Journal of Experimental Psychology:
Learning, Memory, and Cognition
2005, Vol. 31, No. 6, 1433–1458

0278-7393/05/$12.00 DOI: 10.1037/0278-7393.31.6.1433

Category Representation for Classification and Feature Inference

Mark K. Johansen and John K. Kruschke
Indiana University Bloomington

This research's purpose was to contrast the representations resulting from learning of the same categories by either classifying instances or inferring instance features. Prior inference learning research, particularly T. Yamauchi and A. B. Markman (1998), has suggested that feature inference learning fosters prototype representation, whereas classification learning encourages exemplar representation. Experiment 1 supported this hypothesis. Averaged and individual participant data from transfer after inference training were better fit by a prototype than by an exemplar model. However, Experiment 2, with contrasting inference learning conditions, indicated that the prototype model was mimicking a set of label-based bidirectional rules, as determined by the inference learning task demands in Experiment 1. Only the set of rules model accounted for all the inference learning conditions in these experiments.

Keywords: classification and/versus inference learning, categorization, category learning, category representation, predictive inference

Category representation as a result of learning is a fundamental prerequisite of category use and, as such, is an integral part of a basic cognitive capacity. As an area of research, category representation has been strongly associated with formal mathematical modeling, which has allowed precise comparisons of various hypothesized representations. Although a lot of recent categorization research has focused on the need for models with more elaborate, mixed representations (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Love, Medin, & Gureckis, 2004; Nosofsky, Palmeri, & McKinley, 1994), simple exemplar, prototype, and rule-based representations continue to be considered relevant, particularly in the context of research emphasizing different types of category learning and use (Markman & Ross, 2003; Yamauchi & Markman, 1998).

Exemplar models have been widely shown to provide a better account of perceptual category learning than prototype or simple rule models (e.g., Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Nosofsky, 1992). Evidence for simple rules continues to be found (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; Johansen & Palmeri, 2002; Nosofsky, Palmeri, & McKinley, 1994), but usually in conjunction with some other type of representation, such as exemplars or prototypes. Typically, a perceptual category learning task has consisted of training participants to categorize multifeatured stimuli with corrective feedback. Exemplar representation assumes that training and testing stimuli are categorized on the basis of their summed similarity to the known instances of the categories. Prototype representation assumes that stimuli are categorized on the basis of their similarity to the statistical average or modal central tendency of the instances from each category—that is, to the category prototype. Although the concept of rule is infinitely flexible, simple rule representation usually assumes that instances are categorized on the basis of a simple criterion. For example, "If the stimulus is red, then Category A, and if it is blue, Category B." Whether a participant used exemplar, prototype, or rule representation can be (it is hoped) inferred from their generalizations to new stimuli.

Having participants learn perceptual categories by predicting missing features when given the stimulus category and remaining features (Yamauchi & Markman, 1998) is a recent variant of the typical task. Category representation can then be probed using both feature inference and standard classification trials.

The purpose of the present research was to compare and contrast the category representations resulting from feature inference training and standard classification training of the same categories as evaluated with exemplar, prototype, and rule models. In particular, the research purpose was to test the hypothesis that prototype representation is more consistent with the results of feature inference training than is exemplar representation.

This research was motivated by the work of Yamauchi and Markman (1998), who argued that the type of category representation participants form depends on how they learn the categories:

> Although inference and classification are closely related, the two functions require different strategies to be incorporated. The present experiments suggest that these different strategies, which are related

Mark K. Johansen and John K. Kruschke, Department of Psychology, Indiana University Bloomington.

This research was supported by National Institute of Mental Health (NIMH) Grant R01 MH48494 to Indiana University Bloomington and Mark K. Johansen, National Science Foundation Grant BCS-9910720 to Indiana University Bloomington and John K. Kruschke, NIMH Training Grant T32 MH19879 to Indiana University Bloomington, and an Indiana University Cognitive Science Summer Research Fellowship. We thank Robert Goldstone, Tatyana Johansen, David Leake, Robert Nosofsky, Justin Savage, Richard Shiffrin, and Takashi Yamauchi for comments on a previous version of the manuscript. We also thank Nathaniel Blair and Andrew Cohen for helpful discussion and Kotapolis for good company. Finally, we thank Twanna Allen, Colin Campbell, Dan Hall, Kara Kohnen, Indya Watts, and Maxine Weiss for their assistance in conducting this research.

Correspondence concerning this article should be addressed to Mark K. Johansen, who is now at the School of Psychology, Cardiff University, Tower Building Park Place, Cardiff CF10 3AT, Wales, United Kingdom. E-mail: johansenm@cardiff.ac.uk

> to the two functions of categories, give rise to the formation of distinct category representations. (p. 144)

Because we used a similar methodology, we describe Yamauchi and Markman's (1998) paradigm in detail.

Using the simple category structure shown in Table 1, which has four instances in each of two categories, Yamauchi and Markman (1998) trained participants in one of three ways: standard classification, feature inference, or mixed training (which was a combination of the first two). The stimuli were simple geometric figures composed of features from four binary-valued dimensions, with the binary values coded as 1s and 2s, as shown in Table 1. The modal prototype for each category was composed of the most frequent feature value on each dimension for the category instances (see the bottom of Table 1), and all category members deviated from the modal prototype on one and only one feature value, called the *exception* feature. Participants were trained on a series of blocks, each containing a trial for every distinct training item in random order. The prototypes were not presented in the training phase. During a single training trial, participants viewed an instance, chose one of two possible responses, and were then given the correct answer.

The types of training differed by the nature of the question asked on each trial. In the classification training condition, a stimulus composed of four features was presented, and participants had to decide if the stimulus was a member of Category A or B—for example, ? 1 1 1 2, where the *?* indicates the response dimension. In this training condition, there were eight stimuli, four from each category (see Table 1). In the inference training condition, participants were shown the category label for the instance, given the values of the instance on three feature dimensions, and asked to predict the value of the fourth, missing feature—for example, a member of Category A and a large circle on the right: Should its color be green or red? That is, A 1 1 ? 2. Each member of a category was trained in this way on all of its features but the exception feature and the category label. If the category label is assumed to be just another feature, then both training conditions required a response based on four stimulus features. So as to make feature inference training and classification training as symmetrical as possible, the exception features were not trained. (The label dimension cannot sensibly have an exception feature.) Hence, though the underlying category structure was the same as for classification training, participants were trained on 24 distinct trials, 3 for each instance of each category.

Following training, all participants completed two transfer tasks. First, participants viewed each of the eight classification training stimuli and the two prototype stimuli and classified them without feedback. Inference-trained participants were also tested on the classification transfer stimuli. Second, participants were tested on all possible feature inferences, including instances in which the correct answer (see Table 1) was an exception feature, even though none of the participants had been trained to respond with the exception features.

Two findings from this experiment are relevant here: First, on the inference transfer trials for which participants were asked to infer an exception feature, participants from all training conditions tended to incorrectly choose the feature consistent with the category prototype rather than the correct but prototype-incompatible feature. For example, for A 1 1 1 ? (see Table 1), they tended to respond 1 rather than *2*. Second, the degree to which participants incorrectly chose the prototype-consistent feature on exception feature trials varied significantly across the training conditions. In accordance with Yamauchi and Markman's (1998) hypothesis that feature inference training would produce a different type of category representation than classification training, participants were more likely to give prototype-compatible, incorrect responses to the exception feature testing trials following inference training than they were following classification training (see the left side of Table 2).

Yamauchi and Markman (1998) interpreted these results as follows: "We argue that inference promotes the acquisition of category representations characterized with the prototypes of the categories, while classification facilitates the formation of categories consistent with rules and exceptions or concrete exemplars" (p. 129). So their strongly implied hypothesis was and our working hypothesis has been that feature inference training induces categories to be represented by their prototypes, whereas classification training induces exemplar representation.

To further support this hypothesis, Yamauchi and Markman (1998) fit a standard exemplar model, Nosofsky's (1986) generalized context model (GCM), separately to six different sets of data:

Table 1
Category Structure for Yamauchi and Markman's (1998) Experiment 1

Label	Form	Size	Color	Position	Label	Form	Size	Color	Position
				Category members					
A1	1	1	1	***2***	B1	2	2	2	***1***
A2	1	1	***2***	1	B2	2	2	***1***	2
A3	1	***2***	1	1	B3	2	***1***	2	2
A4	***2***	1	1	1	B4	***1***	2	2	2
				Prototypes					
A0	1	1	1	1	B0	2	2	2	2

Note. Nonprototypical (i.e., exception) features are shown in bold italic type. From "Category Learning by Inference and Classification," by T. Yamauchi and A. B. Markman, 1998. *Journal of Memory and Language, 39,* Table 1, p. 128. Copyright 1998 by Elsevier. Adapted with permission.

Table 2
Average Accuracy Memory Testing Results for the 5–4 Category Structure Data From Experiment 1 Compared With the Data From Yamauchi and Markman's (1998) Experiment 1

	Yamauchi & Markman (1998)		5–4 structure	
Result type	Classification training	Inference training	Classification training	Inference training
Participants meeting 90% learning criterion	23/24	22/24	7/59	32/58
Accuracy				
Last training block	—	—	1.00	.98
Classification transfer	.91	.76	.92	.76
Inference transfer				
Prototype-compatible features	.81	.94	.79	.96
Prototype-incompatible (exception) features	.46	.14	.22	.02
Inference				
Ambiguous features/prototype-compatible responses			.43	.97

Note. The structure in Yamauchi and Markman's (1998) Experiment 1 is presented in Table 1; the 5–4 structure of the present Experiment 1 is presented in Table 4. All accuracy data are proportions correct. The data in the last row are proportions of prototype-compatible responding for inference accuracy testing trials without a clear correct answer (see text for details). Dashes indicate that data were not reported.

classification or inference transfer after classification, inference, or mixed training. To allow the GCM to make predictions for inference transfer trials, Yamauchi and Markman (1998) generalized it slightly by treating the category labels as additional features. By adjusting the GCM's parameters to minimize the discrepancy between the model's predictions and the data, they concluded that the GCM provided accurate predictions for the classification transfer data, regardless of the type of training, but that the GCM was not able to account for the data from feature inference transfer for either type of training. This result is not inconsistent with the hypothesis that the two different types of training result in different types of category representation, but it also suggests that the type of transfer trial influenced which representation participants appeared to be using.

