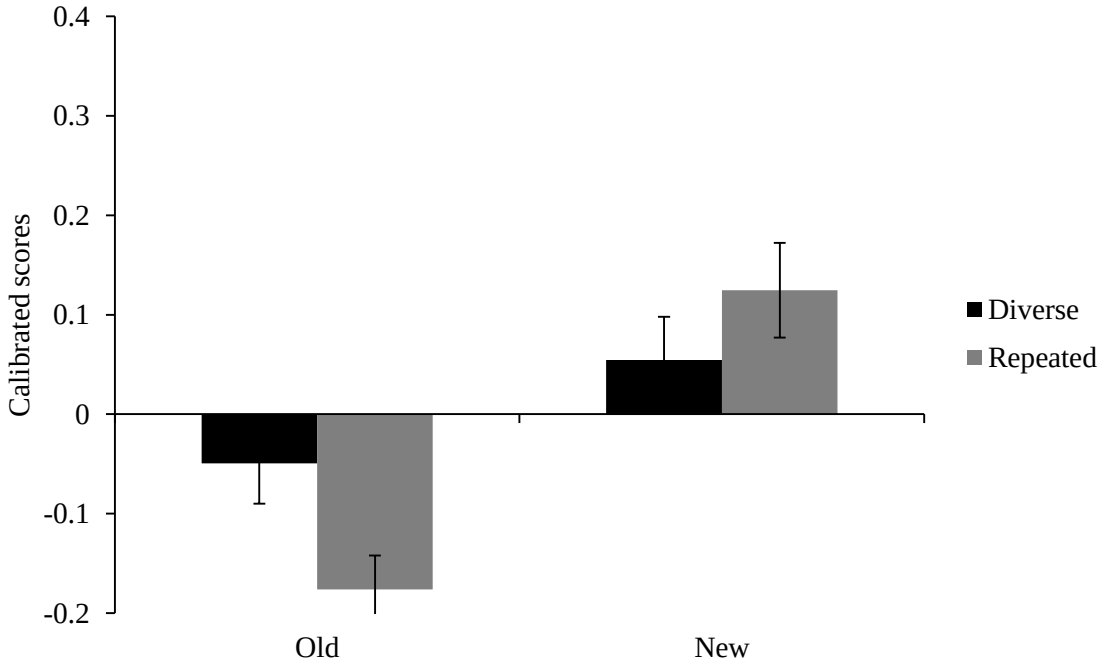

Figure 3.2 Calibration scores (average of last 3 CLJs minus test accuracy) for Diverse and Repeated categories in Experiment 2. *Error bars* represent standard errors of the means.

status: Old vs. New) x 2 (Category type: Diverse vs. Repeated) repeated measures ANOVA. There was a main effect of item status (F (1, 30) = 7.960, MSE = 0.024, η_p^2 = 0.21, $p < 0.008$), but no main effect of category type ($p = 0.512$). There was also a significant interaction (F (1, 30) = 14.802, MSE = 0.02, η_p^2 = 0.33, $p = 0.001$). A one-sample t-test indicated that calibration scores for old items (t (30) = -4.272, $p < 0.001$) in Repeated categories were different from zero (i.e., significantly overconfident), but none of the other calibration scores were different from zero (all ps > 0.417).

