How people learn features in the absence of classification error

Lin Chen[a], Lei Mo[b], Lewis Bott[c]

[a]School of Chinese as a Second Language, Sun Yat-sen University, Guangzhou, China

[b] Center for the Study of Applied Psychology, South China Normal University, Guangzhou, China

[c] School of Psychology, Cardiff University, Cardiff, UK

Correspondence should be addressed to:

Lin Chen, Ph. D.

School of Chinese as a Second Language,

Sun Yat-sen University,

Guangzhou, 510275, China

E-mail: chenlin36@sysu.edu.cn

Phone: 86-20-84134751

Abstract

Models of category learning often assume that exemplar features are learned in proportion to how much they reduce classification error. In contrast, experimental evidence suggests that people continue to learn features even when classification is perfect. We present three experiments that test explanations for how people might learn features in the absence of error. In Experiment 1 we varied the type of feedback participants received. In Experiment 2 we introduced a secondary task during the feedback phase, and in Experiment 3, we restricted the response window and varied the feedback. In all cases we found that participants learn many more features than they need to classify the exemplars. Our results suggest that participants learn the internal correlations between features, rather than directly forming associations between features and the category label. This finding places restrictions on the types of categorization models that can satisfactorily explain learning in the absence of classification error.

Key words: category learning; classification error; error-driven; blocking

A central goal of cognitive psychology is to understand how categories are learned and used. Categories allow us to generalize from past experience and make inductions from observed features of objects to their unknown properties. For example, categorizing an unknown entity as a dog means that we can infer that it might chase sticks, that it may like being petted, but that it is also capable of causing you harm, and so on. But for categories to be useful in this way, two information requirements must be satisfied. First, and most obviously, enough features must be known to allow the object to be correctly classified. Successful use of categories requires successful classification. Second, enough information must be represented about the category for induction processes to predict likely features of the new object. Generally speaking, the more features known about the category, the more useful the categorization will be. For example, classifying an object as a dog is very useful if it is known that dogs may bite, that dogs enjoy being petted, and that dogs eat meat, whereas the classification is less useful if all that is known about dogs is that they like being petted. While these two information requirements are generally considered part of the same process (e.g., Rosch & Mervis, 1975; Nosofsky, 1989; Kruschke, 1992), this article investigates a situation in which the goals of classification and feature induction might differ. In particular, we focus on feature learning in the absence of classification error.

Objects have large numbers of features, and it is important to learn which features to pay attention to. A key assumption of many categorization models is that during category acquisition, people learn how to selectively allocate their attention by weighting stimulus dimensions. Categories described by simple rules tend to corroborate the optimal attention hypothesis. A simple rule-based categorization strategy implies all-or-none attention weighting of diagnostic and nondiagonistic dimensions (Matsuka & Corter, 2008). For exemplar models and prototype models, the crucial difference is how to adapt attention weights. One solution to decide which features to pay attention to is to link feature weighting to classification performance so that features that are very useful in classifying objects are highly weighted. For example, if the presence of fur is useful in distinguishing dogs from other animals, fur should be weighted highly when combining classification cues. This approach to feature learning is known as *error-driven learning* because it reduces the error between predicted and actual classification. The intuitive appeal and the empirical success of the error-driven approach have meant that some version of this algorithm appears in many category learning models. For example, the adaptive network model (Gluck & Bower, 1988) extends the Rescorla-Wagner's (1972) least mean squares model of associative learning to category learning; ALCOVE (Kruschke, 1992) incorporates an exemplar-based representation with error-driven learning; the Rule-Plus-Exception model (RULEX; Nosofsky, Palmeri, & McKinley, 1994) employs classification error to explain rule choice; BAYWATCH (Heit & Bott, 2000) and KRES (Rehder & Murphy, 2003) assume that knowledge effects develop as a function of classification error; and some hybrid models, such as COVIS, predict that classification errors cause changes in the model (Ashby, Alfonso-Reese, Turken, & Waldron, 1998).

While this strategy has proved remarkably successful in accounting for data concerned with classification learning, there are many other reasons why people learn features and it is not clear how minimizing classification error can be reconciled with these other goals of learning (e.g., induction, see Anderson, Ross, & Chin-Parker, 2002; Chin-Parker & Ross, 2002; Erickson, Chin-Parker, & Ross, 2005; Lassaline & Murphy, 1996; Markman & Ross, 2003; Ross 2000; Rehder, Colner, & Hoffman, 2009; Yamauchi, Love & Markman, 2002; Yamauchi & Markman, 1998, 2000a, 2000b). As we suggested above, the more features that people know, the better their inductive abilities, irrespective of classification performance. The apparent discrepancy between goals suggests that separate cognitive processes (or separate error terms) might be involved in classification and feature learning.