Using a psychologically plausible generalization of the original GCM (Maddox & Ashby, 1993; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994) that has now become a standard part of the model, Kruschke, Johansen, and Blair (1999) showed that this standard exemplar model (detailed below) can reasonably account for all of Yamauchi and Markman's (1998) six conditions, as shown in Table 3. The added parameter indexes response confidence and specifies the degree to which a given difference in evidence for one response over another should translate into a difference in response probabilities.

Nevertheless, despite the success of the exemplar model in accounting for Yamauchi and Markman's (1998) data, the hypothesis that classification and feature inference training result in different category representations is still plausible if their simple category structure (see Table 1) does not strongly differentiate exemplar and prototype representation. The fits shown in Table 3 of a comparably parameterized prototype model (detailed below) suggest that this may be the case. Although the prototype model fit the inference transfer data slightly worse than did the exemplar model, particularly for the data from after classification training, the models resulted in almost identical fits to the classification transfer data for all kinds of training. This is particularly telling given that both of these models were developed to account for classification rather than inference learning.

In addition, there are several other reasons why the slight superiority of the exemplar over the prototype model for the inference transfer data may be misleading. The exemplar model has been shown to garner somewhat more flexibility from its free parameters to account for arbitrary data sets than does the prototype model and, hence, sometimes to be marginally more difficult to falsify (Myung, 1997). Further, the exemplar model was fit to data averaged across participants, and some research has indicated that averaged data may tend to favor the exemplar model even when individual participant data do not (Ashby, Maddox, & Lee, 1994; Maddox, 1999). Finally, there is a strong intuitive appeal to the idea that the exemplar model should have trouble accounting for participants' prototype-compatible response errors on exception features: The nearest and most similar exemplar does not predict these errors.

More generally, it is intuitively compelling that a great deal is learned about real-world categories in the context of feature inference. Further, the central tendency of the category seems like a best plausible guess for a missing feature.

Table 3
Exemplar and Prototype Model Fits (Root-Mean-Square Deviations) to the Averaged Data From Each Condition in Yamauchi and Markman's (1998) Experiment 1

Condition	Exemplar	Prototype
Classification transfer		
Classification training	0.055	0.057
Mixed training	0.062	0.067
Inference training	0.041	0.041
Inference transfer		
Classification training	0.084	0.122
Mixed training	0.063	0.109
Inference training	0.040	0.056

In summary, a variety of reasons suggested to us that Yamauchi and Markman's (1998) data were somewhat inconclusive for differentiating the representations resulting from classification versus feature inference training, but the representation-difference hypothesis remained compelling. This suggested that a potential difference in representation resulting from classification and inference training should be reevaluated in the context of a more diagnostic category structure. Experiments 1 and 2 were based on a more diagnostic structure, and the data from these experiments were evaluated with the following models of category representation.

Specification of the Exemplar, Prototype, and Rule Models

Of the three simple representations—exemplar, prototype, and rule—exemplar has arguably been the most successful, particularly in the form of Nosofsky's (1986) GCM. Descended from the context model (Medin & Schaffer, 1978), the GCM, together with its connectionist version, ALCOVE (attention learning covering map; Kruschke, 1992), has been central to the success of exemplar representation. It continues to be relevant (e.g., Nosofsky & Johansen, 2000) despite the recent explosion of more complex, mixed-representation models.

The multiplicative similarity version of the prototype model described below (Estes, 1986; Nosofsky, 1987, 1992) was used so that the similarity calculations and response mechanisms would be identical to those of the exemplar model. Only the representations differed.

The rules in the set of rules model below are simple, single-condition rules relating a single binary-valued stimulus dimension with a binary-valued response dimension. A set of rules, as opposed to a single rule, is needed because the feature inference task required responses on different feature dimensions. This contrasts with the classification learning task, which only required responses on the category-label dimension. Because of this asymmetry, the set of rules model is only applied to data from feature inference learning, not classification learning.

For the sake of simplicity, the set of rules model does not include all possible simple rules but, rather, only the most valid rule for each response dimension. For reasons that should become progressively clear as the results of the experiments are considered, each simple rule is based on the category-label dimension, because the category label is a perfectly valid predictor for the feature inference learning tasks. These simple rules were formalized in the same modeling framework as the exemplar and prototype models. Intuitively, the set of rules model may seem to be very similar to the prototype model, which is sometimes the case. However, the differences between these models become particularly apparent in the context of Experiment 2.

It is worth pointing out that there are a variety of recent categorization models that incorporate rules in addition to other forms of representation. These models include ATRIUM (attention to rules and instances in a unified model; Erickson & Kruschke, 1998), COVIS (competition between verbal and implicit systems; Ashby et al., 1998), and PRAS (parallel rule activation and rule synthesis; Vandierendonck, 1995). Most notably, Nosofsky, Palmeri, and McKinley (1994) applied their RULEX (rule-plus-exception) model, which assumes representation composed of rules and rule exceptions, to the results of learning the main category structure used here (see Table 4) in the classical way (i.e., by classification). RULEX did quite well on these data, but Nosofsky and Johansen (2000), as well as Johansen and Palmeri (2002), presented further evidence that the exemplar model still provides a better account of the results for learning this structure by classification, though both articles acknowledged clear evidence for rules in some participants. It is not our purpose to argue that these models are fundamentally wrong or could not be fruitfully applied to our data or to inference learning in general. However, these models include mechanisms for hypothetical principles that are not required or addressed by the data presented here while requiring, at the same time, the addition of new mechanisms to account for these data.

All three of these models use the same similarity evaluation and decision processes and have comparable free parameterization. Because the GCM has been the most widely used of these models, the specification below of first the prototype and then the rule model is in reference to the GCM, called simply the *exemplar model* from now on.

Applying the Exemplar and Prototype Models to Categorization Transfer Data

Equation 1 is the exemplar model's formula for the similarity of instance i to exemplar j, where k indicates the feature dimensions

Table 4
Abstract Category Structure

	Features on four dimensions			
Category	1	2	3	4
A1	1	1	1	***2***
A2	1	***2***	1	***2***
A3	1	***2***	1	1
A4	1	1	***2***	1
A5	***2***	1	1	1
B1	***1***	***1***	2	2
B2	2	***1***	***1***	2
B3	2	2	2	***1***
B4	2	2	2	2
	Prototypes			
A	1	1	1	1
B[a]	2	2	2	2
	Classification testing items			
T1	1	2	2	1
T2	1	2	2	2
T3	1	1	1	1
T4	2	2	1	2
T5	2	1	2	1
T6	2	2	1	1
T7	2	1	2	2

Note. Nonprototypical (i.e., exception) features are shown in bold italic type. The enumeration of the category instances (i.e., A1, A2, etc.) and the generalization trials (i.e., T1, T2, etc.) are for descriptive convenience only; participants were not shown these numbers.
[a] Note that this prototype is also training instance B4.

composing the stimuli. The feature values of the transfer instance i and the training instance j on dimension k are x_{ik} and x_{jk}, respectively. Each dimension has a free parameter, attention weight (w_k), and c is an overall scaling parameter:

$$\eta_{ij} = \exp(-c\sum_{k} w_k|x_{ik} - x_{jk}|). \tag{1}$$

Intuitively, this equation specifies that similarity is high when features match on dimensions and low when features mismatch, with the importance of a mismatch weighted differently by the dimensional attention-weight parameters.

For two categories, A and B, Equation 2 determines the response probability for Category A given stimulus i. This probability is the summed similarity of i to all of the instances of category A (the numerator) divided by the sum of the summed similarities to both response categories (the denominator) and moderated by a response-confidence or response-determinism parameter, γ. The higher this scaling parameter, the more deterministic the model's responding—that is, the closer the predicted response probabilities are to 0 and 1 for a given set of similarities;

$$p(catA|i) = \frac{(\sum_{j\in catA} \eta_{ij})^{\gamma}}{(\sum_{j\in catA} \eta_{ij})^{\gamma} + (\sum_{j\in catB} \eta_{ij})^{\gamma}}. \tag{2}$$

The prototype model is the same as the exemplar model, with the exception that the exemplars are replaced by the category prototypes in the above equations. So in Equation 1, η_{ij} is the similarity of transfer instance i to the prototype of category j, and Equation 2 reduces to a single similarity for each category—η_{i1} for similarity to the prototype of Category A and η_{i2} for Category B:

$$p(\text{CatA}|i) = \frac{(\eta_{i1})^{\gamma}}{(\eta_{i1})^{\gamma} + (\eta_{i2})^{\gamma}}. \tag{3}$$

The prototype model uses A 1 1 1 1 as the prototype for the first category and B 2 2 2 2 as the prototype for the second category in the category structure used in Experiments 1 and 2 (see Table 4). These are the prototypes assumed for these categories by previous researchers (e.g., Smith & Minda, 2000), and they correspond to the modal category prototypes except for the second feature dimension of Category B. This slight distinction between the specified prototypes—A 1 1 1 1 and B 2 2 2 2—and the modal prototypes—A 1 1 1 1 and B 2 x 2 2, where x means that the modal feature is ambiguous—is relevant to some of the modeling results reported later.