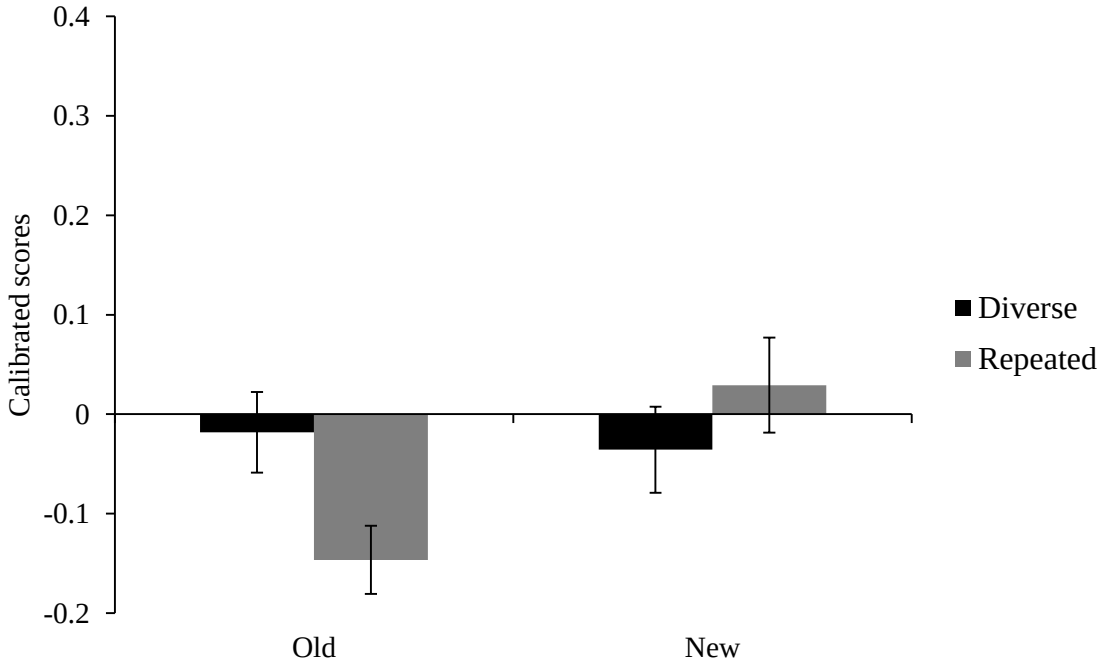

Figure 3.3 Calibration scores (global-level CLJs minus test accuracy) for Diverse and Repeated categories in Experiment 2. *Error bars* represent standard errors of the means.

3.4 Discussion

Experiment 2 was designed to replicate the results of Experiment 1, as well as extend them to a different set of stimuli. Performance during the test phase demonstrated that old items from repeated categories were better classified than old items from diverse categories. There was no difference between category types on new items. As in Experiment 1, this result demonstrates a benefit of repetition on classification accuracy and a non-significant trend for variability (new items from diverse categories were better classified than new items from repeated categories ($p = 0.123$)). CLJs increased as learning went on; they were greater following correct versus incorrect responses, and they did not differ based on category type. Once again, "variability neglect" as described by

Wahlheim et al. (2012) was not found, however calibration scores did replicate the finding that participants were overconfident for repeated categories in terms of classifying new items. The finding that participants were underconfident for old items for repeated categories was not replicated when using global CLJs, which suggests that item CLJs were not appropriate for predicting performance on old items. However, the finding of overconfidence for new items for repeated categories remained, demonstrating a robust effect of repetition on participants' judgments that persists even when participants are explicitly asked to globally judge their future performance on repeated categories.

The trial-by-trial accuracy pattern was similar to Experiment 1. A correct trial was more likely to be followed by another correct trial than by an incorrect trial and an incorrect trial was more likely to be followed by another incorrect trial than by a correct trial. This again suggests that a trial was more likely to be correct if the previous trial (from the same category) was also correct. This is in line with the direction selectivity effect (Histed et al., 2009), but the fact that an incorrect trial was likely to be followed by an incorrect trial suggests that participants were not learning from their mistakes.

The result found in Experiment 1 that changes in CLJ magnitudes differ as a result of feedback was replicated in Experiment 2. This supports the idea that participants actively try and change their CLJs based on the perceptions of their own learning and by the feedback they are given. A more detailed account of why this occurs is discussed further below.

4.1 General Discussion

The purpose of the current thesis was to examine how people perceive their own learning after success compared to after failure, and to determine how metacognitive

assessments of category learning change in response to feedback during learning. The novel contribution of the current study was the analysis of how CLJs change during the process of category learning. An analysis of participants' CLJs revealed that people gave higher CLJs following correct trials than following incorrect trials. Looking at trial-by-trial learning for each category during study showed that a trial was more likely to be correct if the previous trial was correct; therefore success did affect subsequent behaviour. In other words, people identify success, rather than failure, as being a greater determinant of learning. Critically, the way in which CLJs changed across trials speaks to how we perceive our own learning.

