Explanation in Category Learning

Seth Chin-Parker, Olivia Hernandez, and Murray Matens
(contact chinparkers@denison.edu)
Department of Psychology, Denison University
Granville, OH 43023 USA

Abstract
In this study, we examine how explanation can be implicated in category learning. We asked participants to explicitly state what they considered to be important in explaining the category membership of individuals in novel social groups. We examine these explanations, focusing on how prior knowledge was integrated with the empirical information gleaned from the learning examples. We also compare the learning outcomes of explanation learning with those of classification learning. The explanation learners developed an understanding of the categories that was more knowledge-based than the classification learners. We discuss how it is important to consider category-learning paradigms like explanation learning to better understand how prior knowledge affects category learning.

Keywords: category learning; classification; explanation

Introduction
Outside my window, people walk across the academic quad. Some of them are students, some are faculty, and some are staff. Somehow, I have acquired the knowledge necessary to identify and interact with these various categories of individuals. The issues related to how I succeed at such tasks are of interest to philosophers, psychologists, and computer scientists.

In this paper, we consider two points. First, category learning is a rich and multifaceted activity, but this is not necessarily reflected in the basic research being done in cognitive psychology. Several lines of study have investigated unsupervised learning (e.g. Love, 2003), indirect learning (Minda & Ross, 2004), and inference learning (e.g. Yamuachi & Markman, 1998), but the primary focus of the field has been on the classification-learning paradigm (see Markman & Ross, 2003). Second, prior knowledge is important to category learning (see Murphy, 2002, for a review), but the theoretical framework available to understand its influence is not fully developed. These two points are not independent—we argue that in order to study how prior knowledge is implicated in learning we have to utilize different learning paradigms than those typically used. The study presented here is intended to offer another means of assessing how prior knowledge can be implicated in category learning.

The primary focus of this study is on the role that explanation plays in category learning. We certainly do not claim this is a new idea. Developmental psychologists have been working with this idea for decades (e.g. Carey, 1985). Also, roughly 20 years ago, the machine learning literature was teeming with ideas of how explanation-based learning (EBL) could be used to resolve critical problems that intelligent systems face when learning (a useful overview is provided by Ellman, 1989). The basic idea of the approach was that by providing an intelligent system with a set of constraints and connections, the prior knowledge, the system is better able to identify relevant features of an example and thus be able to develop a generalization of the example.

However, this work has not had a strong impact on work examining category learning. Notable exceptions include Schank, Collins, and Hunter’s (1986) attempt to redirect the study of category learning away from strictly inductive learning. Ahn, Brewer, and Mooney (1992) provided a useful and insightful study about how EBL could account for certain learning situations, specifically how people can learn to generalize from a single example of a domain when appropriate prior knowledge is available. Pazzani (1991) possibly went the furthest in integrating the ideas of EBL with evidence from basic psychological studies. However, the role that explanation plays in category learning has not been fully appreciated or developed.

An important paper by Murphy and Medin (1985) inspired some research into the role of prior knowledge in category learning (e.g. Murphy & Allopenna, 1994; Rehder, 2003; Wisniewski, 1995). However, much of this work has been done firmly within parameters established by the dominant classification-learning paradigm (e.g. Medin & Schaffer, 1978). In this learning paradigm, a participant learns about categories by classifying items into (typically) two categories and receiving feedback about those classification decisions. The reliance on this learning paradigm in psychological research has been recently questioned (e.g. Markman & Ross, 2003). In this study, we hope to show the constraints of the classification-learning paradigm can limit the use of prior knowledge during learning.

We introduce the paradigm of explanation learning as a form of category learning. In this learning paradigm, the participant is presented with examples from experimenter-defined categories and is asked to generate an explanation as to why the presented item should be considered a member of the specified category.

This explanation-learning paradigm is interesting since the participants do not receive any feedback during the learning. We predict that prior knowledge will provide sufficient constraint on the learning to guide the development of appropriate representations of the
categories even in the absence of explicit feedback. Our basic idea of the mechanisms and processes involved are similar to Thagard (2000) in that explanations will develop from and be constrained by the knowledge participants bring to the learning task. In the current study, these constraints will affect how the different types of features used to describe the category members are integrated into the explanations and thus the category representations available for making later category-based decisions.