Applying the Exemplar and Prototype Models to Feature Inference

The extension of the GCM to feature inference treats and codes the binary-valued category-label dimension like another feature dimension with its own attention-weight parameter in the similarity calculations (Yamauchi & Markman, 1998). The dimension on which the feature inference is being made (*respdm* in Equation 4) is excluded from the similarity calculation. As before, this equation can be used to calculate similarity to either exemplars or prototypes:

$$\eta_{ij} = \exp(-c\sum_{k\neq respdm} w_k|x_{ik} - x_{jk}|). \tag{4}$$

For a feature inference on a particular dimension, the response-probability calculation is directly analogous to Equation 2 for the exemplar model and to Equation 3 for the prototype model. For the exemplar model, the probability of a Feature 1 response on the response dimension for a particular stimulus is the sum of the similarities to all exemplars in both categories with a 1 feature on the response dimension, $j \in f = 1$, divided by the sum of the similarities to exemplars with a 1 feature and a 2 feature on that dimension, respectively[1] (Kruschke et al., 1999):

$$p(f = 1|i) = \frac{(\sum_{j\in f=1} \eta_{ij})^{\gamma}}{(\sum_{j\in f=1} \eta_{ij})^{\gamma} + (\sum_{j\in f=2} \eta_{ij})^{\gamma}}. \tag{5}$$

The prototype model's response function replaces the summed similarity to exemplars with similarity to the category prototypes. The probability of Feature 1 on the response dimension for given instance i is the similarity of that instance to the category prototype that also has a value of 1 on that feature dimension, η_{i1}, divided by the sum of the similarities to both prototypes:

$$p(f = 1|i) = \frac{(\eta_{i1})^{\gamma}}{(\eta_{i1})^{\gamma} + (\eta_{i2})^{\gamma}}. \tag{6}$$

The Set of Rules Model

The reasons for specifying the set of rules model in the following way should become progressively clearer in the context of the results from Experiments 1 and 2. However, this model can be intuitively motivated by first considering a single, simple rule as a representation for a classification learning task before considering a set of such rules for an inference learning task.

For a classification learning task with binary-valued stimulus feature and response dimensions, this kind of simple rule could be based on a single feature dimension—for example, "If Feature 1 is present, then respond Category A, or if Feature 2 is present, then Category B." Whether a simple rule like this is a sufficient representation for a category learning task depends on the category structure to be learned. For example, the structure in Table 4 cannot be accurately learned by classification with such a simple rule because none of the feature dimensions is perfectly correlated with the category-label dimension.

Moving on to a feature inference task, simple dimensional rules are a sufficient representation when learning the structure in Table 4 by feature inference using the procedure specified by Yamauchi and Markman (1998), as described above. For a given feature response dimension, a simple rule based on the category-label dimension is sufficient for accurate responses on that dimension—

[1] Yamauchi and Markman (1998) specified $P(f = 1|i)$ as the summed similarity to the instances of Category A rather than to the instances in both Categories with a 1 feature on the response dimension divided by the summed similarities to the instances of both categories. We have not used this method because it is not analogous to how the GCM is applied to standard classification.

for example, "If the category label is A, then respond with Feature 1, or if the category label is B, then Feature 2." Because there are multiple response dimensions, a different rule is needed for each. These multiple response dimensions also indicate why only the label dimension is a really useful basis for the formation of these simple rules: When a particular feature dimension is the response dimension, the feature on that dimension is absent by definition and cannot be a sufficient representation to respond about itself.

The set of rules model is formalized in the same framework as the exemplar and prototype models, uses essentially the same similarity and response equations, and is applied to classification and inference transfer items in the same way. The set of rules model differs from the exemplar and prototype models only in terms of its representation. Specifically, the set of rules model replaces the exemplars in the exemplar model with simple rules (see, e.g., Table 5) that are determined by the category structure and inference learning task. That is, the rules are basically exemplars with missing features. The similarity of a transfer instance i to rule j is calculated in the same way as for the exemplar and prototype models (see Equations 1 and 4):

$$\eta_{ij} = \exp(-c \sum_{k \neq respdm} f_k w_k |x_{ik} - x_{jk}|), \qquad (7)$$

where features that are absent from a given rule, f_k, are just dropped from the similarity calculation, $f_k = 0$, and assumed to have no impact.

This way of formalizing the set of rules model allows it to be applied to both feature inference and classification. So, for example, the similarity of the inference transfer instance B 2 2 2 ? (see Table 4) to the rule component A _ _ _ 1 (see Table 5) is just based on the category labels—$\exp(-cw_{\text{label}}|x_{\text{stimulus,label}} - x_{\text{rule,label}}|) = \exp(-cw_{\text{label}} * 1)$—because the fourth feature dimension is the response dimension, and all of the other feature dimensions are absent for A _ _ _ 1. Likewise, the similarity of B 2 2 2 ? to B _ _ _ 2 (see Table 5) is $\exp(-cw_{\text{label}} * 0) = 1$. Because these are the only two rule components that apply when the response dimension is the fourth feature dimension (see Table 5), these are the only two similarities that contribute to the calculation of the response probabilities. The response probabilities are specified by Equation 5, but here, the contributing similarities are to the rule components rather than to exemplars. Correspondingly, consider the classification transfer item T7—? 2 1 2 2—from Table 4 in relation to the rules in Table 5. Its similarity to each rule subcomponent can be calculated with Equation 7, as for feature inference. But unlike feature inference, in which only two rule subcomponents contribute to any response, here all of the rule subcomponents corresponding to the various rules contribute, because the response dimension is the category-label dimension: The similarity of ? 2 1 2 2 to A 1 _ _ _ in Table 5 is $\exp(-cw_1|2 - 1|)$, because only the first feature dimension is present in this rule component; the similarity of ? 2 1 2 2 to A _ 1 _ _ is $\exp(-cw_2|1 - 1|)$, and so forth. All of these similarities then contribute to the response probability in Equation 2.

Table 5
Set of Feature Associations Corresponding to Simple Dimensional Rules Hypothesized to Have Been Formed in the Inference Learning Condition of Experiment 1

	Features on four dimensions			
Category	1	2	3	4
A	1	—	—	—
A	—	1	—	—
A	—	—	1	—
A	—	—	—	1
B	2	—	—	—
B	—	2	—	—
B	—	—	2	—
B	—	—	—	2

Note. Dashes indicate missing features.

To account for feature inference transfer trials on which the label is absent (Experiment 2 only), the set of rules model needs an additional assumption. One way to account for feature inference when the label is *explicitly* absent is to use the features that are present to *implicitly* invoke the category label via the same set of rules and then proceed as for an inference trial on which the label is present. Assuming a two-stage process for the model's handling of the label-absent results, a plausible way to generate the implicit label activation is to treat these trials as exactly the same as classification trials but with three features instead of four (because the feature on the response dimension is by definition not present). First, the set of rules model classifies the instance composed of three features and, as such, generates the probability of a Category A response, which is of course between 0 and 1. This probability is then scaled into a feature value between 1 and 2 by 1 + [1 – p(Label A)], making this implicit label activation more or less similar to rules based on those label values (Label A = 1, Label B = 2). Hence, a high p(Label A) corresponds to a low value on the label dimension, making the instance more similar to rules with a label value of 1 (i.e., Label A), and a low p(Label A) corresponds to high value on the label dimension and makes the instance more similar to rules with a label value of 2 (i.e., Label B). Second, the implicitly invoked value on the label dimension can then be used to predict the missing feature via similarity to the same set of rules used to account for label-present inference and classification data. For example, the label-absent instance _ 2 2 1 ? might result in the classification probability of the A label as $p(\text{Category A}) = 0.2$, which is scaled to $1 + (1 - 0.2) = 1.8$, so the instance would be represented by 1.8 2 2 1 ? and would, consequently, be more similar to rules with label values of 2 (B) than rules with label values of 1 (A).

Modeling Procedure

For fitting the classification transfer experimental results, the exemplar, prototype, and set of rules models all had four attention parameters, w_k in Equations 1, 4, and 7, one for each stimulus dimension (see Tables 1 and/or 4). For fitting the inference transfer results, all three models had an additional attention parameter for the category-label dimension. Also, for all three models, the similarity-scaling parameter—c in Equations 1, 4, and 7—can be absorbed into the w_k dimensional attention parameters by the distributive rule, because all of the attention parameters are free. The response-confidence parameter, γ in Equations 2 and 5, was a free parameter for the exemplar and set of rules models, but as shown by Nosofsky and Zaki (2002), the response-determinism parameter is underconstrained for the prototype model because it can be absorbed into the similarity-scaling parameter. The models were fit to the data using a hill-climbing routine to minimize the discrepancy between the data and the predictions of the model as measured by root-mean-square deviation (RMSD). Local maxima were not likely a problem, because different starting places in the parameter space resulted in similar best fits.

Experiment 1: Classification Versus Feature Inference Learning in the 5–4 Structure

The *5–4* category structure (Smith & Minda's, 2000, terminology) shown in Table 4 was designed by Medin and Schaffer (1978) to contrast exemplar and prototype representation. This category structure, possibly more than any other, has contributed to the dominance of the exemplar model over the prototype (Smith & Minda, 2000), particularly because of the work of Medin, Nosofsky, and colleagues (Medin, Altom, & Murphy, 1984; Medin, Dewey, & Murphy, 1983; Medin & Smith, 1981; Nosofsky, 1992; Nosofsky, Kruschke, & McKinley, 1992; Palmeri & Nosofsky, 1995). This structure has been the focus of an ongoing debate between Smith and Minda (2000) for prototypes and Nosofsky (2000) for exemplars, and it figured in a recent evaluation of shifts in representation over the course of learning (Johansen & Palmeri, 2002). Although the 5–4 structure is not without its critics (Blair & Homa, 2003; Smith & Minda, 2000), its continued prominence was further motivation for its use here.

The purpose of this experiment was to have participants learn a category structure by classification or feature inference training using a structure that has previously been shown to differentiate exemplar and prototype representation. Training was followed by testing without feedback on a transfer block containing both classification and inference generalization and training trials. The purpose of the transfer block was to provide a rich enough data set to differentiate the category representation models formalized above so as to show that inference training on this category structure is better fit by prototype than by exemplar representation.

