Effects of Specific-level Versus Broad-level Training for Broad-Level Category Learning in a Complex Natural Science Domain

Toshiya Miyatsu[1], Robert M. Nosofsky[2], and Mark A. McDaniel[1]

[1]Washington University in St. Louis

[2]Indiana University

Author Note

Correspondence should be addressed to Toshiya Miyatsu, Department of Psychology, Washington University in St. Louis, One Brookings Drive, Campus Box 1125, St. Louis, MO, 631030. E-mail: tmiyatsu@wustl.edu

Acknowledgment

This research was supported by National Science Foundation Grant No. 1534014 (EHR Core Research). The authors thank Harkjoon Lee for his assistance in material preparation, data collection, and data analyses.

Abstract

Category learning is a core component of course curricula in science education. For instance, geology courses teach categorization of rock types. Using the educationally authentic rock categories, the current project examined whether category learning at a broad level (Igneous, Sedimentary, and Metamorphic rocks) could be enhanced by learning category information at a more specific level (e.g., Diorite under Igneous, Breccia under Sedimentary, etc.). Experiments 1 and 2 showed that specific-level training was inferior to broad-level training when participants were required to respond at the broad level regardless of whether broad- and specific- level category labels were presented simultaneously during classification training or specific-level categories were learned initially followed by training on the specific-broad level name associations. However, Experiments 3 and 4 showed that specific-level training was as good as broad-level training when the training was more extensive and participants were allowed to respond at the trained level. By considering confusion matrices (i.e., probabilities that instances in a given category was erroneously classified as belonging to other categories), we conjectured that between-specific-level category similarity, specifically the degree to which similar-looking specific-level categories belong to the same broad-level category, is an important factor in determining the efficacy of specific-level training.

Public Significance Statement

This study suggests that learning broad-level rock categories (Igneous, Sedimentary, and Metamorphic rocks) can be supported through teaching the subtypes of these broad levels (e.g., Diorite under Igneous, Breccia under Sedimentary) and doing so can be as effective as directly teaching the broad-level categories. This subtype-teaching technique might be considered more often in science education because it results in students learning a more expert-like multiple-level structure of categories, even when broad-level training is the initial learning objective.

Keywords: category learning, natural category, geology

Effects of Specific-level Versus Broad-level Training for Broad-Level Category Learning in a Complex Natural Science Domain

Learning of naturally occurring categories can be found at the core of many science disciplines. For example, ornithologists accurately distinguish between different bird species, mycologists identify edible and poisonous mushrooms, and geological scientists must be able to categorize rocks to assess geological features of terrain (Petcovic, Libarkin, & Baker, 2009). To achieve this expertise in classification, students learn from many examples of target categories throughout instruction in their respective discipline. Despite this prevalence in science education, however, studies aimed at identifying factors that can enhance instruction on natural categories have been scarce (for exceptions, see interleaving exemplars from multiple categories: Kornell & Bjork, 2008; Kang & Pashler, 2012; test-enhanced learning of natural categories: Jacoby, Wahlheim, & Coane, 2010; using training exemplars with exaggerated diagnostic features: Pashler & Mozer, 2013; specific-level training: Nosofsky, Sanders, Gerdom, Douglas, & McDaniel, 2017; feature highlighting: Miyatsu, Gouravajhala, Nosofsky, & McDaniel, in press). In the present paper, we focus on learning rock categories, a fundamental objective in geological science courses, and examine the efficacy of a theoretically motivated instructional technique.

Certain characteristics of natural categories, of which rock categories are an instance, make them challenging to learn. Natural categories often have high variability among the instances within a category (Murphy, 2002). That is, although there are instances that can reflect prototypical features of a given category, there are also instances within the category that deviate from such prototypical instances. In addition, this high variability often creates fuzzy boundaries between categories. That is, in some cases, two exemplars from different categories are perceptually more similar than two exemplars from a same category (See Figure 1). Therefore, successful learning of natural categories requires learners to not only identify characteristic (prototypical) features of a category but also to learn the variation in the category features reflected in the unusual or less prototypical instances.

Another important characteristic of natural category taxonomy is the hierarchical organization that creates categories at multiple levels (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). For example, the light grey or reddish rock with darker dots that is often used as kitchen counter tops can be classified as granite at the specific level or igneous rock at the broad level. The three broad-level categories that are ubiquitous in geoscience courses are igneous, sedimentary, and metamorphic. Under each of these three broad categories, there are many specific-level categories, such as andesite, obsidian, and rhyolite under igneous; gneiss, marble, and migmatite under metamorphic; and breccia, chert, and shale under sedimentary. Examination of images of these rock types (see Figure 2), illustrates the high perceptual variability among instances belonging to the same broad category. In the present article, we rely on the hierarchical structure of rock categories to implement a possible technique to assist learning of these highly variable rock categories. Specifically, we examine the interesting possibility that rock-category learning at the broad level might be enhanced by training learners at the specific category level. Note that such a technique would be counter to a typical training sequence in a geological science course, in which the broad rock categories would first be introduced and studied, followed by more in-depth learning of each broad category in terms of the specific-level categories it encompasses (e.g., Busch, & Tasa, 2009: a popular geoscience lab manual in which the rock classification is taught in a typical sequence).

Theoretically, specific-level training could be better at teaching the broad-level categorization than training directly at the broad level for the following reason. In most cases exemplars belonging to the same specific-level category will share more features among each other (i.e., many granite look similar) compared to exemplars belonging to the same broad-level category but coming from different specific-level categories (e.g., granite, basalt, and diorite, all of which are igneous rocks, may look very different from each other). That is, any given specific-level category likely has lower feature variability among its exemplars than any given broad-level category that includes multiple specific-level categories. Consequently, learners should be able to pick up more easily what features are shared among the exemplars belonging to the same category when presented with exemplars from the same specific-level category. Learners could then structure each broad-level category as a collection of these relatively coherent sets of specific-level categories. In contrast, learners trying to directly learn the broad-level categories would be faced with the challenge of accommodating exemplars with widely varying perceptual features (e.g., see Figure 2). That is, the premise is that trying to learn broad-level categories through the highly variable examples that comprise each category might be less effective than learning those broad-level categories as a collection of more coherent specific-level categories.

Specific-level versus broad-level training

In light of the above considerations, a handful of recent studies have compared specific- and broad-level training of natural categories. Tanaka, Curran, and Sheinberg (2005) had participants study pictures of owls or wading birds either with their broad-level category names (i.e., “owl” or “wading bird”) or with their specific-level category names of species (e.g., “great blue crown heron” or “eastern screech owl”). Category learning was then assessed by a sequential matching task (see Gauthier, Curran, Curby, & Collins, 2003) in which the participants judged whether a pair of successively presented bird pictures belonged to the same species. Their results indicated that the 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. That is, even when the participants were discriminating between two specific-level categories that were not studied, the participants who were trained at the specific-level were better able to discriminate them, presumably because the specific-level training equipped them with a more sophisticated perceptual discrimination strategy. This 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 was more suited to studying perceptual expertise, therefore limiting the generalizability of these findings to a broader category learning context. Specifically, in real life category induction situations, one would see a novel exemplar and try to identify which category the observed exemplar belonged to, not just judging several items as belonging to the same or different categories.

Noh, Yan, Vendetti, Castel, and Bjork (2014) more directly examined natural category learning when multiple categorization levels were trained. 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 and to classify others according to their broader category level. 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. The specific-level training disadvantage likely arose from the fact that the venomous and non-venomous distinction could be made through a set of relatively clear-cut rules. 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). Note that these clear-cut rules are less noticeable when the participants were focused on the specific level because these rules are not useful in learning the categorization at that level (i.e., many specific-level categories of venomous snakes share arrow-shaped heads, slit pupils, etc.).

Nosofsky et al. (2017) formalized the theoretical justification for why training at a more specific level can produce superior learning than the training at a broader level by applying 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 memory. GCM also takes into consideration that similarity is context-dependent, such that similarity in the dimension that learners attend to is weighted more, and learners generally learn to focus their attention on highly diagnostic dimensions and ignore less relevant ones (cf. Kruscke, 1992). Based on the GCM, Nosofsky et al. delineated conditions specifying when specific-level training would be more effective than broad-level training. Specifically, specific-level training would be effective when the variability within the broad-level categories is high (and consequently the similarity between some specific-level categories from contrasting broad-level categories is high). In the case of high within-broad-level variability, specific-level training offers storage of exemplars of a category that are more similar than the case of broad-level training. Then, when learners trained at the specific level are presented with an item at test, the similarity signal would be very sharp because the stored exemplars to be compared are based on the specific-level category structure (for which within-category variability is relatively low). On the contrary, the similarity signal would be blurry after broad-level training because the stored exemplars to be compared are based on a highly variable set of within-category exemplars (the broad-level). By contrast, when the within-broad-level variability is low, a few diagnostic dimensions that can separate broad-level categories may be available (like Noh et al.’s, 2014, snake categories). In such cases, broad-level training might allow learners to attend to the diagnostic dimensions, an opportunity that specific-level training would not allow.

Nosofsky et al. (2017) tested their hypothesis using categories of rocks that were artificially structured to exaggerate the within-broad-level variability. They compiled a set of rock pictures that made up either compact (i.e., low variability within the broad-level categories) or dispersed (i.e., high variability within the broad-level categories and high similarity between specific-level categories from contrasting broad-level) category structure organized into the three broad-level categories and three specific-level categories each. Their participants were trained at the broad 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). The results confirmed the hypothesized interaction such that the specific-level training produced better performance than did broad-level training in the dispersed condition, but the broad-level training produced better learning in the compact condition[1].

Nosofsky et al.'s (2017) and Tanaka et al.'s (2005) findings show promise that specific-level training could enhance broader-level category learning, however, some aspects of their methodology limit the applicability of their findings to more educationally authentic situations. First, Nosofsky et al. used a constructed category structure with only a small number of selected exemplars to either exaggerate or diminish the within-broad-level-category variability. Second, they bypassed the challenge of learning the pairing between the broad-level and specific-level category names by explicitly pairing the specific- and broad-level category names so that only the correct pairings were shown as answer options (e.g., "I1", "S2", or "Igneous:Granite", "Sedimentary:Dolomite" in the case of the replication study reported in their Supplemental Material). In a similar vein, Tanaka et al.'s participants were not required to learn the names of each category for the criterial test (indicate whether successively presented examples belonged to the same category), and secondly, participants did not have to discriminate between all of the learned categories which would be the case in most authentic situations.

Accordingly, it remains uncertain whether specific-level training can enhance broad-level categorization when the learning situation is more representative of the challenges present in an authentic training context. In particular, there are several challenges in specific-level training that have not been captured in previous work demonstrating a specific-level advantage. First as just mentioned, in authentic training situations, it is likely that learners given specific-level training would be presented with both the specific-level and the broad-level labels (given that the criterial

[1] Also see their Supplemental Materials for an additional experiment that replicated the specific-level training advantage in the compact condition using the actual rock names (e.g., Igneous:Granite, Metamorphic:Quartzite, Sedimentary:Dolomite).

task is for learners to be able to categorize at the broad level, but are using the specific-level categories to boot strap that learning). Therefore, the associative memory challenge in specific-level training (learn two labels for each exemplar) would arguably be greater in many authentic training situations than has been the case in extant studies (e.g., Tanaka et al., 2005; Nosofsky et al., 2017), and as a consequence specific-level training might lead to worse outcomes that broad-level training (for teaching broad-level categorization).

The above issues notwithstanding, it remains uncertain whether a more naturalistically sampled rock category structure with its many specific-level categories, as found in authentic learning situations, would favor specific-level training. There are a relatively large number of specific-level categories within each broad-level rock category; for instance, in this study we included 24 specific-level categories (eight within each of the three broad-level categories), which is fairly representative of the specific-level categories from introductory geoscience lab-training texts, albeit still not exhaustive (e.g., Busch & Tasa, 2009, a popular lab training text has 51 specific-level categories). In this case, although each specific-level category may reflect a less varied set of features (than the broad-level categories), some of these specific-level categories may introduce a new challenge: that of, discriminating somewhat similar specific-level categories that are members of different broad-level categories. Indeed, an examination of a pictorial representation of all the specific-level categories and their exemplars used in the current study (Appendix A), indicates some shared features among specific-level categories within a broad-level category (e.g., see diorite and nepheline syenite, in the igneous broad-level category) unlike the more idealized set displayed in Figure 2, and for the dispersed condition in Nosofsky et al. (2017). Nevertheless, there are also similarities across specific-level categories that belong in different broad-level categories (e.g., see the phyllite and siltstone in the metamorphic and sedimentary broad-level category respectively), and thus the current use of relatively many specific-level training categories may place the structure of the category space somewhere intermediate between a highly compact and highly dispersed structure (as constructed in Nosofsky et al., 2017). The upshot is that the relatively naturalistic sampling of the present category structures might mitigate or undermine the theoretical advantages of specific-level category learning considered at the outset (after reporting 4 experiments we provide some evidence for this possibility).

To evaluate the possible outcomes of specific-level training outlined above, we conducted a series of four experiments that contrasted specific-level training with broad-level category learning of a natural science category (rock categories). We were interested in doing so in a paradigm that incorporated several features informed by authentic training contexts. Specifically, the training instances were not artificially constrained (to either be highly dispersed or compact within the broad category), and these instances were sampled from a relatively wide range of specific-level categories, as reflected in standard geological sciences labs on learning rock categories.

Experiment 1

One straightforward and authentic way to combine the category learning at the broad and specific level is to present the category information at two levels simultaneously (as did Noh et al., 2014, and Nosofsky et al., 2017). Thus, in Experiment 1, one group of participants studied pictures of rocks with both broad- and specific- level category names (BL + SL condition), and their test performance was compared to a group who studied the rock pictures with just their broad-level category names (BL-only condition).

