Canadian Journal of Experimental Psychology
2007, Vol. 61, No. 2, 109-xxx

DOI: 10.1037/cjep2007009

Prior Knowledge and Correlational Structure in Unsupervised Learning

John P. Clapper, California State University, San Bernardino

Abstract Prior knowledge has been shown to facilitate both supervised and unsupervised category learning, but questions remain about how this facilitation occurs. This article describes two experiments that investigate the effects of prior knowledge on unsupervised learning, using the exemplar-memory task of Clapper and Bower (2002). Experiment 1 demonstrates that prior knowledge facilitates learning in this task, as expected, and that this facilitation extends to both knowledge-relevant and knowledge-irrelevant features of the new categories. Experiment 2 shows that knowledge facilitates learning not only by increasing the probability that people will discover separate categories, but also by making the features of different categories seem less interchangeable, thereby reducing interference and confusion among them. Taken together, these experiments demonstrate that prior knowledge has multiple effects on unsupervised learning and suggests that the exemplar-memory task may provide a useful procedure for disentangling and investigating these effects.

The ability to learn new categories has long been recognized as an essential feature of human cognition. In recent decades, it has become increasingly clear that pre-existing knowledge, commonsense theories, and causal reasoning play an extremely important role in this fundamental human ability (see, e.g., Murphy, 2002 for a review), and that prior knowledge can facilitate category learning in a wide variety of tasks and situations (e.g., Ahn, 1990, 1999; Kaplan & Murphy, 1999; Murphy & Allopenna, 1994; Murphy & Kaplan, 2000; Pazzani, 1991; Spalding & Murphy, 1996; Wattenmaker, Dewey, D.T. Murphy, & Medin, 1986; Wisniewski, 1995). The research described in this article investigates the role of knowledge in observational or "unsupervised" learning, in which people discover new categories without feedback or other assistance from an external "trainer." Unsupervised learning is theoretically interesting and pervasive in everyday life, but it has received much less empirical study than traditional (supervised) classification tasks (e.g., Bruner, Goodnow, & Austin, 1956; Medin & Schaffer, 1978; Posner & Keele, 1968, 1970).

Most demonstrations of knowledge effects on unsupervised learning have occurred in the context of free sorting or "category construction" experiments. This task involves presenting participants with a set of pictures or verbal descriptions (artificial stimuli) printed on cards, and asking them to sort the cards into two or more groups corresponding to their subjectively "natural" categorization. The consistent result is that people prefer to sort the stimuli into categories based on a single attribute dimension, while ignoring their overall similarity or family resemblance structure (e.g., Medin, Wattenmaker, & Hampson, 1987; Regehr & Brooks, 1995). However, people can be induced to sort on the basis of family resemblance if the categories so defined are related to prior knowledge or causal theories. For example, Ahn (1990, 1999) showed that providing learners with an explicit theory that explained why certain combinations of features occurred together in the same category increased the likelihood that they would sort according to family resemblance (overall similarity) rather than based on a single dimension. Spalding and Murphy (1996) demonstrated that people were more likely to sort by family resemblance when the features of the categories were related to coherent themes (e.g., Arctic vehicles vs. jungle vehicles). Kaplan and Murphy (1999) showed that relatedness to a coherent theme would promote learning of category structure even if only one dimension of the stimuli was related to that theme, and that this benefit extended to learning both theme-relevant and theme-irrelevant features of the new category.

Obviously, understanding such knowledge effects should be a major priority for research on unsupervised learning. In particular, it is important to distinguish the different ways in which knowledge could affect learning (i.e., to create a taxonomy of possible knowledge

effects and assess the impact of each upon the overall learning process). This article describes two experiments that provide a first attempt to distinguish different types of knowledge effects within the context of an unsupervised learning task. The task itself differs from category construction in several ways; in particular, it uses recognition memory for the features of individual exemplars (rather than sorting) as the primary measure of learning. However, the exemplar-memory task is similar to category construction in at least one important respect – participants tend to show poor learning of category structure in this task, at least in the absence of relevant knowledge (Clapper, in press; Clapper & Bower, 2002). What has not yet been determined is whether learning will be improved if such knowledge is made available. Thus, the first goal of the present research was simply to demonstrate that prior knowledge improves unsupervised learning in the exemplar-memory task, as it does in category construction. Beyond this initial demonstration, a second goal was to answer more detailed questions about the precise role of prior knowledge in unsupervised learning (i.e., to identify and investigate some of the causal pathways that mediate knowledge effects in these tasks).

The Exemplar-Memory Task

In this task, participants view a series of training instances, each consisting of a list of verbal descriptors displayed on a computer screen. Typically, the instances come from two or more different categories defined by different clusters of correlated (consistently co-occurring) properties (e.g., Category A = 11111111XXX and Category B = 22222222XXX, where each position represents a dimension, each number a particular value on that dimension, and the Xs denote dimensions that have variable values across different instances). The participant studies each instance and is then tested on his/her recognition memory for its features; this study-test cycle continues until all the training instances have been shown. The computer records how long the participant spends studying each feature (the display is designed so that the features must be viewed one at a time), and well as his/her accuracy in remembering each feature.

If participants are learning the categories, their performance should change over trials in several ways. First and most importantly, the consistent (correlated) features associated with each category should become more predictable and recognition accuracy for these features should increase. Second, the greater predictability of the consistent features should enable learners to spend less time studying them and more time studying the variable (uncorrelated) features, which must be memorized separately for each instance (Clapper & Bower, 2002; Metcalfe, 2002). Third, the increased time spent attending to uncorrelated features should cause learners' memory for these features to improve, as well. Plotted over trials, these changes in performance define clear "learning curves" for the correlational patterns underlying each category. None of these changes are observed in the absence of learning (e.g., if no correlational patterns (categories) are present in the training stimuli).

Compared to other types of unsupervised learning tasks, some advantages of the exemplar-memory task are: a) it is *unobtrusive*, in the sense that learning can be observed without forcing participants to intentionally search for categories or feature correlations (in fact, categories or feature correlations need never be mentioned at all), and b) it provides a *continuous* measure of learning over trials, rather than assessing only asymptotic learning (as in standard transfer experiments, in which learning occurs during an earlier "training" phase and is then evaluated during a later "transfer testing" phase, e.g., Billman & Knutson, 1996; Kaplan & Murphy, 1999).

What is Learned?

Before proceeding, it is important to understand what "category learning" actually means for participants in the exemplar-memory task. First, they must notice or recognize the existence of separate categories within the stimulus set (sometimes referred to as the "discovery problem"; see, e.g., Clapper & Bower, 2002; Michalski & Stepp, 1983). Second, once separate categories have been identified, participants must learn the feature distributions or predictive structure within each category (the so-called "characterization problem," e.g., Michalski & Stepp, 1983). Learning the feature distributions within a category (e.g., knowing which dimensions have consistent values, and what these values are) should improve the learner's ability to predict or fill in the unknown (or unremembered) features of individual instances. Thus, improved feature prediction is expected to lead to improved performance (accuracy of recognition) in the exemplar-memory task.

Notice that category learning has a subtly different meaning in this task than in category construction or supervised classification tasks. In category construction, the learner must sort the stimuli into separate categories, but he/she does not necessarily need to learn all the consistent features associated with the categories in order to do this. All he/she needs to learn is the minimum set of features required to assign instances to the appropriate categories; additional (redundant) features can simply be ignored in this task. The same is true of supervised (feedback-based) category learning experiments; again, the only criterion is the ability to

assign instances to the correct categories, and features that are redundant for that purpose are irrelevant to the task (see Markman & Ross, 2003). In the exemplar-memory task, by contrast, people need to use their category knowledge not just to sort or classify the instances but also to reconstruct their features for the immediate memory task. The more predictive structure (consistent features) that people have learned about a category, the better they are likely to perform in this task. Thus, in this task learning a category means learning all (or as many as possible) of the features associated with that category, including those that would be redundant or unnecessary from the perspective of classification alone. This is an important point because prior knowledge could, in principle, facilitate both the classification (discovery) and feature prediction (characterization) aspects of category learning. Ignoring feature prediction and focusing entirely on classification risks ignoring some of the most interesting and important potential benefits of prior knowledge.

The Unsupervised Learning Process

Naturally, the process of discovering and learning about categories in this task must occur in real time (i.e., as a learner encounters successive exemplars and updates their knowledge base (categories) in response to each). The "category invention" framework of Clapper and Bower (1991, 1994, 2002; Clapper, in press) provides a qualitative overview of the trial-by-trial learning process in this task. The essential claims of this framework are that people attempt to assign all the stimuli they encounter to existing categories, and that new categories are created when old ones fail (i.e., when a stimulus is encountered that does not fit into any existing category). In the exemplar-memory task, this implies that the person will create a new category to describe the stimulus presented on the first trial, and then adopt a kind of "win-stay-lose-shift" strategy in which subsequent stimuli will continue to be assigned to this first category if possible, with a new category being created only if this match process fails (i.e., if a new stimulus cannot be assigned to the existing category due to feature mismatch).