During both blocks of the study phase, participants increased their CLJs significantly (relative to the last trial with that category) only if they were correct. Contrary to what was predicted, the largest increases in perceived learning occurred on successful trials that followed a previous incorrect response, not a correct response, as the trial-by-trial performance would suggest. While participants' performance on a given trial was more likely to be correct if the previous trial was correct, within a category, their CLJs did not reflect this pattern. Instead, changes in CLJs indicated that participants view correct feedback as indicative of a substantial increase in learning only when that correct response immediately follows an incorrect response. That is, while CLJs did increase on correct trials when the previous trial had also been correct, the change was much smaller in magnitude than when the previous trial had been incorrect. Participants may view repeated successes with a particular category as reflective of only a small increment in learning (and thus increase CLJs moderately), whereas a correct response that follows an incorrect response may be viewed as resulting from the learner gaining significant insight

into that category. Similarly, CLJs on incorrect trials did not change if the previous trial had also been incorrect, but actually decreased if the previous trial had been correct. This pattern is consistent with the idea that repeated incorrect responses reflect no change in learning, but making an error with a category that had previously been correct may reflect a misunderstanding of category membership. Thus, participants' CLJs were reflective of a sophisticated strategy of assessing one's own learning in response to specific feedback about success versus failure.

Wahlheim et al., (2012) define variability neglect as participants' CLJs not accounting for a variability effect, where high variability leads to better performance on novel items (e.g., Dukes & Bevan, 1967). According to this definition, the results of the test phases in both Experiments 1 and 2 failed to replicate the variability effect in category learning: New items from Repeated categories were classified just as well as new items from Diverse categories. This may have resulted from Diverse categories not being learned well enough during study, as the results from both experiments showed that accuracy in the last bin of the learning phase for Diverse categories was 56% for Experiment 1 and 76% for Experiment 2 (see Figures 1 and 2). Importantly, this was reflected by participants' calibrated CLJs for Diverse items, which were almost identical to actual performance. That is, for both old and new items, participants' CLJs were much closer to their performance on Diverse categories than Repeated categories, for which participants overestimated their ability to classify novel exemplars. Wahlheim et al. did not analyze calibration scores, however they did find that participants' CLJs were overconfident compared to their actual performance on novel items for repeated categories. Additionally, in Experiment 2 of the current study, participants' global level

CLJs for new items did not differ between diverse (M = 64.8, SD = 3.29) and Repeated categories (M = 62.92, SD = 3.05). Therefore, even though the current set of experiments did not show variability neglect as defined by Wahlheim et al., they did show that participants do not account for variability in their item- or global-level CLJs.

Similarly, during the learning phase, participants gave higher average CLJs for Repeated items than for Diverse items, suggesting that they perceived performance on Repeated items as reflecting their overall understanding of the categories. This finding may be explained by the fluency heuristic, which suggests that people give judgments based on how fast information comes to mind (Van Overschelde, 2008). For example, if judging the likelihood of remembering people's names, a higher judgment would be given if the name can be easily recalled at the time the judgment is made. According to this heuristic, CLJs could have been influenced by how easily 'correct' information is retrieved when making a CLJ. After a correct response, the ability to recall an example of the category is much easier than if an incorrect response was made. Given that more incorrect responses were made for Diverse items, lower CLJs for Diverse items might reflect the difficulty of recalling a correct exemplar of that category following incorrect responses.

The fact that CLJs were much greater for Repeated categories throughout the learning phase also suggests that participants prefer repetition to variability when the goal is to classify novel items. Contrary to this prediction, Wahlheim and DeSoto (2016) found that when given the choice of studying a category using high or low variability for a categorization test, most participants chose the high variability option. This finding replicated when the test condition contained only novel items and when participants' pre-

existing familiarity was taken into account. One reason why participants in the current study might have given higher CLJs for Repeated categories than Diverse categories is because they were not aware of how variability would affect performance on the test. In other words, if participants were made aware of the fact that the test would have both old and new items, they may have changed their CLJs accordingly. The results of Wahlheim and DeSoto's (2016) Experiment 4 support this possibility, as they demonstrated that when participants were asked about their preference for studying for old and new items, most participants chose variability over repetition, but when asked about their study preference for new items only, most participants chose repetition over variability. This discrepancy seems to suggest that people understand the benefit of variability (also supported by Experiments 2 and 3 from Wahlheim et al., 2012), but only when they are told that there will be old and new items on the test; when they are not informed about the test composition, they do not adjust their metacognitive judgments for variability.