Participants' learning was assessed through three types of test items: trained, near-transfer, and far-transfer. Trained 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 were included in the study corpus, and far-transfer items were items from a new specific-level category (not included in the study phase). The idea of the far-transfer items was to examine a possible limit of specific-level training for rock-category learning not yet considered. In the BL-only condition, participants are trying to extract general characteristics of a given broad-level category and these would presumably support identification of items even from specific-level categories never studied (cf. Wahlheim, Finn, & Jacoby, 2012), By contrast, in the BL + SL condition, learning would presumably be centered on characteristics of the specific-level categories only, thereby possibly disadvantaging performance for items from an untrained specific-level category.

To recapitulate, if the specific-level training makes the learning of the broad-level categorization easier as suggested by the theoretical ideas and several experiments reviewed in the introduction, the BL + SL condition should show an advantage relative to the BL-only condition in generalization, but possibly limited to near-transfer items. Alternatively, the BL-only condition could perform better than the BL + SL condition across all test types because of several challenges in the BL + SL training that may disadvantage learning relative to the BL-only condition. Notably, there is additional memory demand for the participants in the BL + SL condition: they have to learn the pairings of the specific-level and broad-level names. The broad-specific (B-S) category pairing could be an issue particularly if the participants in the BL + SL condition primarily process the test items at the specific level. That is, one could identify the correct specific-level category of a given test item, but in order to give a correct response at the broad level, the B-S category pairing needs to also have been learned. In addition, it is possible that the participants in the BL + SL condition try to associate training items to both the specific-level and the broad-level names simultaneously. If so, the category learning task in the BL + SL condition would be more complicated than that of the BL-only condition. Finally, for this authentically sampled rock category structure, at least some specific-level categories may not be all that easily distinguished from one another, further compromising category learning in the BL + SL condition.

Method

Participants. Participants were recruited through Amazon Mechanical Turk and were compensated $3 for their time. 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[2]. 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. We determined the sample size (around 30 in each condition) based on Nosofsky et al. (2017).

Materials. Materials were 264 pictures of rocks collected through a web search organized into three broad-level categories, with eight specific-level categories within each broad-level category, and 11 exemplars for each specific-level category (see Appendix A for the list of specific-level categories and the rock pictures; high-resolution version of all the rock images used as well as data from all experiments are available at Open Science Framework: https://osf.io/n2awt). First, the lists of rock categories from websites such as geology.com and Wikipedia were used to compile a list of specific-level rock types for each of the three major rock categories. Then each specific-level rock name was used as a search term in google images to acquire exemplar pictures; most of the pictures were taken from personal websites that were designed to teach others about different rock types (e.g., http://earthphysicsteaching.homestead.com/Anorthosite.html ; http://www.jsjgeology.net/Eclogite.htm). Out of specific-level rock types that had enough number of pictures available, we excluded specific-level categories for which we suspected that

[2] In this and all of the following experiments, we report all measures, conditions, and data exclusions.

participants may have some prior knowledge (granite, rock salt, marble)[3]. The availability of pictures of particular rock types were used as a proxy for how often they appear in educational and recreational situations. This sampling method happened to overlap reasonably well with specific-level categories that students would encounter in introductory geological science courses (17 out of the 24 specific-level categories we used are covered in a popular geology lab manual, Busch & Tassa, 2009), and they are more comprehensive than in previous related research (e.g., in Nosofsky et al.'s, 2017, 9 specific-level categories were used in their dispersed condition, with 6 of those covered in Busch & Tassa, 2009). A very experienced geoscience instructor, who has been teaching an introductory geology lab for decades at one of the author's institutions, confirmed that this set reasonably well encompasses the wide range of features intended to be taught in an introductory geology lab[4] (R. F. Dymek, personal communication, June, 2018).

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 for each participant. Three trained test items were randomly chosen from the training exemplars. The assignment of each specific-level category was counterbalanced so that all of them equally often became studied and non-studied (far-transfer) categories. The data were collected using Collector (http://github.com/gikeymarcia/Collector), a PHP-based open source experiment program.

[3] We retained chalk because in the sedimentary category there were no alternative specific-level rock types with a sufficient number of pictures to serve as training and transfer stimuli.

[4] However, the expert also pointed out that Nepheline Syenite would not be usually included in the geoscience lab, and a few specific-level categories that are almost always included, such as granite and sandstone, were missing.

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. The participants in the BL + SL condition were told that their goal was to learn the categorization at the broad level, but there were several specific-level categories within the three broad-level categories and learning the specific-level categories would help them learn the broad-level categories. Both groups were explicitly told that the final test would be on the broad-level categorization. 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 trained, near-transfer, and far-transfer test item types. Participants were told that the test items included some pictures that they saw during the study phase as well as some pictures that they had not seen before. 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”. After the test, participants completed a post-experimental questionnaire that informally probed their prior knowledge of the rock categories and the presence of logistical problems, such as internet connection difficulty. The entire experiment took about 45 minutes to complete. This experiment’s as well as all following experiments’ protocols were reviewed and approved by the Washington University in St. Louis Institutional Review Board.

Results

Figure 3 shows participants' mean performance on the final test. A 2 X 3 mixed model analysis of variance (ANOVA) with the study condition (BL-only or BL + SL) as the between-subjects variable and the test item type (trained, near-transfer, or far-transfer) as the within-subjects variable was conducted to assess the main effects of condition and test item types as well as their interaction. There was a significant main effect of condition such that the BL-only group ($M = .60$, $SD = .12$) outperformed the BL + SL group ($M = .50$, $SD = .15$), $F(1, 56) = 11.98$, $p = .001$, $\eta p^2 = .18$. There was also a significant main effect of test item types, $F(2, 112) = 37.27$, $p < .001$, $\eta p^2 = .40$. Post-hoc paired-sample t-tests showed that compared to far-transfer items, ($M = .48$, $SD = .13$), the participants performed better on trained ($M = .60$, $SD = .14$), $t(57) = 7.48$, $p < .001$, $d = 0.82$, and near-transfer items ($M = .58$, $SD = .14$), $t(57) = 6.28$, $p < .001$, $d = 0.73$. There was no interaction between the study condition and test item type ($F < 1$, $p > .05$, $\eta p^2 = .00$).

Discussion

Participants in the BL-only condition outperformed those in the BL + SL condition on all three test-item types: trained, near-transfer, and far-transfer. The overall BL + SL impairment could be a reflection of imperfect B-S category pairing learning or participants trying to learn the categorization at both levels simultaneously. As outlined before, even if a participant could correctly identify a specific-level category to which a given test item belongs, correct B-S category pairing knowledge is still required to support a correct broad-level response. It is possible that being presented with the B-S category pairing while simultaneously trying to learn the categorization was not sufficient to teach the pairing perfectly. In addition, the instruction (i.e., the goal is to learn the broad-level categorization, and learning the specific-level categories would help learn the broad-level categories) might not have been direct enough in stating how the participants in the BL + SL condition should approach the task. For instance, some participants might have tried to process the training items at both levels; one could see a training item and assess the feature for igneous as well as for anorthosite. Plausibly, this would increase the complexity of the learning task, contributing to BL+ SL participants' inferior performance relative to BL-only participants. Further, it is possible that some participants in the BL + SL condition may not have consistently attended to one or the other category levels, which would also undercut a potential specific-level training advantage.

It is also possible that for the current materials, which we believe to be a reasonable representative of rock categories as commonly encountered, at least some of the specific-level categories were not reasonably discriminable from each other. If so, this could have produced difficulty for specific but not broad-level training. We return to this point in the final two experiments, but before doing so we next present an experiment addressing other challenges facing specific-level training (e.g., B-S category pairing learning).

Experiment 2

The complex task in the BL + SL condition in Experiment 1 can be broken down into two components, the learning of categorization at the specific level and the learning of the B-S category pairing. In Experiment 2 we attempted to facilitate the B-S category-pairing learning in the specific-level training condition by separating the category learning component from the category-pairing-learning (CPL) component. This way, the participants in the new specific-level training condition could focus on learning the categorization at one level, paralleling the demands in the broad-level training condition. In this new condition, labeled *SL -> CPL* (specific-level training, then category-pairing learning), participants learned specific-level categorization by observing the rock pictures, and then learned the pairing between specific- and broad- level category names[5]. In the separate CPL phase, participants were first presented with

[5] Some may worry that learning the B-S category pairing after studying the categorization might put this condition at a disadvantage because the CPL portion increases the interval between category learning and testing, relative to the BL-only condition. However, in a pilot study we also examined a condition in which participants completed the associative learning first and then learned the lower-order categorization (a CPL -> SL condition) This pilot, using an online sample, showed no statistically significant difference between a SL -> CPL condition and the CPL -> SL condition. Because the longer retention interval for the SL -> CPL condition had minimal effect on final test performance, we chose to use this condition in the current experiment.

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 category-pairing learning. The idea is that at the final categorization test, the participants in this condition could first categorize a given test item at the specific level and then convert the response into its corresponding broad-level category name.

If the global impairment observed in the BL + SL condition in Experiment 1 was primarily because of the B-S category-pairing-learning demand, potentially ambiguous learning instructions, or both, then separating the two components of the task could remedy that deficit. On that reasoning, the predicted outcome is that the SL -> CPL condition will approach or exceed the levels of performance of the BL-only condition on the generalization test. Alternatively, the broad-level training may continue to be superior because some challenges remain for B-S category-pairing learning in specific-level training. First, with total study time between the two conditions held constant, the category-pairing-learning training for the SL->CPL condition reduces the time that can be allocated for category learning. Accordingly, even if the category learning in the SL -> CPL condition were somewhat accelerated, that may not be sufficient to overcome the increased category-learning time afforded to the BL-only condition. Second, some of the category pairings, even with a repeated testing and feedback training protocol, may still be misremembered. For these misremembered pairings, the broad-level category response would be incorrect even when the specific-level category was learned.

Method

Participants. Fifty-two undergraduates (26 in each condition) at Washington University in St. Louis participated in this experiment for a course credit or $10. Again we determined the sample size based on Nosofsky et al. (2017).

Materials. The materials were identical to Experiment 1.

Procedure. Similarly to Experiment 1, Experiment 2 consisted of a study phase and a test phase. During the study phase, participants in the SL -> CPL condition were first told that their goal was to learn the broad-level categorization but they would learn the specific-level categorization first and learn the broad-specific category name associations later. Then, they were presented with 96 pictures of rocks (three broad-level categories, four specific-level categories each, eight exemplars each) 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 -> CPL condition engaged in the category-pairing learning task. First, they were presented with twelve B-S category pairs (i.e., all the specific-level categories they learned in the SL phase) 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 test-plus-feedback cycle, participants were presented with the specific-level category names 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"[6]. 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[7]. The total study time for the two groups was the same (960 seconds in both conditions).

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.

Results and Discussion

[6] We chose observation learning for the category learning but feedback learning for the CPL portion because observation learning is more common in real-life category learning situations, but feedback learning affords more robust learning of the B-S category pairs.

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

Performance during the category-pairing learning. The mean performance of the participants in the SL -> CPL condition during the category-pairing learning was .63 (*SD* = .15) during the first round of testing, .68 (*SD* = .17) during the second round, and .76 (*SD* = .15) for the last round.

Final test performance. Similarly to Experiment 1, a 2 X 3 mixed ANOVA was conducted to assess the main effects of condition and test item types as well as their interaction (see Figure 4 for means). There was a significant main effect of condition such that the BL-only group (*M* = .63, *SD* = .13) outperformed the SL -> CPL group (*M* = .53, *SD* = .15), *F*(1, 50) = 11.55, *p* = .001, ηp^2 = .19. There was also a significant main effect of test item types, *F*(2, 100) = 41.65, *p* < .001, ηp^2 = .45. Post-hoc t-tests showed that there was no significant difference between trained (*M* = .63, *SD* = .15) and near-transfer test items (*M* = .60, *SD* = .14), *t*(51) = 1.88, *p* = .03, *d* = 0.17. Participants performed worse on far-transfer items (*M* = .50, *SD* = .12) compared to trained, *t*(51) = -7.71, *p* < .001, *d* = -0.96, and near-transfer items, *t*(51) = -6.97, *p* < .001, *d* = -0.80. The interaction between the two independent variables was not significant, *F*(2, 100) = 1.27, *p* = .29, ηp^2 = .03.

Though learning of the B-S category-name pairings in the SL -> CPL condition was reasonably high, it still was not perfect. This imperfect learning of B-S name pairings could underlie the observed categorization-performance impairment in the SL -> CPL group relative to the BL-only group. To examine this possibility, we conducted an analysis conditionalized on successful pairing of B-S category names.

Conditional Analysis. In this analysis, the responses for particular test items were considered only if the participant's particular specific-level category name for the test item was successfully paired with the corresponding broad-level category name during the last round of the CPL phase (in the SL -> CPL condition). The far-transfer items were excluded from this analysis because they come from specific-level categories that were not learned during the associative learning phase.

A 2 x 2 mixed ANOVA with the study condition (BL-only or SL -> CPL) treated as the between-subjects variable and the test item type (trained or near-transfer) treated as the within-subjects variable was conducted on the conditionalized categorization performances. The performance of the BL-only group (*M* = .67, *SD* = .16) was not significantly different from that of the SL -> CPL group (*M* = .61, *SD* = .16), *F*(1, 50) = 2.56, *p* = .12, ηp^2 = .049. In addition, there was no significant effect of test item type, (for trained items, *M* = .65, *SD* = .15; for near-transfer, *M* = .62, *SD* = .14), *F*(1, 50) = 3.49, *p* = .07, ηp^2 = .065. The interaction between the two independent variables was not significant, *F*(1, 50) = 1.27, *p* = .16, ηp^2 = .040.

The conditional analysis suggests that when the B-S category-name pairing was successfully learned, there was no significant difference between the SL -> CPL condition and the BL-only condition in terms of learning to correctly categorize rocks at the broad level. Some might wonder if this analysis is challenged by potential selection effects. That is, easier-to-learn rocks (in terms of category) may be represented more so in the SL -> CPL than the BL-only performances. However, the difficulty of learning the B-S category pairing (on which performance was conditionalized) should not reflect the difficulty of learning the category per se of any particular rock, so there should be little systematic bias in terms of the category difficulty in the conditionalized analysis. To confirm this reasoning, we computed the correlation between the learning performance for the B-S category pairing and the categorization performance for each rock token. There was no significant correlation between the correct pairing rates during the CPL phase and the categorization performance on the trained items (*r*(22) = .32, *p* = .13) and on the near-transfer (*r*(22) = .32, *p* = .12) items[8]. Thus, excluding particular rock tokens based

[8] Correlation between the learning performance for the B-S category pairing and the conditionalized categorization performance rock token was also not significant: *r*(22) = .28, *p* = .19 for trained and *r*(22) = .29, *p* = .17 for near-transfer items.

on CPL performance (the conditionalized analysis) would not produce a subset of rocks significantly biased for ease of category learning.