The framework also provides an explanation for why people often have difficulty learning categories based on correlational structure alone (i.e., without the help of relevant knowledge) in unsupervised tasks. A consistent finding in the exemplar-memory task is that people tend to learn poorly if instances of two categories are presented in a mixed (randomly alternating) sequence, but they learn much better if instances of different categories are presented in separate blocks or if one category is well learned prior to encountering the second category (Clapper, in press; Clapper & Bower, 1994, 2002; Medin & Bettger, 1994). The category invention framework argues that this occurs because presenting categories together in a mixed sequence tends to camouflage the contrast between them, resulting in a failure to notice the existence of separate categories within the stimulus set. As a result, instances of both categories are lumped or aggregated together into a single over-generalized category that captures their shared dimensional structure and the base-rate frequency of each value, but not the correlational structure of the set. Once this occurs, both types of instances will continue to be assigned to this aggregated category, resulting in a persistent failure to distinguish between them. On the other hand, seeing a dozen or more examples of one category in succession, prior to seeing any examples of the other category, sets up a strong contrast and greatly increases the likelihood of inventing separate categories.

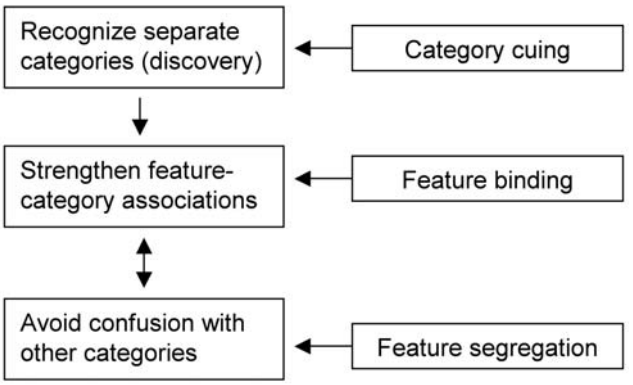

Figure 1. Category learning processes (left) and potential knowledge effects (right) based on the category invention framework.

Knowledge Effects on Learning

The preceding analysis suggests several ways in which prior knowledge might affect the category learning process (see Figure 1). The most obvious would be to increase the probability that a given learner will notice or recognize separate categories in the first place. Prior knowledge could increase the probability of creating separate categories by acting as a cue or signal that different categories are present among the training stimuli (e.g., Clapper & Bower, 1991, 2002; Kaplan & Murphy, 1999; Spalding & Murphy, 1996; Yamauchi & Markman, 1998). For example, a learner studying descriptions of fictitious individuals, some of whom fit his/her pre-existing stereotype of "introverts" and others that of "extroverts," might notice this fact and use it to divide the descriptions into separate categories (Medin, Wattenmaker, & Hampson, 1987). Presumably, this would occur as a kind of "aha" experience in which the learner notices that two general

"themes" are present, and begins to sort stimuli into different mental categories based on these themes.

Once separate categories have been created, prior knowledge could also affect how the predictive structure within each category is learned. Learning a particular feature of a category involves strengthening a feature-to-category association while overcoming possible interference or confusion with the features of other, closely related categories (see Figure 1). Thus, ease of learning should depend on 1) the degree of "fit" or cohesion between the feature and the category, and 2) the amount of interference from other categories currently being learned.

The first factor, henceforth referred to as "feature binding," is assumed to vary as a function of how strongly the features within a category are associated with each other and with the category. It seems reasonable to assume that features related to the same theme or topic would have strong pre-experimental associations in memory (e.g., Heit & Bott, 2000; Rehder & Murphy, 2004), and that as a result they would be relatively easy to learn when encountered together in a new category. From this perspective, prior knowledge functions as a kind of "glue" increasing the coherence or cohesion among relevant features of a category (e.g., Murphy, 2002; Murphy & Medin, 1985). For example, extraverts tend to share characteristics such as "enjoys parties," "spends a large amount of time on the phone," and "has many friends." We would expect these features to be pre-associated in memory due to their common association with the extravert stereotype. Due to this pre-existing link, these features should be relatively easy to learn when they occur together in a new person or group.

The second factor, henceforth referred to as "feature segregation," pertains to the issue of keeping the features of related categories clearly separate and distinct in memory, which in turn determines the potential for interference or confusion among them. To illustrate, consider a pair of categories distinguished on the basis of arbitrary feature correlations. In the absence of any theory or explanation as to why particular combinations of features happen to co-occur, the features of the two categories would seem completely interchangeable; indeed, the only way that they could be segregated into their correct categories would be through rote memorization. However, if the categories happened to be related to prior knowledge or some organizing theme, such as introverted versus extroverted personality types, their features would seem less interchangeable and hence easier to segregate into the appropriate categories. Thus, an introvert would be unlikely to enjoy parties while an extrovert probably would not like to spend a great deal of time alone. The underlying assumption here is that a significant amount of confusion or interference would normally occur between closely related categories, but that relevant knowledge can reduce this interference by reducing the a priori interchangeability of their features.

Notice that category cuing and feature segregation are both related to the problem of keeping different categories separate in the learner's mind; they differ only in the level at which this separation occurs. Category cuing affects separation at the category level (i.e., the probability that the learner will divide the training stimuli into separate categories). Feature segregation refers to separation at the level of individual features (i.e., the ability to avoid confusing or intruding features across category boundaries). Obviously, the issue of feature segregation could arise only if the learner had already created separate categories.

Once these potential knowledge effects have been distinguished at a conceptual level, the next step is to specify how they will affect performance in the exemplar-memory task. First, consider the issue of category cuing. Within the category invention framework, dividing the stimuli into separate categories is a prerequisite for learning which features are consistently associated with each category. Discovering the categories should result in better memory for their consistent features, because these features can be retrieved from the category schema as well as from episodic memory traces of specific instances. Thus, to the extent that prior knowledge prompts or cues a learner to create separate categories, it should improve his/her memory for the consistent features of those categories.

Assuming that separate categories have been created, the next question is how knowledge effects on feature binding and segregation would further affect exemplar-memory performance. To the extent that prior knowledge promotes feature binding and/or segregation, it should tend to result in better memory for thematically relevant compared to thematically irrelevant (neutral) features within each category. For example, if the categories were related to the learner's introvert versus extravert stereotypes, this fact should do more to facilitate learning of relevant dimensions such as "likes/dislikes parties" and "has many/few friends" than of irrelevant dimensions such as "born in New York/California" or "drinks soda/coffee with lunch." In the case of binding, this difference would result from the fact that thematic, but not neutral, features benefit from their pre-experimental association with a common theme or topic. In the case of feature segregation, the difference would be due to the fact that thematic features are less likely to be mixed up or confused across category boundaries (assuming that different categories are related to different themes).

This leaves the issue of distinguishing between knowledge effects that affect feature binding versus those that affect feature segregation. It has already been noted that both category cuing and segregation refer to the problem of keeping (or establishing) different categories and their features as separate in learners' minds. Binding, on the other hand, refers to the problem of learning associations within a given category. Stronger pre-experimental associations among the features within a category should make the binding problem easier. In contrast to such "within-category" factors, segregation (and cuing) are largely dependent on "between-category" factors, in particular, what other categories are present in the same learning task and how these are related to the "target" category in question.

To illustrate, consider the following example. A given category (A) might be learned poorly in the presence of another category (B) if both were unrelated to any particular theme, because in this case people would be likely to lump them together into a single category; even if separate categories were recognized, people might still show a tendency to mix up and confuse their features. On the other hand, the same A category might be learned much better in the presence of a B category related to a strong theme that clearly marked it off as separate from A. In this case, the Bs would be partitioned off from the As in a separate category of their own, and due to the thematic difference there would be little possibility of confusing their features. Here, the learnability of Category A depends not so much on its own properties (a "within-category" factor that is held constant across both conditions), but rather on the properties of Category B and their contrast with those of A (a "between-category" factor that differs across the two conditions). Thus, if a group of "average" personalities was presented with a second group also consisting of average personalities, people might fail to recognize them as separate categories. However, if one of the average groups was shown together with a group of obvious introverts, people would probably find it much easier to tell the two categories apart.

This type of reasoning was used to attempt a fairly precise discrimination between different types of knowledge effects in Experiment 2. However, before attempting such detailed discriminations, a preliminary demonstration was required to show that knowledge does indeed affect category learning in the exemplar-memory task. That was the goal of Experiment 1.

