Category Structure and Recognition Memory

Yasuaki Sakamoto (yasu@psy.utexas.edu)
Bradley C. Love (love@psy.utexas.edu)
Department of Psychology
Austin, TX 78712 USA

Abstract

Current models of human category learning and subsequent recognition are either exemplar-based, rule-based, or some combination of both approaches. We present learning and recognition data that cannot be accounted for by current approaches. The data suggest that the degree to which an item is remembered is determined by the strength of the expectation it violates. In our study, expectations take the form of simple, imperfect rules where the strength of the rules are determined by the number of items that follow the rules in training. Exemplar-based models cannot account for the results because they do not posit organizing knowledge structures that can be violated. The frequency insensitivity of rule-based accounts leads to their failure. We propose a cluster-based approach that is consistent with our findings, as well as findings from schema, stereotype, and basic memory research.

Introduction

Our ability to successfully categorize underlies many of our cognitive abilities. Consequently, there has been a great deal of interest in understanding how we acquire categories from examples. Acquiring new categories necessarily involves changes in memory. The present work asks what is stored in memory as a result of category learning. Specifically, the current work explores the effect of category structure on recognition memory.

While understanding the determinants of recognition is interesting in its own right, data on recognition memory performance also provide constraints for theories of categorization. For example, multiple memory system theories of category learning marshall support from dissociations such as amnesics' above chance categorization performance in the absence of recognition (Knowlton & Squire, 1993; Squire & Knowlton, 1995; see, however, Palmeri & Flanery, 1999). Category learning research plays an important role in such debates. For instance, Nosofsky and Zaki's (1998) simulations of an exemplar model cast some doubt on Knowlton and Squire's (1993) interpretation of their data. Nosofsky and Zaki's (1998) approaches are typical of other work in category learning that posits that the same representations subserve both categorization and recognition.

In this paper, we present new category learning and recognition data that cannot be accounted for by current exemplar- and rule-based models of category learning and recognition. The results are consistent with a cluster-based approach to categorization and recognition. The clustering approach is in accord with work exploring the role of schemas in memory. In the remainder of the paper, we will discuss current approaches to category learning and recognition, present data and model fits that are inconsistent with these approaches, and discuss an alternative cluster-based approach.

Previous Research in Category Learning

Palmeri and Nosofsky's (1995) studies of category learning and subsequent recognition are perhaps the most challenging for models to address. In their studies, an imperfect rule successfully classified the majority of study items (e.g., most small items were in category A, whereas most large items were in category B), but two exception items violated the rule (e.g., a large item that was a member of category A). Their basic finding was that the exceptions were recognized best.

Palmeri and Nosofsky modeled their data with the context model (Medin & Schaffer, 1978) and the RULEX (rule-plus-exception) model of category learning (Nosofsky, Palmeri, & McKinley, 1994). The context model is an exemplar model that stores every studied item in memory as a separate trace. Items are probabilistically classified into category A or B depending on the item's relative similarity to all exemplars belonging to categories A and B. The likelihood of recognizing a stimulus as a studied item is proportional to the sum of similarity to all exemplars (from both categories A and B).

While the context model correctly predicts better recognition for studied items than for novel items, it cannot account for the enhanced recognition of exceptions. This failure arises because the exceptions share the same similarity relations with other items in memory as rule-following items do. Exceptions are distinguished from rule-following items because the exceptions' category assignment runs counter to the rule. According to the context model, this reversal is not germane to recognition.

Palmeri and Nosofsky (1995) had more success with the RULEX model. RULEX is a hypothesis testing model of category learning that constructs rules and stores exceptions to the rules. Rule-following items are not individually stored, but rather are captured by the rule. Information about inconsistent items is explicitly stored. The likelihood of recognizing a test item is determined by summing the response from RULEX's rule system and the items in the exception store. The storage of rule-violating items allows RULEX to predict a memory advantage for exceptions. However, RULEX under-predicts the recognition advantage of rule-following studied items relative to novel items as neither class of items resembles items in the exception store. In order to address this shortcoming, Palmeri and Nosofsky created a combined model that generates recognition responses by summing the responses of RULEX and the context model (which is sensitive to the difference between studied rule-following items and novel items). This combined model does a good job of accounting for the learning and recognition data.