The absence of a statistical difference between the SL -> CPL and BL conditions (in the conditionalized analysis) was observed along with several nontrivial learning benefits conferred by SL training. This condition learned a great deal about the specific-level categories that was not trained for the BL-only group. SL -> CPL participants also learned detailed information about the multi-level organization of rock categories that included the names of specific-level categories and the hierarchical organization of the taxonomy of rocks. Further, it is possible that the specific-level training produced more efficient category learning. To accommodate the CPL training time, the specific-level training group was given less time to study the categorization of the training items relative to the BL-only condition. Indeed, the participants in in the BL-only condition received almost 50% more study time (ten seconds for each rock example and its associated category) than those in the SL -> CPL condition (seven seconds for each item). Even with the reduced study time, the SL -> CPL group showed statistically equivalent conditionalized category learning performance relative to the BL-only group.

Though this pattern suggests that learning the rock categories might be more efficient at the specific than the broad level, the observation learning paradigm (used in this experiment) does not provide an index of learning rate. Further, we do not have a precise sense of how well the specific-level categories per se were learned (as discussed at conclusion of Experiment 1) because categorization responses (during testing) in the specific-level group were at the broad level. Experiment 3 was conducted to examine the generalizability of the findings across standard category-learning paradigms and to evaluate test performances in a paradigm in which the specific-level training condition was not required to undertake learning of the broad-specific category name pairings.

Experiment 3

In this experiment, we used a feedback learning paradigm (as opposed to the observation learning paradigm used in the previous experiments). During training participants made a category response on a trial-by-trial basis, followed by corrective feedback. We did this for two reasons. First, we wanted to see the generalizability of the findings across different learning paradigms. Active (i.e., feedback) versus passive (observational) learning has been shown to modulate the effects of other variables that are associated with category learning performance (e.g., Carvalho & Goldstone, 2015). Second, we wanted to align the current paradigm with the previous demonstration of the specific-level training advantage (Nosofsky et al., 2017).

We also implemented an important change to the specific-level training condition. Following Nosofsky et al. (2017), we had participants in the specific-level training conditions respond at the specific level at final test. We then scored their test responses in terms of broad-level categorization accuracy to directly contrast the specific- and broad-level training conditions. Our aim was to glean further insights into the dynamics of category learning at the specific level versus the broad level, to help inform the initial applied question of the extent to which specific-level training might enhance category learning at the broad level when the training stimuli sets and classification tasks were presumably more educationally authentic than previously examined (cf. Nosofsky et al., 2017; Tanaka et al., 2005). With these changes, we equated the amount of time devoted to category learning across the specific- and broad-level training conditions, and we eliminated the requirement that the specific-level learners had to learn the broad-specific category name pairings. We reasoned that under these conditions, the theoretical advantage of specific-level category learning over broad-level category learning—ease of learning the specific-level categories relative to the broad-level categories which have high within-category variability of exemplars—could come to the fore.

Alternatively, as emphasized earlier in this article, if the performance patterns observed in the specific-level training condition in Experiments 1 and 2 was mostly due to the rock structure in the current experiments not conforming to the properties required to benefit from specific-level training (difficult to learn broad-level categories because of high within-broad-level-category variability and specific-level categories that are relatively easy to learn, presumably because of low within-broad-level category variability and low between-specific-level category similarity), it is possible that the performance in the SL conditions will still be no better (or perhaps worse) relative to the BL conditions. We inform aspects of the present rock structure more directly at the conclusion of Experiments 3 and 4 with confusability analyses.

Finally, we manipulated the number of unique training exemplars (and the number of their repetitions) to compare specific- and broad-level training in a wider variety of conditions to establish the generality (see Wahlheim & DeSoto, 2017; Wahlheim et al., 2012). Specifically, in Experiment 3 half the participants in each learning condition (specific- or broad-level) learned the rock categories with three unique training exemplars with six repetitions per specific-level category (U_3R_6 conditions), and the other half learned with six unique training exemplars with three repetitions (U_6R_3 conditions) per specific-level category. That is, in the U_3R_6 broad-level training condition, each broad-level category was learned through 12 unique training exemplars (three unique exemplars times four specific-level categories) with six repetitions, in the U_6R_3 broad-level training condition, each broad-level category was learned through 24 unique training exemplars (six unique exemplars times four specific-level categories) with three repetitions. It is possible that a potential specific-level training advantage could be accentuated in the condition where a few examples are repeated several times (i.e., U_3R_6), a condition that happens sometimes in authentic learning situations because of a limited availability of the examples. The reasoning is based on consideration of the strength of stored examples and the differential likelihood of similarity of a novel example to the stored examples depending on the type of training (specific- or broad- level). As described before, the variability within a specific-level category is lower than the variability within a broad-level category that includes several specific-level categories. Thus, when a novel example that is similar to the few examples that have been learned well through a greater number of repetitions is presented, it elicits a strong similarity signal favoring that particular specific level category. In contrast, given the higher variability within the broad-level rock categories, it is likely that a novel example at test would be relatively more dissimilar to the stored examples after the broad-level training, and when there are fewer numbers of examples stored, it is even less likely that some of these stored examples would capture similarity to the novel example[9].

Method

Participants. Eighty-one undergraduates (18 in U_3R_6-BL, 21 in U_3R_6-SL, 21 in U_6R_3-BL, and 21 in U_6R_3-SL) at Washington University in St. Louis participated in the experiment for a course credit or $10. We determined the sample size based on Wahlheim et al. (2012), who also manipulated the number of unique training exemplars (for bird categories).

Materials. The materials were identical to the previous experiments.

Procedure. The experiment consisted of two phases, a study phase and a test phase. In the study phase, first, all participants went through one round of observation learning of all the training items (72 in the 6-unique-training-exemplars conditions and 36 of them in the 3-unique-training-exemplars conditions) in a random order. Then, in the 6-unique-training-exemplars conditions, the participants went through 216 feedback learning trials consisting of the 72 unique training exemplars repeated three times. In the 3-unique-training-exemplars conditions, the participants went through 252 feedback learning trials consisting of the 36 training exemplars repeated six times plus one additional time to make up for the differential number of the initial observation learning and equate the total number of training trials[10]. In a given feedback learning

[9] It should be noted that this prediction is not based on formal modeling, and models in which categorization is based on relatively similarity of the novel example to exemplars from different categories (not absolute similarity) might well generate different predictions.

[10] Because our intention was to create as naturalistic of a learning condition as possible within the active learning paradigm, we decided to include the initial observation phase (in naturalistic learning situations, learners are not typically required to categorize something that they have not yet studied). In doing so, we were then faced with the choice either to control the total number of trials between U_3R_6 and U_6R_3 conditions (which we did) or to control the

trial, the participants were presented with a rock picture and clicked an option corresponding to which category they thought the presented rock belonged to from four options in the BL conditions (three broad-level category names plus "I don't know") and 13 options in the SL conditions (12 specific-level category names plus "I don't know"). They were given five seconds to make the response and two seconds to review the correct answer afterwards. Upon finishing the learning phase, participants played tetris as a distractor task for 3 minutes before starting the test phase. The test procedure of the BL conditions was identical to Experiment 1 except the timing was participant-paced rather than 10 seconds per trial. The test procedure of the SL conditions was identical to that of the BL conditions except the classification was made at the specific level by choosing from the 13 options.

Results and Discussion

Training performance. Figure 5 shows the participants' performance during training as a function of the condition and the feedback learning block. Because the U_3R_6 conditions had a greater number of feedback learning trials (due to the initial observation trials including all unique exemplars), the extra feedback learning block for the U_3R_6 conditions was dropped from the analysis below. However, the means for the dropped block are plotted in Figure 5 as block 0. Note that in all the analyses reported below, the responses from the specific-level training groups were converted to corresponding broad-level responses to calculate the accuracy at the broad-level for all groups.

A 2 x 2 x 6 mixed ANOVA, with the number of unique training exemplar (3 or 6) and the training level (BL or SL) as the between-subjects variables and the feedback learning block (1-6)

number of feedback learning trials, in which case the U_6R_3 conditions would have had 36 more training trials in total than the U_3R_6 conditions. We acknowledge that the proportion of observation and feedback learning trials slightly differed between conditions in the current protocol (U_3R_6: 36 observation and 252 feedback trials; U_6R_3: 72 observation and 216 feedback), but we reasoned that it would have little impact on the final test performance given the large number of total trials.

as the within-subjects variable, was conducted. There was a significant main effect of the number of unique training exemplars such that the training performance was higher for the 3-unique-training-exemplar conditions (*M* = .84, *SD* = .11) than for the 6-unique-training-exemplar conditions (*M* = .73, *SD* = .12), *F*(1, 76) = 18.74, *p* < .001, ηp^2 = .20. There was also a significant main effect of the feedback learning block suggesting that performance increased as the training progressed, *F*(5, 380) = 49.47, *p* < .001, ηp^2 = .39. The main effect of the training level was not significant, *F*(1, 76) < 1, ηp^2 = .01. None of the interactions between the independent variables was significant (highest *F* =1.80 for feedback learning block by the number of unique exemplar interaction).

Final test performance. Figure 6 shows the mean performance during the final test as a function of the level of categorization (broad- or specific- level), the number of unique training exemplars (3 or 6), and the test item type (trained, near-transfer, or far-transfer).

A 2 X 2 X 3 mixed ANOVA, with the level of categorization and the number of unique training exemplars treated as the between-subjects variables and the test item type treated as the within-subjects variable, was conducted on these data. Crucially, there was no significant overall difference in performance between specific-level and broad-level training, *F*(1, 76) < 1, *p* = .58, ηp^2 = .004. This pattern was qualified, however, by a significant interaction with the test item type, *F*(2, 152) = 5.22, *p* = .006, ηp^2 = .06. Post-hoc independent-sample t-tests showed that the broad-level training groups performed better than the specific-level training groups on the far-transfer items (*M* = .55, *SD* = .12 vs *M* = .50, *SD* = .12), *t*(78) = 2.11, *p* = .02, *d* = 0.47, but their performance did not differ for trained (*M* = .86, *SD* = .11 vs *M* = .89, *SD* = .11), *t*(78) = 0.94, *p* = .18, *d* = 0.21, and near-transfer items (*M* = .68, *SD* = .12 vs *M* = .69, *SD* = .12), *t*(78) = 0.10, *p* = .46, *d* = 0.02. Neither the interaction between the level of categorization and the number of unique exemplars nor the three-way interaction was significant, (*F*s < 1).

There was a significant main effect of the test item type, *F*(2, 152) = 379.76, *p* < .001, ηp^2 = .83. Post-hoc paired-samples t-tests showed that the participants performed better on trained items (*M* = .87, *SD* = .10) compared to near-transfer items (*M* = .69, *SD* = .11), *t*(79) = 15.88, *p* < .001, *d* = 1.72, and far-transfer items (*M* = .52, *SD* = .13), *t*(79) = 23.08, *p* < .001, *d* = 2.95, and on near-transfer items compared to far-transfer items, *t*(79) = 11.91, *p* < .001, *d* = 1.39. There was a significant interaction between the test item type and the number of unique training exemplars, *F*(2, 152) = 9.40, *p* < .001, ηp^2 = .11, suggesting that the 3-unique-training-exemplars groups performed better than the 6-unique-training-exemplars groups on the trained items (*M* = .92, *SD* = .10 vs *M* = .82, *SD* = .10), but their performance did not differ on near- and far-transfer items (near-transfer: *M* = .68, *SD* = .11 vs *M* = .69, *SD* = .11; far-transfer: *M* = .58, *SD* = .12 vs *M* = .52, *SD* = .12).

In sum, when the challenges associated with learning the broad-specific category-name-pairing were bypassed, specific-level training produced category learning at levels displayed by the broad-level training (for near-transfer). The Experiment 3 training protocol was very similar to the protocol that demonstrated specific-training advantages in a previous study (i.e., feedback learning paradigm, repetition of training exemplars; Nosofsky et al., 2017). Specifically, both our U_3R_6 conditions and the Nosofsky et al. (2017) conditions had three unique training exemplars per category and six repetitions of each exemplar (technically seven repetitions in ours) learned through feedback learning. Yet, our results showed equivalent performance between the BL and SL conditions, rather than an advantage for an SL training condition (but previously found only for the dispersed training items in Nosofsky et al.). These results may suggest that with a more comprehensive set of rock categories and training instances that are haphazardly sampled, the training does not uniformly reflect the properties that favor specific-level training (high within-broad-level variability and easy to learn specific-level categories). In the General Discussion, we address these points in more depth.

Another novel finding related to the far-transfer performance (not tested in Nosofky, et al., 2017), for which broad-level training produced superior performance than specific-level training. This pattern suggests that discriminative aspects of the overarching broad-level category structure were learned with training at the broad level, aspects that were not acquired with specific-level training. Still, specific-level training produced far transfer (with broad-category scoring), as evidenced by above-chance performance. We defer additional discussion pending the replication and extension of this finding (and the near-transfer pattern) in the next experiment.

Experiment 4

In our final experiment we investigated whether implementing a different type of training schedule that might better align with the demands of specific-level training might produce superior category learning with specific-level versus broad level training. Specifically, we employed a blocked training sequence (i.e., examples from the same category are presented consecutively) in Experiment 4, as opposed to an interleaved sequence (i.e., examples from different categories are presented successively) as in Experiments 1 through 3. Presumably, training in an interleaved sequence requires learners to keep track of characteristics in features associated with many categories and be able to retrieve shared features when they encounter another example from the same category. In this case, the greater number of categories to be learned in specific- compared to broad- level training will pose a greater memory demand for specific-level training. By contrast, when the examples from the same categories are blocked, learners do not have to keep track of feature characteristics associated with all of the to-be-learned categories simultaneously; rather, they are able to focus on identifying features that are similar for members of each particular category one by one. Thus, it is possible that with a blocked training sequence, the specific-level training facilitates extracting key features that are shared among exemplars from the each specific-level category, whereas broad-level training remains a difficult situation to determine features that are shared across exemplars from the same broad-level category.