Experiment 1

The primary goal of this experiment was simply to demonstrate that thematic relatedness would facilitate category learning in the exemplar-memory task (as noted above, this has been shown in other tasks, such as category construction, e.g., Spalding & Murphy, 1996, but not in the exemplar-memory task). Participants were shown instances of two categories, defined by contrasting patterns of correlated features. The training instances described fictitious persons in terms of 12 attribute dimensions (e.g., hobbies, preferences, personal characteristics, etc.). In the *Thematic* condition, all correlated values within a category were related to a familiar theme or stereotype. For example, members of one category might enjoy going to the opera and sipping champagne while members of the other category prefer to drink beer and attend professional wrestling matches (stereotypical "highbrows" vs. "lowbrows"). In the *Neutral* condition, the correlated features were chosen to be as neutral as possible so that they would not evoke contrasting themes or stereotypes (e.g., watching movies and drinking cola vs. watching basketball and drinking coffee). The two stimulus sets had the same correlational (statistical) structure, but better learning was expected in the Thematic condition than in the Neutral condition.

A second goal of this experiment was to evaluate whether knowledge facilitates the learning of both relevant and irrelevant features of a new category, and to provide a preliminary evaluation of the relative strength of this facilitation for the two types of features. Thus, a condition was included in which some of the correlated features were related to familiar themes while others were neutral with respect to those themes (referred to as the *Thematic-plus-Neutral* condition). Previous experiments using other tasks (e.g., Kaplan & Murphy, 1999) have shown that the presence of thematic features also facilitates learning the nonthematic (neutral) features of a category. If that is also true in the exemplar-memory task, then the neutral correlated features within the Thematic-plus-Neutral condition should be learned better than the corresponding features within the Neutral condition.

Importantly, having both types of features present in the same category makes it possible to investigate whether prior knowledge produces stronger benefits for thematically relevant than for thematically irrelevant (neutral) features within that category. As noted in the Introduction, enhanced category cuing in the thematic conditions should result in a general improvement in memory for all the consistent features of the categories. A further advantage for thematic over neutral features would provide evidence for additional knowledge effects due to enhanced feature binding and/or segregation.

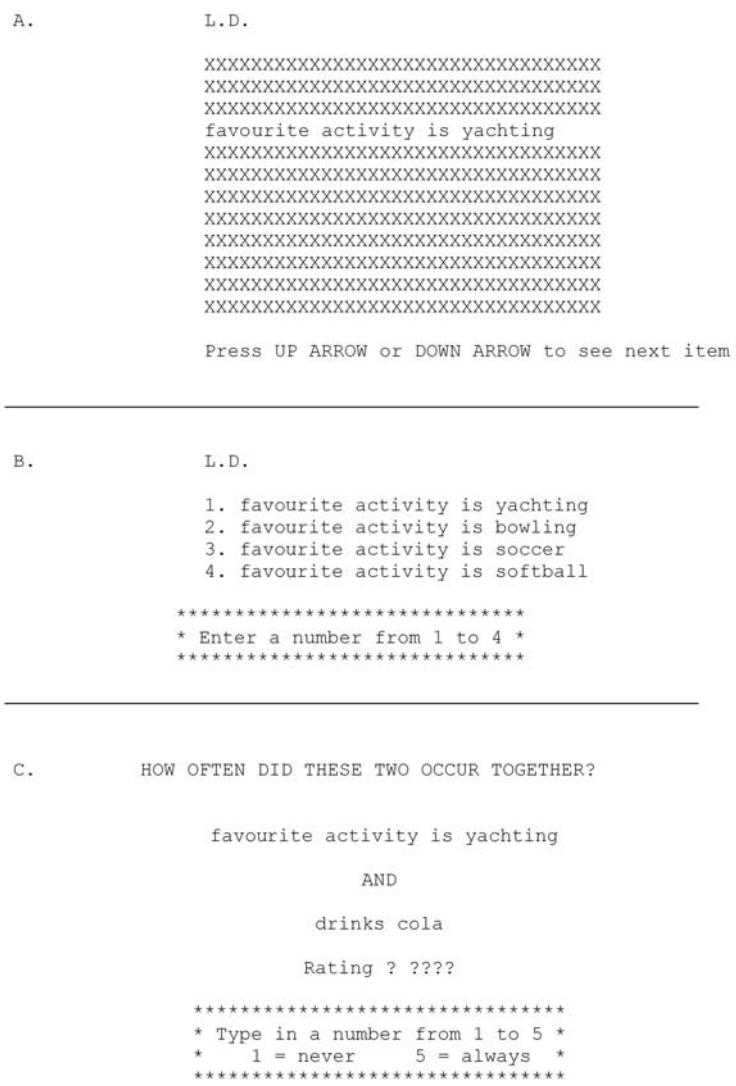

Figure 2. Stimulus displays for the study-test (A and B) and post-test (C) phases of both experiments.

Method

Participants. The participants were 63 undergraduate students of Humboldt State University and the California State University San Bernardino who participated in exchange for extra credit in several psychology classes.

Procedure. Participants were seated at individual microcomputer stations in group laboratories seating up to 20 individuals. The main portion of the experiment consisted of 32 trials. Each trial consisted of two phases, a "study" phase followed by a "test" phase. At the start of the study phase, a verbal list (training instance) was presented in the middle of a computer screen (Figure 2A). This list described a fictitious person. The person's initials were printed in uppercase letters at the top of the list, below which appeared twelve short descriptive phases, each referring to a specific value of a particular attribute of that person (e.g., "member of the Moose Lodge," "favourite hobby is woodworking"). A different attribute was shown on each row (line) of the list display. The position in which the attributes appeared in the list displays remained consistent throughout the experiment for a given participant, but was randomized separately for each participant. Note that a different fictitious person was described on each trial (i.e., a total of 32 different individuals were described over the course of the

experiment).

At the beginning of each trial of the study-test procedure, all the descriptors in the list display were masked by rows of Xs (Figure 2A). The participants could unmask and study each item by pressing a designated "up" or "down" key that exposed the item on the line above or below their current position (the computer chose a different random starting location within the list on each trial). The exposed descriptor was immediately masked again by Xs when the participant uncovered a new line (descriptor). A total of 36 seconds was allocated for the participant to study all the descriptors in the 12-item list, after which the list disappeared and the testing phase began. The computer recorded the total amount of time spent looking at each attribute descriptor during the study portion of the trial.

During the test phase of each trial, participants received a series of forced-choice recognition tests for the attribute values of the just-presented instance. Twelve test questions, one for each attribute dimension, were presented in random order (different on each trial) in a multiple-choice format (see Figure 2B). The person's initials were shown at the top of the display, and four possible answers (different values of the same attribute) were listed below. Participants typed in the number (1-4) of the value they thought had been present in the last instance. The computer then displayed the word "correct" or "incorrect" underneath the multiple-choice display, and, if the response was incorrect, an arrow indicated the correct choice. After all 12 multiple-choice test questions had been completed for a given instance, the participant received summary feedback on his/her memory performance (percent correct on that trial, as well as cumulative percent correct).

Following the 32 study-test trials, a correlation rating "post-test" was used to assess asymptotic learning. In this task, participants answered a series of questions regarding the frequency with which different attribute values had occurred together during the preceding study-test procedure (see Figure 2C). For each pair of attribute values, participants were instructed to rate the frequency with which the values had occurred together in the previous phase on a 5-point scale. A total of 36 such correlation rating trials were shown.

Materials. The stimuli were descriptions of fictitious persons presented in the form of verbal feature lists. Each feature (attribute value) consisted of a brief phrase describing some characteristic of the person such as where he/she was born, where he/she lives now, hobbies, occupation, education, interests, tastes, and so on. A total of 12 attributes were used within a given stimulus set, each of which had four possible values. Two of these were related to coherent themes or stereotypes, while the other two were unrelated to these themes. For example, one attribute might refer to the person's preferred form of entertainment, with the four values being "enjoys watching the opera," "enjoys watching pro wrestling," "enjoys watching basketball," and "enjoys watching movies." Here, the first value suggests an elitist "highbrow" and the second a down-to-earth "lowbrow," while the last two values are relatively neutral with respect to this dichotomy.

Three different stimulus sets were used in this experiment; a given participant only saw one of these. (The stimulus sets are shown in the Appendix). The stimuli in each set were constructed from a different collection of attribute dimensions, and their thematic values suggested different pairs of contrasting themes. In one set, the thematic values suggested either a "highbrow" or "lowbrow" theme, in another "females" versus "males," and in the third, staid "senior citizens" versus trendy "adolescents." These themes were never mentioned explicitly to participants; rather, they had to notice the themes (or not) of their own accord.[1]

The materials for the post-test employed the same attribute dimension descriptors used in the preceding study-test procedure. All four values of both correlated and uncorrelated dimensions were used in different combinations during this phase. Thus, some of the pairs presented values that had occurred together during Phase 1, others showed values that had not occurred together during Phase 1, and still other pairs showed values that had never been presented in Phase 1 (except as foils in the multiple-choice recognition tests).