Related Research in Memory

Palmeri and Nosofsky's (1995) findings suggest that items that stand in opposition to a salient knowledge structure (in this case a rule) are remembered better. Such special status of violating items is also suggested by schema (Rojahn & Pettigrew, 1992) and stereotype (Stangor & McMillan, 1992) research. If violating a known regularity, such as a rule, leads to improved recognition, how does the nature of the regularity affect memory for deviant items, such as exceptions?

One possibility is that the strength or coherency of the knowledge structures determines the degree to which violating items are remembered. In support of this position, Koffka (1935) reported that when there were more anomalous items in a list, the memory advantage for those items was smaller. Similarly, Rojahn and Pettigrew's (1992) meta-analysis suggests that the memory advantage for the schema-inconsistent items is weaker when the proportion of those items is larger. These findings argue against a strict rule-based knowledge organization because a central property of rules is insensitivity to frequency information (Pinker, 1991; Smith, Langston, & Nisbett, 1992).

Experiment

The present experiment tests our prediction that the strength of a regularity determines the degree to which violating items are remembered. Rule strength is manipulated in a within-subjects design by varying the frequency of rule-following items for categories A and B during study. Each category contains a single exception. We predict that the exception of the smaller category (B1 in Table 1) will be remembered better than the exception of the larger category (A1) because the exception of the smaller category (B1) must be differentiated from the membership rule of the larger category (i.e., the stronger of the two "rules"). That is, the exception of the smaller category (B1) must be differentiated from many rule-following items (A2-A9), whereas the exception of the large category (A1) has to be differentiated from only a few rule-following items (B2-B5). Such comparison based on rule dimension values (e.g., comparison between B1 and A2-A9) has to take place in order to master the category memberships and compare items within the same categories (e.g., compare B1 with B2-B5). As a result, the exception of the smaller category (B1) will result in better recognition than the exception of the large category (A1). This prediction conflicts with RULEX (as well as the context model and the combined model). RULEX predicts that the number of rule-following items should not influence memory for exceptions. Exceptions in both the small and the large categories are equally likely to enter the exemplar store and be remembered.

Method

Participants Eighty-two University of Texas undergraduates participated for course credit.

Apparatus The experiment was run on Pentium III computers operating in DOS. Data were collected using an in-house real-time data collection system. The monitors had 15 inch CRT color displays and a refresh rate of 16.67 ms.

Stimuli The stimuli were 24 computer-generated squares varying along five binary-valued dimensions: size (large or small), border color (yellow or white), texture (smooth or dotted), slash (present or absent), and main color (blue or purple). Psychologically, the five dimensions are independent and equally salient as verified by the multi-dimensional scaling of similarity ratings. The abstract structure of the stimuli is displayed in Table 1. The assignment of the abstract dimensions to the physical dimensions, the binary values, and category labels were randomized for each subject. The stimuli can be downloaded at http://love.psy.utexas.edu/stimuli/.

Design and Overview Subjects completed a learning phase consisting of classification learning trials of the items under the heading "Learning Items" in Table 1. Trials were organized in blocks, where each block is a presentation of each stimulus in random order. Subjects completed 20 blocks of learning trials. After the learning phase, subjects completed a filler phase consisting of three arithmetic problems in order to prevent rehearsal of information from the learning phase. Then, subjects completed a recognition phase consisting of forced

Table 1: The abstract structure of the categories. There is an imperfect rule on the first dimension. The Stimuli subsection details the physical dimensions of the stimuli.