This question was addressed in a series of studies by Murphy, Hoffman and colleagues (Bott, Murphy, & Hoffman, 2007; Hoffman & Murphy, 2006; Hoffman, Harris, & Murphy, 2008). These studies taught participants categories using a standard, supervised feedback method, but investigated how many individual features participants learned relative to the number they needed for accurate classification. In all cases they found that participants continued learning features beyond the number that was necessary for perfect classification, in contrast to classification-driven models of feature learning. This article builds on one of these studies, Bott et al., (2007) to establish how it is that people learn features in the absence of classification error.

Blocking in category learning

Bott et al. (2007) investigated blocking in category learning. Blocking is an associative learning phenomenon in which prior learning of one cue prevents, or *blocks*, learning of successive cues that predict the same outcome (Dickinson, Shanks, & Evenden, 1984; Kamin, 1969; Kruschke & Blair, 2000; Rescorla & Wagner, 1972; Shanks, 1985). The effect has often been attributed to the absence of prediction error in the later phases of learning. As an example consider Kamin's original study on rats. There were three phases to the experiment. In the first phase, rats were conditioned to associate a light with a shock. In the second phase, one group of rats learned to associate a shock with a compound stimulus involving a light and a tone. In the final phase, the rats were tested on the tone alone to establish how much they had learned about the association between the tone and the shock. Kamin found that rats who had learned the light-shock association learned less about the tone than a control group who had not experienced the first phase of learning. The explanation provided by Kamin was that the learning of one cue blocked learning of the second cue because there was no error to correct.

Bott et al. (2007) argued that if feature learning in categorization was tied to classification error, blocking effects should be as easy to obtain in categorization paradigms as they are in the associative learning paradigms. Specifically, learning of individual features should be blocked when classification error is zero. Bott et al. tested this hypothesis by adapting a standard category learning experiment to incorporate a blocking manipulation.

We consider Experiment 3 of Bott et al. (2007) in detail because it forms the basis of the experiments presented in this article. Participants learned about two categories of vehicles. One defining feature could perfectly classify all exemplars in a category; this feature corresponds to the light in Kamin (1969). Before the learning phase, participants in the pretraining group learned that defining dimensions could perfectly classify all exemplars into two categories, while participants in a control condition learned to press the keys that corresponded to two category labels. In the learning phase, participants in both conditions learned to classify exemplars into categories, after which they were tested on their individual feature learning. According to classification-driven learning accounts, the existence of defining features should block learning of the other features (the non-defining features), however, the blocking effect was difficult to obtain. Bott et al. did not find any evidence that the pretraining group learned fewer non-defining features than the control condition, and the participants in the pretraining group learned many more features than they needed in order to correctly classify the exemplars.

The experiments of Bott et al. demonstrate that when people learn categories, they learn more than the minimum necessary to classify the exemplars. However, it is not clear how the non-defining features were learned. As Bott et al. (2007) discuss, categorization models that assume people learn features in order to minimize classification error are unable to explain their findings. When there is no error to drive learning, no features should be learned. Other forms of learning exist however. People can learn in unsupervised ways, for example, and there are different ways in which “error” can be defined, some of which are consistent with the results of Bott et al. In the section below we describe three possibilities.

Feature learning in the absence of classification error

First, participants may have been learning direct associations between the features and the category label during the feedback stage. In the feedback stage, participants saw the feature list (the exemplar) and the category label presented at the same time. The category label could have been treated as another feature and participants could have formed the auto-associations necessary to link them all together. Although the non-defining features were not perfectly correlated with the category label, they would have been sufficiently high to lead to better-than-chance performance in the test phase.

A second possibility is that participants were not learning direct associations between the features and the category label, but between the features themselves. Good performance on the non-defining features could have been obtained by recalling that a particular non-defining feature often occurred with a particular defining feature, and that the defining feature predicted the category label. Such an account is consistent with unsupervised category learning.

Finally, participants could have been forming explicit hypotheses about the relationship between each feature and the category label. Even though they were able to perfectly classify the exemplar using the defining feature, they could have believed that knowing a mapping between each feature and the exemplar would also have been useful. During the response phase, participants could have first noted the defining feature but then progressed through the feature list to read the other features and form hypotheses about whether they too predicted the category label. While we do not know of a categorization model that has been proposed exactly along these lines, RULEX (Nosofsky et al. 1994) could reasonably be adapted to accommodate such a strategy. RULEX is a rule-based model that cycles through each dimension trying out hypotheses until it can successfully predict the category label. RULEX could successfully predict the results of Bott et al. if the stopping rule could be removed so people cycle through all of the dimensions.

One of the differences between this type of learning and standard error-driven learning (e.g., ALCOVE) is that the error term is different. The standard error-driven model assumes a single, global error term encoding the difference between prediction and correct (feedback) classification, whereas the explicit hypothesis account suggested here assumes separate error terms for each feature so that learning only ceases when there is zero error for each feature.