Design. Participants were assigned randomly to one of three experimental conditions, referred to as Thematic, Neutral, and Thematic-plus-Neutral (N = 21 per condition). In each condition, 9 of the 12 stimulus dimensions had perfectly correlated values while the remaining three were uncorrelated. The correlated values defined two distinct patterns or categories within the stimulus set shown to each participant. In the Thematic condition, these categories were based on correlated values that related to a pair of familiar, contrasting themes. For example, given a set in which a value of "1" refers to the highbrow theme and a value

[1] The thematic versus neutral values for each attribute were generated by the investigator and a group of nine undergraduate research assistants. In each case, it was the group consensus that our undergraduate participants would recognize the thematic values (but not the neutral values) as referring to the intended themes.

of "2" refers to the lowbrow theme, the categories in this condition could be denoted as Category A = 111111111XXX and Category B = 222222222YYY; the Xs or Ys in the last three positions indicate uncorrelated dimensions, each of which varied independently across three possible values: 1, 3, or 4 for X, and 2, 3, or 4 for Y. (In other words, different instances could have different values on these dimensions). Note that the values of these uncorrelated dimensions were either neutral with respect to the themes (Values 3 and 4), or were related to the same theme as the other features of that category (Value 1 for Category A and Value 2 for Category B). The opposite-theme value was not permitted due to the possible confusion that might result.

In the Neutral condition, the correlated values were not obviously related to any theme (i.e., Category A = 333333333XXX and Category B = 444444444YYY, where values 3 and 4 are neutral with respect to the themes suggested by Values 1 and 2). In the Thematic-plus-Neutral condition, five of the nine correlated values were associated with the themes while the remaining four were neutral (i.e., Category A = 111113333XXX and Category B = 222224444YYY).

A total of 32 instances were presented during the study-test phase, 16 from each category. The instances were presented in a pseudo-random sequence (the same for all participants) in which no more than three instances of a given category were allowed to occur in a row.

Counterbalancing. As noted, three different stimulus sets, each featuring a different pair of contrasting themes (senior/adolescent, male/female, and lowbrow/highbrow) were used to ensure the generality of the results. The assignment of stimulus attributes to particular cells within the experimental design (e.g., correlated vs. uncorrelated), as well as the positions of each attribute within the feature-list displays, was randomized separately for each participant. A different combination of nine dimensions was randomly selected to be consistent and three to be variable for each participant. Each theme (stimulus set) was assigned equally often to the different conditions (Thematic, Thematic plus Neutral, and Neutral).

Results

The three dependent measures in this experiment were the correlation ratings collected during the post-test, plus study times and memory data collected during the study-test portion of the experiment. These will be discussed in turn below.

Correlation ratings. Several types of feature pairs were presented during the post-test. The most informative comparison is between pairs of correlated values from the *same* category versus pairs of correlated values from *different* categories. The more people learned about the correlational structure of the stimulus set, the more their ratings for same-category pairs should exceed those for different-category pairs. By subtracting the mean ratings for different-category pairs from those for same-category pairs, a discrimination score indicating the degree of learning can be obtained for each participant. Since the rating scale ranged from 1 to 5, the maximum discrimination score (absolute value) was 4 and the minimum was zero. The data are shown in Table 1.

TABLE 1
Correlation Rating Data from Experiment 1

Condition	Dimension	Probe type		Discrim. Score
		Same-Category	Different-Categories	
Thematic	Thematic	4.58	1.52	3.06
T + N	Thematic	4.44	1.61	2.83
T + N	Neutral	3.93	2.25	1.68
Neutral	Neutral	3.90	3.22	0.68

TABLE 2
Study Time Data from Experiment 1

Condition	ST (Correlated)	ST(Variable)	Preference Score
Thematic	2.68	3.95	1.27
T + N	2.83	3.51	.68
Neutral	2.90	3.32	.42

Discrimination scores from the Thematic condition significantly exceeded those from the Neutral condition, $t(40) = 6.67$, $SE = .36$, $p = .000$; however, they did not exceed scores for thematic dimensions from the Thematic-plus-Neutral condition, $t(40) = .55$, $SE = .41$, $p > .50$. Within the Thematic-plus-Neutral condition, discrimination was significantly higher for thematic than for neutral dimensions, $t(20) = 3.56$, $SE = .32$, $p = .002$. Nevertheless, neutral dimensions from the Thematic-plus-Neutral condition had significantly higher scores than did those from the all-Neutral condition, $t(40) = 2.49$, $SE = .40$, $p = .017$.

Study times. The main index of learning derived from the study-time data was a "preference score" computed by subtracting the average study time for correlated dimensions from that of uncorrelated dimensions, (i.e., mean ST(uncorrelated) – mean ST(correlated), where "ST" denotes "study time"). Since category learn-

TABLE 3
Recognition Memory Data from Experiment 1

Condition	Dimension Type		
	Thematic	Neutral	Uncorrelated
Thematic	.967	*	.905
T+N	.925	.896	.839
Neutral	*	.831	.796

* not present in that condition

ing was expected to cause people to attend more to uncorrelated than correlated features, a positive preference score is interpreted as evidence of learning. The study time data for this experiment (averaged over all instances following the sixth instance of each category) are displayed in Table 2.

Study times did not differ significantly for neutral versus thematic dimensions in the Thematic-plus-Neutral condition, $t(19) = 1.67$, $SE = .08$, $p = .112$, so these have been averaged together in Table 2 and in the comparisons to follow.[2] Preference scores in the Thematic condition were significantly greater than those in the Neutral condition, $t(40) = 2.77$, $SE = .31$, $p = .008$, and marginally greater than those in the Thematic-plus-Neutral condition, $t(39) = 1.80$, $SE = .33$, $p = .081$. However, the comparison between the Thematic-plus-Neutral versus Neutral conditions was not significant, $t(39) = 1.29$, $SE = .16$, $p = .21$. Thus, contrary to the correlation ratings, the study time data provided no useful data indicating that learning in the Thematic-plus-Neutral condition exceeded that in the Neutral condition.

Memory. The recognition memory data from this experiment are shown in Table 3. In addition, the data for correlated features are shown plotted by trials in Figure 3.

Memory was greater in the Thematic condition than in the Neutral condition for both correlated dimensions (as indicated in Table 3, these were all thematic in the

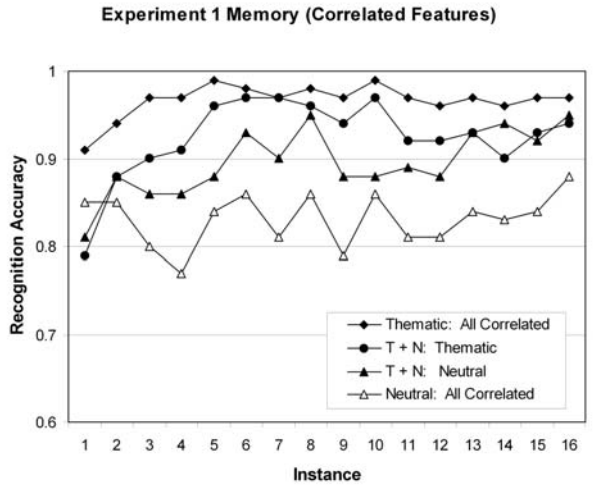

Figure 3. Recognition memory data for Experiment 1 plotted over successive category instances for the three conditions; note that both thematic and neutral correlated features are displayed for the Thematic-plus-Neutral condition.

Thematic condition and neutral in the Neutral condition, $t(40) = 5.02$, $SE = .03$, $p = .000$, and uncorrelated dimensions, $t(40) = 3.66$, $SE = .03$, $p = .001$. Thematic dimensions within the Thematic-plus-Neutral condition were remembered slightly more poorly than those from the Thematic condition, $t(40) = 1.88$, $SE = .02$, $p = .067$. Uncorrelated dimensions were also remembered more poorly in the Thematic-plus-Neutral condition than in the Thematic condition, $t(40) = 2.07$, $SE = .03$, $p = .045$. Within the Thematic-plus-Neutral condition itself, thematic dimensions were remembered better than neutral dimensions averaged over all trials, $t(20) = 3.46$, $SE = .01$, $p = .002$, but this difference disappeared by the end of training, $t(20) = 0.24$, $SE = .01$, $p = .816$ over the last five instances of each category. Most importantly, neutral dimensions were remembered significantly better in the Thematic-plus-Neutral condition than in the Neutral condition, $t(40) = 2.03$, $SE = .03$, $p = .049$. Another interesting result was that memory for correlated dimensions showed an immediate advantage on the first trial of the Thematic condition (see Figure 3), compared to the pooled data from the other two conditions, $t(61) = 2.20$, $SE = .04$, $p = .032$.

Discussion

It is clear that prior knowledge facilitated unsupervised learning in this experiment. Participants in the Thematic condition had higher preference scores on the study time measure and better memory for both correlated and uncorrelated features than did those in the Neutral condition. Direct correlation ratings collected in the post-test confirmed that learning was better in

[2] Note that one participant was excluded from these and the following study-time analyses due to a very unusual pattern of data (he/she spent nearly all of his/her study period on one particular attribute, which happened to be one of the neutral correlated attributes, throughout most of the experiment, suggesting that he/she was simply "resting" on that attribute until the fixed study period expired and the list disappeared). Although this single participant did not affect the outcome of any of the statistical analyses, he/she did slightly alter the visual appearance of the data.

the Thematic than in the Neutral condition.

