Summer 8-2016

Utilizing Multiple-level Category Information to Enhance Category Learning: Theoretical and practical considerations in application to authentic natural categories

Toshiya Miyatsu
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychological and Brain Sciences

Thesis Examination Committee:
Mark McDaniel, Chair
Henry Roediger
Ian Dobbins

Utilizing Multiple-level Category Information to Enhance Category Learning:

Theoretical and practical considerations in application to authentic natural categories

by
Toshiya Miyatsu

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Art

August 2016
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am grateful to my adviser Mark McDaniel for chairing the committee and patiently guiding me through this process and to Roddy Roediger and Ian dobbins for generously serving in my committee. I am also thankful to my labmates Khuyen Nguyen, Jihae Lee, Sharda Umanath, and Carolina Kupper-Tetzel, and most of all, my wife, Rose, for their invaluable support without which I would not be writing this acknowledgement.

Toshiya Miyatsu

Washington University in St. Louis

August 2016

ABSTRACT

Utilizing Multiple-level Category Information to Enhance Category Learning:

Theoretical and practical considerations in application to authentic natural categories

by

Toshiya Miyatsu

Master of Art in Psychological and Brain Sciences

Washington University in St. Louis, 2016

Professor Mark McDaniel, Chair

Category learning is ubiquitous in science education. From botany courses teaching to identify different types of plants to geology courses teaching to distinguish various types of rocks, examples of category learning being a core of a course curriculum can be easily found. Using the educationally authentic rock categories, the current project examined whether category learning at a broad level (i.e., distinguishing between Igneous, Sedimentary, and Metamorphic rocks) could be enhanced by using category information at a more specific level (e.g., Harzburgite under Igneous, Breccia under Sedimentary, etc.). Both in Experiment 1 (broad- and specific-level information were presented simultaneously) and Experiment 2 (specific-level categories were learned separately from specific- and broad- level category names association), direct training at the broad level was superior to the protocol combining the specific-level category information.

Introduction

Learning of naturally occurring categories can be found in the core of many science disciplines. For example, ornithologists can rapidly distinguish between different bird species, mycologists are characterized by their ability to identify kinds of mushrooms, and geological scientists can categorize various types of rocks. Despite this prevalence in science education, studies aimed at identifying elements that can be utilized to enhance instruction aimed at teaching natural categories have been scarce. These few published studies have identified the beneficial effect of interleaving exemplars from multiple categories rather than blocking exemplars from one category (Kornell & Bjork, 2008; also see Kang & Pashler, 2012), test-enhanced learning of natural categories (Jacoby, Wahlheim, & Coane, 2010), and when using training exemplars with exaggerated diagnostic features (i.e., fading) is effective (Pashler & Mozer, 2013).

However, the effects of any identified variable depends on the characteristics of the categories to be learned (e.g., Carvalho & Goldstone, 2014) and a certain characteristic of natural categories makes them extremely challenging to learn. Namely, natural categories have a high variability among the exemplars of categories (Murphy, 2002). That is, although there are many instances that appear prototypical of a given category, there are also many instances that deviate from such prototypical representation yet still belong to the same category. In addition, this high variability often creates fuzzy boundaries between categories. In some cases, two exemplars from different categories are perceptually more similar than two exemplars from a same category. Therefore successful learning of natural categories requires learners to not only identify prototypical instances but also anticipate the allowable variation in each category to accommodate exceptional instances.

My approach thus takes this challenging natural category characteristic into consideration and appeals to another one of the characteristics of natural category taxonomy, multi-level categorization. Virtually all natural categories have a taxonomic organization such that they are organized hierarchically at multiple levels. For example, the national bird of the United States can be identified as an accipitriformes (i.e., birds of prey) at the order level, an accipitridae (i.e., birds with strongly hooked bills) at the family level, or a bald eagle at the species level.

The key idea that I will examine throughout the current study is that category learning at a target level can be enhanced by training learners at a subordinate, more specific level. This is because learning categories at a more specific level orients the learners towards identifying the characteristic features. In perceptual category learning, all exemplars differ in a number of features. In the case of the bird categories, those features include their body shape, size, color, and the shape of their bills. Exemplars belonging to the same category would show a very small variation in some of those features (i.e., characteristic features) while other features vary widely. Importantly, in most cases exemplars belonging to the same specific-level category will share more features among each other compared to exemplars belonging to the same broad-level category but coming from different specific-level categories. That is any given specific-level category has a lower feature variability among its exemplars than any given broad-level category that includes multiple specific-level categories. For example, exemplars of the specific category of bald eagle have very similar shape, color, and bills, whereas exemplars of the more broad accipitridae that includes various species of eagles and hawks will vary more widely so that the identification of the shape of their bill (i.e., the diagnostic feature) is more difficult. Then, learners will be able to more easily pick up what features are shared among the exemplars belonging to the same category when presented with exemplars from the same specific-level category, whereas they would have a harder time doing so if presented with exemplars from the same broad-level category with its exemplars' features varying more widely. If the learning of specific-level categories is easier than the learning of broad-level categories, then learning broad-level categories as a collection of more coherent specific-level categories would produce better learning than directly learning less coherent broad-level categories. Figure 1 is a pictorial illustration of this specific-level training advantage using the categories of rocks that I will use in the current study. The top of the figure shows 12 exemplars from four specific-level categories in the broad category of igneous rocks as a broad-level training may show. The high variability of the exemplars presented as members of the same category makes it very difficult to see what the characteristics of this category is. In the bottom of the Figure 1, the same exemplars are shown by their specific-level category in each row as learners trained at the specific-level category may organize the categories. In this case, the characteristics of each of the specific-level categories are relatively clear so that learning this broad-level category as a collection of these coherent specific-level categories would be easier than the case of broad-level training as shown in the top of the figure.

1.1 Abstraction-based versus Exemplar-based approach

A long-standing debate in the category learning literature is on learners' internal representation of categories, whether it's *abstraction-based* or *exemplar-based*. According to an abstraction-based approach, learners develop some sort of a summary representation of the learned categories. For example, learners may develop a set of rules that determines categorization (*rule-based*) or prototypes of the learned categories (*prototype-based*), and the rules or the prototypes are applied to classify exemplars that they subsequently encounter (e.g., Allen & Brooks, 1991; Erickson & Kruschke, 1998; Keele, 1968; Keele, 1970; Regehr & Brooks, 1993; Trabasso &

Bower, 1968). On the other hand, an exemplar-based approach assumes that learners store exemplars from each category and the subsequent categorization judgement is made according to the similarity of the newly encountered exemplars and the stored exemplars (e.g., Brooks, 1978; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986; Nosofsky & Kruschke, 1992). In laboratory studies of category learning using artificial categories (e.g., a combination of dots, lines, and circles), the abstraction-based approaches have better explained data obtained using high-structure categories with logical rules (e.g., Ashby & Ell, 2001; Smith &Minda, 1998) while the exemplar-based approaches have been generally favored in low-structure categories without clear-cut rules (e.g., Bourne, 1974; Little, Nosofsky, & Denton, 2011; Nosofsky, Palmeri, & McKinley, 1994).

The literature reviewed above suggests that the category structure to be learned may determine learners' approach. However, the investigation on this issue has not been done with realistic natural category structures and learning situations. Here I draw on evidence from some key laboratory studies in which abstraction-based models described the learning patterns better than exemplar-based models. Further, I review a limited number of studies that showed that the structure of the categories to be learned influence the learning approaches. Taking these studies as evidence, I argue that learners tend to take an abstraction-based approach in learning natural perceptual categories with many unique training exemplars.

In a classic paper by Posner and Keele (1968), their participants learned a series of dot patterns created by distorting several original dot patterns (i.e., prototypes). As the participants learned exemplars with varying degree of distortion, their ability to identify the prototypes as a member of the category increased although the prototypes were never actually presented. Importantly, the prototypes were better identified as the members of a category than other novel items that were more similar to some of the presented members of the category. They interpreted this result as learners abstracting a prototypical representation as they go through exemplars and learn the category. This was one of the first formal variations of the abstraction-based approach. In their follow up study, Posner and Keele (1970) also showed that whereas the memory for specific items fades quickly with time, prototypical representation endures delay. In real life natural category learning situations, the time interval between the presentations as well as between the presentation and test are likely to be longer than typical laboratory studies of category learning outlined above. Therefore, learners will likely rely more on abstraction-based approaches in these situations.