The three explanations above attribute acquisition of the non-defining features to different learning processes. Importantly for our purposes, they also differ with respect to which stage in the trial learning occurs. Identifying associations between features and category label was likely to have occurred during the feedback stage, since this was the only time in which the category label was presented alongside the exemplar features. Learning within-category correlations between features, however, could have occurred during the response stage or the feedback stage, since in both cases all of the features were visible on the screen. Finally, generating explicit hypotheses about the relationship between each feature and the category label must have occurred during the response stage. Here, participants were not aware of the category label with certainty and could therefore generate hypotheses regarding the relationship between individual features and the category label. In the feedback stage, however, participants saw the exemplar features and the category label at the same time. Hypothesis generation would have been nonsensical because the relationship between a given feature and the exemplar was known; participants would not have hypothesized about a relationship that they could read on the screen (in other words there was no error to drive a hypothesis testing mechanism).

The preceding paragraph describes how different forms of learning are likely to occur at different stages in the learning trial. In the studies described below we use these assumptions to investigate how the non-defining features were learned in Bott et al. (2007). Our approach was to vary the form of the response and the feedback stages in order to establish which parts of the learning trial were necessary for feature learning.

Experiment 1

In Experiments 1 and 2 participants completed a standard category learning task similar to Bott et al. (2007; Experiment 3, categorization condition). Participants learned about two different types of vehicles by classifying exemplars and receiving feedback on their responses. The category structure was such that a single dimension was sufficient to classify all of the exemplars correctly (the defining dimension) but there were many other features that were also diagnostic of the category, although not perfectly so (the non-defining dimensions). There were three phases of the experiment. First, participants completed a pretraining phase in which they learned that the defining feature was perfectly predictive of the category label. Next came the learning phase in which they classified whole exemplars and received feedback. The final phase was an individual feature testing phase where they classified single features. Of primary interest was how participants performed on the non-defining features.

In Experiment 1 we tested whether participants were learning in the feedback stage. Participants saw one of two types of feedback. One group received feedback that involved the presentation of the exemplar and the category label (the *exemplar feedback* condition) and the other group saw feedback that only involved the category label (the *no-exemplar feedback* condition). We hypothesized that if participants in Bott et al. (2007) were learning associations between the non-defining features and the category label, removing the exemplar features from the feedback would lead to reduction in learning of the non-defining features. Conversely, if participants were learning in the response phase, either using internal correlation between features or rule testing, no difference would be observed between the two conditions.

Method

Participants. Forty-eight students from Cardiff University participated for course credit. Participants were randomly assigned to the exemplar feedback condition or the no-exemplar feedback condition.

Stimuli and Design. There were two categories of vehicles, called *A* and *B*. Exemplars in each category were constructed from eight binary dimensions. The eight typical features of Category A are represented by 1s, while the eight typical features of Category B are represented by 0s. The prototype of Category A was therefore 1 1 1

1 1 1 1 1; and the prototype of Category B was 0 0 0 0 0 0 0 0 (see Table 1 for the list of features). Table 2 shows that knowing Dimension 1, the *defining dimension*, was sufficient to perfectly categorize all of the exemplars. However, the other dimensions, the *non-defining dimensions*, were also highly diagnostic of the categories but attending to only one of them was insufficient to perfectly categorize all of the exemplars. The precise defining dimension was rotated across participants.

Insert Table 1 and Table 2 here.

Procedure. The experiment involved a pretraining phase, a learning phase and a single-feature test phase. In the pretraining phase, participants learned to assign category labels to exemplars with a single feature (the defining feature for that category). Participants viewed the feature in the centre of the screen and pressed a button corresponding to the category they believed was appropriate. They then received feedback on their choice. For example, if the air bags dimension was the defining dimension, they would see the sentence, “Has airbags” and respond with the A or B category button. Each exemplar was presented 7 times, so there were 14 trials in the pretraining phase.

In the learning phase, participants learned the 16 exemplars shown schematically in Table 1. One block consisted of a presentation of all the exemplars and exemplars were presented in a random order in each block. Exemplars were shown as a list of 8 features on the screen. The features appeared in a different random order for each exemplar.

Exemplars remained on the screen until the participant classified the exemplar. Feedback was provided immediately after the participant had made their response. In the exemplar feedback condition, participants saw the whole exemplar together with the category label and information about whether they were correct. In the no-exemplar feedback condition, the feedback reported “correct” or “incorrect” and the correct category label. Feedback in both conditions lasted 5 s.