Overall, participants in the Neutral condition seemed to have learned surprisingly little about the correlational structure of the stimuli. Their post-test discrimination scores were only slightly greater than zero (mean of 0.68 out of a possible 4.00, compared to 3.01 in the Thematic condition). Neither the study times nor the memory data from this condition showed strong evidence of learning (e.g., a clear pattern of increasing performance over trials, see Figure 3). The relative lack of learning in this condition seems somewhat counterintuitive given the very strong statistical structure underlying the categories, each distinguished by nine perfectly correlated dimensions. However, this result is consistent with other data discussed in the Introduction (e.g., Clapper & Bower, 2002; Medin et al., 1987; Regrehr and Brooks, 1995) showing that people often have difficulty extracting separate categories purely on the basis of statistical feature correlations within a stimulus set.

The present results also show that if some features of a new category are related to familiar themes, this not only facilitates learning those thematic features but also facilitates learning other, neutral features of that category. Neutral correlated features may have been learned more poorly than thematic features in the Thematic-plus-Neutral condition, but they were learned significantly better than correlated features from the Neutral condition (according to both correlation-rating and memory measures). Thus, it is clear that the presence of thematic features enhanced learning of neutral features in this experiment. This is reminiscent of earlier results by Kaplan and Murphy (1999) showing that thematic features can facilitate learning of both thematic and nonthematic features in category construction (sorting) tasks.

What do these results imply in terms of the three types of knowledge effects discussed in the Introduction? First, the category cuing idea finds support from the fact that both thematic and neutral features of the categories were learned better in the thematic conditions of these experiments. The overall pattern of results suggests that relatively few people noticed the presence of separate categories in the Neutral condition, whereas many or most participants in the Thematic conditions must have done so. Another result consistent with a category cuing effect is that memory for thematic features was elevated from the first trial in the Thematic condition. This suggests that the themes were noticed immediately, which in turn implies that they would have been readily available (and highly salient) as a basis for creating separate categories.

In addition to being more likely to divide the stimuli into separate categories in the thematic conditions, people also showed evidence of knowledge effects on their learning of individual features within those categories. In particular, thematic features were learned somewhat better than neutral features in this experiment (the memory measure showed faster learning of thematic features, although final performance was the same for both, while the correlation ratings showed a significant difference in final learning). Such a thematic-feature advantage could have occurred as a result of knowledge effects on either binding and/or segregation. Stronger binding, for example, should directly facilitate learning of thematic features relative to neutral features, because the former but not the latter should be strongly preassociated in the learner's memory. Enhanced segregation would also tend to improve learning, in this case by reducing interference between the two categories. Of course, this effect should benefit thematic but not neutral features, as there is no reason to expect contrasting themes to prevent confusion among neutral features unrelated to those themes.

Experiment 2

The previous experiment provided evidence not only that thematic features promoted the learning of separate categories, but also that such features were learned better than neutral features due to feature binding and/or segregation effects. However, that experiment did not provide any information about which of the latter two effects was actually responsible for the observed thematic-feature advantage. In Experiment 1, both categories were either neutral or related to opposite themes. By contrast, Experiment 2 included a "mixed" condition in which one category was thematic while the other was neutral. This provided a situation in which it was possible to tease apart the different knowledge effects and assess their individual contributions to unsupervised learning.

There were two conditions in this experiment. In the first, both categories were similar to the neutral categories used in the last experiment. This will be referred to as the Neutral-Neutral (NN) condition. The categories in this condition were not expected to evoke contrasting themes, hence learning was expected to be relatively poor (as in the Neutral condition of Experiment 1). In the second condition, one category was neutral while the other was related to a specific theme. This will be referred to as the Thematic-Neutral (TN) condition. The contrast between neutral and thematic instances was expected to provide an effective cue indicating that separate categories were present in this condition. Thus, both categories should have been learned better in the TN condition than in the NN condition.

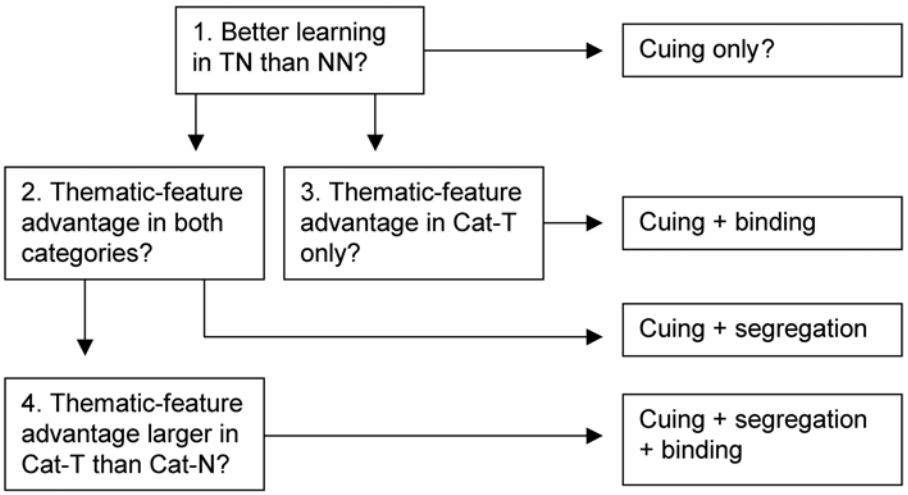

Figure 4. Possible outcomes (left) and interpretations (right) for Experiment 2.

Note that the neutral category was identical in both conditions of this experiment; only the nature of the *other* category was manipulated (neutral in the NN condition, thematic in the TN condition). Thus, the central independent variable in this experiment was the particular *context* in which the neutral category was encountered. Obviously, this manipulation could not directly affect feature binding within the neutral category itself; it could only affect how easily that category could be distinguished and its features segregated from those of the other category.

The detailed predictions for this experiment are illustrated in Figure 4. As noted above, better overall learning of both categories in the TN condition, without significant differences between neutral and thematic features (Outcome 1 in Figure 4), could be explained by simply assuming that people are more likely to notice separate categories in that condition (a category cuing effect). However, better learning of thematic compared to neutral dimensions within the TN condition would provide evidence for additional segregation and/or binding effects.

A segregation effect implies that the presence of the theme in the TN condition would reduce the confusability of the categories' features, but only along dimensions that have thematic values in one of the categories (i.e., the thematic category, henceforth referred to as Cat-T). This reduced confusability should benefit both categories equally; thus, a significant thematic-feature advantage for both categories (Outcome 2) can be taken as evidence for a segregation effect. By contrast, the theme would only imply stronger binding among actual thematic values (i.e., within the thematic category itself). Thus, a thematic-feature advantage within the thematic category, but not within the neutral category (henceforth referred to as Cat-N; see Outcome 3), can be taken as evidence for such a binding effect.

Of course, it is entirely possible that prior knowledge might simultaneously strengthen feature binding within the thematic category while also enhancing feature segregation between both categories. If both categories show a thematic-feature advantage, but this advantage is larger in the thematic category than in the neutral category (Outcome 4), the most likely explanation is that knowledge facilitates both segregation and binding. Enhanced segregation would tend to benefit both categories equally, while enhanced binding would provide an additional benefit, but only within the thematic category.

Method

Participants. Fifty-three undergraduate students of California State University San Bernardino participated in exchange for extra credit in several psychology classes.

Procedure. The procedure was the same as that of Experiment 1. The first phase consisted of 32 study-test trials while the second phase consisted of 30 pairwise correlation ratings using the same 5-point scale as before.

Materials and design. The same person descriptions were used as in Experiment 1. Only one theme from each contrasting set (highbrow/lowbrow, female/male, senior citizen/adolescent) was used as a basis for categories in this experiment. In each case, the theme judged by the experimenter and his research group to be the more obvious or salient within its contrast set was the one used (highbrows, females, senior citizens).

TABLE 4
Correlation Rating Data from Experiment 2

Condition	Dimension	Probe Type: Same-Category	Probe Type: Different-Category	Discrim. Score
TN	Thematic	4.44	1.60	2.84
TN	Neutral	4.24	2.10	2.14
NN	Neutral	4.14	2.25	1.89

TABLE 5
Recognition Memory Data from Experiment 2

Condition	Dimension Type: Thematic	Neutral	Uncorrelated
TN (Pooled)	.945	.917	.850
TN (Cat-T)	.944	.920	.852
TN (Cat-N)	.947	.913	.848
Neutral	*	.889	.872

* not present in that condition

Participants were randomly assigned to two conditions. In the "Thematic-Neutral" (TN) condition (N = 28), one category (the thematic category, here referred to as Cat-T) had values 111111333XXX and the other (the neutral category, referred to as Cat-N) had values 444444444YYY. Value 1 was related to a specific theme while Values 3 and 4 were neutral. Note that six out of nine correlated values were related to a specific theme in Cat-T, while Cat-N had all neutral correlated values. The dimensions represented by the last three positions were uncorrelated, with "X" denoting values of 1, 3, or 4 and "Y" denoting values of 2, 3, or 4 (as in Experiment 1). In the second, "Neutral-Neutral" (NN) condition (N = 25), one category had values 333333333XXX and the other had values 444444444YYY (i.e., all the correlated values of both categories were neutral).