The current results regarding variability demonstrate support for both an exemplar model and a prototype model. As was mentioned previously, the focus on natural categories led to the formation of exemplar based models (Rosch & Mervis, 1975), because natural categories are more difficult to learn using specific rules for categorizing exemplars. This is true as well for the set of stimuli used in Experiments 1 (birds) and 2 (paintings) of the current thesis. However the current set of findings cannot rule out the possibility that participants were forming a prototype from a set of criterion rules throughout the learning phase. One finding that would argue against this idea is that old items were categorized better for Repeated categories than Diverse categories. This means that people were better at categorizing studied categories if the exemplars were

presented repeatedly rather than if a greater number of exemplars were presented. It is unclear whether prototype models would be able to account for this finding, whereas exemplar models specifically mention memory traces for exemplars as being integral to category learning (e.g., Hintzman's Minerva 2 model, 1986).

Generally, prototype models are supported by the finding that new items for diverse categories are better categorized than new items for repeated categories; this result was not replicated in the present study. A diverse category would lead to a better prototype representation, which would explain why novels items would be easier to categorize. Given that the current study and two experiments by Wahlheim and DeSoto (2016) also did not find a benefit for variability for novel items, future studies are needed to understand the conditions needed for variability to have an effect on classification of new exemplars in natural categories.

A procedural difference between present study and that of Wahlheim et al. (2012) was that in the present study, participants were provided with immediate feedback following each classification choice, whereas Wahlheim et al.'s experiments did not provide immediate feedback. Their experiment involved participants studying bird families (with restricted study time per bird), followed by confidence judgments given on each bird family individually and finally, participants' learning was tested with a recognition of 'old' vs. 'new' items. The present study gave participants unrestricted time to study each bird before giving a response and providing feedback. Confidence judgments were collected after each bird was shown and participants were tested on their ability to actually identify category membership for both old and new birds. It is possible that artefacts, such as the use of unrestricted time during learning, or when CLJs were

collected, could explain the difference in results between the present study and those of Wahlheim et al. (2012). Clapper (2015) examined a similar procedural difference in sequencing studies, which he referred to as *supervised* versus *unsupervised* learning; the former involves having participants categorize items and giving them feedback, while the latter simply gives them a certain amount of time to study items for each category. Clapper found that the usual spacing effect (spaced study results in better learning than massed study) was reversed when participants were given a supervised learning period than when they were unsupervised. While the current study did not find completely different results than previous studies on variability by using supervised rather than unsupervised learning, Clapper's study demonstrates that using these two different procedures may have some unexpected effects (such as changing the level of categorization difficulty).

Considering the learning phase of the current study in more detail, the analysis of trial-by-trial learning indicated that, for the repeated categories, a trial was more likely to be correct if the previous trial was correct, but trials for diverse categories did not follow this pattern. These results were partially consistent with the study by Histed et al. (2009), in which the researchers used an association task to show that a correct trial directly influenced the subsequent trial. One possible explanation for the current study`s results might be that only repeated categories were similar to the association task used by Histed et al., which might be why they replicated their results.

An alternative interpretation of the trial-by-trial learning data is that participants may have been using an error discounting method during learning. Error discounting is when people slow down their rate of learning in order to account for a certain amount of

errors (Craig, Lewandowsky, & Little, 2011). In both of the experiments by Craig et al., participants had to categorize four items into two categories (A and B) on a probabilistic basis (items belonging to category A changed periodically from probabilities of .8 to .6 to .4). An analysis of trial-by-trial response patterns showed that for trials following an incorrect trial, where stimuli were the same, participants were more likely to make a different response, but following incorrect trials where the stimuli were different they were likely to make the same response. Craig and his colleagues interpreted their findings as indicating that when a certain amount of error is unavoidable, people gradually discount errors as they contribute less to learning. In terms of the present study, error discounting may explain why classification of repeated categories were more likely to be correct following a correct trial, but the same finding was not true for diverse categories. When participants made a wrong classification for an item in a repeated category, then the next time it may have been more likely that they changed their response. On the other hand, after making a wrong classification for an item in a diverse category, participants may have been more likely to make the same response. Future studies may look into whether increasing the number of study trials affects the response patterns of diverse categories. If error discounting does explain the different results found for repeated and diverse categories, then future studies should show the same results regardless of the amount of trials people are given.

The goal of the present thesis was to examine how people think they learn, following success versus failure. Participants tended to rate their category knowledge based on whether they had just classified a member correctly, regardless of whether the category type was diverse or repeated. Participants also tended to give higher CLJs for

repeated categories, which suggests that they believed repetition helps category learning more than it actually did. Critically, the overall pattern of changes in CLJs was responsive to changes in classification accuracy.