Method

Participants

One hundred and seventeen volunteers participated for partial credit in an introductory psychology course at Indiana University Bloomington.

Stimuli

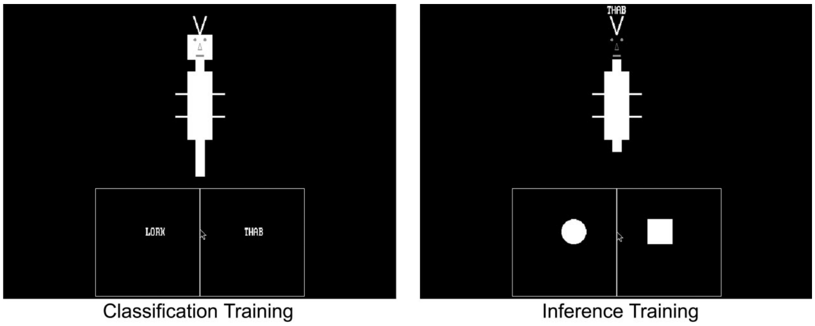
The stimuli were diagrams of fictitious alien insects (see Figure 1). The insects varied on four binary-valued feature dimensions that were chosen randomly from the following five dimensions (with the fifth dimension fixed for all stimuli for a given participant): head shape (round or square), nose direction (pointing up or down), tail length (long or short), antenna shape (straight or curved), and number of legs (four or eight). The stimuli were assigned to one of two categories labeled *THAB* or *LORK*. The abstract category structure used in this experiment (Medin & Schaffer, 1978, Experiment 3) is shown in Table 4 and has five training instances in Category A and four in Category B. The assignment of physical to abstract feature dimensions was randomized for each participant, as was the assignment of physical to abstract features within each dimension. The stimuli were presented on a computer screen, and participants responded by clicking the mouse in the box containing their chosen answer, as shown in Figure 1.

Procedure

Participants were assigned alternatingly to either classification or feature inference training on the basis of their order of arrival. All participants read identical instructions.

In the classification training condition, each trial had a stimulus with features from the four stimulus dimensions (e.g., Figure 1, left panel). The participant assigned it to one of two categories and then received corrective feedback. If participants made no response within 20 s, the message *FASTER* was briefly displayed, along with corrective feedback. Corrective feedback was a tone and the message *WRONG!* for an incorrect response and *CORRECT!* without a tone for a correct response. In addition, the correct response was displayed as part of the stimulus, and the participant had up to 30 s to study it. Each training block contained one trial for each of the nine instances in Table 4 in random order.

In the inference training condition (e.g., Figure 1, right panel), a stimulus with three out of four features and the category label was presented, the participant predicted the missing feature, and then received corrective feedback. It is important to note that during feature inference training, participants were *not* trained to infer every feature for all of the category instances. Rather, as in the procedure used by Yamauchi and Markman (1998), participants were only trained to infer prototype-compatible features, not the prototype-incompatible exception features (in bold italics in Table 4) so as to be symmetric with the classification training task (see the introduction). Note that the modal prototype for Category A is 1 1 1 1, and the prototype for Category B is traditionally assumed to be the opposite, 2 2 2 2 (see the above model specifications). So, for the first member of Category A—A 1 1 1 2 in Table 4—participants were trained to infer the first feature by trials of the form A ? 1 1 2, the second feature by trials of the form A 1 ? 1 2, and so forth. But because the fourth feature of A 1 1



Classification Training

Inference Training

Figure 1. Types of training trials.

1 *2* does not match the Category A prototype (A 1 1 1 1), there were no trials of the form A 1 1 1 ? during inference training. Altogether, there were 25 prototype-compatible features in the instances of the two categories (see Table 4), so each inference training block consisted of 25 randomly ordered feature inference trials—one for each of these. Participants in both training conditions received 225 trials of training.

Both training conditions were followed by a transfer block composed of the same set of memory testing and generalization trials in random order. No corrective feedback was given during the transfer block. Each response was followed by this message: *After you have studied this case (up to 30 seconds), click here to see the next one.*

In more detail, all participants, including those in the inference training condition, were asked to classify all of the classification training instances (see Table 4) and to classify new stimuli corresponding to the generalization instances at the bottom of Table 4. In addition, all participants in both conditions were given a feature inference trial for every possible feature inference on the instances in Table 4, including the prototype-incompatible exception features. For example, for the first instance of Category A, an exception feature inference trial was of the form A 1 1 1 ?. Finally, participants were given trials corresponding to a subset of every possible feature inference generalization trial. Classification transfer and inference transfer trials were randomly intermixed in the transfer block, which consisted of 80 trials—16 classification transfer trials and 64 feature inference transfer trials (11 of which were inadvertently redundant, so there were 53 unique inference transfer trials).

Terminology

The term *memory testing trial* is used to refer to any transfer block trial that either required the participant to classify one of the category instances (A1–B4 in Table 4) or perform a feature inference on one of those instances. Note that the set of memory testing trials for Experiment 1 includes the training items from both the classification and the inference conditions as well as exception-feature inference trials, even though no one was required to infer the exception features in either condition's training. The term *generalization trial* is used to refer to all the remaining transfer trials, both classification and feature inference, that were not part of the training for either condition and were not exception-feature inferences for the category members. For example, instances T1–T7 (see the bottom of Table 4) are classification generalization trials.

Results and Discussion

Assessing Performance in the Learning Tasks

The category structure used in this experiment (see Table 4) was harder to learn than the one used in Yamauchi and Markman's (1998) Experiment 1 (see Table 1), presumably because the structure was more complicated. Use of the same learning criterion as Yamauchi and Markman, 90% accuracy in the last training block, resulted in 7 out of 59 participants for classification training, compared with 23 out of 24 for Yamauchi and Markman's (1998) Experiment 1, as shown in Table 2. Likewise, 32 out of 58 participants achieved this criterion for inference training, compared with 22 out of 24 for Yamauchi and Markman's (1998) Experiment 1.

The learning criterion of 90% accuracy in the last block of training is quite strict, particularly for classification training that has only nine trials per block, because a single wrong answer in the last block is sufficient to fail the criterion (8/9 = .88 < .90). In addition, the 5–4 category structure in Table 4 has been historically difficult for participants to learn. For example, Medin and Schaffer (1978) used the criterion of 1 errorless block, with up to 32 blocks to achieve this criterion, and eliminated 18 out of 32 participants. Also, Johansen and Palmeri (2002) used the criterion 75% correct in the last 4 blocks of training and still eliminated more than one third of their participants (68 out of 198), even after training for 32 blocks. We used this 75% criterion in the last block of training except when directly comparing our results with Yamauchi and Markman's (1998), and it left 26 out of 59 participants from classification training. Because the primary focus of this research is feature inference learning, and the classification results for this structure have been widely replicated, we are not overly concerned about eliminating so many participants. This criterion left 41 out of 58 participants from the feature inference condition. It is important to note that the same qualitative trends occurred even when the poorer learners were included in the analyses.

Averaged Testing Data

The average accuracy results for the memory testing trials were similar to the results from Yamauchi and Markman (1998), as can be seen in Table 2. Most important, participants were significantly more likely to respond incorrectly when inferring prototype-incompatible features after inference training than they were after classification training (.02 vs. .22 correct), $t(37) = 9.94$, $p < .0001$. As converging evidence, a statistical resampling procedure produced similar statistical results.

In contrast to Yamauchi and Markman's (1998) experiment, six of the memory testing trials for this experiment were ambiguous in terms of correct answer, because some of the exception-feature testing trials were equivalent to some of the nonexception-feature testing trials. For example, in inference training, participants were trained to infer the second feature of Item A1 in Table 4—denoted by A 1 (?1) 1 2—and to infer the prototype-compatible features of Item A2—denoted by A (?1) 2 1 2 and A 1 2 (?1) 2. In the subsequent transfer phase, participants were tested on all features of Items A1 and A2. In the context of A1, participants were tested with A 1 (?) 1 2, for which the "correct" answer is 1, but in the context of A2, this case can be construed as a test for the second feature, so the "correct" answer is *2*. Because the correct inference for this case is ambiguous, participants' responses were coded as prototype-compatible or prototype-incompatible.

The results for these ambiguous trials (see Table 2) strongly indicate that participants were significantly more likely to infer prototype-compatible features after inference training than after classification training (.97 vs. .43 of prototype-compatible responding), $t(37) = 12.37$, $p < .0001$, and a statistical resampling procedure produced a similar statistical conclusion. This result further supports the hypothesis that feature inference training encourages prototype-compatible responding.

The classification transfer results for classification and inference training were also quite different from each other, as shown in Figure 2, which plots the Category A response proportions for classification transfer from after classification and inference training against each other. The largest differences were for the cases with an equal number of features from each of the category prototypes, as specified in Table 4—that is, A2 1 2 1 2, B1 1 1 2 2, B2 2 1 1 2, T1 1 2 2 1, T5 2 1 2 1, and T6 2 2 1 1. If participants were using prototype representation, the response proportions for these cases might be expected to near .50 because these have an equal number of features from both category prototypes and,

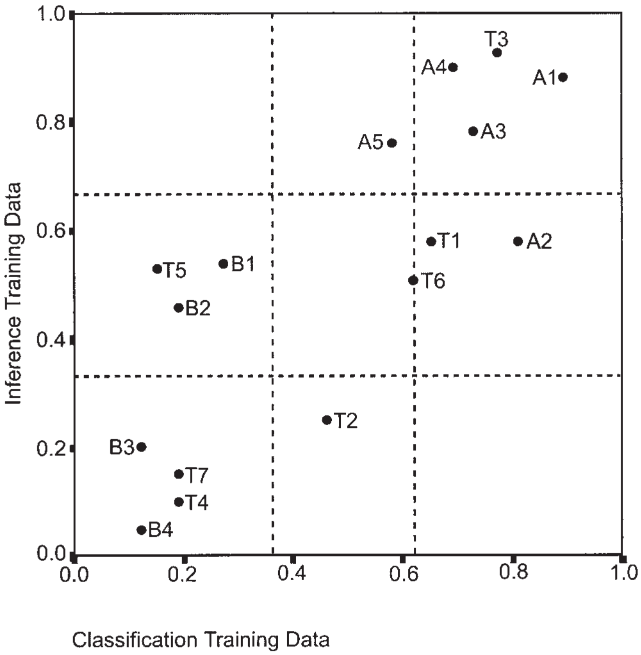


Figure 2. Classification transfer data from classification training versus classification transfer data from inference training. Each data point is the proportion of Category A responses. The dashed lines represent the approximate 90% confidence interval for the .50 population response proportion (based on $n = 35$ for simplicity). The individual data points are identified using the instance labels from Table 4.

hence, might be considered somewhat ambiguous. As can be seen in Figure 2, where the dashed lines indicate a rough 90% confidence interval around .50, all of these cases were near .50 following inference training, but they were all much more extreme—closer to 0 or 1—following classification training. These results are also consistent with the hypothesis that inference training induces prototype representation.