Instances of the two categories were presented in a pseudo-random sequence in both conditions, with no more that 3 instances of a given category permitted to occur in a row. The ordering of instances was the same in the two conditions, the only difference being that each instance of Cat-T from the Thematic condition was replaced by an instance of a second neutral category in the Neutral condition. All other counterbalancing measures were the same as in Experiment 1.

Results

Correlation ratings. As in Experiment 1, discrimination scores were computed by subtracting ratings for pairs of correlated features from different categories from those for pairs from the same category.[3] Mean

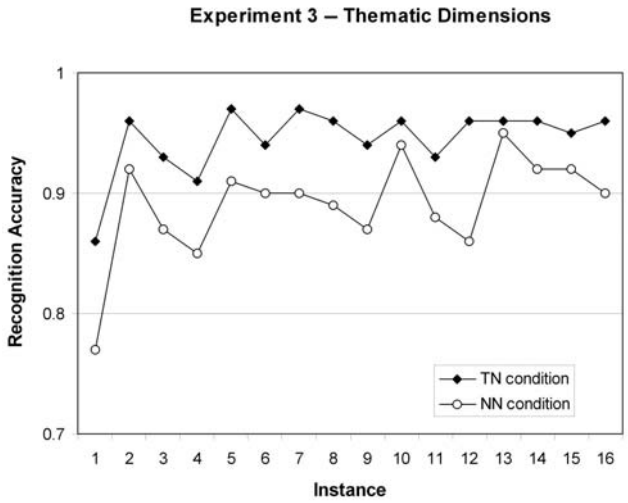

Figure 5. Recognition memory data for Experiment 2, plotted over successive category instances within both conditions (thematic dimensions only).

[3] All the analyses reported here were based on data from the top 75 % of learners within each condition, as estimated by discrimination scores for thematic attributes. As the same criterion was applied in both conditions, this is a fair procedure and should not bias the results in any way. In fact, it did not alter the pattern of results in any case, but eliminating the (presumably) less motivated and attentive participants did tend to increase the power of the statistical tests in this experiment.
As in Experiment 1, a single participant in this experiment showed a deviant pattern of data in the study time task, spending almost all his/her study period on a single neutral correlated dimension. This participant is excluded from the following graphs and analyses, although in no case did this alter the statistical significance or nonsignificance of any comparison.

[4] Note that the statistical power (e.g., Cohen, 1992) of these tests exceeded the conventional level of .80 for effect sizes greater than 1.18 points in between-groups tests and 0.86 points in within-group tests, assuming a standard deviation of 1.3 points (estimated from Experiment 1). For comparison, the overall difference for correlated features between Cat-T versus Cat-N within the TN condition was only about 0.04 points. Thus, it seems unlikely that the lack of any significant differences between these categories was merely due to a lack of statistical power. The other theoretically interesting null result from this data was the lack of difference between the TN versus NN conditions for neutral correlated dimensions. At 0.25 points, this difference was considerably larger than the 0.04 point Cat-T versus Cat-N difference within the TN condition, but considerably smaller than the 1.0 point between-groups difference for neutral dimensions from Experiment 1. It is conceivable that the relatively small difference observed in the present experiment might have been significant had a larger number of participants been used.

scores for the relevant probe types are shown in Table 4. There were no significant differences[4] between the Cat-N versus Cat-T within the TN condition ($t < 1.0$ for all comparisons) and hence the two categories are pooled in Table 4.

Discrimination scores were higher for thematic than for neutral dimensions in both categories of the TN condition, $t(20) = 2.33$, $SE = .30$, $p = .031$. Discrimination scores for thematic dimensions from the TN condition were also significantly higher than scores from the NN condition, $t(38) = 2.38$, $SE = .40$, $p = .022$. This difference remained significant when the data from the TN condition was restricted to either Cat-T, $t(38) = 2.39$, $SE = .42$, $p = .022$, or Cat-N, $t(38) = 2.29$, $SE = .40$, $p = .028$. Neutral dimensions were also rated slightly higher in the TN condition than in the NN condition, but this difference failed to reach statistical significance, $t(38) < 1.0$.

Study times. Within both the TN and NN conditions, correlated dimensions had lower average study times than did uncorrelated dimensions, $t(20) = 2.82$, $SE = .18$, $p = .011$ and $t(18) = 3.65$, $SE = .17$, $p = .002$, respectively. The between-groups analyses showed no significant differences between the TN and NN conditions for either thematic, neutral, or uncorrelated dimensions (all t-values less than 1.0, p-values greater than .50). Thus, as in Experiment 1, the study time data from this experiment failed to reveal any significant differences in learning between the two conditions.

Memory. The recognition memory data from this experiment are shown in Table 5. Figure 5 shows recognition memory plotted over trials for thematic dimensions from both conditions. Since there were no significant differences between the two categories in the TN condition, they are pooled in most of the analyses described below.[5]

Within the TN condition, thematic dimensions were remembered better than either neutral, $t(20) = 3.21$, $SE = .01$, $p = .004$ or uncorrelated dimensions, $t(20) = 3.66$, $SE = .03$, $p = .001$, and neutral dimensions were remembered better than uncorrelated dimensions, $t(20) = 3.01$, $SE = .02$, $p = .006$. In the NN condition, memory for correlated (i.e., neutral) dimensions marginally exceeded that for uncorrelated dimensions, $t(18) = 1.77$, $SE = .01$, $p = .093$.

In between-groups comparisons, correlated (thematic) dimensions were remembered better in the TN condition than were correlated (neutral) dimensions in the NN condition, $t(39) = 2.21$, $SE = .03$, $p = .033$. This comparison remained significant when restricted to either Cat-T, $t(39) = 2.13$, $SE = .03$, $p = .04$, or Cat-N, $t(39) = 2.21$, $SE = .06$, $p = .033$, from the TN condition. There were no significant differences between conditions in memory for neutral, $t(39) = 1.12$, $SE = .03$, $p = .27$, or uncorrelated dimensions, $t(39) = .57$, $SE = .04$, $p = .57$.

Discussion

As expected, people in the TN condition showed better learning of the thematic dimensions of both categories than did those in the NN condition, as well as better learning of thematic compared to neutral dimensions within both categories of the TN condition itself. There were no significant differences of any kind between the two categories within the TN condition.

As illustrated in Figure 4, each type of knowledge effect should be identifiable by its own distinctive "signature" in this experiment. First, enhanced cuing should increase the probability of creating separate categories and therefore result in better overall learning of both categories within the TN condition. However, this cuing effect should not, by itself, favour learning one type of consistent feature over another; it would simply make it possible for people to capture the correlational patterns in the stimulus set by dividing the stimuli into separate categories. Second, enhanced feature segregation should increase learning of thematically relevant dimensions in both categories, due to reduced confusability of the values along these dimensions. The result should be a thematic-feature advantage for both categories in the TN condition. Third, stronger binding should also result in a thematic-feature advantage in the TN condition, but only within Cat-T; due to its lack of thematic values, Cat-N could not directly benefit from this effect of within-category structure. If a thematic-feature advantage is found in Cat-N as well, then the size of the binding effect would be indicated by the degree to which this advantage is larger in Cat-T than in Cat-N.

Overall, the data from this experiment provide evi-

[5] The power of the following tests exceeded .80 for effect sizes of 8% in between-groups tests and 6% in within-group tests, given a standard deviation of 9% on the recognition memory measure (estimated from Experiment 1). The overall difference in memory (averaging all dimensions together) between Cat-T and Cat-N within the TN condition was only about 0.1%; thus, as with the correlation ratings, the lack of significance in this comparison is unlikely to be due to inadequate statistical power. The other critical null difference in this data was that between the TN versus NN conditions in memory for neutral correlated dimensions. The observed difference of 2.8% was obviously much larger than the 0.1% difference between Cat-T and Cat-N within the TN condition, but considerably smaller than the corresponding 6.5% effect from Experiment 1. As for the correlation rating data from this experiment, it is at least conceivable that this difference might have been significant given a larger sample size.

dence for cuing and segregation effects, but not for any additional binding effect. Thus, the fact that both categories were learned better in the TN condition than in the NN condition suggests that prior knowledge increased the probability of creating separate categories in the TN condition. The results for Cat-N are particularly informative because the only way that thematic knowledge could improve learning of this thematically unrelated category would be by increasing its perceived contrast with Cat-T. In addition, the data showed a significant advantage for "thematic" dimensions within Cat-N compared to the corresponding dimensions in the NN condition and also compared to the "neutral" dimensions within the same category. This indicates a significant, knowledge-based enhancement of feature segregation along the thematically relevant dimensions of the two categories.