Another factor that influences learners' category learning approaches is the set size. Generally, as the set size increases, so does learners' tendency to take abstraction-based approaches. For example, Little and McDaniel (under review) had their participants learn categories of geometric shapes that could be well accommodated by both rule- and exemplar- based approaches. Their participants showed stronger tendency towards rule-based approach when they learned the categories with 24 training exemplars compared to eight training exemplars, as indicated by their performance on ambiguous test items that could be categorized either according to an abstracted rule or the similarity to the presented exemplars. In addition, Knapp and Anderson (1984) reported an experiment using categories of dot patterns that showed abstraction as indicated by the correct identification of prototypes that were not presented was more prominent when the category was learned through 24 training exemplars compared to six and one training exemplars. Of course, in real life natural category learning, the number of exemplars learners go through per category is likely much greater than 24.

Lastly, an influential model created to explain complex patterns of results obtained in ill-defined categories with fuzzy boundaries suggests that rule-based abstraction processing persists even when learners are presented with category structures with clear exceptions to their plausible defining rules (RULEX; Nosofsky, Palmeri, & McKinley, 1994). The basic assumptions of the model are that learners first look for rules of varying simplicity and certainty, such as a perfect rule according to one dimension, a fairly well-defining rule according to one dimension, and a fairly well-defined rule according to multiple dimensions, and if all fails, they rely on their memory on exception items of the learned categories. This model has been able to account for a wide variety of established phenomena in category learning, such as prototype and specific exemplar effects, sensitivity to correlational information, difficulty of learning linearly separable versus nonlinearly separable categories, selective attention effects, and difficulty of learning concepts with rules of differing complexity (Nosofsky et al., 1994).

Of course processing based on both abstraction- and exemplar- based approaches is possible and the individual differences on this learning tendency may be pervasive, such that some learners almost always take one approach versus the other (Little & McDaniel, 2015; Little & McDaniel, under review). However, I suggest that learners are more likely to take an abstraction-based approach, whether rule- or prototype- based, in authentic natural category learning situations as well as in my task. Therefore, when I discuss the current experiments, I will describe the learners' processing from an abstraction-based perspective as a plausible possibility for brevity. Importantly, it should also be noted that exemplar-based approaches also make the same prediction regarding my hypothesis that specific-level training enhances broad-level categorization (Nosofsky, Sanders, Gerdom, Douglas, and McDaniel, under review), as discussed in details later in this section.

1.2 Review of recent literature on specific- and broad- level training

To my knowledge, the first study directly comparing specific- and broad- level training and their effectiveness on later categorization using natural categories was published by Tanaka, Curran, and Sheinberg (2005). Their participants were shown pictures of owls or wading birds with their broad-level category names (i.e., “owl” or “wading bird”) and the other with their specific-level category names of species (e.g., “great blue crown heron” or “eastern screech owl”). The participants’ category learning was later assessed by a sequential matching task (Gauthier, Curran, Curby, & Collins, 2003) in which they judged whether a pair of pictures of birds that were sequentially presented belonged to the same species. Their results indicated that specific-level training enabled the participants to better transfer to novel exemplars from the learned specific-level categories as well as to exemplars from novel specific-level categories. Their study is important in demonstrating that learners’ category learning is affected by the perceptual categorization experience, as manipulated by the labeling of the exemplars, not just the perceptual exposure per se. However, the form of assessment used that is more suited in studying perceptual expertise limits the generalization of their findings to a broader category learning context. Specifically, in real life category induction situations, one would see an exemplar and try to identify which category the observed exemplar belongs to, not just referring to the previous item and judge same or different.

Noh, Yan, Vendetti, Castel, and Bjork (2014) recently published a study examining the identification of a natural category at multiple levels. In their experiments, participants were presented with a series of pictures of snakes with their genus (e.g., Throp, Arix) and broader category information (i.e., either venomous/non-venomous or tropical/non-tropical depending on the condition) simultaneously and were instructed to focus on either the genus or broader category information. Later the participants were presented with a new set of pictures of snakes and asked to classify some of them according to their genus while being asked to classify some others according to their broader category information. Their main finding was the interaction between the level of categorization participants were instructed to focus on and the perceived value of categorization. When the broader categories were presented as tropical or non-tropical (i.e., low value), participants were able to better learn the classification at the level they were instructed to attend to (i.e., genus or broader category). In contrast, when the broader categories were presented as venomous or non-venomous (i.e., high value), the participants focused on the broader category information regardless of the instruction. Thus the more valuable categorization "hijacked" the learning of categorization at another level.

Noh et al.'s (2014) additional findings on their instruction to focus on the more specific genus level or the broad level goes contrary to my idea of broad-level categorization being enhanced by specific-level training. Specifically, participants' broad-level classification performance was better overall when they were told to focus on the broad level than when they were told to focus on the specific level. In other words, focusing on the specific level did not help learn the broad categories. I speculate that the learning of specific-level categories didn't help the learning of broad-level categories in the snake category because the venomous and non-venomous distinction can be made through a set of relatively clear cut rules, which is rather rare in natural categories. Specifically, while venomous snakes have arrow-shaped heads, slit pupils, and thicker and shorter bodies with patchy patterns, non-venomous snakes have spoon-shaped heads, round pupils, and longer bodies with more defined patterns (see Noh et al., 2014, for examples). However, rules underlying many natural categories with high variability are usually more complex and multi-dimensional. Further, these natural categories often have fuzzy boundaries, such that some exemplars from another category look more similar to the prototype of a particular category than some atypical exemplars that actually belong to that particular category (Murphy, 2002). In some cases, there may not be a single verbalizable rule that can encompass all the instances, and the specific-level training might be particularly effective in these cases because the high variability makes it especially difficult to identify the common characteristics among the exemplars belonging to the same broad-level category.

Nosofsky et al. (under review) laid out a theoretical justification for why the training at a more specific level can produce superior learning than the training at a broader level from a perspective of a highly influential variant of an exemplar-based perspective, the *generalized context model* (GCM; Medin & Schaffer, 1978; Nosofsky, 1984, 1986). GCM, like other exemplar-based models, assumes that learners classify exemplars according to their similarity to all the exemplars of different categories that are stored in their memory. However, unlike many other models, GCM takes into consideration that similarity is context-dependent, such that similarity in the dimension that learners attend to is weighed more, and learners generally learn to focus their attention on highly diagnostic dimensions and ignore less relevant ones.

Using the GCM, Nosofsky et al. (under review) derived specific hypotheses on when the specific-level training would be effective depending on the category structure. In the case of specific-level training, the similarity signal would be very sharp when learners are presented with an item at test because the stored exemplars to be compared are restricted to one specific-level category that presented test item belonged to. On the other hand, in the case of broad-level training, the similarity signal would be blurry because the stored exemplars to be compared would include several specific-level categories. This advantage of specific-level training should be especially pronounced when the variability within the broad-level categories is high because the similarity signal at test would be even blurrier in such cases. However, when the within broad-level variability is low, a few diagnostic dimensions that can separate broad-level categories may be available. In such cases, a broad-level training would allow learners to attend to the diagnostic dimensions, an opportunity that a specific-level training would not allow.

Nosofsky et al. tested this hypothesis using the categories of rocks. The three main rock categories (i.e., the broad level) are *igneous*, *sedimentary*, and *metamorphic*. However, each of these broad categories have several more specific categories such as anorthosite, basalt, and carbonatite under igneous. First, they compiled a set of rock pictures that made up either compact (i.e., low variability within a broader category) or dispersed (i.e., high variability within a broader category) category structure organized into three broad level categories and three specific level categories each. Then their participants were trained with twenty seven items either at the broader level by learning the pictures of rocks paired with their broad-level labeling or trained at the specific level by learning the pictures of rocks paired with their broad-level labeling and specific-level number (e.g., S1, S2, and S3 for the specific-level categories under sedimentary). After going through six rounds of feedback learning in which the participants were asked to identify which category these training items belonged to followed by corrective feedback on which category they actually belonged to, the participants were tested on the training items as well as the transfer items, the new items from the trained category. The results confirmed the hypothesized interaction such that the broader-level training produced better learning in the compact condition while the specific-level condition produced better performance in the dispersed condition in which the within broader level category variability was set to be high.