We compared the blocked specific-level training to two versions of blocked broad-level training. The first version, simply referred to as the broad-level (BL) condition is what commonly is thought of as blocking, where the participants see examples from several different specific-level categories successively with their unifying broad-level labeling within a training block. For example, one may see examples from anorthosite, basalt, diorite, and dunite successively, all labeled as igneous rocks. The second version, referred to as broad-level-labeled—blocked by specific level (BL-blocked specific) condition, is essentially what a specific-level training looks like except the examples are labeled according to their broad-level categories. For example, one may see examples of anorthosite successively labeled as igneous rocks. This condition was necessary to tease apart the contribution from the effect of categorization experience (i.e., labeling: encoding the examples at a different level of categorization) and the effect of perceptual experience (i.e., successively encoding similarly looking examples) because the categorization experience on its own has been shown to enhance category learning (Tanaka et al., 2005).

Because of the nature of the feedback learning paradigm (i.e., guessing which category the presented exemplar belongs to), a strict implementation of blocking was not possible (because participants could simply keep on responding with the same category). Accordingly, we implemented a probabilistic blocking (e.g., Carvalho & Goldstone, 2014) in which a particular category appears at a high probability within one feedback learning block. In the current experiment, out of 4 trials within one feedback learning mini-block (see below), three of them belonged to a blocked category.

Method

Participants. Eighty-six undergraduates (31 in BL, 26 in BL- blocked specific, and 29 in SL) at Washington University in St. Louis participated in the experiment for a course credit or $10. We determined the sample size based on Nosofsky et al. (2017).

Materials. The materials were identical to the previous experiments.

Procedure. The experiment consisted of three phases, an observation learning phase, a feedback learning phase, and a test phase. In the study phase, first, all participants went through one round of observation learning of all the training items (48 total training exemplars: four exemplars per one specific-level category, four specific-level categories per one broad-level category, and three broad-level categories to be learned). The training exemplars were randomly selected from the set for each participant. In all three conditions (BL, BL- blocked specific, and SL as described above), the order of the presentation was blocked by the broad-level categories, such that the participants saw all 16 training exemplars from the same broad-level category and moved on to the next broad-level category although in the SL condition the exemplars were labeled according to their specific-level categories. The order of the broad-level categories during the observation learning was counter-balanced. The participants were given seven seconds per observation trial.

Upon completing the observation learning, the feedback learning followed. At any given feedback learning trial, the participants made their response by pressing a key representing the category they thought the given picture of rock belonged to (out of four options in the BL and BL- blocked specific conditions, and out of 13 options in the SL condition similar to the previous experiments), and feedback with a correct answer was presented afterwards. The participants had five seconds to make a response and a feedback was presented for two seconds. The participants went through the 48 training exemplars three times (i.e., three feedback learning blocks). At any given feedback learning mini-block of four trials, three exemplars came from the same category.

The order of feedback learning trials was block-randomized for each participant, such that the order of categories was randomized anew for each feedback learning block.

After the feedback learning phase, the participants played tetris for three minutes as a distractor task, and then completed the final test. The procedure of the final test was identical to Experiment 3 (36 each of trained, near-transfer, and far-transfer items were tested in a unique randomized order for each participant, and it was self-paced).

Results and Discussion

Training performance. Figure 7 shows the participants' performance as a function of the condition and the feedback learning block. Note that in all the analyses reported below, the responses from the specific-level training groups were converted to corresponding broad-level responses to calculate the accuracy at the broad-level for all groups. A 3 X 3 mixed ANOVA, with the training condition treated as the between-subjects variable and the feedback learning block treated as the within-subjects variable, was conducted on these data. There was a main effect of the feedback learning block suggesting the participants' performance increased across the blocks (Block 1: *M* = .67, *SD* = .13; Block 2: *M* = .78, *SD* = .12; Block 3: *M* = .84, *SD* = .10), *F*(2, 166) = 215.54, *p* > .001, ηp^2 = .72. The main effect of conditions was not significant, *F*(2, 83) < 1, *p* = .74, ηp^2 = .007. The interaction between the two variables was not significant, *F*(4, 166) = 2.36, *p* = .06, ηp^2 = .05.

Final test performance. Figure 8 shows the participants' final test performance as a function of the training condition (BL, BL-blocked specific, or SL) and the test item type. A 3 X 3 mixed ANOVA, with the training condition as the between-subjects variable and the test item type as the within-subjects variable, was conducted on these data. There was a significant main effect of test item type, *F*(2, 166) = 269.31, *p* > .001, ηp^2 = .76. Post-hoc paired-samples t-tests showed that the participants performed better on trained items (*M* = .85, *SD* = .11) than near-transfer (*M* = .68, *SD* = .11) , *t*(85) = 14.50, *p* < .001, *d* = 1.56, and far-transfer items (*M* = .54, *SD* = .11) , *t*(85) = 22.98, *p* < .001, *d* = 2.79, and on near-transfer than far-transfer items, *t*(85) = 9.71, *p* < .001, *d* = 1.30. Neither the main effect of training condition, *F*(2, 83) < 1, *p* = .48, ηp^2 = .02, nor the interaction between this variable and test item type was significant, *F*(4, 166) < 1, *p* = .61, ηp^2 = .02.

In sum, with a blocked training sequence in Experiment 4, final test performance was again equivalent across BL and SL conditions (even for far-transfer items in this experiment). This pattern generally converges with the results from an interleaved training sequence (Experiment 3). Therefore, it appears training sequence is not critical for the equivalent performances across specific-level and broad-level training; rather the natural structure of the rock categories may not afford an advantage from specific-level training. We address this point in the next section.

Confusion Matrix Analyses

As outlined in the introduction, the generalized context model formally predicts that specific-level training benefits category learning at the broad level when the within-broad-level variability is high and the between-specific-level similarity across contrasting broad-level categories is relatively high (Nosofsky et al., 2017). Figure 9 presents a schematic illustration of the dispersed broad-level category structure that is theorized to benefit from specific-level training. An extreme case of dispersed structure illustrated on the right panel of Figure 9 shows similar-looking specific-level categories from contrasting broad-level categories (e.g., I1, M1, and S1) forming a cluster in a psychological similarity space. On the other hand, a compact structure illustrated in the left panel has low within-broad-level variability and high between-broad-level variability, such that all specific-level categories belonging to one broad-level category (e.g., I1, I2, and I3) are situated close to each other. In this section, we report analyses on the pattern of errors from the responses from Experiments 3 and 4 to gauge the degree of the between-specific-level similarity across contrasting broad-level category in the current category structure. Figures 10 and 11 present confusion matrices for Experiment 3 and 4 final test performance on the near-transfer items[11]. We focused on the near-transfer items in these analyses because the trained items are inappropriate in assessing category learning (contaminated by memory for the specific item) and far-transfer items disfavors specific-level training conditions (no explicit mention of broad-level structure that encompass the far-transfer items). Further, we focused on Experiments 3 and 4 results because these provide the contrast between the items when categorized at the specific level versus at the broad level.

The rows and columns are the names of the 24 specific-level categories used in the current experiments organized in the alphabetical order within each broad-level category. The rows represent the proportion at which a near-transfer test item was classified as a member of a given specific-level category. For instance, in Figure 10 as you move through the top row from left to right, you can see the proportion at which an instance from anorthosite was classified as each specific-level category (i.e., P(response|stimulus); e.g., correctly classified as anorthosite at .29, often mis-classified as Granulite at .22 and as three other igneous specific-level categories at .08-.11). Note that the darker shades of green highlight higher proportions.

For a number of reasons, the confusion rates based on the data from the specific-level training conditions (Figures 10 & 11) appear to represent systematic errors based on the between-specific-level-category similarity rather than random errors. First, in the majority of cases the confusion is bidirectional such that if category A was incorrectly classified as category B at a high rate, then category B was also incorrectly classified as category A at a high rate. For example, dunite was incorrectly categorized as lherzolite 32% of the time while lherzolite was

[11] Note that the sum of some rows does not add up to 1 because the proportions shown in the matrix tables do not include the “I don’t know” responses.

incorrectly categorized as dunite at 24 % of the time (Figure 10). Second, the confusions align with the similarity in the visual features for the specific-level categories that were confused as each other (see Appendix A to compare anorthosite and granulite or dunite and lherzolite for examples). Third, the pattern of confusions is highly consistent across Experiment 3 and 4 (Figures 10 and 11). A cell-by-cell correlational analysis of the confusion matrices from Experiments 3 and 4[12] shows that they are indeed extremely highly correlated, $r(550) = .81$, $p < .001$.

Critically, an examination of these figures reveals that the between-specific-level similarity across contrasting broad-level categories in the current category structure was neither very high nor very low. One way to quantify the between-specific-level similarity across broad-level categories from the confusion matrices is to see for a given specific-level category, if the specific-level category that was incorrectly classified most often belonged to the same or different broad-level category. Out of 24 specific-level categories, 14 of them in Experiment 3 (excluding two ties, multiple specific-level categories from within and between broad-level categories had the identical confusion rates) and 8 of them in Experiment 4 (excluding four ties) had their most confusable counterparts in other broad-level categories. In other words, about half the specific-level categories were most similar to another specific-level category in a different broad-level category, whereas about half were more similar to a specific-level category in the same broad-level category. Importantly, if the present broad-level categories had a clear dispersed structure (as did in Nosofsky et al., 2017), a structure that has been shown in one previous study to afford a specific-level training advantage, the above analyses would have been expected to show that most of the specific-level categories were most confusable with a specific-

[12] This analysis excluded the diagonal cells (i.e., the correct responses) to examine whether the pattern of confusion errors aligned across experiments.

level category from a different broad-level category. By the same token, if the present broad-level category structure were generally compact, then it would be expected that most of the specific-level categories were confusable with other specific-level categories from the same broad-level category.   But neither pattern emerged; instead approximately half of the specific-level categories were most often misclassified within the same broad-level category and half more often misclassified between broad-level categories.

Figure 12 shows a summary table of the confusion matrices shown in Figures 10 and 11 in terms of the correct rates, the sum of the within-broad-category confusion, and the sum of the between-broad-level confusion, for each specific-level category. This table shows that sedimentary rocks are easier to learn, and Igneous and Metamorphic rocks have higher between-broad-category confusion rates than Sedimentary. This observation combined with how the confusable categories are distributed across the broad-level categories in Figure 10 and 11 (i.e., between-broad-level confusion happens more often between igneous and metamorphic) suggest that igneous and metamorphic rocks are more overlapping and sedimentary rocks are more distinct. Given that categories with high between-broad-category confusion rates are more likely to benefit from specific-level training, one may wonder if there is a more pronounced specific-level training advantage for igneous and metamorphic rocks than for sedimentary rocks. Somewhat in line with this possibility, for the near-transfer items plotted in Figures 10 and 11, a greater number of specific-level categories in igneous and metamorphic show a specific-level training advantage than sedimentary (out of 8 specific-level categories within each broad category, in Experiment 3, 4 igneous and 4 metamorphic specific-level categories showed a specific-level training advantage compared to 2 sedimentary specific-level categories; in Experiment 4, 4 igneous and 4 metamorphic specific-level categories showed a specific-level training advantage compared to 3 sedimentary specific-level categories). Importantly, the summary table tells a similar story to that of Figures 10 and 11 in that the number of specific-level categories showing higher confusion rates for between-broad-level categories than within-broad-level categories, which signals a high between-specific-level similarity across contrasting broad-level categories, is neither very high or very low (16 in Experiment 3 and 13 in Experiment 4). Thus, the current category structure appears to be somewhat intermediate between extremes of compactness and dispersion.

General Discussion

In the current experiments we examined the theoretically motivated hypothesis that the broad-level categorization of rock category might be enhanced by training at the specific level compared to training at the broad level. Testing this hypothesis with more naturalistic training material (i.e., not artificially constrained like the previous demonstration of specific-level training benefit: Nosofsky et al., 2017) and training procedure (e.g., learn at the specific level and respond at the broad level) across four experiments, we found no specific-level training advantage. Indeed, in some cases, broad-level training was better than specific-level training (Experiments 1 & 2). A specific-level advantage was absent even when normal challenges associated with specific-level training was relaxed, such that the participants could respond at the specific level, the level they were trained at (Experiments 3 and 4). Further, the absence of a specific-level advantage was shown to hold across a number of important training parameters that have been shown to influence category learning: observation learning (Experiments 1 and 2), learning through active response with feedback (Experiments 3 and 4), interleaved sequence (Experiments 1-3), blocked sequence (Experiment 4), learning through small number of unique training exemplars with several repetitions (Experiment 3), learning through a greater number of unique training exemplars with fewer or no repetitions (Experiments 1-4). Below we discuss theoretical and practical implications of these findings.

The results from Experiments 3 and 4 strongly suggest that the naturalistic category structure we used in the current experiments did not entirely have the properties for which specific-level training could be advantageous. This conclusion is based on several observations. First, there was no specific-level training advantage even when the training protocol was very similar to the protocol that showed a specific-level training advantage previously (Nosofsky, et al., 2017; Experiment 3). A specific-level training advantage also did not emerge when the training protocol was presumably designed to favor specific-level training through blocked presentation of training items within each specific level (Experiment 4). In light of the results from Experiments 1-3, we had thought it possible that the interleaved sequence that was implemented in those experiments could have disadvantaged specific-level training because making connections between examples from a same category, a process necessary for identifying characteristic features, was harder in specific-level training. (Specifically, while there were examples from two other categories in between two examples from a same category in the broad-level training, there were examples from eleven other categories in the specific-level training.) However, this possibility was not borne out, as noted above (Experiment 4 result). In light of (a) the similarities between the Nosofsky et al. (2017) training conditions and Experiments 3 and 4, and (b) the equivalent performance between the SL and BL conditions in Experiments 3 and 4, it appears that the current category structure, with an authentically generous number of specific-level categories (8 specific-level categories within each of the 3 broad-level categories) within which the training items were sampled somewhat haphazardly, did not reflect the extreme characteristics of either the compact or dispersed structures created by Nosofsky et al. (2017).