Given the existence of this segregation effect, any additional binding effect should have further increased the size of the thematic-feature advantage within Cat-T. However, there were no significant differences between the two categories in the TN condition. Thus, there was no clear evidence for a direct feature binding effect in this experiment

General Discussion

Both of these experiments showed robust knowledge effects on unsupervised learning. Experiment 1 showed that prior knowledge facilitates learning of both thematic and non thematic features of a new category, and provided evidence that this effect was somewhat greater for thematic features. Experiment 2 showed that a neutral category, not specifically related to a particular theme, was learned better in the presence of a thematic category than in the presence of another neutral category. Moreover, Experiment 2 replicated the thematic-feature advantage shown in Experiment 1 (the benefit for neutral features obtained in Experiment 1 was not repeated in Experiment 2, but the general pattern of means was in the expected direction, see Tables 4 and 5). It is clear that prior knowledge must have enhanced learning indirectly in Experiment 2, by highlighting the contrast between the two categories rather than by strengthening the connections among features within Cat-N.

Both experiments showed a thematic-feature advantage when categories contained both thematic and neutral correlated features. One question that might legitimately be asked, however, is whether this thematic-feature advantage could have been due to a guessing bias in favour of thematic features rather than a real difference in learning. For example, participants could have adopted a strategy dictating that, when in doubt, they should select the alternative consistent with the present theme in the recognition task and rate same-theme pairs higher than other pairs in the correlation rating task. However, it is important to recall that half or fewer of the dimensions had thematic values in the Thematic-plus-Neutral condition of Experiment 1 and the TN condition of Experiment 2. This would have rendered such a guessing strategy rather costly and impractical; in particular, it would have reduced memory accuracy for neutral and variable dimensions relative to the all-Neutral control conditions, contrary to the actual data. A similar argument applies to the correlation ratings – a blanket strategy of rating same-theme pairs higher than other pairs would have had significant costs because many dimensions did not have consistent thematic values.

Perhaps the most convincing rebuttal to the guessing-strategy argument is the fact that people also showed a thematic-feature advantage in Cat-N of Experiment 2. Of course, the "thematic" features in that category were not really thematic at all, and were labeled as such for convenience only – they were actually neutral values of dimensions that happened to have thematic values in the opposing category. There is no way that a thematic guessing strategy could have produced a memory benefit for these features. Thus, the "thematic" feature advantage within this condition is probably the strongest and most unambiguous evidence for feature segregation effects obtained from these two experiments.

Overall, these experiments provide strong evidence for "between-category" knowledge effects (cuing, segregation), but no direct evidence for "within-category" (binding) effects. In other words, the primary factor that seemed to affect the probability of detecting a given category, as well as the learnability of the individual features within that category, was the specific *context* in which the category was encountered. If the *other* category being learned pertained to a clear theme that marked it as different from the target category in question, people were not only more likely to recognize the existence of the target category itself, but also showed better learning of dimensions that had thematic values in the other category. The overall picture that emerges is one in which the primary benefits of prior knowledge were related to enhanced discrimination between different categories, rather than enhanced binding or cohesion within a given category. Another way to say this is that knowledge effects in this experiment appeared to operate at the level of category *systems*, making different categories within such a system more or less discriminable, rather than affecting the learnability of individual categories independent of context.

The absence of a direct binding effect might seem

surprising given the importance generally attached to prior knowledge in specifying relationships among features and increasing overall coherence within a category (e.g., Ahn, 1999; Kaplan & Murphy, 1999; Murphy, 2002; Murphy & Medin, 1985). However, the failure to detect a binding effect in these experiments does not imply that binding effects do not exist or that they played no role in participants' actual learning process. On the contrary, easier segregation in the TN condition might be an indirect outcome of stronger binding within Cat-T in that condition. In other words, the fact that thematic features are bound together by a common theme should make it easier to distinguish them from the features of Cat-N, and vice-versa (i.e., greater within-category cohesion may result in greater between-category discriminability). If so, improved segregation between the two categories might be viewed as a side-effect of stronger binding within the thematic category.

In addition, the lack of binding effects could have been due in part to the way the materials in this experiment were designed. Thus, if a participant remembered that the last person he/she studied was a wealthy highbrow, he/she could be reasonably certain that this person would not have been described as driving a battered pickup truck. Knowing that someone is wealthy can tell us with a fairly high degree of confidence what he/she does *not* drive, but it does not tell us with equal confidence what he/she does drive. A rich person might drive a Rolls Royce, a Ferrari, a Lamborghini, or a wide variety of other expensive cars. The point is that the thematic values in this experiment are, for the most part, *consistent* with the themes but not directly *implied* by those themes (i.e., they are not thematic *defaults*). For this reason, the present themes might have been more useful for rejecting features that did *not* belong to a category than for remembering the actual features that it did possess. If the themes had specified pre-existing default values for thematically relevant dimensions, then direct binding effects might have been observed in this experiment.

Implications for Models of Category Learning

An important question about these results concerns their implications for specific computational models of category learning. Most existing models of category learning are concerned with supervised rather than unsupervised learning, and so are not directly relevant to this discussion. Among those models that are concerned with unsupervised learning, knowledge effects have so far received relatively little attention. The strategy taken in this article has been to adapt the category invention framework of Clapper and Bower (1991, 1994, 2002; Clapper, in press) to highlight the various roles that knowledge might play in unsupervised learning. In previous work, the rational model of J. R. Anderson (1990, 1991) has been cited as a computational example of this category invention framework, and computer simulations of that model have been used make detailed predictions regarding performance in the present task (Clapper, in press; Clapper & Bower, 2002). Therefore, it might seem appropriate to apply the rational model in an attempt to account for the present results.

Unfortunately, the rational model as currently formulated does not provide any completely satisfactory way to account for these results. For example, there is no obvious mechanism within that model by which thematic relevance could affect the probability of creating separate categories or the learning of features within those categories. In addition, the model appears unable to account for between-category feature segregation effects. Once separate categories are created by the rational model, there is little opportunity for confusion or interference in learning the features of such categories. For example, assuming that the model created separate categories in the TN condition of Experiment 2, the fact that certain dimensions had thematic values in one category (Cat-T) would have no influence on the ease of learning the nonthematic values of these dimensions in the other category (Cat-N).

Two models that have been formulated to deal explicitly with the effects of prior knowledge are the Baywatch model of Heit and Bott (2000) and the KRES model of Rehder and Murphy (2004). Both of these are formulated as models of supervised classification learning tasks and would require some elaboration before they could be applied to the present unsupervised task. However, they do provide a way to think about the "characterization" stage of learning that must occur following the invention of separate categories. In both models, a collection of feature units are prewired to a set of category units, and exposure modifies the strengths of these connections. The models explain better learning of thematic features in supervised learning experiments by assuming that the thematic features activate a pre-existing "theme" or "prior knowledge" node, and that this node becomes more strongly connected to the category with experience, thereby facilitating memory for the thematic features associated with it. (KRES also assumes that connections between related features might be stronger than those between neutral values at the start of the model run, providing a second way in which prior knowledge could facilitate learning thematic features; see Rehder & Murphy, 2004).

The main difficulty with these models as presently formulated is that their account of knowledge effects is, in essence, an implementation of the feature binding principle – thematic features are assumed to be learned

better due to pre-existing associations with the theme and/or each other. Thus, knowledge effects are framed as within-category effects, rather than between-category context effects as implied by the present results. It seems likely that the models could be elaborated to account for feature-level interference between related categories and to show how knowledge could play a role in reducing that interference. So far, however, those issues have not been explored within these models.

Another model that deals with the role of knowledge in the learning process is the Integration model of Heit (1994, 1998). This model assumes that prior knowledge has an effect that is equivalent to having extra instances of a category stored in memory. In the present experiment, one could imagine that thematic features might be learned better than neutral features due to this kind of integration effect – as if the learner had already been exposed to instances with thematic features prior to the experiment. Again, however, this provides no obvious explanation for the segregation effects observed for Cat-N in the TN condition of Experiment 2. Nor does the integration model explain how or on what basis people would create separate categories in the exemplar-memory task.

It is apparent from this discussion that a major blind spot of all these models is their exclusive focus on within-category knowledge effects and their failure to consider possible between-category (context) effects – that is, knowledge effects are treated as a function or property of individual categories rather than of the larger category system (stimulus set) of which they are a part. From this perspective, the primary contribution of the present results is to provide a clear demonstration of the importance of these between-category (in particular, feature segregation) effects. In that sense, the present results represent a significant challenge to existing models.