Although the findings from Nosofsky et al. (under review) suggests that multi-level category information can be successfully utilized to enhance the category learning at a broader target level, the applicability of their findings to more authentic educational situations is still questionable. This is primarily because they used a category structure with only a small number of selected exemplars to exaggerate the within category variability in both directions rather than testing their hypothesis using an existing category structure. Specifically, exemplars in their compact condition were intentionally compiled to have low variability whereas the exemplars in their dispersed condition were complied with the opposite intention. Therefore, in the current study, I will examine this issue with a more comprehensive category structure representative of what exists in nature using the rock categories.

Experiment 1

One straightforward way to combine the category learning at the broad and specific level is to present the category information at two levels simultaneously as Noh et al. (2014) did. In my first experiment, a group of participants who studied the pictures of rocks accompanied by their broad-level category names (BL Only condition) was compared to a group in which the pictures were presented with both broad- and specific- level category names (BL + SL condition).

Participants' performance was assessed through three types of test items: memory, near-transfer, and far-transfer. Memory items were the ones that participants studied during the study phase, near-transfer items were new items that belonged to one of the specific-level categories that participants studied, and far-transfer items were new items from a new specific-level category. The idea of the far-transfer items is that although participants never studied the particular specific-level categories, if they were able to extract the general characteristics of a given broad-level category, they should be able to identify which broad-level category a new item belongs to (Wahlheim, Finn, & Jacoby, 2012).

The predictions on each test item type in terms of the study conditions are as follows. If the specific-level category learning makes the characteristic features more salient, then participants in the BL + SL condition should process the presented exemplars focusing on those features. Such focus aimed at identifying the rule or developing a prototype for each specific-level category may cost the processing of given individual exemplars resulting in reduced memory for training exemplars (Smith & Minda, 1998). Therefore, I predict the BL + SL group to perform worse on the memory test items compared to the BL Only group. Conversely, the rule-abstraction processing that is more prominent in the BL + SL group should make the participants in this condition better able to identify the near-transfer test exemplars. Therefore I predict the BL + SL group to perform better on the near-transfer test items compared to the BL Only group. The hypothesis about the enhanced diagnostic feature identification that I laid out does not make a strong prediction regarding the far-transfer test items. Nevertheless, I predict that identification of different sets of diagnostic features for each specific-level category in the BL + SL condition will make the participants in this condition better able to anticipate the allowable variation. Therefore, I predict that the BL + SL condition to perform better in the far-transfer test items.

Alternatively, the simultaneous presentation of the broad- and the specific- level category information in the BL + SL condition might exceed the processing capacity of learners (i.e., *cognitive overload*, Sweller, 1988) because this condition forces the participants to keep track of the variations of features associated with both broad- and specific- level categorization at the same time. This is similar to several reported situations of cognitive overload hindering learning. For example, students struggle to grasp biological concepts in the classroom when they have to simultaneously learn new terms that are used to describe those concepts (McDonnell, Barker, & Wieman, in press). Similarly, learning the rules of particular problem solving as learners solve mathematical problems reduces their performance (Cooper & Sweller, 1987), and simultaneously presenting information in texts and a diagram that are both processed visually compromises the effectiveness of multimedia learning of science lessons (Mayer & Moreno, 2003). In sum, tasks that force learners to simultaneously engage in processes that utilize the same cognitive resource often compromise performance. Thus we may see a global impairment in the BL + SL condition resulting in the BL Only condition to perform better in all three test item types.

Another well-established theory that predicts that the BL Only condition will outperform the BL + SL condition is transfer-appropriate processing (e.g., Blaxton, 1989; Lockhart, 2002). This theory states that the task performance is optimized when the processing required by the test matches the processing required at encoding. Because the participants are asked to classify the test items into the three broad-level categories at the final test, the test of the current task requires the processing at the broad-level categorization. In the BL Only condition, participants learn the categories by processing each training exemplar strictly at the broad-level categorization. However, the BL + SL condition requires the participants to simultaneously process the training exemplars at the broad- and specific- level categorization. Therefore, the transfer-appropriate processing perspective also predicts that in the BL Only condition will perform better in all three test item types.

2.1 Method

2.1.1 Participants

Participants were recruited through Amazon Mechanical Turk and were compensated $3 for their time. Seventy six participants logged into the experiment and 60 of them completed it. Out of the

60 participants who completed the entire experiment, two of them were excluded for indicating in the post-experimental questionnaire that they had known more than 75 percent of the presented information on the rock categories previous to their experimental participation. Fifty eight participants (25 in BL Only and 33 in BL + SL condition: the differing numbers of participants arising through random incompletions) were included in the analysis.

2.1.2 Materials

Materials were pictures of rocks collected through a web search organized into three broad-level categories, eight specific-level categories under each broad-level category, and 11 exemplars under each specific-level category (see Table 1 for the list of specific-level categories and Figure 2 for examples of rock pictures), 264 exemplars in total. Of the entire set of 264 exemplars, 168 of them were used for any given participant. 96 of them (three broad-level categories, four specific-level categories each, eight exemplars each) served as the training exemplars. The remaining three exemplars in each of the studied specific-level categories served as the near-transfer test items, and three exemplars from each of the remaining four specific-level categories in each broad-level category served as the far-transfer test items. The assignment of exemplars in the studied specific-level categories to the training exemplars or the near-transfer items was randomized so that all exemplars had equal chance of being the training exemplars and the near-transfer items. Three memory items were randomly chosen from the training exemplars. The assignment of each lower-order category was counterbalanced so that all specific-level categories equally often became studied and non-studied (far-transfer) categories during the study phase. The data were collected using Collector (http://github.com/gikeymarcia/Collector), a PHP-based open source experiment program.

2.1.3 Procedure

The experiment consisted of a study phase and a test phase. During the study phase, all participants were presented with 96 pictures of rocks one at a time for seven seconds in a randomized order. The pictures of rocks were presented with their broad-level category names in the BL Only condition, whereas the pictures of rocks were presented with both their broad- and specific- level category names simultaneously in the BL + SL condition (see Figure 3 for examples of the training trials). The participants in the BL + SL condition were told that there were several specific-level categories within the three broad-level categories and learning the specific-level categories may help them learn the broad-level categories. Upon finishing the study phase, participants played tetris for three minutes as a distractor task before starting the test phase.

During the test phase, all participants were presented with 108 pictures of rocks consisting of the three test item types discussed above. There were 36 items (three broad-level categories, four specific-level categories each, three items each) of each of the memory, near-transfer, and far-transfer test types. Participants were presented with these items one by one for 10 seconds in a random order and asked to indicate which broad-level category the given item belonged to by clicking on one of four choices: Igneous, Sedimentary, Metamorphic, or ”I don’t know” condition (see Figure 3 for examples of the test trials). After the test, participants completed a post-experimental questionnaire that measured their prior knowledge on the rock categories and the presence of logistical problems, such as internet connection difficulty. The entire experiment took about 45 minutes to complete.

2.2 Results

A 2 X 3 mixed repeated measures analysis of variance (ANOVA) with the study condition (BL Only or BL + SL) treated as the between-subjects variable and the test item type (memory, near-transfer, or far-transfer) treated as the within-subjects variable was conducted to assess the main effects of condition and test item types as well as their interaction (see Figure 4). There was a significant main effect of condition such that the BL Only group (M = .598, SD = .12) outperformed the BL + SL group (M = .495, SD = .15), $F(1, 56) = 11.98$, $p < .01$, $\eta p^2 = .18$. There was also a significant main effect of test item types, $F(2, 112) = 37.27$, $p < .001$, $\eta p^2 = .40$. Post-hoc t-tests showed that while there was no significant difference between memory (M = .597, SD = .14) and near-transfer (M = .581, SD = .14) test items, $t(57) = 1.44$, $p > .05$, $d = 0.11$, participants performed worse on far-transfer items (M = .484, SD = .13) compared to both memory, $t(57) = -7.48$, $p < .001$, $d = -0.82$, and near-transfer items, $t(57) = -6.28$, $p < .001$, $d = -0.73$. The interaction between the two independent variables was not significant, $F(2, 112) = 0.020$, $p > .05$, $\eta p^2 = .00$. In addition, a one-sample t-test was conducted to see if the participants' performance on the novel far-transfer test items were significantly different from chance, and it was, $t(57) = 8.99$, $p < .001$ (chance = .33).