Clustering Analysis of the Inference Generalization Data

Although the regularities in the inference memory testing data are well summarized in Table 2, the regularities in the inference generalization data were harder to see. Figure 3 shows the results of an exploratory hierarchical clustering analysis for the inference generalization data by training condition. Each participant's responding to the 20 inference generalization trials was represented in the analysis by a 20-dimensional vector. Individual participants are on the x-axis and the distance between participants on the y-axis indicates how dissimilar their response patterns were, as indicated by the lowest horizontal line connecting them.

Inspection of the generalization results from inference transfer following classification training (see Figure 3, top panel) reveals no strong clustering and indicates idiosyncratic responding. Participants did relatively poorly on the inference transfer memory testing items following classification training (see Table 2), suggesting that they were guessing on these generalization trials. Apparently, classification training did not support consistent feature inference.

The bottom panel of Figure 3 shows the inference generalization results from after inference training. Inspection reveals a large subset of participants who responded identically, as indicated by the horizontal line going from Participant 1 to Participant 86 at the bottom left of the clustering results. Altogether, 18 participants responded identically to the 20 feature inference generalization trials, and the remaining participants were apparently noisy deviations from this core pattern. There are no correspondingly clear clusters in the results for inference generalization after classification training (see Figure 3, top panel), further supporting the hypothesis of a difference in representation for classification and inference training.

Closer inspection of the data for the 18 participants who responded identically indicated that these participants were probably responding to the inference generalization trials from after inference training by using a set of rules based on the category labels (see Table 5) and completely ignoring all of the other features. That is, they responded with the prototypical feature for the missing feature as determined by the label, hence the origin of the set of rules model. So, for example, if the inference generalization trial was A ? 2 2 1, these participants ignored the other features and responded to the missing feature on the basis of its value in the Category A prototype 1 1 1 1 (see Table 4). So the rule for this response dimension would be as follows: "If Label A, respond with Feature 1, and if Label B, Feature 2." The set of rules, then, had a rule for each response dimension specifying the prototype-compatible feature, as determined by the category label. This pattern of responding was inserted into the clustering analysis as an additional participant vector with the tag *label* in the clustering diagram. It matches the responding for the 18 core participants exactly. Note that this set of rules is a sufficient learning strategy for inference training only because participants were never asked to infer nonprototypical/exception features. Finally, to put the deviation from this pattern of responding in perspective, Participant 11, who was the farthest from the label in the clustering, produced only 8 responses out of 20 inconsistent with this set of rules.

To summarize the results, the differences in the transfer data for participants in the two training conditions are consistent with the hypothesis that classification and feature inference training produce different types of category representation and, in particular, that feature inference training results in prototype representation. However, these results are only suggestive, and mathematical modeling was used to formally evaluate these claims about category representation with the exemplar, prototype, and set of rules models specified above. Modeling is particularly important because the clustering results suggest a set of rules, and also the exemplar model has been shown to sometimes account for prototype effects—that is, for high transfer accuracy on untrained category prototypes (e.g., Nosofsky & Kruschke, 1992).

Mathematical Modeling of the Results From Experiment 1

The following subsections report several different kinds of modeling analyses that provide converging evidence for the same general conclusions. The exemplars, prototypes, and sets of rules used by the corresponding models are listed in Table 4, Table 4, and Table 5, respectively, and several representation variants are also discussed.

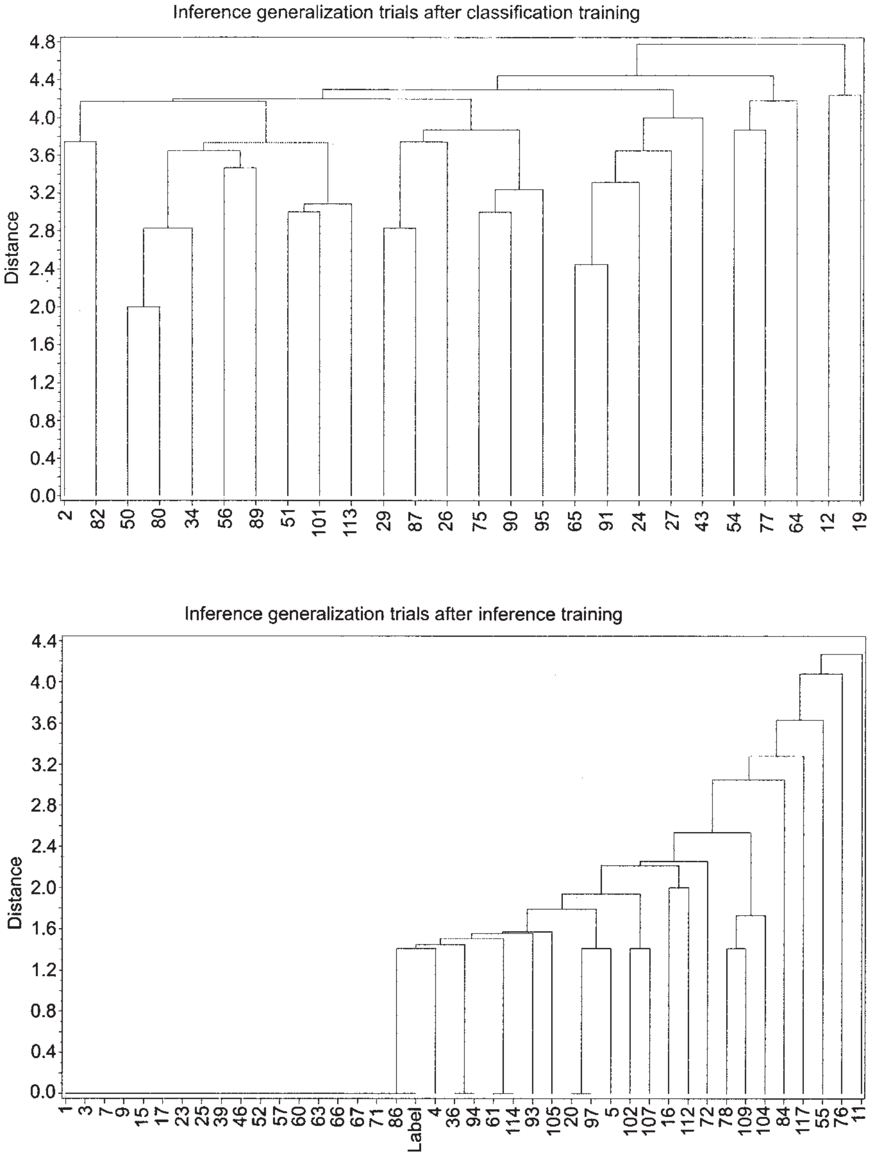


Figure 3. Experiment 1 hierarchical clustering results for inference generalization transfer trials after classification and inference training. Distance is Euclidean (see text for details).

Modeling Results for Averaged Data

Table 6 shows results for the models fit separately to four subsets of group data averaged across participants: classification transfer after (a) classification or (b) inference training and inference transfer after (c) classification or (d) inference training. The numbers in parentheses are the average values of the fits to individual participant data (discussed later).

The exemplar model produced marginally better fits than the prototype model to both transfer data subsets from after classification training: 0.086 RMSD versus 0.096 RMSD for classification transfer and 0.108 RMSD versus 0.116 RMSD for inference transfer. However, the prototype model produced clearly better fits than the exemplar model to both classification and inference transfer data after inference training: 0.039 RMSD versus 0.135 RMSD for classification transfer and 0.030 RMSD versus 0.140 RMSD for inference transfer. Finally, the fits of the set of rules and prototype models were essentially identical for the averaged data.

Clearly, the model fits summarized in Table 6 support the hypothesis that classification and inference training result in different types of category representation and, in particular, that inference training may induce prototype representation. In more detail, the superiority of the prototype over the exemplar model for the data from classification transfer after inference training can be seen in the middle-top panels of Figure 4, in which the predictions of the models are plotted against the data. The corresponding plots for the models' predictions for inference transfer data are shown in the bottom panels of Figure 4. The set of rules model is not shown because it is extremely similar to the prototype model for these data.

Table 6
Model Fit Results for Averaged Data (and Individual Participant Data) as Measured by Root-Mean-Square Deviation for Each Condition in Experiment 1

Type of training	Model Exemplar	 Prototype	 Set of rules
	Classification transfer		
Classification (n = 26)	0.086 (0.259)	0.096 (0.294)	
Inference (n = 41)	0.135 (0.287)	0.039 (0.229)	0.039 (0.229)
	Inference transfer		
Classification (n = 26)	0.108 (0.401)	0.116 (0.413)	
Inference (n = 41)	0.140 (0.238)	0.030 (0.175)	0.030 (0.184)

Note. Numbers in parentheses are averages of the fits to the individual participant data. The set of rules model fits for classification training are not reported for reasons discussed in the main text.