Conclusions

These experiments provide strong evidence regarding the beneficial role of prior knowledge in unsupervised learning, adding to earlier results by Ahn (1990, 1999), Spalding and Murphy (1996), Kaplan and Murphy (1999), and others. Further, the present data suggest that knowledge affects learning in specific ways – by helping people to distinguish different patterns (category cuing) and avoid confusion among corresponding features of related patterns (feature segregation). Although no direct feature binding effect was observed, enhanced binding may to some extent underlie the segregation effect shown in Experiment 2.

More broadly, these experiments demonstrate the multifaceted role of prior knowledge in new learning and show the usefulness of the exemplar-memory task for investigating how such knowledge effects occur. The present experiments represent an early attempt to begin disentangling and investigating the various influences of prior knowledge on unsupervised learning. Hopefully, later research will go further in terms of cleanly separating these different effects, estimating their relative importance as a function of different variations of tasks and materials, and integrating empirical data with formal and computational modeling.

I thank Briana Boyd, Noelle Camarena, Jenna DeVoid, Joe Hernandez, Brady Kafer, Karissa Rasdal, Sonja Seglin, Kristina Schmukler, and Sam Tafoya for their help in evaluating the stimulus materials and testing the participants in this research.

Correspondence concerning this article should be addressed to: John P. Clapper, Department of Psychology, California State University, San Bernardino, 5500 University Parkway, San Bernardino, CA 92407-2397 (Tel: (909) 537-3843; Fax (909) 880-7003; E-mail jclappe@csusb.edu).

References

Ahn, W. (1990) Effects of background knowledge on family resemblance sorting. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 149-156). Hillsdale, NJ: Erlbaum.

Ahn, W. (1999). Effect of causal structure on category construction. *Memory & Cognition, 27*, 1008-1023.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409-429.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aides learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 458-475.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956*). A study of thinking*. New York: Wiley.

Clapper, J. P. (in press). When more is less: Negative exposure effects in unsupervised learning. *Memory & Cognition*.

Clapper, J. P., & Bower, G. H. (1991). Learning and applying category knowledge in unsupervised domains. In G. H. Bower (Ed.), *The psychology of learning and motivation, Vol. 27*. New York: Academic Press.

Clapper. J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 443-460.

Clapper. J. P., & Bower, G. H. (2002). Adaptive categorization in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 908-923.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1264-1282.

Heit, E. (1998). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 712-731.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 39, pp. 163-199). San Diego, CA: Academic Press.

Kaplan, A. S., & Murphy, G. L. (1999). The acquisition of category structure in unsupervised learning. *Memory & Cognition, 27*, 699-712.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129*, 592-613.

Medin, D. L., & Bettger, J. G. (1994). Presentation order and categorically related examples. *Psychonomic Bulletin and Review, 1*, 250-254.

Medin, D. L., & Shaffer, M. M. (1978). A context theory of classification learning. *Psychological Review, 85*, 207-238.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*, 242-279.

Metcalfe, J. (2002). Is study-time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, *131*, 349-363.

Michalski, R. S., & Stepp, R. E. (1983). Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. Palo Alto, CA: Tioga Publishing.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus on knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 904-919.

Murphy, G. L., & Kaplan, A. S. (2000). Feature distribution and background knowledge in category learning. *The Quarterly Journal of Experimental Psychology*, *53A*, 962-982.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 416-432.

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. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 347-363.

Rehder B., & Murphy, G. L. (2004). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review*, 10, 759-784.

Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 525-538.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology, 18*, 158-194.

Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 449-468.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124-148.

Appendix

Stimulus Materials

Highbrow/Lowbrow Theme

Dimension	Values
1. *drinks*:	sherry; beer; coffee; cola
2. *favourite activity is*:	yachting; bowling; soccer; softball
3. *enjoys watching*:	the opera; pro wrestling; basketball; movies
4. *drives*:	a Mercedes; an old Pickup; a Honda; a Toyota
5. *lives in*:	Beverly Hills; Detroit; Sacramento; Tucson
6. *is*:	a lawyer; unemployed; an accountant; a technician
7. *favourite food is*:	Russian caviar; pizza; steak; fish
8. *last vacation was in*:	Paris; Tijuana; New Orleans; Disney World
9. *favourite music by*:	Mozart; Aerosmith; the Beatles; BB King
10. *clothes by*:	Armani; Discount-Mart; Old Navy; The Gap
11. *favourite TV show is*:	Masterpiece Theater; the Jerry Springer Show; the evening news; ER
12. *graduated from*:	Harvard; dropped out of high school; community college; state university

Senior/Adolescent Theme

Dimension	Values
1. *listens to*:	Frank Sinatra; Eminem; BB King; Paul Simon
2. *had _____ for breakfast today*:	oatmeal; cold pizza; a bagel; eggs
3. *favourite hobby is*:	playing checkers; playing video games; woodworking; hiking
4. *favourite TV show is*:	Murder, She Wrote; MTV Real World; ER; Friends
5. *wears*:	cardigan sweaters; hooded sweat shirts; business suits; cotton slacks
6. *is*:	retired; a student; an executive; a teacher
7. *would like to buy a*:	Cadillac; Ferrari; Volvo; BMW
8. *drink of choice*:	prune juice; Dr. Pepper; coffee; Calistoga water
9. *health*:	poor, often ill; excellent, never ill; average, occasionally ill; moderate, rarely ill
10. *member of*:	Bingo Club; local Skateboarder's Club; Sierra Club; Volunteer Fire Department
11. *favourite movie star is*:	John Wayne; Jennifer Lopez; Meryl Streep; Robert DeNiro
12. *goes to*:	shuffleboard games; "rave" parties; movies; live theatre

Female/Male Theme

Dimension	Values
1. *favourite sport is:*	figure skating; boxing; skiing; swimming
2. *hobby is:*	shopping; fixing cars; tennis; painting
3. *favourite TV show is:*	Martha Stewart; Monday Night Football; ER; Friends
4. *favourite movie is:*	the Ya-Ya Sisterhood; The Terminator; Henry the Fifth; Rear Window
5. *favourite books are:*	romance novels; science fiction novels; biographies of famous people; humorous or satirical
6. *typical clothing accessory:*	silk scarf; tie; jacket; vest
7. *employed as:*	a model; an engineer; a teacher; a realtor
8. *dream vacation:*	Paris fashion tour; Superbowl road trip; Disney Land; New Orleans
9. *favourite store is:*	Greystone Jewelry; Pierson's Hardware; Northtown Books; Sjaak's Chocolates
10. *perfect gift:*	diamond earrings; power tools; clothes; money
11. *reads the _____first:*	health and beauty section; sports section; front page; classified ads
12. *favourite magazine is:*	Vogue; Popular Mechanics; Newsweek; Rolling Stone

Sommaire

People often learn new categories without external guidance or feedback. The two experiments described in this article attempted to distinguish some of the ways in which a learner's prior knowledge might facilitate such unsupervised learning.

Both experiments used the exemplar-memory task of Clapper and Bower (2002). Participants were shown a series of instances from two different categories, with instructions to memorize the features of each instance (categories were never explicitly mentioned). The stimulus displays were designed so that the computer could track the amount of time people spent attending to each feature. In addition, recognition memory for the features of each instance was tested immediately after that instance was shown. If people detected the categories, memory for their consistent features was expected to improve steadily over trials. Study times were also expected to decrease over trials as these features became easier to learn and remember.

Participants in Experiment 1 were assigned to three conditions. In the Thematic condition, the consistent features of the two categories were related to familiar contrasting themes (social stereotypes such as highbrow vs. lowbrow, male vs. female, or senior vs. youth). In the Neutral condition, the consistent features of both categories were unrelated to any specific themes. In the Thematic-plus-Neutral condition, some features of each category were thematically-relevant while others were thematically neutral. Learning was better in the two thematic conditions than in the Neutral condition, indicating that the themes helped people notice the existence of separate categories (a "category cuing" effect). Both neutral and thematically-relevant features of thematic categories were learned better than the corresponding features of neutral categories. In addition, thematic features were learned better than neutral features within the thematic categories. However, it was not clear whether the latter difference occurred because thematic features were more strongly pre-associated in memory than neutral features (a "feature binding" effect), or because they were less confusable across category boundaries (a "feature segregation" effect).

Experiment 2 was designed to resolve this issue. In the Thematic-Neutral (TN) condition, one of the categories was thematically-relevant while the other was thematically neutral. In the Neutral-Neutral (NN) condition, both categories were neutral. Both categories in the TN condition (including the neutral category) were learned better than the categories in the NN condition. Within the TN condition, thematically relevant dimensions were learned better than thematically neutral dimensions; this was equally true for both neutral and thematic categories. This pattern of results suggests that prior knowledge improved learning via enhanced feature segregation, but provides no evidence for an additional feature binding effect. Existing models generally cannot account for these data because they focus exclusively on within-category coherence (feature binding) while ignoring the impact of prior knowledge on between-category discriminability (feature segregation).