2.3 Discussion

Participants in the BL + SL condition showed inferior performance on all three measures: memory, near-transfer, and far-transfer test items. This global impairment confirms the cognitive overload hypothesis, the transfer-appropriate processing hypothesis, or the combination of them. The cognitive overload hypothesis explains the results as the demand to process complex information in the BL + SL condition exceeding the cognitive capacity of the participants. More specifically, while the participants in the BL Only condition only had to learn the categorization at the broad level, the participants in the BL + SL condition had to process categorization at the broad and the specific level at the same time. That is, speaking from an abstraction-based perspective, the participants in the BL + SL condition needed to keep track of change in various features associated with not only three broad-level categories but also twelve specific-level categories (four in each broad-level category) in search of defining rules (i.e., diagnostic features) or prototypes for all twelve category labels presented. Similarly, the transfer-appropriate processing hypothesis explains the results as the performance of the participants in the BL + SL condition being compromised by the mismatch in the processing at their encoding and the test.

In addition, it is also possible that the variability in the rock category is distributed differently from how I initially thought. Our hypothesis regarding the high variability and the better diagnostic feature identification in the specific-level category training assumes that the variability is high between the specific-level categories (e.g., the dispersed condition in Nosofsky et al.'s study). In other words, I assumed that specific-level categories belonging to the same broad-level category look very different. However, it might be the case that rock categories have high variability within the specific-level categories but their between-specific-level-category variability is not as high as I expected. That is, exemplars in each specific-level category look very different from each other and that may create fuzzy boundaries and make the categorization challenging, but the specific-level categories belonging to the same broad-level category as a whole look relatively similar to each other. I will discuss more about the variability distribution in the rock categories in the general discussion.

It should also be noted that the participants' performance on the far-transfer test (48%) items were significantly different from chance (33%) indicating that participants were able to extend their learning to specific-level categories that they did not encounter during the learning phase. This finding replicates the past study that used similar test item type (Wahlheim, Finn, & Jacoby, 2012) and shows that broad-level training can be extended to novel specific-level categories. Further, it shows that specific-level training of broad-level categories can also be extended to novel specific-level categories.

Experiment 2

The complex task in the BL + SL condition can be broken down to two components, the learning of categorization at the specific level and the learning of the association between given broad- and specific- level category names. Given the results from Experiment 1, Experiment 2 attempted to reduce the cognitive load of the learners by separating the category learning component from the associative learning (AL) component. The BL Only condition (identical to Experiment 1) was compared to SL -> AL condition in which participants learned specific-level categorization by observing the rock pictures, and then learned the association between specific- and broad- level category names[1]. In the separate associative learning phase, participants were first presented with all the twelve pairs of broad- and specific- category names and then went through three rounds of test-plus-feedback cycle to ensure a robust associative learning.

If the global impairment observed in the BL + SL condition in Experiment 1 was primarily because of the cognitive overload, separating the two components of the task should eliminate the deficit and make the benefit of specific-level training apparent. In the previously mentioned studies examining the hindering effect of cognitive overload on learning, dividing up the task and freeing learners' cognitive resource from a competition from multiple demands alleviated the negative effect (Cooper & Sweller, 1987; Mayer & Moreno, 2003; McDonnell et al., in press). In addition, the mismatch in the processing at encoding and test as suggest by the transfer-appropriate processing theory should be reduced or eliminated in Experiment 2 because the participants in the SL -> AL condition are processing the categorization at the final test at the specific level, at the level they initially encoded, before converting the answer to the corresponding broad-level category name. Therefore, I predict the SL -> AL condition should outperform the BL Only condition.

On the other hand, it is also possible that the global impairment in the BL + SL condition in Experiment 1 was caused by the rock category structure having a different variability distribution than initially suspected. Namely, whereas I assumed the rock categories to have a high between-specific-level-category variability, they might actually have a relatively low between-specific-level-category variability and a high within-specific-level-category variability. If this is the case, the BL Only condition should outperform the SL -> AL condition because the category structure stayed the same between Experiment 1 and 2.

3.1 Method

3.1.1 Participants

Fifty two undergraduates (26 in each condition) at Washington University in St. Louis participated in this experiment for a course credit or $10. The post experimental questionnaire indicated that none of the participants had known more than 50 percent of the presented information on the rock categories previous to their experimental participation.

3.1.2 Materials

The same materials were used as Experiment 1.

3.1.3 Procedure

Similarly to Experiment 1, Experiment 2 consisted of a study phase and a test phase. During the study phase, participants in the SL -> AL condition were first presented with 96 pictures of rocks (same breakdown as experiment 1) accompanied by their specific-level category names. The pictures were presented one by one for seven seconds in a randomized order. Upon finishing studying the pictures of rocks, participants in the SL -> AL condition engaged in the associative learning. First they were presented with twelve broad- and specific- level category names pairs (all four names of the studied specific-level categories under each broad-level category; e.g., Harzburgite-Igneous, Breccia-Sedimentary) one at a time for five seconds in a randomized order. Then they went through three rounds of test-plus-feedback cycles going over the 12 pairs three times in a randomized order. During the first round of the test-plus-feedback cycle, participants were presented with specific-level category name one at a time and asked to identify which broad-level category they belonged to by clicking on one of four choices: Igneous, Sedimentary, Metamorphic, or "I don't know". Participants were given five seconds to answer during the first round, and four seconds to answer during the second and third round. During all three rounds, correct feedback was provided immediately after each trial for two seconds. Participants in the BL Only condition studied the same 96 pictures of rocks but with only their broad-level category names, one at a time for 10 seconds[2]. The total study time for the two groups was exactly the same (960 seconds in both condition).

Upon finishing the study phase, participants played tetris as a distractor task for 3 minutes before starting the test phase which was identical to experiment 1.

3.2 Results

3.2.1 Performance during the associative learning

The mean performance of the participants in the SL -> AL condition during the associative learning was .69 (*SD* = .14) and the mean for the last round was .76 (*SD* = .15). Although participants' performance in the last round was not perfect, they seem to have learned the association reasonably well especially considering the facilitative effect of the feedback after the last round.

3.2.2 Final test performance

Similarly to Experiment 1, a 2 X 3 mixed repeated measures ANOVA was conducted to assess the main effects of condition and test item types as well as their interaction (see Figure 5). There was a significant main effect of condition such that BL Only group (M = .629, SD = .13) outperformed SL -> AL group (M = .525, SD = .15), $F(1, 50) = 11.55$, $p < .01$, $\eta p^2 = .19$. There was also a significant main effect of test item types, $F(2, 100) = 41.65$, $p < .001$,$\eta p^2 = .45$. Post-hoc t-tests showed that while there was no significant difference between memory (M = .629, SD = .15) and near-transfer test items (M = .604, SD = .14), $t(51) = 1.88$, $p > .05$, $d = 0.17$, participants performed worse on far-transfer items (M = .498, SD = .12) compared to memory, $t(51) = -7.71$, $p < .001$, $d = -0.96$, and near-transfer items, $t(51) = -6.97$, $p < .001$, $d = -0.80$. However, the interaction between the two independent variables were not significant, $F(2, 100) = 1.270$, $p > .05$, $\eta p^2 = .03$.

3.2.3 Conditional Analysis

The identification at the final test in the SL -> AL group can fail either at the specific-level categorization or at the retrieval of the association between the specific- and broad- level category names. For instance, one might correctly identify a picture of a rock as anorthosite but fail to remember that anorthosite belongs to Igneous. In order to compare the performance difference on the category learning portion between the study conditions more directly, I conducted a conditional analysis only looking at the responses in the SL -> AL condition at the final test for the specific-level categories that were correctly associated with their corresponding broad-level categories at the last round of the associative learning phase. This analysis can give a better sense of how well participants in the SL + AL condition did in the category learning portion of their task, which may be a more fair comparison to the participants in the BL Only condition whose solo task was the category learning at the broad level.

A two by two mixed ANOVA with the study condition (BL Only or SL -> AL) treated as the between-subjects variable and the test item type (memory or near-transfer) treated as the within-subjects variable was conducted (the far-transfer items were excluded from this analysis because they come from the specific-level categories that were not learned during the associative learning phase) to assess the main effects of condition and test item types as well as their interaction. The main effect of condition was not significant such that the performance of BL Only group (M = .666, SD = .16) was not significantly different from that of the SL -> AL group (M = .607, SD = .16), $F(1, 50) = 2.56$, $p > .05$, $\eta p^2 = .049$. There was a marginally significant main effect of test item types, such that participants performed better on memory (M = .650, SD = .15) then near-transfer (M = .623, SD = .14), $F(1, 50) = 3.49$, $p < .01$, $\eta p^2 = .065$. The interaction between the two independent variables was not significant, $F(1, 50) = 1.270$, $p > .05$, $\eta p^2 = .040$.