Learning Items	Dimension Values	Novel Items	Dimension Values
Category A			
A1	21112	N1	11221
A2	12122	N2	12112
A3	11211	N3	12221
A4	12211	N4	12212
A5	11122	N5	12222
A6	12111	N6	21221
A7	11222	N7	22112
A8	11212	N8	22221
A9	12121	N9	22212
Category B		N10	22222
B1	11121		
B2	22122		
B3	21211		
B4	22211		
B5	21122		

choice recognition judgments involving items from the learning phase and novel stimuli. Finally, subjects completed a transfer phase in which they classified both the items from the learning phase and the novel items presented in the recognition phase without corrective feedback.

The variables of primary interest were the item type (either rule-following or exception) and the category size (either small or large) in the learning phase. The rule-following items (A2-A9 and B2-B5) followed an imperfect category rule (see Table 1 for the imperfect rule on the first dimension), and two exception items (A1 and B1), one from each category, violated the rule. Following Medin and Smith (1981) and Palmeri and Nosofsky's (1995) Experiment 1, subjects were provided with a hint to attend to the first dimension. The hint was provided in order to ensure that subjects engaged in the rule-plus-exception strategy.

The recognition phase involved two-alternative forced choice (2AFC) judgments on 50 pairs of stimuli presented in a random order. Each pair consisted of a studied item from the learning phase and a novel item they had never seen before. Ten studied items, five items from category A (A1-A5) and the other five from category B (B1-B5), and ten novel items displayed under Novel Items in Table 1 were used. Because items with value 1 on the first dimension were more frequent than items with value 2 in the learning phase, the false alarm rate for recognizing the items with value 1 on the first dimension would

be higher than items with value 2. One way to overcome this bias was to pair the items with the same value on the first dimension. Thus, each of the five studied items with value 1 on the first dimension (i.e., A2-A5 and B1) was paired with each of the five novel items with value 1 on the first dimension (i.e., N1-N5), which resulted in 25 pairs. Another set of 25 pairs was created in the same manner using the items with value 2 on the first dimension.

In the transfer phase, subjects classified the same 20 stimuli used in the recognition phase without corrective feedback. Subjects completed two blocks of transfer trials.

Procedure The general introduction and instructions were presented on the monitor at the beginning of the experiment. The specific instructions for the learning, filler, recognition, and transfer phases were displayed on the monitor immediately before subjects started each phase. The background color was black during the entire experiment. The stimuli were presented in a different random order for each subject in all of the phases.

On each trial in the learning phase, one stimulus appeared at the center of the monitor, and the text "Category A or B?" was displayed above the stimulus. If size was the rule dimension, the hint "Look whether the size is small or large." appeared above the text. Subjects indicated their category membership judgment by pressing the A or the B key. After responding, the text and the hint above the stimulus were replaced with visual (e.g., "Right! The correct answer is A.", "Wrong! The correct answer is B.") and auditory corrective feedback (i.e., a low pitch tone for errors and a high pitch tone for correct responses). The stimulus and the visual feedback was displayed for 2501 ms (150 screen refreshes) after responding. Then, a blank screen was displayed for 834 ms (50 screen refreshes) and the next trial began.

After completing the learning phase, subjects were presented with a series of three arithmetic problems. Each problem consisted of two integers (randomly generated between 10 and 49) presented side by side (e.g., 22 + 34 = ?) and the problem remained displayed until subjects responded. The subjects received both auditory and visual feedback indicating whether they added the numbers correctly.

A pair of stimuli was presented on each trial in the recognition phase. Each pair consisted of a studied item and a novel item as described earlier. The two stimuli were displayed side by side at the center of the monitor together with the text "Old: left (Q) or right (P)?" above the stimuli. Subjects pressed the Q key if they thought the studied item was on the left. They pressed the P key if they thought the studied item was on the right. For each pair, the studied and novel items were randomly assigned to the left or the right position. No corrective feedback

was given to the subjects. Instead, a high pitch tone was presented after the subjects responded together with the text "Thank You" below the stimulus. The pair of stimuli and the texts were displayed for 2501 ms (150 screen refreshes) after responding. Then, a blank screen was displayed for 834 ms (50 screen refreshes) and the next trial began.