Research into metacognitive judgments and learning are applicable to students and also education in general. Being aware that variability neglect occurs during learning can make the difference between accurately and inaccurately assessing your understanding of a given topic. Although variability neglect was not replicated in the current study, overconfidence in the benefits of repetition for classifying new items was observed; therefore, study strategies for a large amount of diverse material should attempt to minimize or eliminate the use of repetition. In a similar way, the finding that correct responses led to a greater chance of another correct response occurring can be applied to allocation of study time. In particular, the focus should be on learning from correct responses rather than paying attention to mistakes. While it is important to learn from your mistakes, the implication of the present study is that success is a better indication of how much one has learned.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81,* 126-131. doi: http://dx.doi.org/10.1037/h0027455.

Ashby, G. F., & Maddox, T. W. (2005). Human category learning. *Annual Review of Psychology, 56,* 147-178. doi: 10.1146/annurev.psych.56.091103.070217.

Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior, 8,* 323-343. doi: 10.1016/S0022-5371(69)80124-6.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2,* 331-350. doi: 10.1080/17470218.2014.981553.

Brown, R. (1958). How shall a thing be called? *Psychological review, 65,* 14. doi: http://dx.doi.org.qe2a-proxy.mun.ca/10.1037/h0041727.

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 216-226. doi: http://dx.doi.org/10.1037/0278-7393.30.1.216.

Clapper, J. P. (2015). The impact of training sequence and between-category similarity on unsupervised induction. *The Quarterly Journal of Experimental Psychology, 68,* 1370-1390. doi: http://dx.doi.org/10.1080/17470218.2014.981553.

Close, J., Hahn, U., Hodgetts, C. J., & Pothos, E. M. (2010). Rules and similarity in adult concept learning. In D. Mareschal, C.P. Quinn, S.E.G. Lea, (Eds.), *The making of human concepts* (29-51). Oxford, NY: Oxford University Press Inc.

Craig, S., Lewandowsky, S., & Little, D. R. (2011). Error discounting in probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 673-687. doi: 10.1037/a0022473.

Dukes, W. F., & Bevan, W. (1967). Stimulus variation and repetition in the acquisition of naming responses. *Journal of Experimental Psychology*, *74*, 178-181. doi: http://dx.doi.org/10.1037/h0024575.

Dunlosky, J., & Metcalfe, J. (2009). Chapter 3: Methods and Analyses. In J. Dunlosky & J. Metcalfe, *Metacognition* (pp. 37-59). Thousand Oaks, CA: Sage Publications, Inc.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*, 906-911. doi: http://dx.doi.org/10.1037/0003-066X.34.10.906.

Ganguli, S., Huh, D., & Sompolinsky, H. (2008). Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, *105*, 18970-18975. doi: 10.1073/pnas.0804451105.

Goldman, D., & Homa, D. (1977). Integrative and metric properties of abstracted information as a function of category discriminability, instance variability, and experience. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 375-385. doi: http://dx.doi.org/10.1037/0278-7393.3.4.375.

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1388-1393). Austin, TX: Cognitive Science Society.

Hintzman, D. (1986). “Schema abstraction” in a multiple-trace memory model. *Psychological Review, 93*, 411-428. doi: http://dx.doi.org.qe2a-proxy.mun.ca/10.1037/0033-295X.93.4.411.

Histed, M. H., Pasupathy, A., & Miller, E. K. (2009). Learning substrates in the primate prefrontal cortex and striatum: sustained activity related to successful actions. *Neuron*, *63*, 244-253. doi: 10.1016/j.neuron.2009.06.019.

Homa, D. (1978). Abstraction of ill-defined form. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 407-416. doi: http://dx.doi.org/10.1037/0278-7393.4.5.407.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 418-439. doi: http://dx.doi.org/10.1037/0278-7393.7.6.418.

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 322-330. doi: http://dx.doi.org/10.1037/0278-7393.2.3.322.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1441-1442. doi: http://dx.doi.org/10.1037/a0020636.

Kellogg, R. T., Bourne Jr, L. E., & Ekstrand, B. R. (1978). Feature frequency and the acquisition of natural concepts. *The American Journal of Psychology*, 211-222. doi: 10.2307/1421532.