We developed categories comprised of features that varied in their relevance to the underlying sense of the categories as well as how often the specific feature value occurred across the category members. For instance, the relevant features are informative with regard to prior knowledge and are relevant to the category membership, but the specific instantiation of these features is unique to each category member. Thus, they should allow the learner to integrate prior knowledge into his or her representation of the category, and this integration should be illustrated by explanations that use more abstract components to explain the specific feature values of each individual. The meaningful irrelevant features are informative, but they are not relevant to the category membership. So, they may be mentioned in the explanations, possibly even abstracted to some extent because of their informativeness, but we predict they will not become as central to the participant’s understanding of the category. Finally, the diagnostic irrelevant features are not related to prior knowledge within the domain of social groups, but the features are perfectly predictive of the category membership of an item. These features may be mentioned in the explanations, but there should not be any integration with prior knowledge because of their lack of informativeness. As a result, we predict these features will not be central to the participants’ understandings of the categories.

For the participants that learn about the categories through the classification-learning paradigm, we make very different predictions. Since the diagnostic irrelevant features are perfectly predictive of the category membership, they should become the focus during the learning. Also, since the other features vary in their specific instantiations across the category members, the classification learners should have difficulty appreciating their relationship to the underlying category structure (Yamauchi & Markman, 2000).

The current study is intended to provide two contributions to the study of human category learning. First, we examine explanations used by participants as they learn about categories of social groups, a domain in which they have prior knowledge. Second, we provide a comparison of the learning outcomes of explanation learning and classification learning.

Experiment

Participants Twenty-six undergraduates from a small Midwestern liberal arts college participated in the study in exchange for $10. Participants were randomly assigned to a learning condition. All participants signed an informed consent and were fully debriefed.

Design The study consisted of two parts. The first part of the study was an examination of explanations generated during the course of learning about two novel social categories. The purpose was to determine what information was being used in the explanations. The second part of the study consisted of a comparison of the learning outcomes between participants that learned by explanation and those that learned by classification.

Materials The learning items consisted of descriptions of fictitious individuals that belonged to two experimenter-defined social clubs. The social clubs were labeled as the “Blue Club” and the “Purple Club.” The Blue Club consisted of four individuals that could be described as “social.” The Purple Club consisted of four individuals that could be described as “caring.”

Each individual was presented to the participants as a description that consisted of four features. The specific values that instantiated the features were chosen based on pretesting. One group of undergraduates came up with characteristics that they associated with several different personality types. These features were then rated by a different group of undergraduates in order to determine the general strength of the associations between the characteristics and the personality types and the level of agreement of those ratings across individuals. For instance, the feature value “visits a friend in the hospital” was rated as being strongly associated with a “caring” person, and there was perfect agreement across raters. The same feature value was rated as being weakly related to a “social” personality type, and negatively related to an “aggressive” type.

Two of the features within each description were relevant features (see Table 1 for an abstract rendering of the category structure). These features were directly related to the underlying category essence, either social or caring. The specific values for the relevant features were unique to each individual. Even though each feature value appeared only once across the category members, the relevant features all pointed to the category essence. The other two feature values were not related to the category essence, but they appeared more consistently within the category members. The meaningful irrelevant features were related to a cautious personality type in both of the categories. Each meaningful irrelevant feature value was shared by two of the four items within each category. The final feature type was the diagnostic irrelevant feature; prior rating found the feature values that instantiated these features, “Drinks Pepsi” and “Owns a laptop”, to be unrelated to any specific personality type. The diagnostic irrelevant feature value was the same for all four items within each category.

The descriptions of the individuals in each category were printed on 3” by 5” note cards. Four sets of the note
The experimenter provided feedback and placed the card onto the sheet that the participant could use to study the relevant and diagnostic irrelevant features. The recorded explanations were coded as to what information was used by the participant. The mention of a specific feature value (whether relevant, meaningful irrelevant, or diagnostic irrelevant) was noted. These will be referred to as concrete components of the explanations. The generated explanations were coded for both hierarchical and abstract components (adapted from Wisniewski and Medin, 1994). A hierarchical component is one that relates to a specific feature value (a concrete component), but goes beyond it in scope. For example, the item on a given trial might contain the feature value, “volunteers at the hospital,” and the participant might say in his or her explanation, “This person helps those in need.” Since helping is directly related to volunteering, and we often think of those in a hospital as being “in need,” this component was considered hierarchical. An abstract component of an explanation is a general personality trait that is not directly tied to the concrete feature values present during a given trial. For instance, if the participant said, “This person is caring,” the explanation was coded as having an abstract component.