Participants saw repeated blocks of exemplars until they successfully classified all the exemplars of a block or they had experienced 14 blocks. They then proceeded to the testing phase. In each trial of the testing phase, individual exemplar features were presented and participants judged which category the feature was most likely to belong to. Feedback was not provided. All 16 features were tested and each feature was tested twice. There were thus 32 trials in the testing phase.

Results[1]

Participants in both conditions achieved ceiling accuracy on the defining features in the pretraining phase. The accuracy of classifying defining features was 0.91(*SD* = 0.10) and 0.95 (*SD* = 0.03) in the exemplar feedback and no-exemplar feedback conditions respectively, which was not a significant difference, *t* (46) = 1.73, *p*>.05, Cohen’s *d* = 0.54. In the learning phase, participants in both conditions required an average of 2.0 blocks (*SD* exemplar feedback = 1.43 and *SD* no-exemplar feedback = 1.87) to accurately categorize all the exemplars, *t* (46) = 0.087, *p*>.5, Cohen’s *d* =

[1] The raw data for all the experiments shown in this paper is available on the Open Science Framework, (https://osf.io/).

0.03.

We also analyzed the response times (RTs) of categorizing an exemplar into a category during the learning phase. RTs in the two conditions were not significantly different, however (*M* = 3084ms, *SD* = 973ms vs. *M* = 3584ms, *SD* = 2061ms), *t* (46) = 1.08, *p* > .1, Cohen's *d* = 0.31.

Of principal interest however, was performance on the non-defining features in the test phase. If participants needed feedback to learn the non-defining features, performance in the no-exemplar feedback condition should be at chance. This was not the case, however. Participants in the exemplar and the no-exemplar feedback conditions scored significantly above chance on the non-defining features (*M* =.63, *SD*=.18; *M* = .67, *SD*=.18), *t* (23)'s >3.44, *p*'s < .005. Indeed, according to the guessing correction, participants would get half of the features correct simply by chance. Therefore, we measured the number of dimensions learned by subtracting the incorrect proportion from the correct proportion and multiplied by the number of learning dimensions. Participants learned on average 1.8 (*SD*=2.6) and 2.4 (*SD*=2.6) non-defining dimensions in the exemplar feedback and the no-exemplar conditions respectively. There was no significant difference between the two conditions in learning of non-defining features, *t* (46) < 1. A Baysian analysis (see Kruschke, 2013; implemented by Baath, 2012) also demonstrated no reliable difference between the means, N burn-in samples = 20000, N samples = 20000, M_{diff} = -.042, 95% HDI (-0.159, 0.065).

Discussion

Participants who saw the exemplar features during feedback learned more features than they needed. Just as in Bott et al. (2007), participants could have perfectly classified the exemplar by focusing entirely on the defining feature, but instead, they learned an average of 1.8 non-defining features. More originally, participants who saw only the category label during feedback also learned the non-defining features, and to the same extent as those who saw the category label and exemplar presented concurrently. Seeing an exemplar concurrently with its label is clearly not necessary for learning the non-defining features.

In the Introduction we argued that the feature to category associations were most likely to be learned in the feedback stage of Bott et al. (2007), in which whole exemplars and category labels were used. Our demonstration that non-defining features can be learned in the absence of exemplar feedback suggests that participants were learning non-defining features using strategies other than directly associating features to category labels (note that we are not suggesting that participants are unable to learn features in this way; only that they are must have other strategies in addition to feature-category associations).

One argument against this conclusion is that participants in the no-exemplar feedback condition might have been remembering the feature lists and associating the category label with the memory trace of the features. Thus, learning could have occurred by direct association, but the association was with a memory trace and not a visual display. Some evidence is provided against this by our failure to find a difference between learning of the non-defining features across the two conditions, however. If it were the case that participants were remembering the features, the memory trace of the features in the no-exemplar condition would have been weaker than the visual list of the features in the exemplar feedback condition. A weaker representation would have led to lower accuracy in the no-exemplar feedback condition, which we did not find. Nonetheless, our argument against the memory explanation rests on a null difference between the two conditions. In Experiment 2 we present a more direct test of whether participants were actively memorizing the non-defining features in the feedback phase.

Experiment 2

Participants in Experiment 2 completed a task similar to the no-exemplar-feedback condition of Experiment 1. The crucial difference, however, was that during the feedback phase, participants had to count backwards out loud in threes from a number presented randomly on each trial (e.g., “297, 294, 291…”). This task is widely used to disrupt working memory (e.g., Allen, Baddely & Hitch, 2006; Baddely, Hitch & Allen, 2009; Bartz & Selhei, 1970). If learning of the non-defining features is due solely to participants memorizing the exemplar and associating it with the category label during feedback, the working memory task would eliminate learning of the non-defining features.

Method

Participants. Twenty-four students from the University of Pittsburgh were paid for their participation in this experiment.