The transfer phase followed the recognition phase. The procedure for the transfer phase was identical to that for the learning phase except that subjects received no hint or feedback. After responding A or B, a high pitch tone was presented and the text "Thank You" appeared below the stimulus.

Results

One subject who performed at the chance level of 50% in the learning phase was excluded from further analysis. Including this subject does not change the pattern of the results. Table 2 displays the subjects' performance on different items in the learning, recognition, and transfer phases.

Subjects' overall accuracy in the learning phase (.84) was significantly better than the chance (.50), $t(80) = 48.87$, $p < .001$. The overall recognition accuracy (.72) was significantly better than the chance (.50), $t(80) = 14.53$, $p < .001$. Although no feedback was provided in the transfer phase, for the purposes of analyses novel items were considered to be in the category for which they satisfied the imperfect rule. The overall accuracy in the transfer phase (.83) was significantly better than the chance (.50), $t(80) = 13.45$, $p < .001$, which suggests a high occurrence of rule application in classifying stimuli.

Table 2: Mean accuracies in the learning, recognition, and transfer phases are shown. Xcept S is the exception of the small category, Xcept L is the exception of the large category, Rules S are the rule-following items of the small category, and Rules L are the rule-following items of the large category.

Item Types	Learning	Recognition	Transfer
Xcept S	.44	.87	.64
Xcept L	.46	.79	.56
Rules S	.86	.69	.85
Rules L	.91	.70	.86

Our main interest was whether the size of the categories influences the recognition memory for the different types of items. A factorial category size by item type analysis of variance (ANOVA) was performed on 2AFC recognition accuracy. Subjects were more accurate (.78 vs. .74) with items in the small category than with items in the large category, $F(1, 80) = 5.28$, MSe $= .02$, $p < .05$. As predicted, the exceptions were better remembered (.83 vs. .70) than the rule-following items, $F(1, 80) = 39.31$,

MSe $= .04$, $p < .001$. As predicted, there was a significant category size by item type interaction, $F(1, 80) = 7.25$, MSe $= .02$, $p < .01$. For the exceptions, recognition was 8% higher for the small category than for the large category (.87 vs. .79). In contrast, recognition for the rule-following items was 1% lower for the small category than for the large category (.69 vs. .70).

Consistent with our main prediction, subjects remembered the exception from the small category better (.87 vs. .79) than the exception from the large category, $t(80) = 2.72$, $p < .01$. The difference between the rule-following items from the small (.69) and large (.70) categories was not significant, $t < 1$.

Model Fits

The context model, RULEX, and the combined model were tested on the recognition performance and the transfer phase classification performance as in palmeri and Nosofsky (1995). For all the models, the fit was measured by mean squared deviations ($RMSD$). The best fitting parameters (i.e., minimizing $RMSD$) were found by searching the parameter space using a genetic algorithm for 1000 generations. Each parameter evaluation for the context model involved a single run because the context model generates the same response probabilities on each run. Because RULEX is stochastic, each parameter evaluation was determined by averaging over 5000 model runs. The combined model was evaluated in the same manner as RULEX. The model fits are displayed in Table 3. For the formal descriptions of the models, see Palmeri and Nosofsky (1995).

Table 3: The recognition and the transfer classification performances observed in the experiment and predicted by the models. Obs, Cont, RUL, and Comb stand for observed, context model, RULEX, and combined model, respectively.