We also wanted to assess the focus of the participant during the explanation process. We counted the number of inferences that the participants made about the individual, as well as about the group. These inferences were related to the hierarchical and abstract components mentioned above. The purpose of this coding was to allow us to ascertain whether the participant was using the explanation to develop an understanding of the individual or the group. We also coded whether the participants were comparing the information about an individual with others in the same group or between the two groups.

There were several questions that we hoped to answer by means of evaluating the explanations. Did the meaningfulness of the features have an effect on their use in the explanations? How were the participants using the specific feature values to create more hierarchical features? What kind of abstract components were included? Were the participants using the explanations to map the specific feature values to the individual or the social group? Were comparisons being made between members of the same group, or were the comparisons being made between the
two groups? The analyses that follow provide some answers to these questions.

Table 2: Explanation Components by Feature Type

<table>
<thead>
<tr>
<th>Explanation Component</th>
<th>Concrete/Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>0.47 (0.23)</td>
</tr>
<tr>
<td>Meaningful Irrelevant</td>
<td>0.38 (0.31)</td>
</tr>
<tr>
<td>Diagnostic Irrelevant</td>
<td>0.55 (0.32)</td>
</tr>
</tbody>
</table>

Means: Mean proportion of trials (and standard deviation) that included these components reported.

Did the meaningfulness of the features have an effect on their use in the explanations? We determined the mean number of times within each trial the explanations contained the various concrete components (see Table 2). We adjusted the relevant feature proportion, dividing it by two, since there were two of those feature values in each item and one each of the meaningful irrelevant and diagnostic irrelevant feature values. We used a repeated-measures ANOVA to compare the proportional use of relevant, meaningful irrelevant, and diagnostic irrelevant features in the concrete components. There was a non-significant difference in the number of times the participants used the different types of features as concrete components, *F*(2, 22) = 2.836, *MSE* = 0.009, *p* = 0.08, *η*^2^ = 0.21.

How were the participants using the specific feature values to create more hierarchical features? We determined the use of the different feature types as the basis for hierarchical components within the explanations. We again adjusted the calculated proportion of relevant features. A repeated-measures ANOVA revealed a significant difference in which feature types were used as the basis for hierarchical components, *F*(2, 22) = 7.699, *MSE* = 0.186, *p* < 0.01, *η*^2^ = 0.84. Post-hoc comparisons between the three feature types showed no difference between the use of the relevant and meaningful irrelevant features, *t*(11) = 0.22, *p* > 0.20, but large differences between the use of relevant and diagnostic irrelevant features, *t*(11) = 5.18, *p* < 0.01, and meaningful irrelevant and diagnostic irrelevant features, *t*(11) = 3.65, *p* < 0.01.

What kind of abstract components were included? The explanations included many abstract components, with varying degrees of relationship to the personality traits we used to construct the two groups. The analysis here focuses on how the abstract components relate to the types of features used in the items. We found no evidence that any abstract component was related to a diagnostic irrelevant feature in any explanations generated. However, 41.70% of the trials (*SD* = 22.80) contained an abstract component related to the relevant characteristics, and 7.30% of the trials (*SD* = 13.30) contained one related to the meaningful irrelevant feature. The use of these feature types to generate abstract components was significantly different, *t*(11) = 4.89, *p* < 0.01. We did not adjust the relevant feature use for this analysis because it is impossible to determine whether an abstract component was related to one or both of the relevant feature values present. The adjusted use of the relevant feature (*M* = 0.83%, *SD* = 11.40), is still significantly larger than the meaningful irrelevant feature.