Stimuli, Design and Procedure. The stimuli were the same with Experiment 1.

According to American English, some words of stimuli were changed, “gears” and “boot” were replaced by “transmission” and “trunk”.

The procedure was similar to the no-exemplar feedback condition of Experiment 1. The only difference was the feedback stage. For 1s, participants saw feedback of “correct” or “incorrect” together with the correct category label. Immediately after the offset of the feedback, a three-digit number was presented for 4s. During the 4s, participants had to count backwards in threes from this number. For example, if the number was 309, participants had to say out loud 306, 303, 300 etc. Number counting was recorded by a separate recorder. The total time of the feedback stage was the same between Experiment 2 and the no-exemplar condition of Experiment 1.

Results

Participants were generally very accurate in the pretraining phase, scoring a mean of 0.93 (*SD*=0.06). In the learning phase, participants required an average of 2.3 (*SD*=2.4) blocks to accurately categorize all the exemplars. One participant who learned more than 15 blocks to achieve perfect classification performance was excluded from further analysis.

Crucially, participants learned significantly more features than necessary to classify the exemplars, *M*=0.63 (*SD*=0.16), and significantly more than chance, *t* (22) =3.77, *p* < .005. Participants learned on average 1.8 (*SD*=2.3) of the 7 non-defining dimensions. Thus, preventing participants from associating the category label with the memory trace of the features in the feedback stage did not eliminate non-defining feature learning. A comparison with the no-exemplar feedback condition of Experiment 1 revealed that the interference task did not even reduce learning of the non-defining features (although the usual caveats of cross-experimental comparisons apply here), M = 0.63 *vs* M =0.67, t(45) < 1, Baysian analysis: M_{diff} = -.043, 95% HDI (-0.152, 0.059).

Discussion

Participants who performed a memory interference task during the feedback phase were still able to learn the non-defining features. This result, in combination with our findings from Experiment 1, suggests that participants in the no-feedback condition were learning the non-defining features in the response stage.

Recall from the introduction that there were two possible forms of response-stage learning. The first was that participants could have been learning internal associations among the features of the exemplar, and the second was that they were forming hypotheses about individual features and the category labels. In the next experiment we test between these possibilities.

Experiment 3

Learning the internal correlations of features could have occurred in the feedback stage or in the response stage of Bott et al. (2007). In both stages, exemplar features were presented on the screen and participants could have noted that certain features often occur with one of the defining features. Rule-based learning on the other hand, could only occur during the response stage. It was only during the response stage that participants would realistically make predictions about the category label because during the feedback stage, the category information was shown to participants

(participants would have nothing to form hypotheses about). One way of testing between these accounts would be to restrict the response stage to a minimum but use a long feedback stage. If participants were using a hypothesis testing strategy, they should learn relatively few non-defining features in this condition. In contrast, if they were learning internal feature correlations they should be able learn features during the feedback stage and consequently, learn many non-defining features.

Participants were assigned to one of three conditions. In the long feedback condition, participants were given a restricted response stage (3s), but a long feedback stage (10s). Feedback included the whole exemplar and the category label. In the short feedback condition, participants were given a long response stage (10s) but only short feedback (3s) that did not include the exemplar. The total time that the exemplar features were on screen was equal across the long and the short feedback conditions (13s). Finally, in the control condition, participants had a short response stage (3s) and a short and restricted feedback stage (3s with only the category label). Table 3 illustrates the design.

Participants in the long feedback condition would only have been able to use an unsupervised, correlation strategy to learn the non-defining features because the restricted response stage would not allow them the opportunity to form hypotheses. Participants in the short feedback condition could learn using either a rule-based or a correlation strategy. While the short feedback condition was not strictly necessary given the results of Experiment 1, we were concerned that we may observe no learning in any condition if we only included the long feedback and the control condition. Finally, the control condition was included because we wanted to verify that a 3s response stage was sufficient to allow correct exemplar classification but to prevent the formation of rules (or any other type of non-defining feature learning).

Insert Table 3 here.

Method

Participants. Seventy-two participants from Cardiff University participated for course credit. Participants were randomly assigned to each condition.

Stimuli, Design and Procedure. The stimuli and general procedure was the same as for the previous experiment. Participants first went through a prelearning phase in which they learned the defining feature. They then progressed onto a training phase and finally a feature test phase. Conditions differed in the length and type of feedback during the training phase, but were identical in the prelearning and feature test phase. In the response stage, participants in the long feedback and control conditions saw exemplars for 3s whereas those in the short feedback condition saw exemplars for 10s. After the exemplar disappeared, one sentence on the screen indicated participants to make a choice of the category. After the response, feedback was provided. In the feedback stage, participants in the long feedback condition saw feedback for 10s, and participants in the short feedback and the control conditions saw feedback for 3s. In the long feedback condition, the feedback consisted of the exemplar, the category label, and the corrective message, whereas in the short feedback and control conditions, the feedback involved only the category label and corrective message.