The attention parameters for the fit of the prototype model to the data from classification transfer after inference training for Dimensions 1–4 were 0.912, 0.732, 0.680, and 0.776 (the exemplar model's parameters were 3.032, 0.868, 1.844, and 35.260), respectively. The prototype model's parameters approximate an equal allocation of attention, which is conceptually consistent with a set of rules model and explains why the fits of these two models are virtually identical.

The category-label dimension had the largest attention parameter by an order of magnitude for the best fits of all the models to data from inference transfer, regardless of training condition. This indicates that participants based their responding to inference transfer trials almost exclusively on the category-label dimension, as is discussed more extensively below.

Further Constraining the Model Fits to the Averaged Data

In contrast to the modeling results from Yamauchi and Markman (1998), here the type of training rather than the type of transfer seems to have determined whether the exemplar or prototype model (and set of rules model) provides a better account of the data (see Table 6). Consequently, it seemed reasonable to fit the models to both classification and inference transfer data simultaneously. Figure 5 plots the models' predictions against the data by training condition. Again, the exemplar model provided slightly better predictions for the transfer data following classification training (0.126 RMSD vs. 0.154 RMSD, respectively), but the prototype model provided clearly better predictions for the transfer data following inference training (0.060 RMSD vs. 0.171 RMSD, respectively), despite the fact that the prototype model has one less free parameter. Finally, the set of rules model fit was actually somewhat better than the prototype model fit (0.033 RMSD vs. 0.060 RMSD, respectively), indicating that these two models are not completely equivalent when constrained to account for all of the data.

The modeling results for the averaged data support the hypothesis that classification and feature inference training produce different types of category representation and, in particular, that inference training is more consistent with prototype than with exemplar representation, at least for this category structure and learning procedure. Nevertheless, data averaging can be misleading, particularly when different participants using fundamentally different representations are averaged together. Hence, there follows modeling analysis of individual participant data.

Modeling Results for Individual Participant Data

The values in parentheses in Table 6 are the results of fitting exemplar, prototype, and set of rules models separately to each individual participant's classification and inference transfer data and then taking the average of the resulting RMSDs across all participants. The procedure and numbers of free parameters were the same as for the fits to the averaged data.

All of the averages of the fits to individual participant data in parentheses in Table 6 are higher than the corresponding fits to the group averaged data, probably because individual participant data are far noisier. Each participant responded to each transfer trial type only once (except for a few redundant trials), so the response "proportions" in the individual participant data are all 1 or 0. Hence, even a single "random" response nontrivially worsened a model's fit.

Despite the worse quantitative fits, the same qualitative trends occurred in the fits to individual participant data as in the averaged data (see Table 6), and the fits to independent participants allow for methodologically sound statistical conclusions. The exemplar model fit data from classification and inference transfer after classification training marginally better than the prototype model, $t(25) = -2.039$, $p < .052$, or $z = -1.185$ (n = 17 nonties), $p = .236$, with a Wilcoxon signed-ranks test for classification transfer, and $t(25) = -1.597$, $p = .123$, or Wilcoxon $z = -2.207$ (n = 25 nonties), $p = .027$, for inference transfer. More important, the prototype model fit data from classification and inference transfer after inference training better than the exemplar model, $t(40) = 2.851$, $p = .008$, or Wilcoxon $z = -2.430$ (n = 33 nonties), $p = .015$, for classification transfer, and $t(40) = 7.476$, $p < .0001$, or Wilcoxon $z = -5.106$ (n = 41 nonties), $p = .000$, for inference transfer. The prototype and set of rules model fits were only microscopically different.

Analyzing the Contribution of the Category Labels to the Model Fits

The central role of the category labels was clearly suggested by the clustering results for inference transfer after inference training (see Figure 3, bottom panel). Further, the prototype and set of rules models performed very similarly. Together, these results suggested evaluating the importance of the category label for both the exemplar and prototype model fits by fixing all of the other dimensional attention parameters to 0. Allowing only attention to the category-label dimension in this way resulted in only a slight decline in the average of the individual participant fits to inference transfer after inference training by both the prototype (0.184 RMSD from 0.175 RMSD) and exemplar models (0.258 RMSD from 0.238 RMSD). This again indicates that inference training was mediated almost exclusively by a set of label-based rules that facilitated accurate responding.

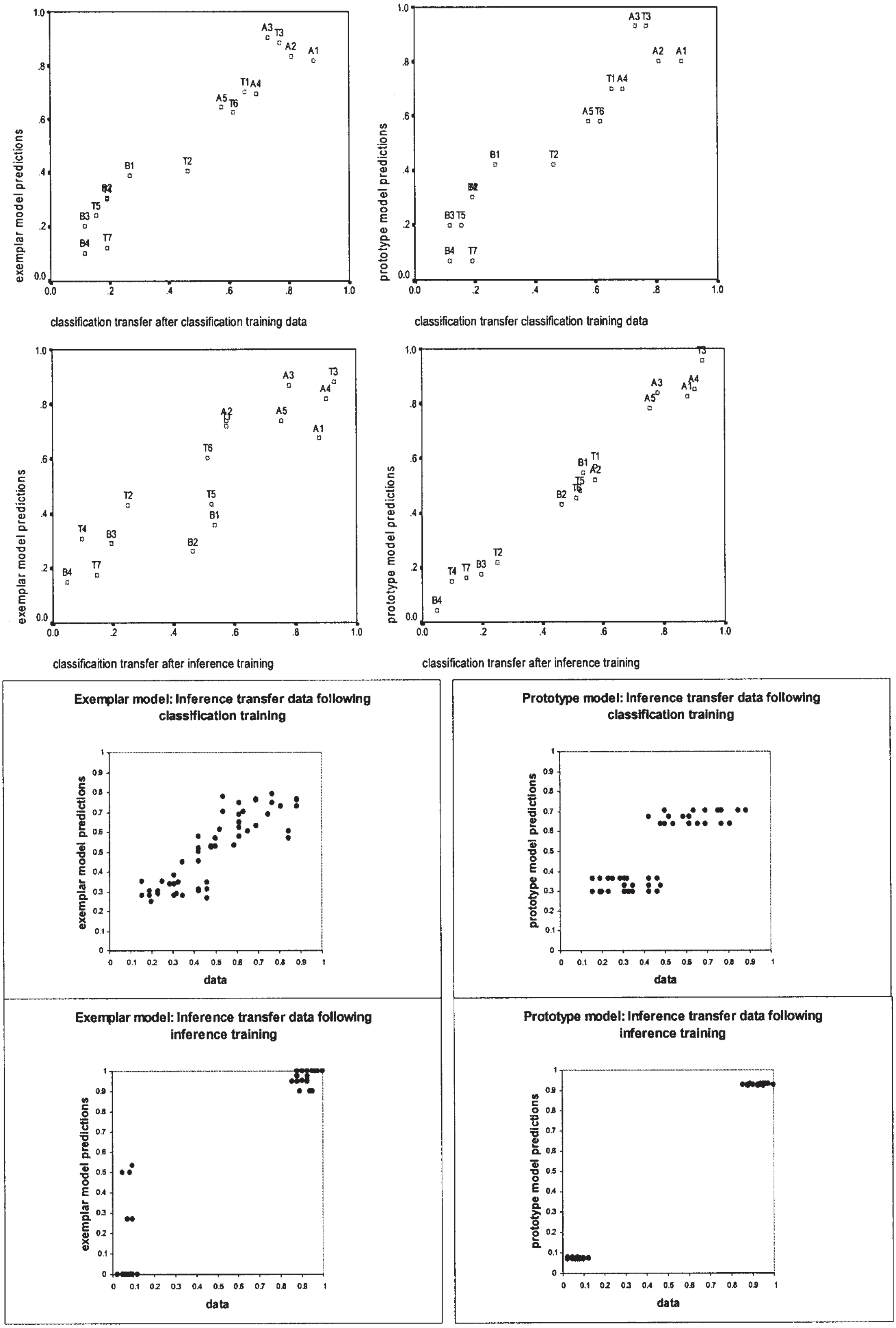


Figure 4. Predications of the exemplar model and the prototype model plotted against classification or inference data by training condition. Each data point represents the probability of a response consistent with the Category A prototype on each transfer trial, and the classification trials are labeled by the trial types from Table 4.