3.3 Discussion

The separate training on the specific-level category learning and the associative learning in the SL -> AL condition produced inferior performance compared to the direct training on the broad-level category learning in the BL Only condition in all three measures. However, because the performance on the associative learning was not perfect, a conditional analysis only including the specific-level categories that had been successfully associated with their broad-level category was conducted. This analysis was aimed at more directly comparing the specific-level categorization in the SL + AL condition and the broad-level categorization in the BL Only condition independent from errors stemming from the associative reasoning, and it revealed that the categorization performances in these conditions were actually not significantly different. One may argue that this comparison was underpowered (N = 52) and the six percent point difference (67% in BL Only vs 61 % in SL + AL, $\eta p^2 = .049$) would be reliable if assessed with an increased number of participants. A post-hoc power analysis using the obtained effect size ($\eta p^2 = .049$) revealed that observed power in detecting a significant main effect of the condition was .35.

It is possible that the similarity in the way information was processed at the encoding and the test as suggested by the transfer-appropriate processing still contributed to the performance difference. Specifically, although the separation of the specific-level category learning and the associative learning reduced the degree of mismatch in the SL -> AL condition in Experiment 2 compared to the BL + SL condition in Experiment 1, the training and the test tasks were more similar in the BL Only group both in their appearance and the processing of information (see Figure 3). Therefore, the participants in this condition might have been benefited from transfer-appropriate processing relative to their counterparts in the SL -> AL condition.

Perhaps more crucially, another possibility of cognitive overload posed by the increased complexity of the task in the SL -> AL condition needs to be considered. Unlike the participants in the BL + SL condition in Experiment 1, the participants in the SL -> AL condition in Experiment 2 did not have to process categorization at multiple levels. However, they still had to learn higher number of categories. Specifically, whereas the participants in the BL Only condition needed to learn three categories, the participants in the SL -> AL condition needed to learn twelve categories (i.e., four specific-level categories under each broad-level category). This simple increase in the number of categories to keep track of certainly increased the cognitive load. Implications of this factor for category learning protocols attempting to utilize the specific-level training will be discussed further in the general discussion.

General Discussion

The current experiments used authentic natural category of rocks and examined the hypothesis that in learning of categories with a high variability that are organized at multiple levels, the training at a more specific level promotes better category learning than the direct training at a broad level. In Experiment 1, simultaneously presenting the specific- and broad- level category information produced inferior performance compared to presenting only the broad-level category information. Considering the cognitive overload hypothesis, in Experiment 2, the specific-level category learning and the associative learning components of the task were learned separately. Although this separate training procedure still produced inferior performance relative to direct training at the broad level, a closer examination of the category learning portion through a conditional analysis suggested that the specific-level categorization might have been learned to a similar degree to the broad-level categorization.

Using natural category exemplars with an artificial structure, Nosofsky, Sanders, Gerdom, Douglas, & McDaniel (under review) demonstrated that the training at the specific level produced better learning than the training at the broad level when the variability was high (i.e., dispersed). However, in my experiments using the same natural category but with an increased number of exemplars and a category structure that is more representative of what exists in nature, the training at the specific level was outperformed by the training at the broad level. In the following, I will discuss several differences in experimental design between my and Nosofsky et al.'s experiments that likely contributed to the discrepancy of the results, assess the current state of the research investigating specific-level training of category learning, and make recommendations for future investigations.

The first factor to be considered is the cognitive overload hypothesis. I interpreted the relative impairment in the BL + SL condition in Experiment 1 to be caused by the simultaneous presentation of the broad- and the specific- level category information that needed to utilize the same cognitive resources (e.g., working memory). Therefore, in Experiment 2, I attempted to reduce the cognitive load by separating the specific-level category learning and the associative learning in the SL -> AL condition. However, as discussed previously, the participants in the SL -> AL condition in Experiment 2 still certainly had a higher cognitive load because of the higher number of the categories to be learned (i.e., 12 versus three). A highly cited study in working memory (called short term memory back then) suggests that people can keep five to nine things in their working memory (Miller, 1956). Category learning involves not only working memory but also other multiple memory systems, such as explicit long-term memory and implicit memory (Smith & Grossman, 2008), so whether learning 12 categories exceeded most of our participants' cognitive capacity is difficult to determine. However it is clear that an increased number of categories to be learned pose a challenge to learners' processing capacity.

This factor has implications to any category learning protocol that attempts to utilize specific-level training because specific-level training by definition requires learning an increased number of categories compared to broad-level training. From an instruction perspective, one needs to be careful that the increased number of specific-level categories to be learned doesn't overwhelm learners' cognitive processing. A potential remedy is to learn a selected number of specific-level categories at the beginning, so that learners could take an advantage of specific-level training without being overwhelmed by a high number of categories. The selection of the specific-level categories in this case also demands careful consideration. For example, including only more prototypical specific-level categories might slow down the acquisition of more atypical specific-level categories later.

Relatedly, the associative learning of broad- and specific- level category names is another issue that consistently increases the cognitive load and makes the task more complex when implementing specific-level training. Nosofsky et. al. (under review) bypassed the issue of additional associative learning in their specific-level training condition by having their participants learn the specific-level categories as their broad-level category initial plus a number. For example, conglomerate, coquina, and breccia under sedimentary were presented as S1, S2, and S3. Although this setting helped Nosofsky et al. isolate the effect of their training manipulation on category learning, it also made their task less authentic from an educational perspective. One way to bypass this issue in realistic situations is to offload the associative learning component of the task to an external device. For example, an information sheet that shows the association between the specific- and broad- level categories can be provided after the learning of specific-level categorization. Offloading memory tasks is a long-standing practice that is both common and effective in enhancing cognitive performance (Nestojko, Finley, & Roediger, 2013). Although giving such a device may not seem authentic, it may actually fit well with how the field geologists are trained. Because learning a dozen or so paired associates is far easier than learning the same number of categories, when a field geologist in training is perfecting her ability to classify the type of rocks, it's likely that she already knows which specific-level category belongs to which broad-level category.

Another major factor that needs to be considered is the distribution of variability within the category structure to be learned. As outlined in the introduction, the hypothesis about the specific-level training providing better category learning rests upon the assumption that the variability is high between the specific-level categories, from the perspectives of both abstraction-based and the exemplar-based approaches. The abstraction-based approaches, particularly rule-based approaches, predict that diagnostic and characteristic feature identification would be easier in the case of specific-level training because the high variability between the specific-level categories would make it extremely difficult to extract the commonalities among their exemplars in broad-level training. The exemplar-based approaches would also predict the specific-level training advantage in a category structure with a high between-specific-level-category variability because the similarity signal when comparing the test items to the stored exemplars of a given category would be sharper when the comparison was restricted to only the exemplars from one specific-level category the presented test items belong to. However, in a category structure with a low between-specific-level-category variability, the similarity signal would be still sharp in the case of broad-level training because all the stored exemplars associated with a broad-level category would be relatively similar.

As discussed earlier, the variability distribution in the rock category might be not exactly as we assumed initially or as Nosofsky et al. (under review) set up their dispersed category structure. Specifically, it might be the case that variability in authentic rock category structure is high within the specific-level categories but lower than expected between the specific-level categories. However, a formal examination of the variability within the rock category structure using the *multi-dimensional scaling* (MDS; Shepard, 1962; see Nosofsky, 1992, for a review of this approach) refutes this possibility. In a MDS, each exemplar is rated its similarity to other exemplars according to several dimensions (e.g., lightness, grain size, and uniformity of grain, Nosofsky et al., under review), and then mapped out in a psychological space with the distance between exemplars signifying their similarity (i.e., the more similar two exemplars are the closer they are represented in the psychological space). Nosofsky and his colleagues recently completed MDS analysis of the rock category (Nosofsky et al., under review). In this analysis, they subjectively selected the most prototypical exemplar from 10 specific-level categories in each broad-level category including the majority of the specific-level categories I used in the current experiment, and their participants rated each exemplar's similarity to other exemplar according to their saturation, lightness, and grain size, that are identified by an expert geologist in their team as the most likely diagnostic features. Figure 6 shows the mapping of the 30 specific-level categories according to the three-dimensional similarity rating. Notice that each cluster (i.e., the middle, the left, and the top ones) contains specific-level categories from all three broad-level categories indicated by the different shapes, meaning these categories look very similar to each other in terms of features that are likely diagnostic. Figure 7 shows a pictorial illustration of two-dimensional mapping according to their saturation and lightness. On the bottom left, a cluster of four dark and relatively smooth rocks can be seen. They are obsidian from igneous, shale and bituminous coal from sedimentary, and anthracite from metamorphic.