Results

Participants were generally very accurate in the pretraining phase, *M* = 0.95 (*SD*=0.04), *M* = 0.93 (*SD*=0.08), and *M* = 0.95 (*SD*=0.05) in the short feedback, long feedback and control conditions respectively. Similarly, participants learned the exemplars quickly in the learning phase. After two blocks of learning, participants in all three conditions could categorize the exemplars correctly. Interestingly, we observed no differences in learning rates across conditions. Participants needed 2.1 (*SD*=1.1) blocks to learn in the long feedback condition, and M = 1.5 (*SD*=0.8) and M = 1.6 (*SD*=1.0) in the short feedback and control conditions respectively, *F* (2, 69) =2.15, *p*>0.1, η^2=0.06. Most participants could correctly categorize exemplars in one block.

We observed significant differences in the proportion of features correctly classified in the testing phase, however. In the control condition, participants learned no more features than chance, *M*=0.55 (*SD*=0.15), *t* (23) =1.68, *p* > .1. In the long feedback and short feedback conditions participants learned significantly more than was necessary to classify the exemplars, *M*=0.65 (*SD*=0.19), and *M*=0.71 (*SD*=0.13), and learning was above chance in both conditions, *t*'*s* (23) > 3.8, *p*'*s* < .005. Participants learned on average 0.7 (*SD*=2.1), 2.1(SD=2.6) and 3.0 (*SD*=1.9) non-defining dimensions in control, long feedback and short feedback conditions respectively. An ANOVA revealed significant overall differences between groups, *F* (2, 69) = 6.47, *p*<0.005, η^2=0.16. Both long feedback and short feedback conditions were significantly greater than the control condition: *p's* <.01. However, no significant differences were observed between the short feedback and long feedback conditions, *t*'s < 1. Baysian analysis: *M* = 0.066, 95% HDI (-0.040, 0.166).

Discussion

Participants in the long feedback stage learned significantly more non-defining features than chance and more than the control condition, who learned none at all. Since the minimal response stage was sufficient to prevent learning of the non-defining features in the control condition, participants in the long feedback condition must have been learning the non-defining features during the feedback stage. Based on the logic presented in the Introduction, we conclude that participants were learning the correlations between features and not forming hypotheses about how individual features predict the category label.

Consistent with the results of Experiment 1, participants who received a long response stage but a short feedback stage (the short feedback condition) learned the non-defining features at above chance levels. We also found no difference in non-defining feature learning between the long feedback and short feedback conditions, suggesting that participants in all conditions acquired knowledge of the non-defining features in the same manner.

General Discussion

Categories are more useful when more features are known about the category. More features means that category induction is more effective, and the more features that are known, the more likely classification is to be successful (Anderson, Ross, & Chin-Parker, 2002; Yamauchi & Markman, 1998, 2000). Indeed Murphy and colleagues (Bott, et al. 2007; Hoffman & Murphy, 2006; Hoffman, Harries, & Murphy, 2008) have demonstrated that people are motivated to learn as many features as possible even when they can perfectly classify the exemplars. The paradox however, is that most models of category learning predict that feature learning ceases when classification-error is eliminated. In an attempt to reconcile the classification driven learning of successful categorization models with the empirical observations of Bott et al., we tested three explanations for how people learn features in the absence of classification error. Our results suggest people learn non-defining features by learning within-category feature correlations, rather than forming explicit hypotheses about the link between individual features and category labels or by learning the feature-to-outcome correlations.

Implications for models of category learning

Bott et al. (2007) described in detail how the current forms of many error-driven models of category learning could not explain feature learning in the absence of classification error (e.g., ALCOVE, RULEX). These arguments apply equally to learning of the non-defining features that we observed in our experiments. Our studies build on these findings in several ways, however.

First, our results show that feature learning in the absence of error is not restricted to certain parts of the learning trial. Without our data, one explanation of Bott et al. (2007) could have been that classification learning occurred in one part of the trial, but feature learning occurred in another (e.g., classification learning occurred in the response phase but feature learning in the feedback stage). This would have been consistent with multiple systems theories (e.g., Anderson & Betz, 2001; Ashby & Maddox, 2005; Ashby & O'Brien, 2005; Patalano, Smith, Jonides & Koeppe, 2001; Smith & Grossman, 2008; Smith, Patalano, & Jonides, 1998), which could allocate different modules or systems could to different stages in the learning trial, or even that standard classification-driven models might be applied during one part of the trial, and unsupervised systems could apply in another. In contrast to these suggestions, our experiments suggest that learning of the non-defining features and classification learning all occur during the same stage, and consequently, are most parsimoniously explained by a single learning system (although multiple learning systems operating in parallel are equally consistent with our results).