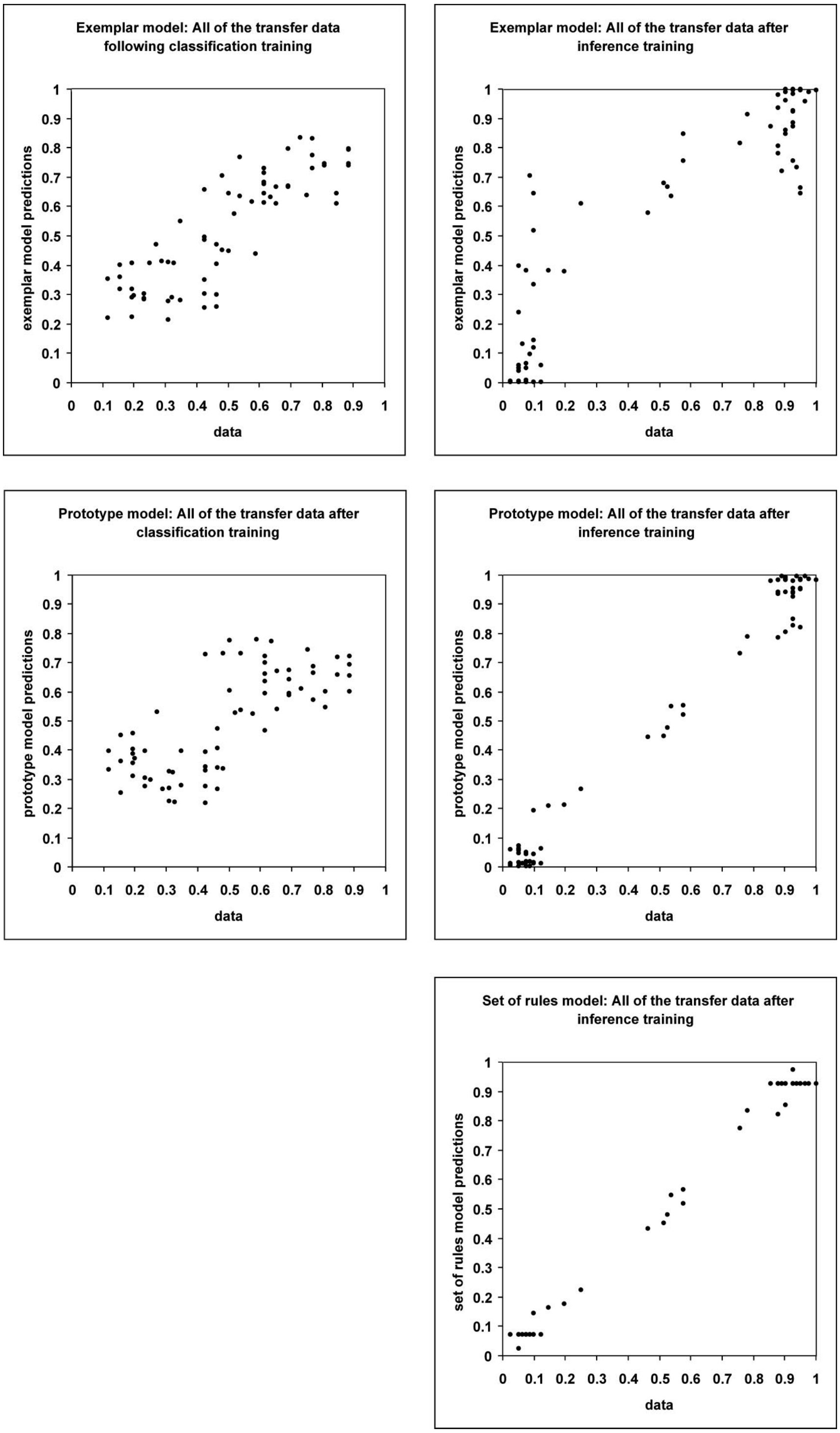


Figure 5. Predictions of the exemplar, prototype, and set of rules models plotted against all of the averaged transfer data by training condition. Each data point represents the proportion of responses consistent with the Category A prototype. The set of rules model predictions for the data from classification training are not reported for reasons discussed in the main text.

An Alternative Prototype Model Provides Evidence for the Set of Rules Model

The above result then raises this question: Why do the exemplar and prototype models not fit the inference transfer data equally well (see Figure 4, lower panels) given that both are capable of attending exclusively to the category-label dimension? Closer inspection of the fits reveals that the problem for the exemplar model is that it has trouble making certain feature inferences on Feature Dimension 2 of the category structure in Table 4. Specifically, there are an equal number of 1 and 2 features in the instances of Category B. Informally, on an inference transfer trial, when the label indicates Category B, the exemplar model has equal evidence for both features, with half indicating 1 and half indicating 2. However, during inference training, participants were only trained on the nonexception/prototype-compatible features (2s), based on the prototypes as traditionally defined for this category structure. During inference transfer after inference training, participants strongly responded with the prototype-compatible features for inference memory testing trials on the second dimension for members of Category B, even when the correct answer was an exception feature—B 1 (1) 2 2, where the *1* in parentheses indicates the correct answer—participants' proportion correct was only .085, and for B 2 (1) 1 2, their proportion correct was .098. This is why the prototype model does relatively well on these results compared with the exemplar model. When the prototype model is reformulated to reflect this ambiguity in its definition (see Appendix A), it does as poorly as the exemplar model and, in particular, worse than the set of rules model, suggesting that feature inference training, in general, may not always induce prototype representation. Rather, the standard prototype model does well on this data set precisely because it is mimicking a set of rules model.

An Alternative Exemplar Model

Exemplar representation in the model fits described above was based on the assumption that during training, the responses become part of the stored exemplars for purposes of representing the category. That is, the exemplars used by the exemplar model were the nine exemplars listed in Table 4. These exemplars are certainly plausible for data from classification training. Further, Yamauchi and Markman (1998) used similar exemplars for the inference training condition, as we did above, but there is an alternative set of plausible exemplars for the inference training condition that does not assume that the responses become part of the stored exemplars. Consider the first instance from Category A in Table 4: A 1 1 1 2. This single category member corresponds to three types of trials during inference training: A ? 1 1 2, A 1 ? 1 2, and A 1 1 ? 2. Because there were 25 distinct inference trials, an alternative conceptualization of exemplar representation can use these 25 exemplars rather than the nine exemplars in Table 4 and, hence, not include the response dimension in the representation. However, this alternative exemplar model did even worse than the version of the exemplar model described above for classification after inference training (0.156 RMSD vs. 0.135 RMSD for the averaged data and 0.351 RMSD vs. 0.287 RMSD for the individual participant data), so it was not considered further.

The Prototype Model and the Set of Rules Model

The results of Experiment 1 are interesting because a prototype model fit a perceptual category learning data set better than an exemplar model. More important, although the results of this experiment are consistent with prototype representation providing a better account of feature inference learning than exemplar representation, they do not support the conclusion that feature inference training *in general* induces prototype representation. Rather, the results suggest that the prototype model is mimicking a set of rules for these data, leading to this working hypothesis: Inference learning results in a set of label-based rules rather than prototype representation.

Experiment 2: Distinguishing Prototype Representation From a Set of Rules

Although prototype representation can be conceived of as integration across a set of simple rules, there are many sets of simple rules that are incompatible with prototype representation. In this experiment, our purpose was to show that feature inference tasks in general do not induce prototype representation by having participants learn a feature inference task that was likely to induce rule representation that was incompatible with both prototype and exemplar representation.

For simplicity, the category structure used in this experiment was the same as that used in Experiment 1 (see Table 4). In this experiment, however, instead of learning the category structure by either classification or feature inference, all participants learned the structure by inference, but they did so in two different inference conditions. In the first condition, participants learned the category structure by inferring the prototype-compatible/nonexception features, as in the inference learning condition in Experiment 1. In the second, more important condition, participants learned the category structure by inferring the prototype-incompatible/exception features (in bold italics in Table 4).

If participants represent inference learning tasks by forming label-based rules, as hypothesized above, then the feature-label rules for these two conditions should be mirror opposites (see Tables 5 and 7, respectively). Correspondingly, the transfer data sets for these two training conditions should be very dissimilar

Table 7
Set of Feature Associations Corresponding to Simple Dimensional Rules Hypothesized to Have Been Formed in the Exception Learning Condition of Experiment 2

	Features on four dimensions			
Category	1	2	3	4
A	2	—	—	—
A	—	2	—	—
A	—	—	2	—
A	—	—	—	2
B	1	—	—	—
B	—	1	—	—
B	—	—	1	—
B	—	—	—	1

Note. Dashes indicate missing features.

and, in some senses, also mirror opposites. In addition, if feature inference training encourages category representation in terms of category prototypes, then participants should find the exception-feature inference condition extremely difficult (if not impossible) to learn compared with the nonexception-feature inference condition. The exception condition cannot be mediated by the modal category prototypes, by definition, but the nonexception condition can. However, if the representation participants form in the exception condition is still a set of label-based rules, similar numbers of participants should reach the learning criterion in both conditions.

Method

Participants

Ninety-seven volunteers participated for partial credit in an introductory psychology course at Indiana University Bloomington.

Stimuli

The stimuli were identical to the alien insects used in Experiment 1.

Procedure

Participants were assigned alternatingly to either exception inference training or nonexception inference training on the basis of their arrival order. In the nonexception condition, participants learned about the abstract category structure in Table 4 by feature inference on the nonexception/prototype-compatible features, as in the inference condition in Experiment 1. Participants in the exception condition learned about the category structure by feature inference on the exception/prototype-incompatible features (in bold italics in Table 4). Participants in the nonexception condition received 225 trials of training in 9 randomly ordered blocks, and participants in the exception condition received 220 trials in 20 blocks.

The testing phase was changed from Experiment 1 in two ways: (a) Thirty-two unique feature inference trials were added on which the category label was neither present nor queried for. On these trials, participants were given three features and asked to predict the missing feature even though no category label was present. (b) The 11 feature inference trials that were inadvertently redundant in the testing phase for Experiment 1 were eliminated and replaced by the 11 unique cases that complete the set of possible inference trials with the label present.

Results and Discussion

Assessing Performance in the Learning Tasks

The learning criterion was 90% accuracy in the last training block and was reached by 29 out of 50 participants in the nonexception condition (similar to the 32 out of 58 participants for the inference condition in Experiment 1) and 26 out of 47 participants in the exception condition. Note also that the average accuracy in the last block of training was essentially identical in both conditions (see Table 8), as would be expected for a set of rules representation.

Averaged Testing Data

The average accuracy results for the memory testing trials are shown in Table 8. In qualitative overview, there are two main results: (a) The averaged testing block data for the nonexception training condition from Experiment 2 are qualitatively the same as those for the inference training condition from Experiment 1, as expected (see Table 2). (b) More important, the averaged testing block data for the exception and nonexception inference training conditions from Experiment 2 are qualitative mirror opposites of each other above and below .50, where .50 corresponds to guessing for two possible responses. The participants' accuracy on the exception features following training in the nonexception condition was only .13, but it was .88 after training in the exception condition. Correspondingly, accuracy on the prototype-compatible features was low following exception training (.28) but high following nonexception training (.88). Finally, the ambiguous feature inference trials (defined in Experiment 1) also showed a large difference between the two conditions: The proportion of prototype-compatible responding was high following nonexception training (.91) and low following exception training (.14). Overall, this pattern of results is highly consistent with the hypothesis that participants learned these tasks by forming a set of label-feature rules, because the hypothesized rules for the exception and nonexception conditions (see Tables 7 and 5) are mirror opposites.