Somewhat direct support for the above conclusions was provided by the confusion-pattern analyses. As discussed in the previous section, the range of specific-level categories trained herein and the haphazard sampling of training items (that may be reflective of an introductory geoscience classroom environment) seemed to be intermediate between a highly dispersed category structure and a compact structure (structures that Nosofsky et al., 2017, constructed for their experiments). The critical finding in this regard was that a fair number of the specific-level categories were most similar to other specific-level categories belonging to the *same* broad-level category, whereas others were most similar to specific-level categories belonging to a *different* broad-level category. Consequently, specific-level training would not be heavily favored but neither would broad-level training. An experimental demonstration of how these factors of category structure (i.e., within-broad-category variability and between-broad-category similarity of specific categories) modulate the benefit of specific-level training has been reported previously (Nosofsky et al., 2017). The current study explored the extent to which the category-learning effects of those experimentally constrained category structures might be reflected with a more naturalistic category structure, encompassing a wider sampling of rock categories (potentially representative of geoscience education) and offering a more practical, accessible means of gauging the category similarities (i.e., the confusion matrix which can be easily computed from the response data from a category learning task).

One novel aspect of the current experiments was the test of far-transfer test items (generalization to new instances from untrained specific-level categories). A straightforward prediction was that broad-level training should be favored for this type of test because there is no need to extract overarching broad-level features in specific-level training. This prediction was generally supported; there was a statistically significant advantage for broad-level training on far-transfer items in Experiments 1-3 and a numerical advantage in Experiment 4. When presented with a far-transfer item, a learner trained in specific-level training is presumably making the classification decision according to the similarity to the specific-level categories that they studied, which would not necessarily encompass broad level features needed to classify the far transfer items. In some cases, far-transfer items looked relatively different from any of the specific-level categories that were studied, resulting in participants in the specific-level training conditions choosing the “I don’t know” option at a greater proportion (*M* = .06, range: .01-.09 across Experiments 1-4) than participants in the broad-level training conditions this option (*M* = .02, range: .00-.04). Presumably the overarching broad-level features that were extracted through the broad-level training provided these participants with a basis to select a broad-level category in which a given far-transfer item might belong (see Appendix B for the results and discussion pertaining to the “I don’t know” responses). This broad-level advantage on far-transfer items has clear practical implication. If the goal of the instruction is to prepare learners for a situation where broad-level identification of instances from several unknown specific-level categories is expected, broad-level training should be favored (or broad-level training could be incorporated along with specific-level training as a hybrid).

Clearly, however, for near-transfer test items (generalization to new instances from a learned category), specific-level training was as good as broad-level training across several training protocols: when participants could respond at the trained level for both interleaved (Experiment 3) and blocked training (Experiment 4) sequences and when all participants were required to respond at the broad level, provided that the specific-broad level label associations were trained separately and responding was conditionalized on correct learning of those associations (Experiment 2).   This performance equivalence between broad- and specific- level training has several important practical implications. Consider first the protocol for Experiments 3 and 4. The experimenter conversion of the specific-level responses to broad-level responses is not as artificial as it might first appear because the specific-broad category-name pairing could be externalized easily (e.g., providing a table in a textbook). If this option was not favored, separately training broad-specific category pairings would be another possibility. Although the degree of the broad-specific category pairing learning was not perfect (.76 at the end of the training) in Experiment 2, these simple associations should not be difficult to train if multiple sessions were spread across several days (e.g., studying with flashcards; Kornell, 2009).

In either case, from a practical standpoint, one might argue that the specific-level training is preferred because it taught a great deal more than just the broad-level categories; specific-level training teaches a finer classification that characterizes the experts in a given domain (e.g., Tanaka, 2001; Tanaka & Taylor, 1991). In fact, in geo-science and other domains, the eventual training objective for the students is to know not only broad-level but also specific-level categories. For example, in many introductory geology lab courses, a lab section that is aimed at teaching the broad-level rock classification is often followed by lab sections that are aimed at teaching the specific-level classification (e.g., Busch, & Tasa, 2009). Given this trajectory toward expertise that accompanies specific-level training, as well as the eventual training objective of teaching specific-level categorization, proceeding with specific-level training initially or concurrently with broad-level training might be considered in educational settings when teaching natural science categories.

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.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition, 41*(3), 392-402.

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.

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.

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.

Busch, R., & Tasa, D. (2009). *Laboratory Manual in Physical Geology* (8th ed.). Upper Saddle River, NJ: Pearson.

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, *22*(1), 281-288.

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. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*(9), 1297-1317.

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.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review, 111*(2), 309.

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.

Miyatsu, T., Gouravajhala, R., Nosofsky, R. M., & McDaniel, M. A. (2019). Feature Highlighting Enhances Learning of a Complex Natural-Science Category. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *45*(1), 1-16.

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. (2011). The generalized context model: An exemplar model of classification. In E. Pothos & A. Wills (Eds.), *Formal Approaches in Categorization* (pp. 18–39). New York, NY: Cambridge University Press.

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. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, *50*(2), 530-556.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017). On Learning Natural-Science Categories That Violate the Family-Resemblance Principle. *Psychological Science 28*(1), 104-114.

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

Petcovic, H. L., Libarkin, J. C., & Baker, K. M. (2009). An empirical methodology for investigating geocognition in the field. *Journal of Geoscience Education, 57*(4), 316-328. Chicago

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.

Reiser, B., & Tabak, I. (2014). Scaffolding. In R. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences, Second Edition* (pp. 44-62). Cambridge: Cambridge University Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382-439.

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.

Soderstrom, N. C., & Bjork, R. A. (2013). Learning versus performance. In D. S. Dunn (Ed.), *Oxford Bibliographies Online: Psychology*. New York: Oxford University Press.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176-199.

Tanaka, J. W. (2001). The entry point of face recognition: evidence for face expertise. *Journal of Experimental Psychology: General*, *130*(3), 534-543.

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.

Vlach, H. A., Sandhofer, C. M., & Bjork, R. A. (2014). Equal spacing and expanding schedules in children's categorization and generalization. *Journal of Experimental Child Psychology*, *123*, 129-137.

Wahlheim, C. N., & DeSoto, K. A. (2017). Study preferences for exemplar variability in self-regulated category learning. *Memory*, *25*(2), 231-243.

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.

Figure 1. A couple of examples for each broad-level category showing the fuzzy boundaries – some examples from the different broad-level categories are perceptually more similar than some examples from the same broad-level category.

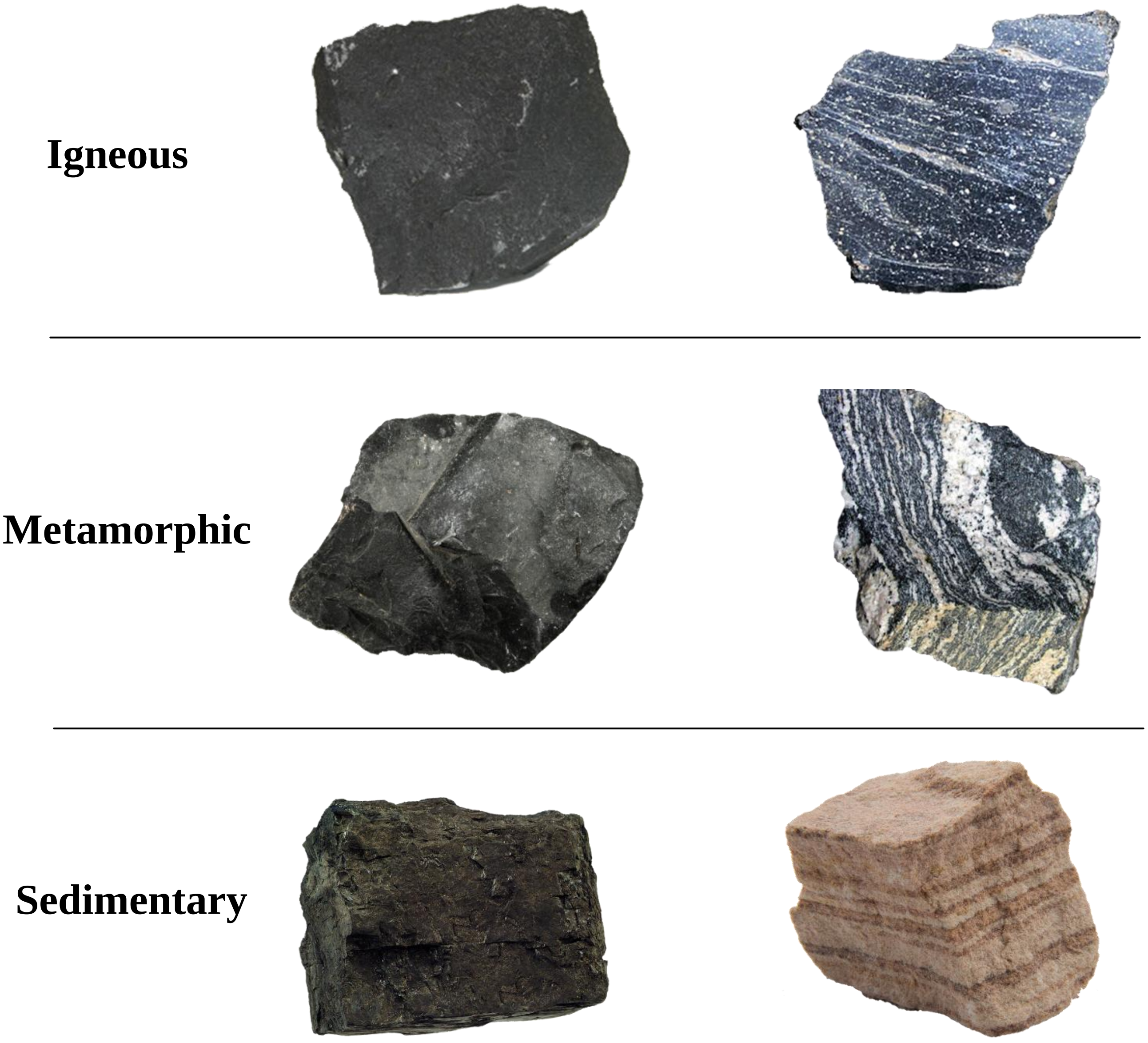


Figure 2. Examples of a few specific-level categories within the three broad-level categories illustrating high variability among them.

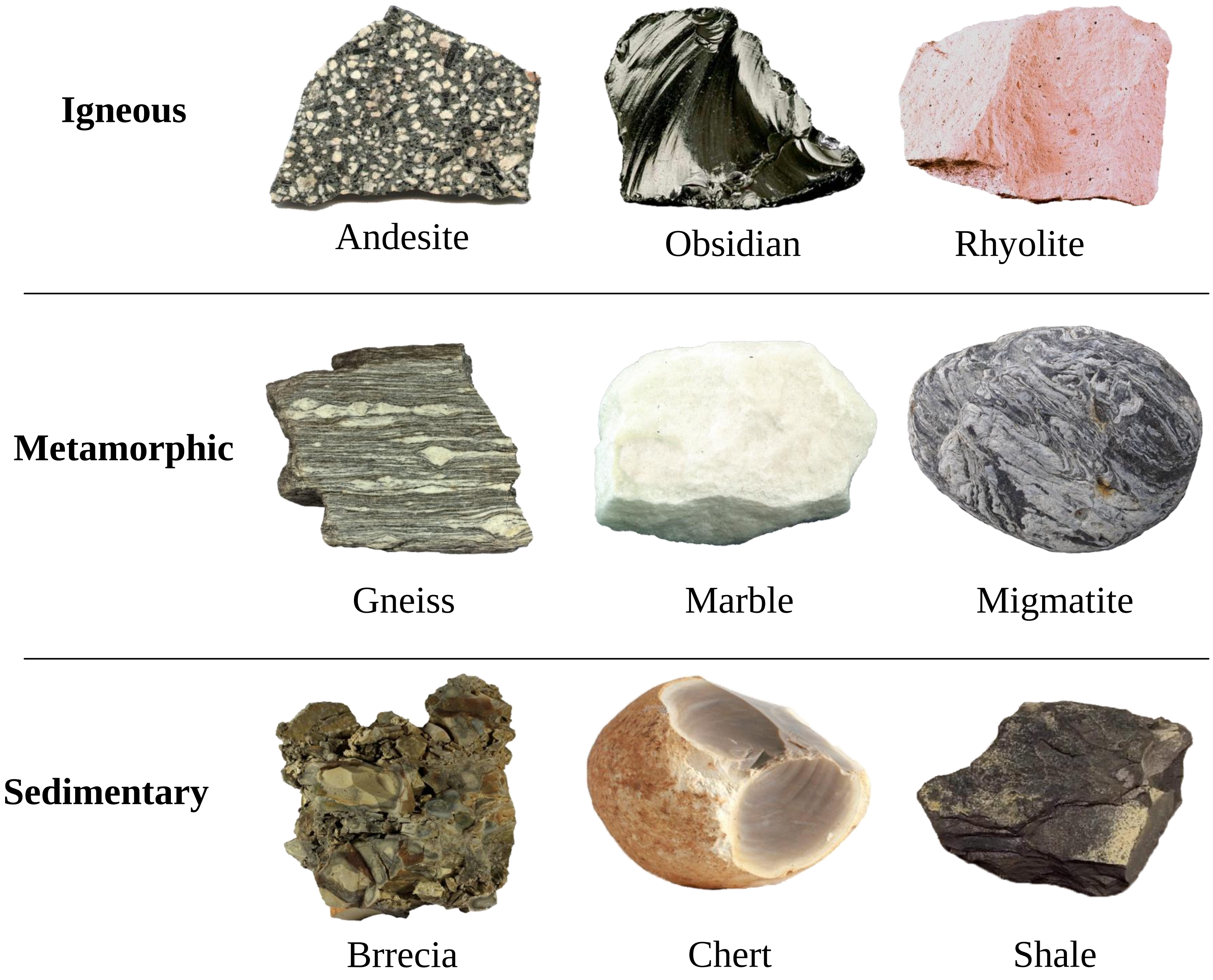


*Figure 3***.** Participants' mean performance on the final test in Experiment 1 plotted by conditions and test item types. Error bars denote ± 1 standard error.