Second, while Bott et al. presented good arguments why models that were driven by global classification error terms were unable to explain their results, they were more agnostic about models that were driven by other classification error signals, such as feature-to-outcome error. A feature to outcome error signal would continue to drive learning even after classification error was zero, provided there was some discrepancy between each feature's predicted outcome and the actual outcome, as we described in the Introduction. Our experiments eliminate at least one version of such a model. Participants cannot have been forming hypotheses about individual feature to outcome associations during the response stage, because learning of the non-defining features occurred when the response stage was minimal (i.e., too short to allow any learning, as shown by the control condition of Experiment 3). Thus, a rule-based model that assumes feature to outcome error terms (or any kind of classification error term) would not be able to explain our results.

A question that naturally arises, however, is whether any classification driven model can explain our data. While we have argued that a hypothesis testing model with a feature-to-outcome error term cannot account for our findings, other types of classification driven model incorporating similar error terms might be able to (e.g., Kruschke, 2001, based on EXIT, Kruschke & Johansen, 1999; and Jones, Maddox & Love, 2006). For example, Jones et al. present an exemplar-based model that can use either global classification error or feature-to-outcome error. Such a model might account for the results of Experiment 3, long feedback condition, by maintaining that the error term was parameterized to learn feature-to-outcome associations, and that participants were learning those associations in the feedback stage. We cannot completely eliminate this possibility but we feel it is unlikely, for the following reasons. First, in Experiments 1 & 2, no-exemplar feedback conditions, and Experiment 3, short feedback condition, we found that removing the concurrent presentation of the exemplar and the category label did not eliminate learning of the non-defining features. If the only way that participants were learning the non-defining features was in the feedback stage, then removing the feedback should have eliminated this learning. Second, classification driven models assume that people learn categories by making predictions and learning from their errors. It seems implausible that people would make category predictions if the true category label was visible during the prediction process (they would have nothing to predict). Admittedly it is not clear how literally the notions of "prediction" and "error-driven" should be taken in non-hypothesis driven models (unlike hypothesis driven learning in which there is an emphasis on explicit rule-formation) but our paradigm involves language-based stimuli with few features and it seems likely that participants engaged in high-level learning strategies.

The discussion above suggests that pure classification-driven models are unlikely to be able to account for our findings. Models with an unsupervised component, like SUSTAIN (Love, Medin, & Gureckis, 2004), might therefore provide a suitable explanation for our results. SUSTAIN learns by forming clusters of exemplars that are similar and mapping those clusters onto category labels. New clusters can be formed if feedback indicates that an incorrect cluster assignment has been made (supervised learning) or if a new exemplar is sufficiently different from the existing clusters (unsupervised learning). Adjustable weights connect input dimensions to the clusters. As Bott et al (2007) write, SUSTAIN can explain learning in the absence of error because the dimension weights are not optimized to minimize classification error, but to minimize the distance between the cluster centre and the presented exemplar. Since higher weights on the non-defining dimensions would bring exemplars closer to the cluster centres, SUSTAIN would continue to increase weight on the non-defining dimensions even when classification error was zero (see Equations 12 and 13 of Love et al.). The unsupervised component of the model would apply during the response stage or the feedback stage, which would make it consistent with our experimental demonstrations that learning occurs in either stage. Interestingly, while the input dimensions are not adjusted in proportion to classification error, the cluster to category units are (see Equation 14 of Love et al.). This means that, when classification error was zero, the model would not obtain any benefit from having the exemplar presented concurrently with the category label, compared to the exemplar being presented on its own. This is very similar to the results of Experiment 1, in which we showed that learning of non-defining features was approximately equal in the exemplar and the no-exemplar feedback conditions.

One issue with SUSTAIN is that in its current form it would not be able account for the results of the control condition in Experiment 3, in which participants learned no non-defining features but still managed to classify the exemplars. Of course, this results is exactly what a standard classification-driven model would predict (a blocking effect). One solution would be to add feature sampling as a "front-end" to SUSTAIN. With a 3s response time, participants were only able to sample a limited number of features. They therefore read and processed the most diagnostic dimensions first (see Lamberts, 1995, 1998), that is, the defining feature, and based their classification on this information. In its current form, however, SUSTAIN does not make any processing assumptions and so it is difficult to say for sure whether the parameter range exists that will account for our pattern of data. We look forward to possible extensions of SUSTAIN and simulations that might test this hypothesis and the predictions above.