The analysis described above also helps us reinterpret Noh et al.'s (2014) results in which a protocol that required their participants to focus on specific-level category information hindered their learning. As described in the introduction, unlike many natural categories, their broad-level categories (i.e., venomous versus non-venomous) had a clearly verbalizable rule to distinguish them according to the shapes and patters of the snakes' head, eye, and body. What this meant in terms of their category structure is that their broad-level categories clustered fairly separately. That is their between-specific-level-category variability was low. As described in the above section, both abstraction-based and exemplar-based approaches predict that specific-level training does not help in such a case.

Lastly, providing the participants with an opportunity to integrate their knowledge of the specific-level categories to form an understanding of the broad-level categories might be necessary for taking an advantage of the specific-level training. From the perspective of the rule-based approaches, learning which specific-level categories collectively form a broad-level category may help them integrate these information and come up with a probabilistic rule to determine the broad-level categorization and anticipate allowable variation. Similarly, from the perspective of the prototype-based approaches, this integration allows the learners to come up with a prototypical representation for each broad-level category. This may be particularly important in cases such as far-transfer in the current experiments in which learners generalize their learning to specific-level categories that they haven't yet encountered. On the other hand, the exemplar approaches predict that this integration is not necessary because the categorization advantage of specific-level training solely depends on the sharper similarity signal at the test. In my Experiment 1, the participants' opportunity to engage in this integration was likely overshadowed by the cognitive overload of learning the broad-and the specific- level categorization simultaneously. Even in Experiment 2, it is highly doubtful that the participants were able to engage in integrating the specific-level category information because when they were introduced with the broad-level category information for the first time during the associative learning phase, they were busy learning the semantic association of the category names rather than thinking about the category structure. Nosofsky et al. (under review) potentially bypassed issue by the use of the simplified and artificial specific-level category names (e.g., I1, I2, S1, S2 etc.). In this naming scheme, the specific- and broad- level category association was apparent and easily understood, so that it's possible that their participants could learn the specific-level categorization while they simultaneously formed an understanding of broad-level category structure. A further investigation is needed to clarify this issue. For example, it would be informative to conduct an experiment using a similar paradigm of Nosofsky et al. but also including far-transfer test items and manipulate the presence or absence of the integration opportunity by initially labeling the specific-level categories as Nosofsky et al. did or simply labeling them as one through nine and giving the association of each category to their broad-level category right before the test.

The current experiments failed to obtain the specific-level training advantage using authentic rock categories with an increased number of exemplars and more realistic category structure compared to Nosofsky et al.'s (under review) experiment that showed the specific-level training advantage. The analysis of the results and further examination of the category structure illuminated the importance of considering the cognitive load, the within-broad-level-category variability, and potential integration of specific-level category information to broad-level categories. Regarding the Experiment 2 finding that the degree of category learning itself was similar between the specific- and the broad- level training, there is one more appeal for the specific-level training protocol that has not been mentioned. If the category learning at the broad and the specific level can be achieved to the same degree in the same training time, one may argue that learners should be trained at the specific level. That is because the specific-level categorization offers more sophisticated understanding of the category structure allowing the learners a finer categorization of exemplars. In fact, experts show a tendency to identify objects at a more specific level of categorization (Johnson & Mervis, 1997; Tanaka & Taylor, 1991). After all, the goal of scientific categorization training is hardly ever just to categorize observed objects into their broad-level categories, but often to categorize them at the specific level so that more information can be instantly inferred. For example, geologists can infer much more scientifically valuable information such as how fast the magma cooled down to form a particular igneous rock or how much pressure was put to form a particular sedimentary rock. Given this inherent sophistication of the specific-level categorization and the factors identified through the current investigation, specific-level training of scientific categories remains promising. Future investigations should consider other natural scientific categories as there are many other categories that can be taught more efficiently incorporating the specific-level training.

References

Allen, S., & Brooks, L. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120,* 3–19.

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. Trends in Cognitive Sciences, 5, 204–210.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual. Review of Psychology*, *56*, 149-178.

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 657.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Bourne, L. E., Jr. (1974). An inference model of conceptual rule learning. In R. L. Solso (Ed.), Theories in cognitive psychology: The Loyola symposium (pp. 231–256). Potomac, MD: Erlbaum.

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & cognition*, *42*(3), 481-495.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of educational psychology*, *79*(4), 347.

Daniel, L.H., Hacket, J., Moyer, R.H., Vasquez, J. (2005). *MacMillan McGraw-Hill Science.* New York, NY: MacMillan McGraw-Hill.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

Gauthier, I., Curran, T., Curby, K.M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. Nature Neuroscience, 6, 428–432.

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,* 1441-1451.

Johnson, K., & Mervis, C. (1997). Effects of varying levels of expertise on the basic level of categorization. Journal of Experimental Psychology: General, 126, 248–277.

Kang, S.H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26,* 97-103.

Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 616.

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

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Little, J. L., & McDaniel, M. A. (2015). Some learners abstract, others memorize examples: Implications for education. *Translational Issues in Psychological Science*, *1*(2), 158.

Little, J. L., & McDaniel, M. A. (under revision). A Contextual Approach to Category Learning: The Contribution of Set Size, Prior Experience, and Individual Differences.

Little, D. R., Nosofsky, R. M., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 1–27.

Lockhart, R. S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding. Memory, 10(5-6), 397-403.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 43-52.

McDonnell, L., Barker, M., & Wieman, C. (in press) Concepts First, Jargon Second Improves Student Articulation of Understanding. *Biochemistry and Molecular Biology Education*.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review 85,* 207–238.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Murphy, G.L. (2002). *The Big Book of Concepts.* The MIT Press; Cambridge, Massachusetts.

Nestojko, J. F., Finley, J. R., & Roediger III, H. L. (2013). Extending Cognition to External Agents. Psychological Inquiry, 24(4), 321-325.

Noh, S. M., Yan, V. X., Vendetti, M. S., Castel, A. D., & Bjork, R. A. (2014). Multilevel Induction of Categories Venomous Snakes Hijack the Learning of Lower Category Levels. *Psychological science, 25*(8), 1592-1599.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, *10*(1), 104.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39 –57.

Nosofsky. R.M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*, 25-53.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 28, pp. 207–250). San Diego, CA: Academic Press.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. (1994). Rule-plus-exception model of classification learning. Psychological Review, 101, 53–79.

Nosofsky, R. M., Sanders, C., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (under review). *An Exemplar-Similarity Model Account of Learning at Low and High Levels of a Real-World Category Hierarchy.*

Pashler, H., & Mozer, M.C. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology: Learning, Memory and Cognition, 39,* 1162-1173.

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 77, 353-363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. Journal of Experimental Psychology, 83, 304-308.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General, 122,* 92–114.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125-140.

Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience & Biobehavioral Reviews*, *32*(2), 249-264.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: the early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257-285.

Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real-world perceptual expertise. *Psychological Science*, *16*(2), 145-151.

Tanaka, J.W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? Cognitive Psychology, 23, 457–482.

Trabasso, T., & Bower, G. H. (1968). *Attention in learning: Theory and research*. New York, NY: Wiley.

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.

Endnotes

[1] I also considered a condition in which participants complete the associative learning first and then learn the lower-order categorization (AL -> LO condition). However, a pilot study using an online sample showed no statistically significant difference between the LO -> AL condition and the AL -> LO condition, therefore we chose to use the LO -> AL condition which was numerically better in the current experiment.

[2] The study time in HO Only condition was increased to equate the total study time between the groups.

Table 1. List of specific-level categories under each broad-level category used in the current experiments

Broad-level Category Name	Specific-level Category Name
Igneous	Anorthosite
	Basalt
	Diorite
	Dunite
	Gabbro
	Lherzolite
	Nepheline Syenite
	Peridotite
Metamorphic	Blueschist
	Eclogite
	Gneiss
	Granulite
	Phyllite
	Quartzite
	Schist
	Slate
Sedimentary	Breccia
	Chalk
	Conglomerate
	Coquina
	Diatomite
	Mudstone
	Oolite
	Siltstone

Figure 1. A pictorial illustration of the perceived cohesiveness in broad- (top) and specific- (bottom) level training.