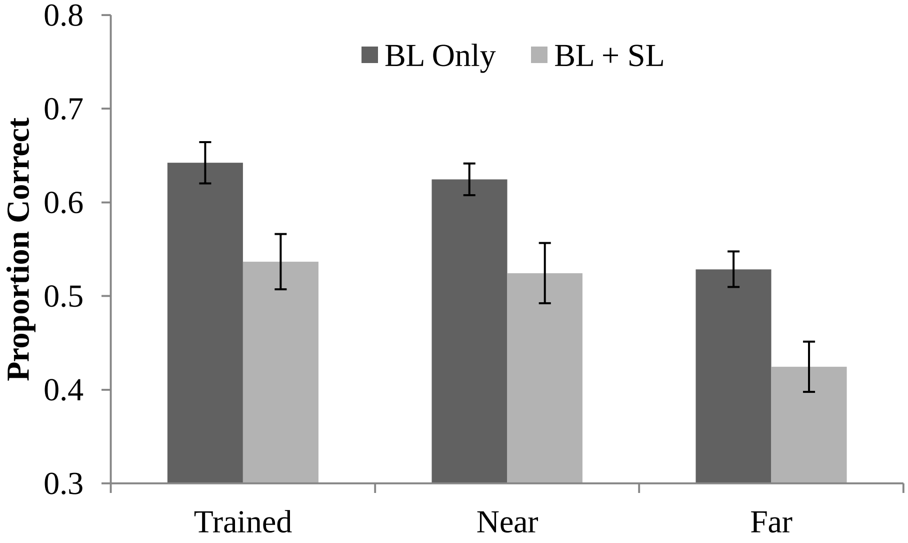


*Figure 4***.** Participants’ mean performance on the final test in Experiment 2 plotted by conditions and test item types. Error bars denote ± 1 standard error.

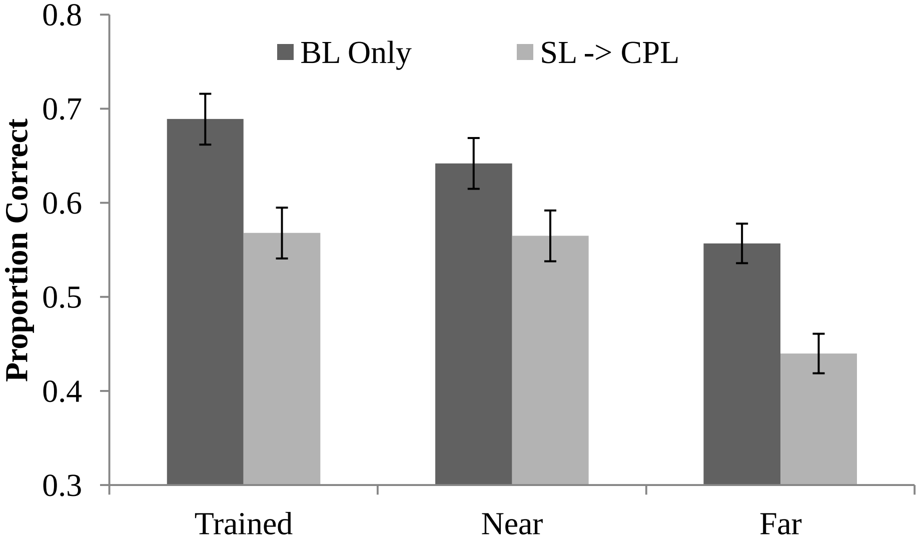


Figure 5. Learning curve in Experiment 3 as a function of conditions and feedback learning block. The performance from each of the three blocks in the U_6R_3 conditions was divided at the midpoint and presented as two blocks (i.e., 1st block into 1 and 2, 2nd block into 3 and 4, etc.), so that the performance in all conditions could be plotted on the same axis.

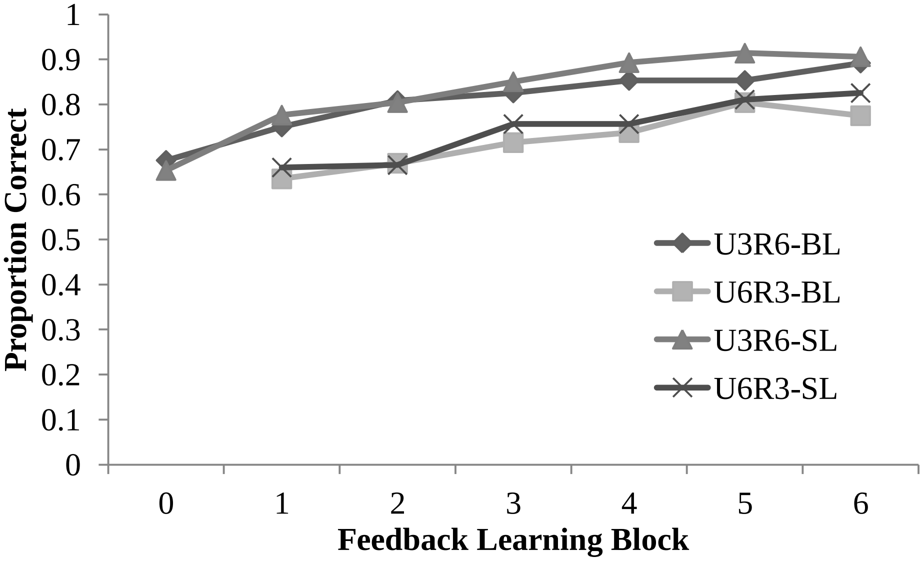


Figure 6. Participants' mean performance on the final test in Experiment 3 plotted by conditions and test item types. Error bars denote ± 1 standard error.

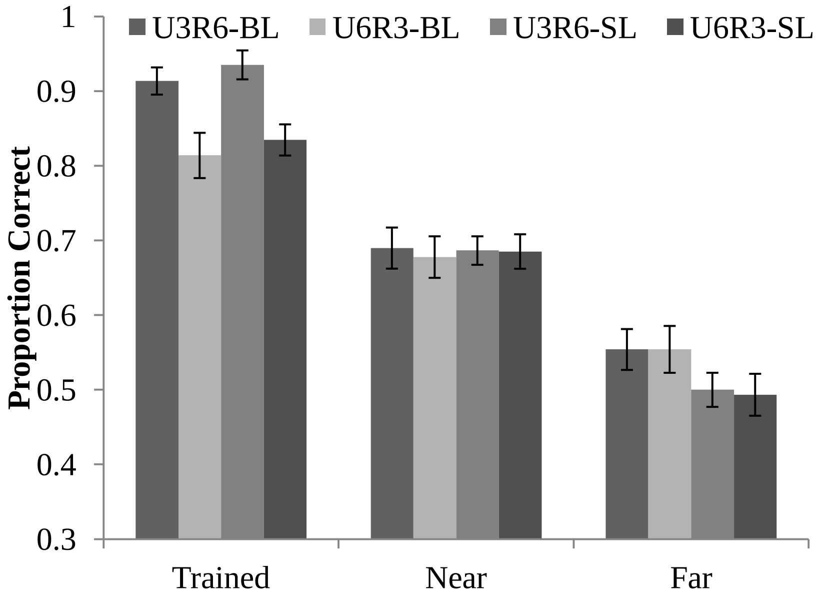


Figure 7. Learning curve in Experiment 4 as a function of conditions and feedback learning block.

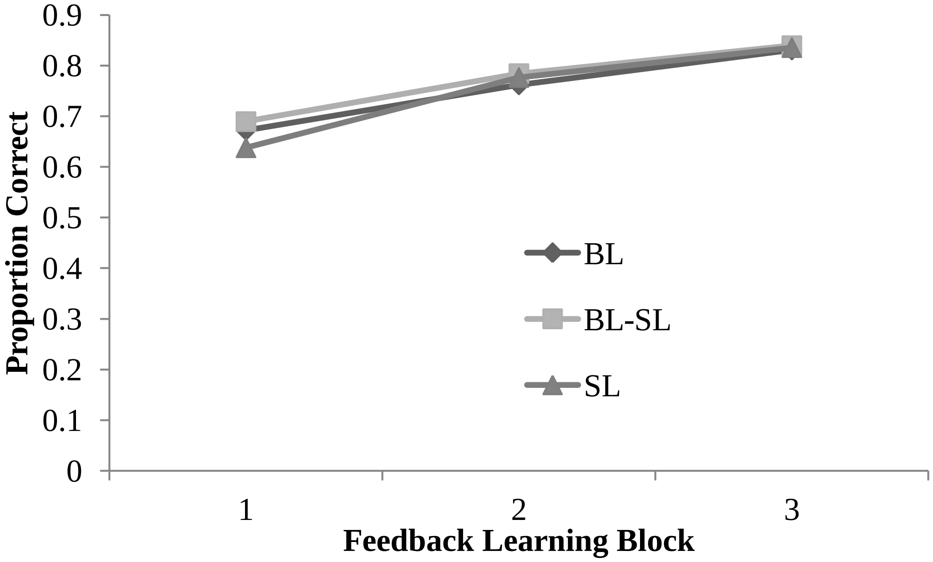


Figure 8. Participants' mean performance on the final test in Experiment 4 plotted by conditions and test item types. Error bars denote ± 1 standard error.

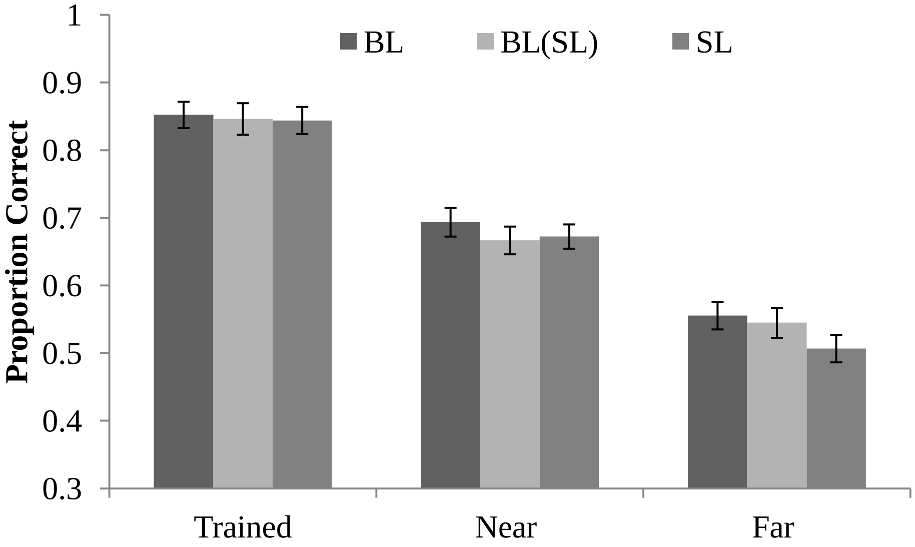


Figure 9. Schematic illustration of compact and dispersed category structures, using the example of specific-level categories of igneous (I), metamorphic (M), and sedimentary (S). Reprinted from “On learning natural-science categories that violate the family-resemblance principle.” by Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017), *Psychological Science, 28(1), 104-114*.

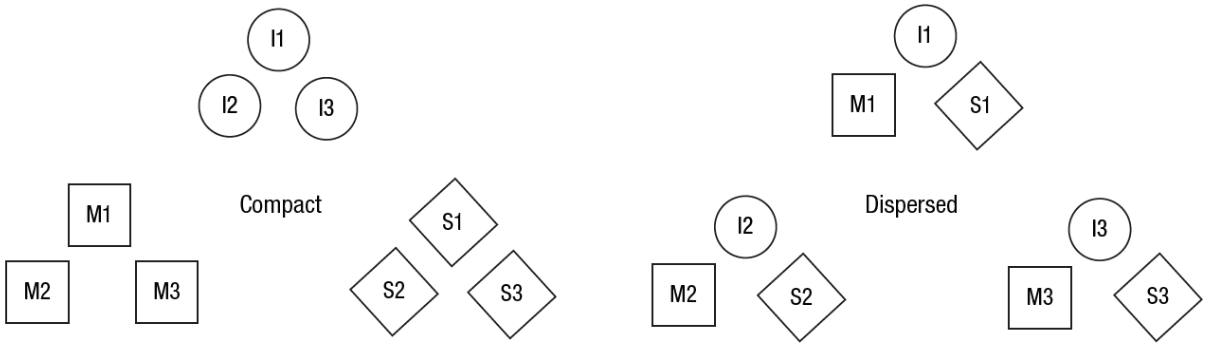


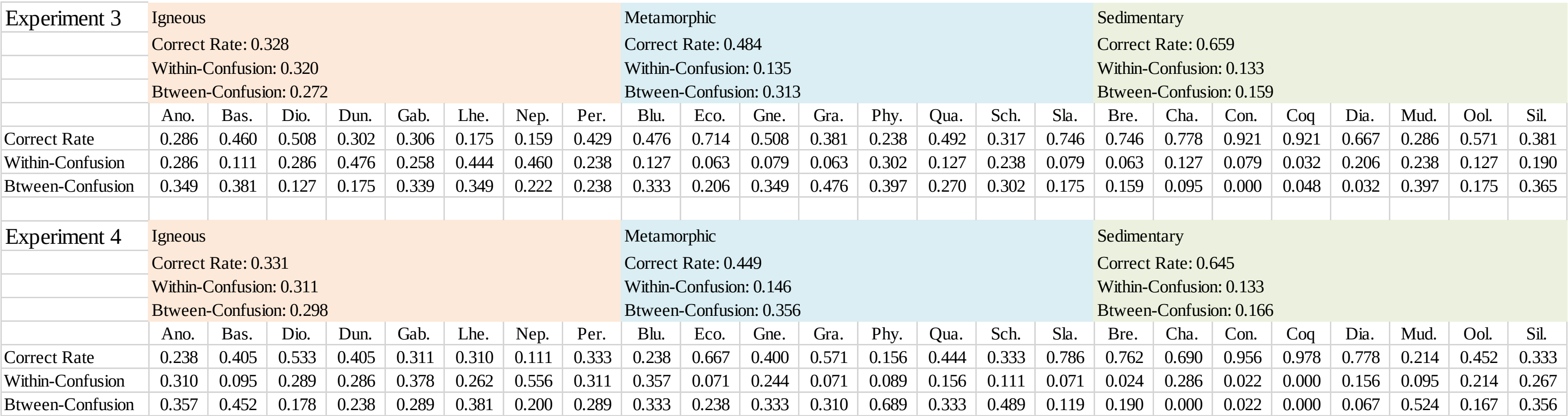
Figure 10. Confusion matrix based on the specific-level responses on the near-transfer items from the specific-level conditions in Experiment 3.