Were the participants using the explanations to map the specific feature values to the individual or the social group? The participants could generate their explanations with regard to the individual or the group being considered. For instance, a participant could say, “She is kind for volunteering at the hospital,” or “She is in the blue group because they are kind.” We determined the number of times per trial on average each participant made the two types of connections. The participants used both connections to the individual (*M* = 1.52, *SD* = 0.71) and the group (*M* = 0.36, *SD* = 0.34) in their explanations. However, the individual was more often the focus of the explanation, *t*(11) = 5.43, *p* < 0.01.

Were comparisons being made between members of the same group, or were the comparisons being made between the two groups? For each participant, we determined the number of times per trial there was a comparison made to another “person” within the social group being considered, and the number of times per trial the comparison was to the other social group. For instance, the participant might say, “She volunteers at the hospital. That’s like the one who helped sell Girl Scout cookies,” or “She volunteers at the hospital unlike the Purple Group people who just party.” On average, participants made more within-group comparisons (*M* = 0.48, *SD* = 0.34) than between-group comparisons (*M* = 0.17, *SD* = 0.15), *t*(11) = 2.95, *p* = 0.01, but did make some of each type.

**Explanation versus Classification Learning Conditions**

There were three measures of interest in this portion of the study: classification of old (learning) items, classification of conflict feature pairings, and classification of novel items. For the old items, we determined each participant’s accuracy, and then calculated the mean of both groups. For the conflict feature pairings, we determined the proportion of responses by each participant that indicated he or she was classifying the feature pairing according to the relevant feature value. We separated these data also by whether the conflicting feature was a meaningful irrelevant or diagnostic irrelevant feature and determined the mean for each type of item for each condition. The novel items were analyzed in the same manner as the old items.

There was no difference in the ability of the groups to classify the old items. Every participant in the study perfectly classified the set of old items.

The data from the conflict feature pairings (see Table 3) were analyzed using a mixed 2 (classification/explanation learning) X 2 (meaningful/diagnostic irrelevant) ANOVA. There was a main effect of learning, *F*(1, 24) = 91.421, *MSE* = 4.674, *p* < 0.01, *η*^2^ = 0.79, showing that the explanation learners classified the pairings more often.
The pattern of results found across the three classification tests provides an interesting glimpse into the learning outcomes of both classification and explanation learning. Both groups were perfect when classifying the items they encountered during learning. However, when the results of the conflict feature pairings are considered, it is obvious that this equivalence in performance is not due to the fact that the participants in the two groups acquired the same knowledge about the categories. The explanation learners focused on the relevant feature to guide their classification.
of the feature pairs. The classification learners did not. The classification learners showed a strong preference for classifying the feature pairs according to the diagnostic irrelevant feature when it was available, and then showing no real preference as a group otherwise. This would seem to indicate that the classification learners were focused on the highly diagnostic feature to the exclusion of other information (see Chin-Parker & Ross, 2004). This would make sense since it would be the most efficient learning strategy available to the classification learners, However, one could argue that focusing on the diagnostic irrelevant feature would also be the most efficient strategy for the explanation learners (“This person is in the blue group because she owns a laptop.” – could be the explanation for every member of that category.) Interestingly, the explanation learners showed a slight avoidance of the diagnostic irrelevant feature. Possibly, the explanation learners realized those feature values were uninformative (irrelevant) with regards to their understanding of the categories and so they were purposely not used during this task. The explanation learners mentioned the diagnostic irrelevant features during the learning, so it is not the case that they simply didn’t know about them. The forced-choice nature of the task also would have emphasized the differential reliance on the relevant features.

The results of the novel item classification task underscore a critical difference in the knowledge acquired from the two learning tasks. The explanation learners were very accurate when classifying the items that shared the abstract sense of the categories specified by the learning items but none of the specific feature values. The explanation-learning task allowed the participants to develop an understanding of the categories that went beyond the instantiations of the features seen during learning. This indicates that the explanation learners were able to develop a more knowledge-based representation of the categories. The classification learners were unable to classify the novel items because they had been focused on the occurrence of specific features during the learning. Their representations of the categories were more sparse, less connected to available prior knowledge, and, we would argue, less useful for later tasks such as predicting missing features or communicating about the categories.

Acknowledgments

This work was supported by a Hughes Early Research in the Sciences summer scholarship awarded to O. Hernandez and an Anderson Summer Research scholarship awarded to M. Matens.

References