Stimuli	Obs	Cont	RUL	Comb
Recognition				
Xcept S	.869	.754	.783	.836
Xcept L	.788	.754	.794	.841
Rules S	.692	.755	.619	.690
Rules L	.699	.755	.619	.691
Classification				
Xcept S	.636	.513	.636	.586
Xcept L	.562	.683	.661	.625
Rules S	.853	.862	.893	.838
Rules L	.863	.906	.881	.852

Context model A three-parameter version of the context model fit poorly ($RMSD = 0.08$) and failed to capture the qualitative patterns in the data. The

context model made uniform predictions across item types for recognition (see Table 3). The three parameters with the best fitting values were similarities of mismatches along the first dimension, $s_1 = 0.055$, and along the second through the fifth dimensions, $s_x = 0.291$, and the decision parameter for the forced choice tasks, $D = 1.229$. The lower setting of s_1 indicates that the context model is devoting more attention to the rule-relevant dimension than to the other dimensions. This focus allows the context model to correctly predict that exceptions result in more errors than rule-following items during classification, but is not sufficient to explain the recognition advantage for exceptions.

RULEX A six-parameter version of RULEX also fit poorly ($RMSD = 0.06$) and did not capture the qualitative patterns in the data. As predicted, RULEX failed to predict the memory advantage for the exception in the small category over the exception in the large category, although it correctly predicted that the exception items were remembered better than the rule-following items (see Table 3). The six parameters with the best fitting values were the salience parameter for the rule dimension, $W_1 = 0.61$ (with $W_2 = W_3 = W_4 = W_5 = 0.0975$; the W_is are constrained to sum to 1.0), the criterion for accepting a permanent rule, $scrit = 0.65$, the exception storage probability, $pstor = 0.85$, the similarity parameter for the dimensions that were not sampled, $s_w = 1.0$, the similarity parameter for the mismatching dimensions $s_s = 0.80$, and the decision parameter for the forced choice tasks, $D = 26.57$. The higher saliency of the rule-relevant dimension is sensible and psychologically plausible given that this was the most diagnostic dimension and was cued by the hint provided to subjects. The saliency for dimensions two through five were set to a common value because nothing in the experimental design distinguishes one dimension from another.

Combined model An eight-parameter version of the combined model fit poorly ($RMSD = 0.04$) and failed to capture the qualitative patterns in the data. As displayed in Table 3, the combined model failed to predict that the exception in the small category was remembered better than the exception in the large category. Like RULEX, the combined model correctly predicted that the exception items were remembered better than the rule-following items. The eight parameters were the same six parameters used in RULEX plus two additional parameters, s and ω. The parameter, s, determined the residual exemplar-similarity and the parameter ω weighted how much exceptions or exemplars contributed to the familiarity of a given stimulus. The best fitting parameters were $W_1 = 0.47$ (with $W_2 = W_3 = W_4 = W_5 = 0.1325$; the W_is sum to 1.0), $scrit = 0.68$, $pstor = 0.86$, $s_w = 0.65$, $s_s = 0.56$, $D = 14.83$, $s = 0.83$, and $\omega = 0.67$.

General Discussion

As predicted, the memory trace for the exception from the small category was stronger than the memory trace for the exception from the large category. Recognition memory for violating item is better when the violated knowledge structure is stronger. Our results are consistent with basic work in memory (Koffka, 1935) and schema research (Rojahn & Pettigrew, 1992; Stangor & McMillan, 1992). The parallels suggest that exploring further connections between these literatures and the categorization literature could be fruitful.

Existing models that utilize strict rules cannot account for the frequency manipulation in our experiment. Likewise, exemplar models cannot account for our results because they do not posit organizing knowledge structures that can be violated. Previous experiments and simulations by Palmeri and Nosofsky (1995) in support of the rule-based account did not manipulate the strength of rules and thus did not disconfirm the rule account of mental representation.