BL	Classified →	Igneous								Metamorphic								Sedimentary							
Actual ↓	SL	Ano.	Bas.	Dio.	Dun.	Gab.	Lhe.	Nep.	Per.	Blu.	Eco.	Gne.	Gra.	Phy.	Qua.	Sch.	Sla.	Bre.	Cha.	Con.	Coq	Dia.	Mud.	Ool.	Sil.
Igneous	Anorthosite	0.286	0.079	0	0.111	0	0.095	0	0	0.016	0.032	0	0.222	0	0	0	0	0.048	0.016	0	0	0	0.016	0	0
	Basalt	0.016	0.46	0	0.063	0	0.032	0	0	0.111	0	0	0.095	0	0	0	0.079	0	0	0	0	0	0.032	0.063	0
	Diorite	0	0	0.508	0	0.063	0	0.127	0.095	0	0	0.032	0	0.016	0	0.032	0	0	0	0	0	0.048	0	0	0
	Dunite	0.111	0.048	0	0.302	0	0.317	0	0	0.016	0.048	0	0.048	0	0	0	0	0	0	0	0	0	0.048	0.016	0
	Gabbro	0	0	0.097	0	0.306	0	0.113	0.048	0	0	0.032	0	0.048	0.016	0.194	0	0	0	0	0	0.016	0	0	0.032
	Lherzolite	0.095	0.111	0	0.238	0	0.175	0	0	0.079	0.079	0	0.032	0	0	0	0.016	0.079	0	0	0	0	0	0.063	0
	Nepheline S.	0	0	0.286	0	0.032	0	0.159	0.143	0	0	0.016	0	0.032	0.048	0.032	0	0	0	0	0	0.095	0	0	0
	Peridotite	0	0	0.048	0	0.048	0	0.143	0.429	0	0	0.016	0	0.016	0.032	0.048	0	0	0	0.095	0.032	0	0	0	0
Metamorphic	Blueschist	0.048	0.238	0	0.048	0	0	0	0	0.476	0.016	0	0.048	0	0	0	0.063	0	0	0	0	0	0	0	0
	Eclogite	0.048	0	0	0.032	0	0.127	0	0	0.016	0.714	0	0.048	0	0	0	0	0	0	0	0	0	0	0	0
	Gneiss	0	0	0.095	0	0.032	0	0.095	0.063	0	0	0.508	0	0.032	0.032	0.016	0	0	0	0.016	0	0.032	0	0	0.016
	Granulite	0.238	0.095	0	0.063	0	0.048	0	0	0.016	0.032	0	0.381	0	0	0	0.016	0	0	0	0	0	0.016	0.016	0
	Phyllite	0	0	0	0	0.079	0	0.032	0	0	0	0.048	0	0.238	0.032	0.222	0	0	0	0	0	0.016	0	0	0.27
	Quartzite	0	0	0	0	0.016	0	0.063	0.063	0	0	0.016	0	0.079	0.492	0.032	0	0	0	0	0	0.032	0	0	0.095
	Schist	0	0	0.063	0	0.095	0	0.048	0.063	0	0	0.095	0	0.063	0.079	0.317	0	0	0	0	0	0	0	0	0.032
	Slate	0.016	0.095	0	0.016	0	0	0	0	0.032	0	0	0.048	0	0	0	0.746	0	0	0	0	0	0.048	0	0
Sedimentary	Breccia	0.048	0.016	0	0	0	0.016	0	0	0.016	0.016	0	0.032	0	0	0	0.016	0.746	0	0	0	0	0.032	0.032	0
	Chalk	0	0.048	0	0.032	0	0	0	0	0	0	0	0	0	0	0	0.016	0	0.778	0	0	0	0.063	0.063	0
	Conglomerate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.921	0.079	0	0	0	0
	Coquina	0	0	0.032	0	0	0	0	0	0	0	0	0	0	0.016	0	0	0	0	0.032	0.921	0	0	0	0
	Diatomite	0	0	0	0	0	0	0	0	0	0	0	0	0	0.032	0	0	0	0	0	0.048	0.667	0	0	0.159
	Mudstone	0.016	0.079	0	0.063	0	0	0	0	0	0	0	0.032	0	0	0	0.206	0.048	0.095	0	0	0	0.286	0.095	0
	Oolite	0.016	0.048	0	0.111	0	0	0	0	0	0	0	0	0	0	0	0	0	0.048	0	0	0	0.079	0.571	0
	Siltstone	0	0	0	0	0.111	0	0	0.016	0	0	0.016	0	0.111	0.079	0.032	0	0	0	0	0	0.19	0	0	0.381

Figure 11. Confusion matrix based on the specific-level responses on the near-transfer items from the specific-level conditions in Experiment 4.

BL	Classified →	Igneous								Metamorphic								Sedimentary							
Actual ↓	SL	Ano.	Bas.	Dio.	Dun.	Gab.	Lhe.	Nep.	Per.	Blu.	Eco.	Gne.	Gra.	Phy.	Qua.	Sch.	Sla.	Bre.	Cha.	Con.	Coq	Dia.	Mud.	Ool.	Sil.
Igneous	Anorthosite	0.238	0.119	0	0.048	0	0.143	0	0	0.095	0	0	0.214	0	0	0	0	0.048	0	0	0	0	0	0	0
	Basalt	0.024	0.405	0	0.048	0	0.024	0	0	0.024	0	0	0.048	0	0	0	0.238	0	0.024	0	0	0	0.024	0.095	0
	Diorite	0	0	0.533	0	0.067	0	0.111	0.111	0	0	0.067	0	0.044	0	0.022	0	0	0	0	0	0.022	0	0	0.022
	Dunite	0.048	0	0	0.405	0	0.238	0	0	0.024	0.024	0	0.048	0	0	0	0	0	0.024	0	0	0	0.024	0.095	0
	Gabbro	0	0	0.222	0	0.311	0	0.111	0.044	0	0	0.022	0	0.089	0.022	0.156	0	0	0	0	0	0	0	0	0
	Lherzolite	0	0.119	0	0.143	0	0.31	0	0	0	0.095	0	0.119	0	0	0	0	0.095	0	0	0	0	0	0.071	0
	Nepheline S.	0	0	0.267	0	0.089	0	0.111	0.2	0	0	0	0	0.044	0.111	0.022	0	0	0	0	0	0.022	0	0	0
	Peridotite	0	0	0.067	0	0.156	0	0.089	0.333	0	0	0.022	0	0.067	0.044	0.067	0	0	0	0.044	0.044	0	0	0	0
Metamorphic	Blueschist	0.048	0.167	0	0.024	0	0.048	0	0	0.238	0.048	0	0.143	0	0	0	0.167	0	0	0	0	0	0.024	0.024	0
	Eclogite	0.024	0	0	0	0	0.19	0	0	0	0.667	0	0.071	0	0	0	0	0.024	0	0	0	0	0	0	0
	Gneiss	0	0	0.067	0	0.044	0	0.111	0.111	0	0	0.4	0	0.111	0.067	0.067	0	0	0	0	0	0	0	0	0
	Granulite	0.143	0.071	0	0.048	0	0.024	0	0	0.024	0.048	0	0.571	0	0	0	0	0.024	0	0	0	0	0	0	0
	Phyllite	0	0	0.044	0	0.133	0	0.022	0.022	0	0	0.022	0	0.156	0.022	0.044	0	0	0	0	0	0.022	0	0	0.444
	Quartzite	0	0	0	0	0.044	0	0.067	0.022	0	0	0	0	0.133	0.444	0.022	0	0	0	0	0	0.089	0	0	0.111
	Schist	0	0	0.089	0	0.267	0	0.044	0.067	0	0	0.067	0	0	0.044	0.333	0	0	0	0	0	0	0	0	0.022
	Slate	0.048	0.048	0	0	0	0.024	0	0	0.024	0.048	0	0	0	0	0	0.786	0	0	0	0	0	0	0	0
Sedimentary	Breccia	0.024	0.024	0	0	0	0.048	0	0	0	0	0	0.095	0	0	0	0	0.762	0	0	0	0	0.024	0	0
	Chalk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.69	0	0	0	0.214	0.071	0
	Conglomerate	0	0	0	0	0	0	0	0.022	0	0	0	0	0	0	0	0	0	0	0.956	0.022	0	0	0	0
	Coquina	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.978	0	0	0	0
	Diatomite	0	0	0	0	0	0	0.044	0	0	0	0	0	0.022	0	0	0	0	0	0	0.022	0.778	0	0	0.133
	Mudstone	0.024	0.071	0	0.071	0	0.024	0	0	0	0.024	0	0.048	0	0	0	0.262	0.024	0.071	0	0	0	0.214	0	0
	Oolite	0	0.024	0	0.048	0	0.048	0	0	0	0	0	0.024	0	0	0	0.024	0.024	0.143	0	0	0	0.048	0.452	0
	Siltstone	0	0	0.022	0	0.089	0	0	0	0	0	0.022	0	0.089	0.067	0.067	0	0	0	0.022	0.022	0.222	0	0	0.333

Figure 12. Summary of the confusion matrices shown in Figures 10 and 11 in terms of the correct rate, the sum of within-broad-category confusion, and the sum of between-broad-category confusion, for each specific-level category.



Experiment 3	Igneous								Metamorphic								Sedimentary							
	Correct Rate: 0.328								Correct Rate: 0.484								Correct Rate: 0.659							
	Within-Confusion: 0.320								Within-Confusion: 0.135								Within-Confusion: 0.133							
	Btween-Confusion: 0.272								Btween-Confusion: 0.313								Btween-Confusion: 0.159							
	Ano.	Bas.	Dio.	Dun.	Gab.	Lhe.	Nep.	Per.	Blu.	Eco.	Gne.	Gra.	Phy.	Qua.	Sch.	Sla.	Bre.	Cha.	Con.	Coq	Dia.	Mud.	Ool.	Sil.
Correct Rate	0.286	0.460	0.508	0.302	0.306	0.175	0.159	0.429	0.476	0.714	0.508	0.381	0.238	0.492	0.317	0.746	0.746	0.778	0.921	0.921	0.667	0.286	0.571	0.381
Within-Confusion	0.286	0.111	0.286	0.476	0.258	0.444	0.460	0.238	0.127	0.063	0.079	0.063	0.302	0.127	0.238	0.079	0.063	0.127	0.079	0.032	0.206	0.238	0.127	0.190
Btween-Confusion	0.349	0.381	0.127	0.175	0.339	0.349	0.222	0.238	0.333	0.206	0.349	0.476	0.397	0.270	0.302	0.175	0.159	0.095	0.000	0.048	0.032	0.397	0.175	0.365
Experiment 4	Igneous								Metamorphic								Sedimentary							
	Correct Rate: 0.331								Correct Rate: 0.449								Correct Rate: 0.645							
	Within-Confusion: 0.311								Within-Confusion: 0.146								Within-Confusion: 0.133							
	Btween-Confusion: 0.298								Btween-Confusion: 0.356								Btween-Confusion: 0.166							
	Ano.	Bas.	Dio.	Dun.	Gab.	Lhe.	Nep.	Per.	Blu.	Eco.	Gne.	Gra.	Phy.	Qua.	Sch.	Sla.	Bre.	Cha.	Con.	Coq	Dia.	Mud.	Ool.	Sil.
Correct Rate	0.238	0.405	0.533	0.405	0.311	0.310	0.111	0.333	0.238	0.667	0.400	0.571	0.156	0.444	0.333	0.786	0.762	0.690	0.956	0.978	0.778	0.214	0.452	0.333
Within-Confusion	0.310	0.095	0.289	0.286	0.378	0.262	0.556	0.311	0.357	0.071	0.244	0.071	0.089	0.156	0.111	0.071	0.024	0.286	0.022	0.000	0.156	0.095	0.214	0.267
Btween-Confusion	0.357	0.452	0.178	0.238	0.289	0.381	0.200	0.289	0.333	0.238	0.333	0.310	0.689	0.333	0.489	0.119	0.190	0.000	0.022	0.000	0.067	0.524	0.167	0.356

Appendix A

The rock pictures used in the current experiments. Visit Open Science Framework (https://osf.io/n2awt) for high resolution version of these images.