Kornell, N. & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*, 585-592. doi: 10.1111/j.1467-9280.2008.02127.x.

Luque, D., López, F. J., Marco-Pallares, J., Càmara, E., & Rodríguez-Fornells, A. (2012). Feedback-related brain potential activity complies with basic assumptions of associative learning theory. *Journal of Cognitive Neuroscience*, *24*, 794-808. doi: 10.1162/jocn_a_00145.

Mandler, G., Pearlstone, Z., & Koopmans, H. S. (1969). Effects of organization and semantic similarity on recall and recognition. *Journal of Verbal Learning and Verbal Behavior, 8*, 410-423. doi: 10.1016/S0022-5371(69)80134-9.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238. doi: http://dx.doi.org/10.1037/0033-295X.85.3.207.

Medin, D. L. & Smith, E. E. (1981). Strategies and Classification Learning. *Journal of Experimental Psychology: Human, Learning and Memory, 7*, 241-253. doi: http://dx.doi.org/10.1037/0278-7393.7.4.241.

Medlock, L. S. (2015). Don't fear failure: Nine powerful lessons we can learn from out mistakes. Retrieved from http://www.huffingtonpost.com/lisabeth-saunders-medlock-phd/dont-fear-failure-9-powerful-lessons-we-can-learn-from-our-mistakes_b_6058380.html

Meuwese, J. D. I., van Loon, A. M., Lamme, V. A. F., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: comparing metacognition for object detection and object categorization. *Attention, Perception, & Psychophysics*, *76*, 1057-1068. doi: 10.3758/s13414-014-0643-1.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57. doi: http://dx.doi.org/10.1037/0096-3445.115.1.39.

Pasupathy, A., & Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, *433*, 873-876. doi: 10.1038/nature03287.

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological Science*, *21*, 1894-1902. doi: 10.1177/0956797610389189.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353-363. doi: http://dx.doi.org/10.1037/h0025953.

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605. doi: 10.1016/0010-0285(75)90024-9.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439. doi: 10.1016/0010-0285(76)90013-X.

Seo, H., Barraclough, D. J., & Lee, D. (2007). Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cerebral Cortex, 17,* i110-i117. doi: 10.1093/cercor/bhm064.

Seo, H., & Lee, D. (2009). Behavioral and neural changes after gains and losses of conditioned reinforcers. *The Journal of Neuroscience, 29,* 3627-3641. doi: 10.1523/JNEUROSCI.4726-08.2009.

Smith, L. B. (2003). Learning to recognize objects. *Psychological Science, 14,* 244-250. doi: 10.1111/1467-9280.03439.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81,* 214-241. doi: http://dx.doi.org/10.1037/h0036351.

Tauber, S. K., Dunlosky, J. (2015). Monitoring of learning at the category level when learning a natural concept: Will task experience improve resolution? *Acta Psychologica, 155,* 8-18. doi: 10.1016/j.actpsy.2014.11.011.

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychonomin Bulletin & Review, 20,* 356-363. doi: 10.3758/s13423-012-0319-6.

Thomas, R. C., Finn, B., & Jacoby, L. L. (2015). Prior experience shapes metacognitive judgments at the category level: the role of testing and category difficulty. *Metacognition and Learning,* 1-18. doi: 10.1007/s11409-015-9144-4.

Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2011, November). *A metacognitive illusion in category learning.* Poster presented at the annual meeting of the Psychonomic Society, Seattle, WA.

Van Overschelde, J. P. (2008). Metacognition: Knowing about knowing. In J. Dunlosky, R.A. Bjork (Eds.), *Handbook of metamemory and memory* (47-71). New York, NY: Psychology Press.

Wahlheim, C. N., & DeSoto, K. A. (2016). Study preferences for exemplar variability in self-regulated category learning. *Memory,* 1-13. doi: 10.1080/09658211.2016.1152378.

Wahlheim, C. N., Dunlosky, J., Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39,* 750-763. doi: 10.3758/s13421-010-0063-y.

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: evidence for variability neglect. *Memory & Cognition*, *40,* 703-716. doi: 10.3758/s13421-011-0180-2.

Wang, X., Schiner, T., & Yao, X. (2008). Automatic feature-queried bird identification system based on entropy and fuzzy similarity. *Expert Systems with Applications*, *34,* 2879-2884. doi: 10.1016/j.eswa.2007.05.045.