Interestingly, analysis of RULEX's performance suggests an alternative explanation of the results closely related to the frequency explanation. Increasing the number of items following a rule typically leads to increase in the diversity of rule-following items. One possibility is that the strength of the large category's imperfect rule was partially attributable to diversity rather than to frequency alone. On this view, the exception from the small category is better remembered because it must be differentiated by a number of different, but related, items. Although exception storage processes in RULEX are sensitive to diversity (see Palmeri & Nosofsky, 1995), RULEX was unable to fit the observed pattern in the data. If this proves to be a critical theoretical issue, future experiments should tease apart these two closely related explanations. The answer partially lies in understanding how human learners draw type/token distinctions (cf., Barsalou, Huttenlocher, & Lamberts, 1998).

Given the disconfirmation of existing models, one key question is what mechanism would give rise to our results. Love (2002) presented a clustering model related to the SUSTAIN model (Love & Medin, 1998; Love, Medin, & Gureckis, in press) that accounted for Palmeri and Nosofsky's (1995) results. The model stored rule-violating items in their own cluster. Importantly, the model developed sharper tunings (related to memory distinctiveness) for exception clusters, which, in part, allowed the model to predict improved recognition memory for exceptions. The model should be able to account for our results because it predicts that the tuning of the cluster encoding the exception from the small category should be sharper than the tuning of the cluster encoding the exception from the large category.

The dynamics that drive this outcome are con-

sistent with our explanation of the results. Each cluster's tuning is adjusted on each learning trial in order to minimize prediction errors. The cluster encoding the exception from the small category tends to be activated by the presentation of rule-following items from the large category because this cluster matches these items on the rule-relevant dimension. In order to avoid these unwanted intrusions, the cluster becomes highly tuned and specific, which minimizes activation by items other than the exception from the small category. The increased distinctiveness of the cluster enhances its recognition. The same dynamics govern the cluster encoding the exception of the large category, but this cluster does not become as distinct as the cluster encoding the exception from the small category because of the frequency manipulation (i.e., fewer trials in which its tuning is sharpened).

Modeling work along these lines is currently being carried out. Importantly, this modeling endeavor seeks to unite previous work in the memory literature with work in categorization. In addition to accounting for findings in category learning and schema research, the same mechanisms should allow this clustering model to account for basic memory phenomena like the list strength effect (Ratcliff, Clark, & Shiffrin, 1990; Tulving & Hastie, 1972) in which increasing the study of certain items can actually increase the recognition of other items as well.

References

Barsalou, L. W., Huttenlocher, J., & Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology, 36*, 203–272.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science, 262*, 1747–1749.

Koffka, K. (1935). *Principles of Gestalt Psychology* New York: Harcourt, Brace.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 565–572.

Love, B. C. (2002). Two systems or just one? *J. S. McDonnell Foundation Cognitive Neuroscience of Category Learning Consortium Workshop.*

Love, B. C. & Medin, D. L. (1998). SUSTAIN: A Model of Human Category Learning. *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 671-676). Cambridge, MA: MIT Press.

Love, B. C., Medin, D. L., & Gureckis, T. M. (in press). SUSTAIN: A Network Model of Human Category Learning. *Psychological Review.*

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

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 241–253.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-Plus-Exception Model of Classification Learning. *Psychological Review, 101(1)*, 53–79.

Nosofsky, R. M., & Zaki, S. F. (1998). Dissociations between categorization and recognition in amnesic and normal individuals. *Psychological Science, 9*, 247–255.

Palmeri, T. J., & Flanery, R. M. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science, 10*, 526–530.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 548–568.

Pinker, S. (1991). Rules of language. *Science, 253*, 530–535.

Rojahn, K., & Pettigrew, T. F. (1992). Memory for schema-relevant information: A meta-analytic resolution. *British Journal of Social Psychology, 31(2)*, 81–109.

Ratcliff, R., Clark, S., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 163–178.

Smith, E. E., & Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science, 16*, 1–40.

Squire, L. R., & Knowlton, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Science, USA, 92*, 12470–12474.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and developmental literatures. *Psychological Bulletin, 111*, 42–61.

Tulving, E., & Hastie, E. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology, 92*, 297–304.