Igneous

Anorthosite

Basalt

Diorite

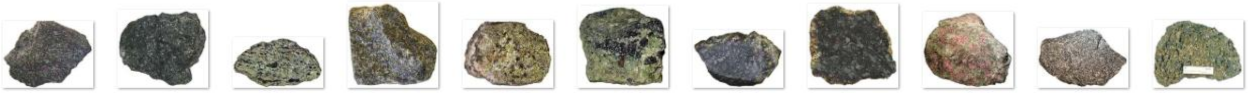
Dunite

Gabbro

Lherzolite

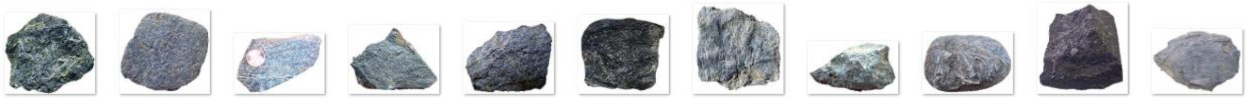
Nepheline Syenite

Peridotite

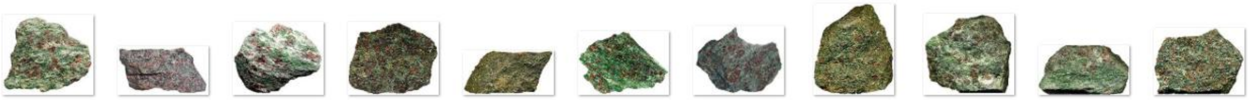


Metamorphic

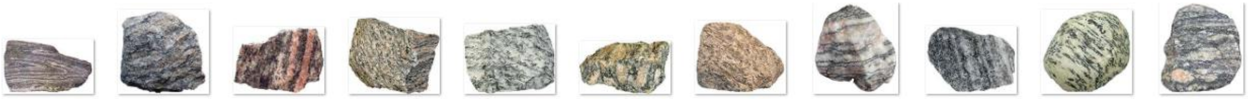
Blueschist



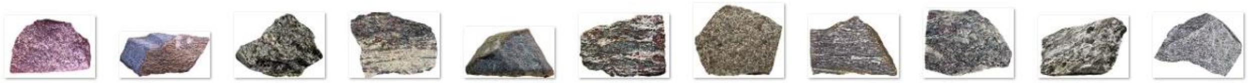
Eclogite



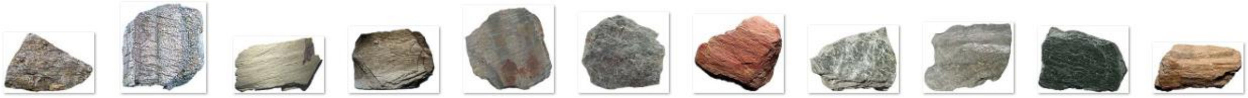
Gneiss



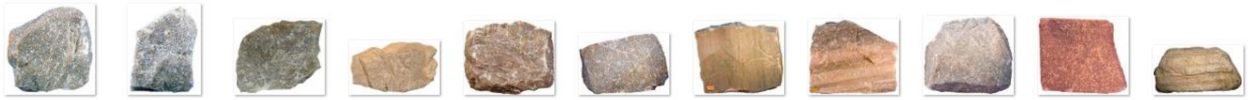
Granulite



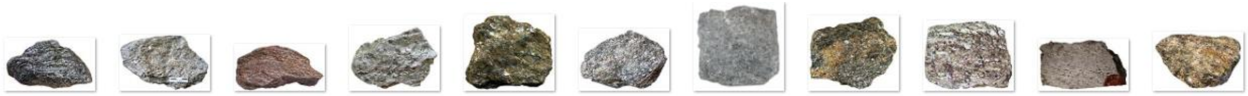
Phyllite



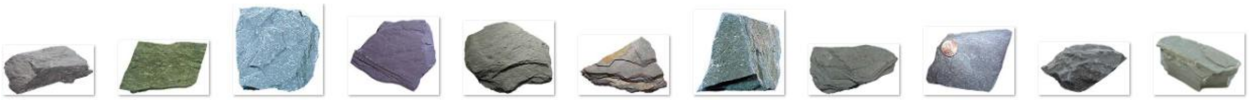
Quartzite



Schist

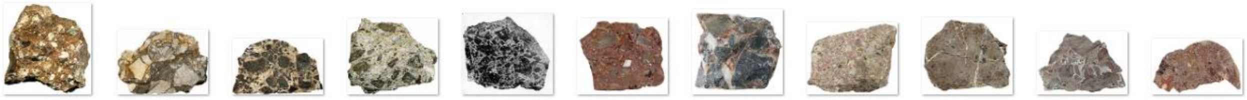


Slate

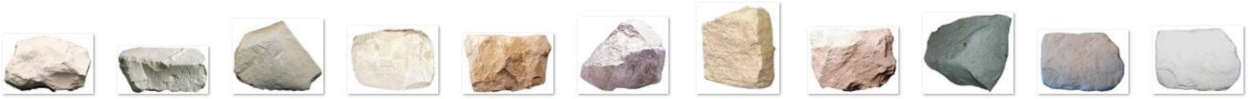


Sedimentary

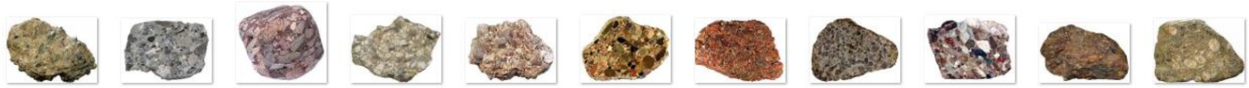
Breccia



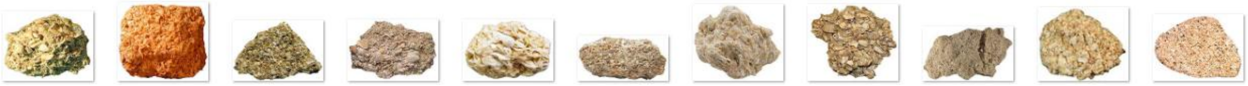
Chalk



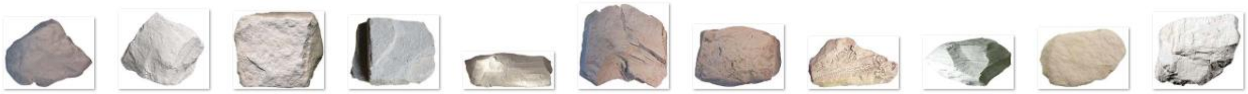
Conglomerate



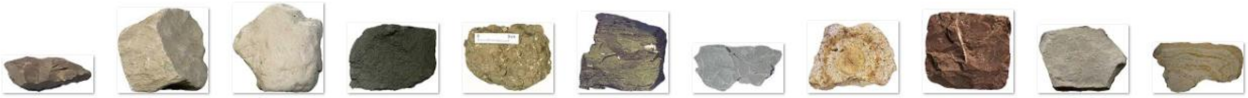
Coquina



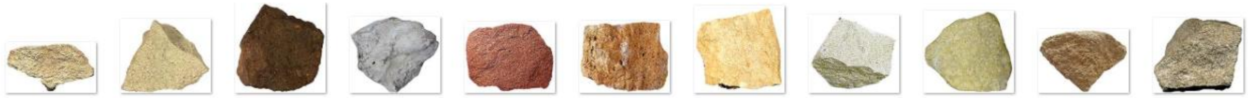
Diatomite



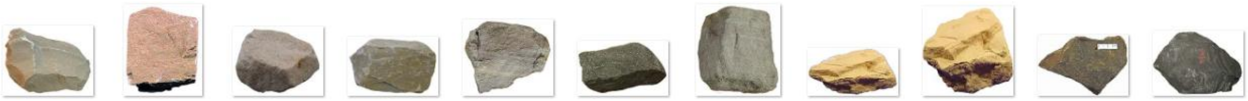
Mudstone



Oolite



Siltstone



Appendix B

Results and discussion regarding the “I don’t know” responses for all four experiments.

Experiment 1

Table B1 shows the proportion of “I don’t know” responses out of the total response in each condition. A 2 X 3 mixed ANOVA with the study condition (BL-only or BL + SL) as the between-subjects variable and the test item type (trained, near-transfer, or far-transfer) as the within-subjects variable was conducted on these data. The proportion of “I don’t know” responses did not differ as a function of the study condition, *F*(1, 56) < 1, or item type *F*(2, 112) < 1, and the interaction between the two variables was not significant, *F*(2, 112) = 2.89, *p* > .05, ηp^2 = .05.

Experiment 2

Table B2 shows the proportion of “I don’t know” responses out of the total response in each condition. Similarly to Experiment 1, a 2 X 3 mixed ANOVA with the study condition (BL-only or SL -> CPL) as the between-subjects variable and the test item type (trained, near-transfer, or far-transfer) as the within-subjects variable was conducted on these data. There was a significant main effect of item type, *F*(2, 100) = 4.12, *p* < .05, ηp^2 = .08. Post-hoc t-tests revealed that there were greater number of “I don’t know” responses for near-transfer (*M* = .04, *SD* = .08) and far-transfer items (*M* = .05, *SD* = .09) than in trained items (*M* = .02, *SD* = .05), *t*(51) = 2.32, *p* > .05, *d* = 0.23; *t*(51) = 2.62, *p* > .05, *d* = 0.29, while near- and far- transfer items did not significantly differ.

Experiment 3

Table B3 shows the proportion of “I don’t know” responses out of the total response in each condition. A 2 X 2 X 3 mixed ANOVA, with the level of categorization (BL or SL) and the number of unique training exemplars (3 or 6) treated as the between-subjects variables and the test item type (trained, near-transfer, or far-transfer) treated as the within-subjects variable, was conducted on these data. There was a significant main effect of item type, *F*(2, 152) = 13.81, *p* < .001, ηp^2 = .15, such that there were greater number of “I don’t know” responses for far-transfer items (*M* = .06, *SD* = .13) than near-transfer (*M* = .04, *SD* = .10), *t*(79) = 2.65, *p* < .05, *d* = 0.11 and trained items (*M* = .01, *SD* = .04), *t*(79) = 3.81, *p* < .001, *d* = 0.46, and for near-transfer items than trained items *t*(79) = 3.81, *p* < .001, *d* = 0.42. In addition, there was a significant main effect of the level of categorization, such that the number of “I don’t know” responses was greater in the specific-level training conditions (*M* = .06, *SD* = .10) than in the broad-level conditions (*M* = .01, *SD* = .11), *F*(1, 76) = 8.54, *p* < .01, ηp^2 = .10. Lastly, there was a significant training level by item type interaction suggesting that the difference between the item types were more prominent in the specific-level conditions than in the broad-level conditions, *F*(2, 152) = 8.85, *p* < .01, ηp^2 = .10. The main effect of the number of unique training exemplars, the interaction between the number of unique training exemplars and the level of categorization, and the three-way interaction was not significant (*F*s < 1).

Experiment 4

Table B4 shows the proportion of “I don’t know” responses out of the total response in each condition. Similarly to Experiment 1, a 3 X 3 mixed ANOVA with the training condition (BL, BL-blocked specific, or SL) as the between-subjects variable and the test item type (trained, near-transfer, or far-transfer) as the within-subjects variable was conducted on these data. There was a significant main effect of item type, *F*(2, 166) = 11.96, *p* < .001, ηp^2 = .13. Post-hoc t-tests revealed that there were greater number of “I don’t know” responses for near-transfer (*M* = .02, *SD* = .05) and far-transfer items (*M* = .04, *SD* = .10) than in trained items (*M* = .01, *SD* = .03), *t*(85) = 2.26, *p* < .05, *d* = 0.28; *t*(85) = 3.74, *p* < .001, *d* = 0.44; there were also greater number of “I don’t know” responses for far-transfer than near-transfer items, *t*(85) = 3.12, *p* < .01, *d* = 0.26. The main effect of the training condition was also significant, *F*(2, 83) = 6.35, *p* < .01, ηp^2 = .13. Post-hoc Tukey tests showed that the number of “I don’t know” responses was greater in the SL condition (*M* = .05, *SD* = .10) than in the BL (*M* = .01, *SD* = .04), *p* < .01, *d* = 0.26, and BL-blocked specific conditions (*M* = .01, *SD* = .05), *p* < .01, *d* = 0.26, while there was no difference between the BL and the BL-blocked specific conditions, *p* > .05. Lastly, there was a significant study condition by item type interaction, *F*(4, 166) = 6.95, *p* < .001, ηp^2 = .14, suggesting that the difference between the item types were more prominent in the SL condition than in the broad-level conditions.

In summary, the proportion of “I don’t know” responses tended to be greater in far-transfer items (Experiments 2-4) and in specific-level training conditions when feedback learning paradigm was employed (Experiments 3 & 4). In addition, the effect of item type (i.e., more “I don’t know” responses in far-transfer) was more prominent in specific-level training conditions when feedback learning paradigm was employed (Experiments 3 & 4) as indicated by the significant item type by training level interactions. The frequency of “I don’t know” responses roughly reflected the difficulty of the task in the given condition, such that there were greater number of “I don’t know” responses when the classification accuracy was low (i.e., for far-transfer items and in specific-level conditions). Finally, the high standard deviations observed across all experiments suggest that the participants considerably differed in the way they used the “I don’t know” responses.

Table B1. Proportion of “I don’t know” responses as a function of conditions and item types in Experiment 1. The value inside of the parentheses denotes standard deviations.

Conditions	Trained	Near	Far
BL Only	.02 (.05)	.01 (.04)	.03 (.06)
BL + SL	.01 (.02)	.02 (.04)	.01 (.03)

Table B2. Proportion of “I don’t know” responses as a function of conditions and item types in Experiment 2. The value inside of the parentheses denotes standard deviations.

Conditions	Trained	Near	Far
BL Only	.02 (.04)	.04 (.08)	.04 (.07)
SL -> CPL	.03 (.06)	.04 (.08)	.06 (.11)

Table B3. Proportion of “I don’t know” responses as a function of conditions and item types in Experiment 3. The value inside of the parentheses denotes standard deviations.

Training Level	# of Exemplars	Trained	Near	Far
BL	U_3R_6	.00 (.00)	.00 (.01)	.00 (.00)
BL	U_6R_3	.00 (.01)	.01 (.04)	.02 (.06)
SL	U_3R_6	.01 (.04)	.07 (.11)	.09 (.15)
SL	U_6R_3	.02 (.05)	.06 (.12)	.08 (.16)

Table B4. Proportion of “I don’t know” responses as a function of conditions and item types in Experiment 4. The value inside of the parentheses denotes standard deviations.

Conditions	Trained	Near	Far
BL	.00 (.02)	.01 (.03)	.02 (.07)
BL – blocked specific	.01 (.05)	.00 (.01)	.01 (.06)
SL	.01 (.03)	.05 (.08)	.09 (.14)