Finally, SPEED (Ashby, Ennis & Spiering, 2007), might also be able to model our findings. SPEED is a neuroscientifically plausible model that is designed to explain how categorization becomes automatic in procedural learning. In SPEED, categorization initially depends on a slow, subcortical pathway, but as more experience is gained, responses become dominated by a separate, cortical-cortical pathway. For our purposes, the important point is that cortical-cortical pathway involves an (unsupervised) Hebbian learning component, and could thus acquire knowledge of the non-defining features in the absence of an error signal. As expertise begins to be acquired therefore, knowledge of the non-defining features might also be acquired. Of course, as with SUSTAIN, it is difficult to know how this model would be able to account for the blocking effect we observed in Experiment 3.

Conclusion

People learn features in order to help them classify exemplars. Nonetheless, people also learn features when they can perfectly classify the exemplars, as shown by Bott et al (2007) and the present results. Our contribution is to show how people learn the features. Rather than explicitly learning more and more features to better predict classification directly, participants seem to learn the correlations between features, which lets them classify exemplars if they need to.

Acknowledgments

This work was supported by the Ministry of Education in China under Grant (11YJC190005). The authors wish to thank Greg Murphy and Wolf Vanpaemel for comments on an earlier version of the paper.

References

Allen, R. J., Baddeley, A. D., & Hitch, G. J. (2006). Is the binding of visual features in working memory resource-demanding*? Journal of Experimental Psychology. General*, 135, 298–313.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review, 8,* 629–647.

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition, 30,* 119–128.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review, 114,* 632-656.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *The psychology of learning and motivation, 2,* 89-195.

Ashby, F.G., Maddox, W.T. (2005) Human category learning. *Annual Review of Psychology, 56,* 149-78.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences, 9,* 83–89.

Baath, R. (2012). Bayesian Estimation Supersedes the t-test (BEST) – online. http://sumsar.net/best_online/

Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language*, 61, 438–456

Bartz, W.H, & Salehi, M. (1970). Interference in short- and long-term memory. *Journal of Experimental Psychology,* 84(2), 380-382.

Bott, L., Hoffman, A. B., & Murphy, G. L. (2007). Blocking in category learning. *Journal of Experimental Psychology: General, 136,* 685–699.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition*, 30, 353-362.

Dickinson, A., Shanks, D. R., & Evenden, J. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A,* 29-50.

Erickson, J. E., Chin-Parker, S., & Ross, B. H. (2005). Inference and classification learning of abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 86–99.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *Psychology of learning and motivation* (pp. 163–199). San Diego: Academic Press.

Hoffman, A. B., Harris, H. D., & Murphy, G. L. (2008). Prior knowledge enhances the category dimensionality effect. *Memory & Cognition, 36,* 256-270.

Hoffman, A. B., & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 301–315.

Jones, M., Maddox, W. T., & Love, B. C. (2006). The role of similarity in generalization. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society,* 405-410.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In Church, R. M. & Campbell, B. A. (Eds.), *Punishment and aversive behavior* (pp. 279-296). New York: Appleton-Century Crofts.

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

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45,* 812-863.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573-603.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review, 7,* 636– 645.

Kruschke, J. K., & Johansen, M. K. (1999). A Model of Probabilistic Category Learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25,* 1083-1119.

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, *124*, 161-180.

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 695-711.

Lassaline, M. E., & Murphy, G. L. (1996). Induction and category coherence. *Psychonomic Bulletin & Review, 3,* 95–99.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004) SUSTAIN: A network model of category learning. *Psychological Review, 11,* 309-332.

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

Matsuka, T., & Corter, J. E. (2008). Observed attention allocation processes in category learning. *The Quarterly Journal of Experimental Psychology, 61(7),* 1067-1097.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics, 45,* 279-290.

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

Patalano, A. L., Smith, E. E., Jonides, J., & Koeppe, R. E. (2001). PET evidence for multiple strategies of categorization. *Cognitive, Affective, and Behavioral Neuroscience, 1,* 360-370.

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of experimental psychology, 58(3),* 193-198.

Rehder, B., Colner, R. M., & Hoffman, A. B. (2009). Feature inference learning and eyetracking. *Journal of Memory & Language, 60,* 394-419.

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning, *Psychonomic Bulletin & Review, 10,* 759-784.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current research and theory* (pp. 64–99). New York: Appleton- Century-Crofts.

Rosch, E. & Mervis, C. B. (1975), Family Resemblances: Studies in the Internal Structure of Categories, *Cognitive Psychology, 7,* 573–605.

Ross, B. H. (2000). The effects of category use on learned categories. *Memory & Cognition, 28,* 51-63.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology, 37B,* 1–21.

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

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative mechanisms of categorization. *Cognition, 65,* 167-196.

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

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 3,* 585–593.

Yamauchi, T., & Markman, A. B. (2000a). Inference using categories. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 26, 776–795.

Yamauchi, T., & Markman, A. B. (2000b). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory & Cognition, 28,* 64–78.