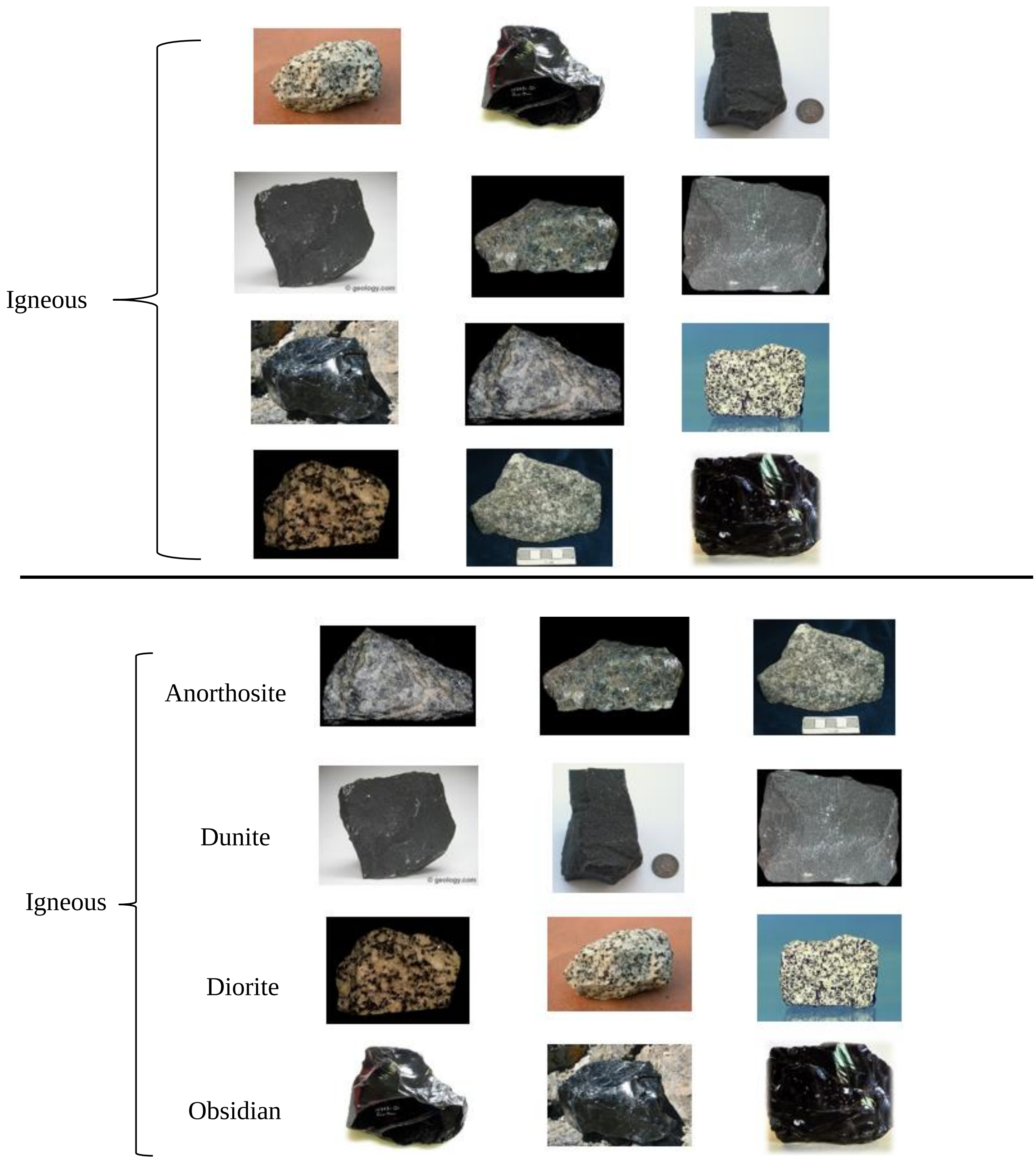

Figure 2. Examples of each specific-level category exemplars under each broad-level category

Figure 3. Examples of training trials

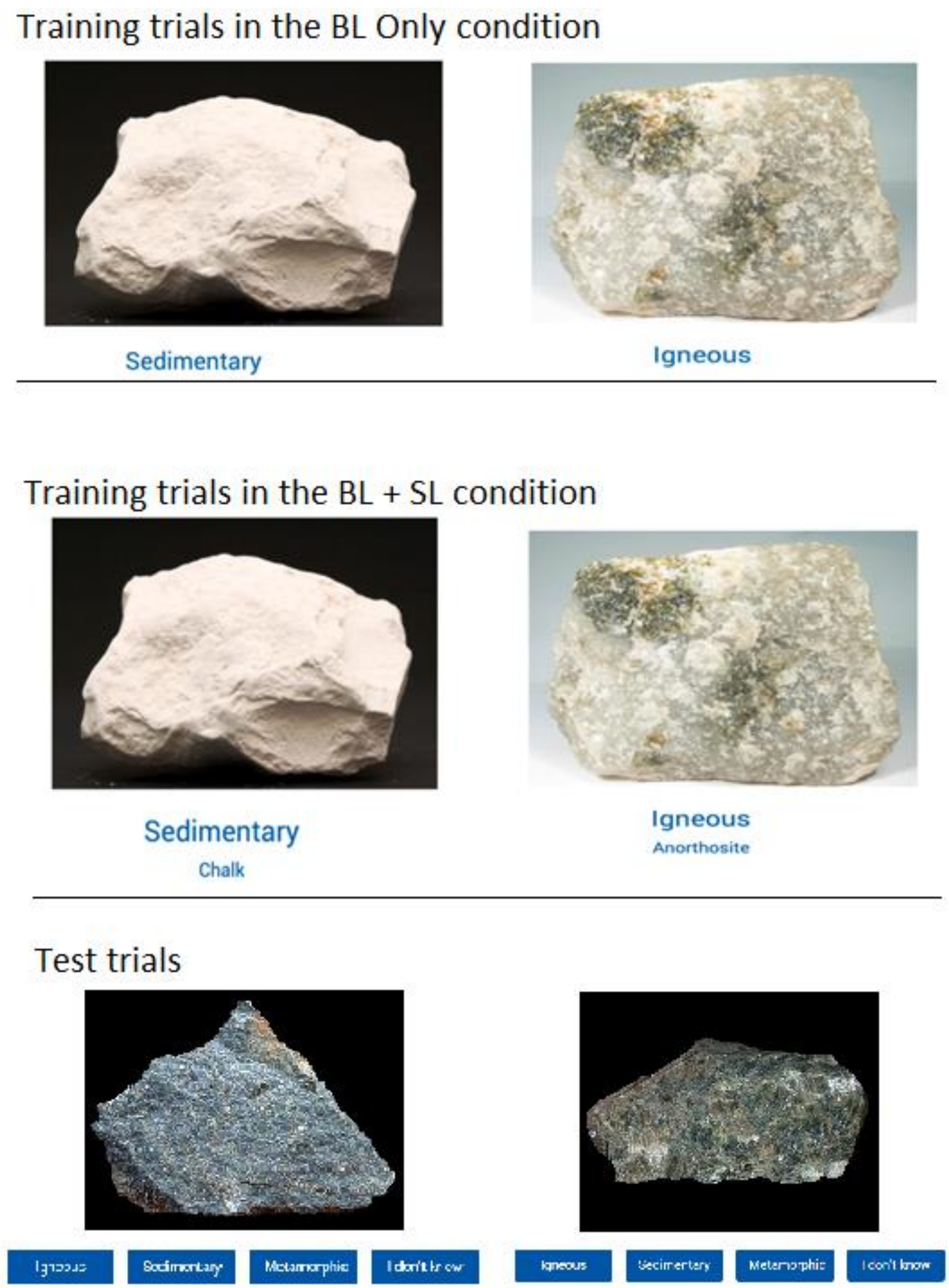

Figure 4. Participants' mean performance on the final test in experiment 1 plotted by conditions and test item types. Error bars denote ± 1 standard error.

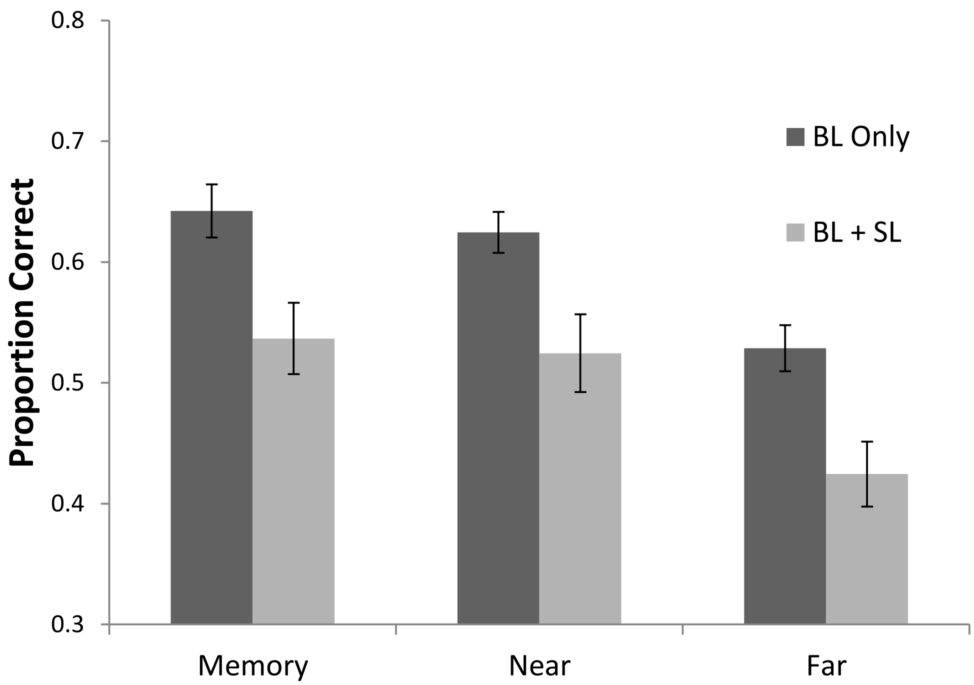

Figure 5. Participants' mean performance on the final test in experiment 2 plotted by conditions and test item types. Error bars denote ± 1 standard error.

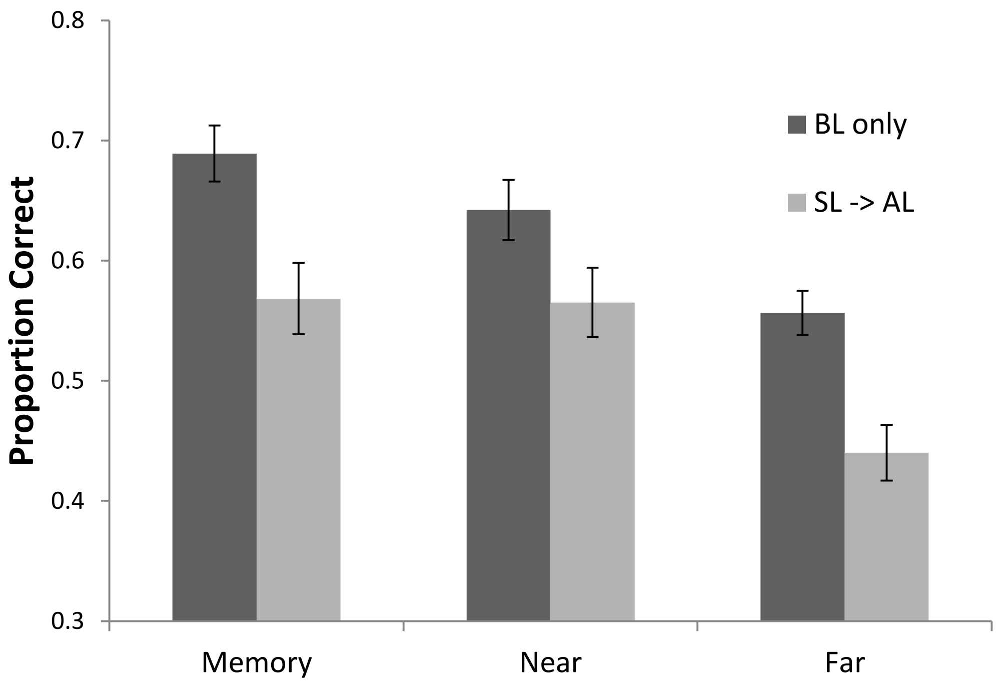

Figure 6. Multi-dimensional solution by Nosofsky et al. (under review) showing the clustering of the specific-level categories according to their saturation, lightness, and grain size. Each object represents a specific-level category and different shapes of the objects indicate different broad-level categories they belong to.

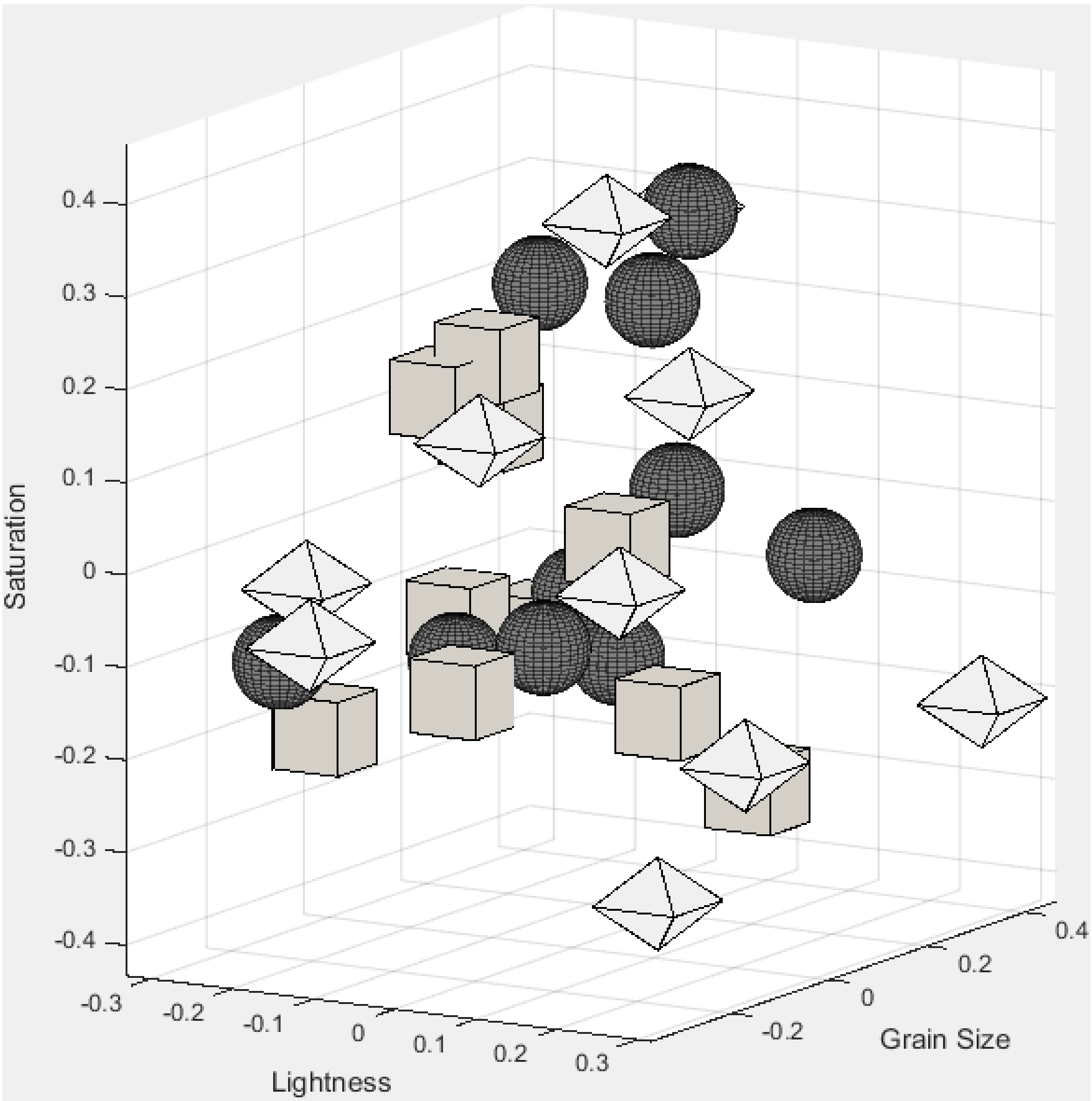

Figure 7. Multi-dimensional solution by Nosofsky et al. (under review) showing the clustering of the specific-level categories according to their saturation and lightness. Each picture of rock represents a different specific-level category.