Table 8
Average Accuracy Memory Testing Results by Training Condition for Experiment 2

	Type of training	
Result type	Exception	Nonexception
Accuracy criterion	90%	90%
Participants meeting criterion	26/47	29/50
Accuracy		
Last training block	.99	.99
Classification transfer	.40	.72
Inference transfer		
Prototype-compatible features	.28	.88
Prototype-incompatible (exception) features	.88	.13
Inference		
Ambiguous features/prototype-compatible responses	.14	.91

Note. All accuracy data are proportions correct. The data in the last row are proportions of prototype-compatible responding for inference accuracy testing trials without a clear correct answer (see text for details).

The qualitative differences between the exception and nonexception inference training conditions for the averaged memory testing data are very large indeed, so an elaborate statistical comparison is not undertaken here. However, the critical difference was extremely significant: The accuracy for prototype-compatible inference trials in transfer was .877 in the nonexception condition but only .280 in the exception condition, $t(53) = 11.974$, $p < .001$.

The average results for the classification transfer trials for the two conditions are plotted against each other in Figure 6. Note that these results and those reported below are for participants who met a lower learning criterion, 75% in the last block of training, so as to be comparable to the corresponding results from Experiment 1. As with the inference transfer results, the results for the two conditions tend to be mirror opposites of each other above and below .50 and are highly compatible with the hypothesis that participants form opposite sets of label-feature rules in the two training conditions (see Tables 7 and 5).

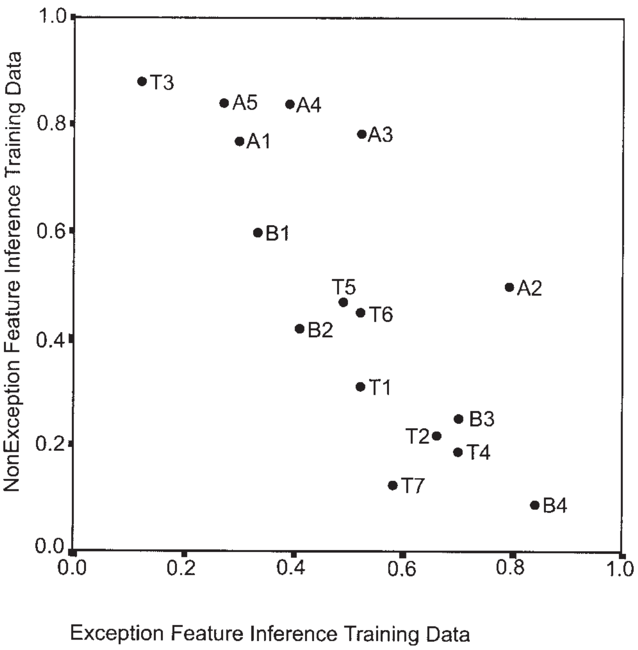


Figure 6. Classification transfer data from exception feature inference training versus classification data from nonexception feature inference training. Each data point represents the proportion of Category A responses.

Clustering Analysis of the Inference Generalization Data

The clustering results based on the inference generalization trials in the transfer phase for both conditions are given in Figure 7 and are consistent with the formation of a set of rules based on the category labels. The clustering for the nonexception condition is qualitatively similar to the clustering for the inference training condition from Experiment 1 (cf. the bottom panels of Figures 7 and 3): A number of participants responded identically to a set of rules based on the category label (represented by *label* in the clustering diagram, as explained in Experiment 1), and the remainder were noisy deviations from this core pattern. The nonexception condition data are noisier than the inference condition data from Experiment 1 (see Figure 3, bottom panel), as indicated by fewer participants responding identically and a greater maximum distance between clusters. This is probably attributable to the introduction of label-absent inference trials during transfer in Experiment 2. Clearly, the absence of the label makes using a rule based on the label difficult, though one possibility (formalized below) is that the labels are implicitly invoked by the other features via the set of rules used for label-present inference. The clustering results for the exception condition also indicate some tendency toward noisy deviation from a set of rules based on the category labels (indicated by *label* in the clustering results) but not as strongly as in the nonexception results. Some possible reasons for this are discussed below.

Label-Absent Testing Trials

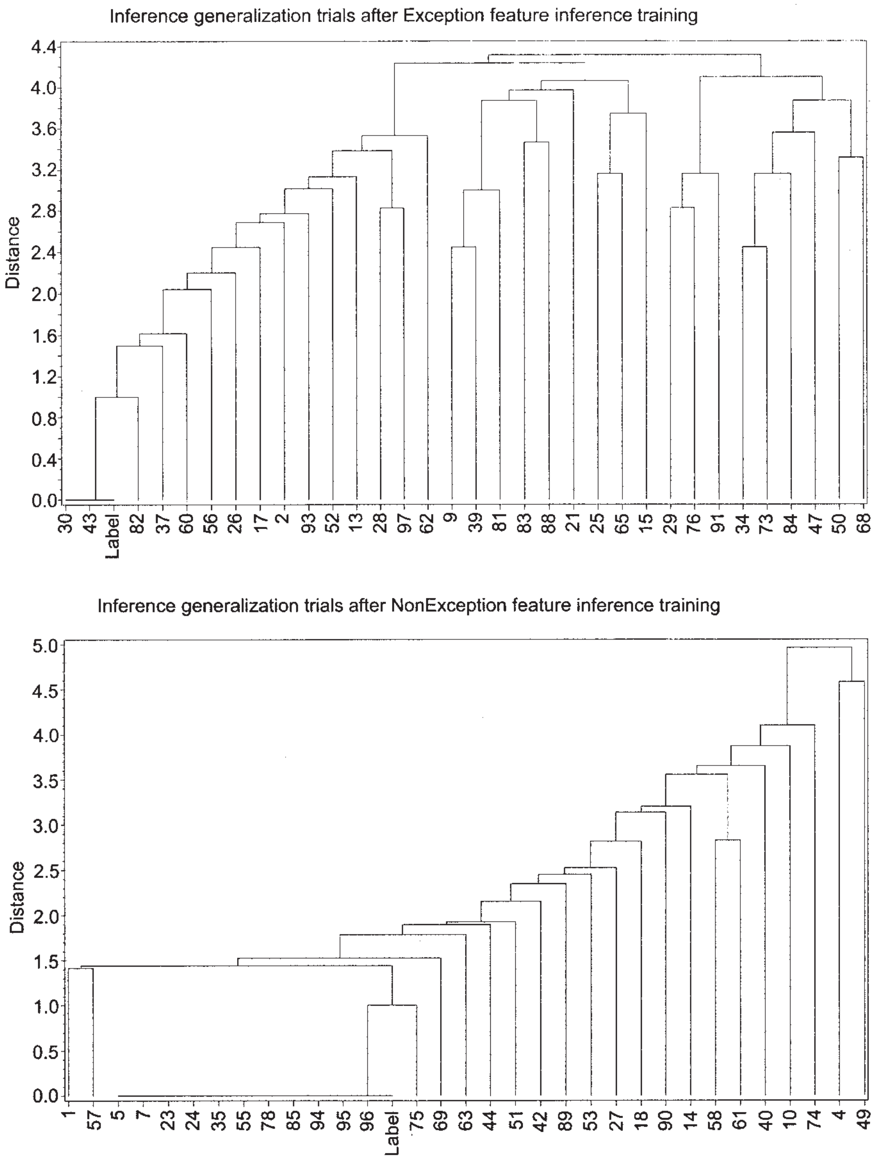
Table 9 summarizes the label-absent results in terms of proportions of responses consistent with the prototype for Category A (A 1 1 1 1) by training condition and collapsed across all trials with a given number of features consistent with the prototype for Category A. For example, the label-absent trial _ 1 1 2 ?, where the underscore indicates the absence of the label and *?* indicates the response dimension, has two features consistent with the prototype for Category A. A 1 response would be consistent with the A prototype, and a 2 response would not. These results are summarized in relation to the prototypes because the hypothesized rules for the two training conditions are opposites. For both training conditions, the proportion of responding consistent with the Category A prototype varied with the number of features from the A prototype, but this tendency was much stronger for participants in the nonexception condition. A random response strategy would result in response proportions near .50: The exception condition results (see Table 9) are all considerably closer to .50 than are the corresponding nonexception results, suggesting a stronger tendency for guessing after the exception condition.

At first glance, the label-absent results are counterintuitive and do not seem consistent with the formation of mirror-opposite sets of rules in the two conditions or, for that matter, with any label-based rules: The label-absent results are qualitatively similar for both training conditions. However, despite the explicit absence of the category labels, the label-absent results are consistent with a set of rules representation if the present features implicitly invoke the category labels. The implicitly invoked labels then predict the missing feature via the appropriate rule in the rule set applied in reverse. For example, a participant in the exception condition could have formed a set of rules including the Category A label and the 2 features on each dimension (see Table 4). So, for the label-absent trial _ 2 2 1 ?, the first two feature dimensions match the A-to-2 rules for the first two feature dimensions, providing more evidence for the Category A label than the single 1 feature on the third dimension that matches the B-to-1 rule for the third dimension. The implicitly invoked Category A label and the A-to-2 rule for the fourth dimension can then be used to predict the missing 2 feature on the fourth dimension. So, the mirror-opposite sets of rules can still predict similar patterns of label-absent results, because although they both implicitly invoke the category labels, they tend to invoke the *opposite* labels.

Assessing Noise in the Data

Finally, an overall difference between the two conditions is that responding was noisier following exception training than following nonexception training. Consistent with the clustering results difference for the two conditions (see Figure 7), the label-absent results (see Table 9) and classification transfer results (see Figure 6) both suggest that there was considerably more variability in responding following the exception condition than following the nonexception condition, as indicated by response proportions that are *closer* to .50. Counterintuitively, response proportions less extremely different from .50 actually indicate greater response variability at the individual participant level, because they correspond to roughly half of the participants making each response.

One possible reason for this pattern of results is that the nonexception condition had a greater diversity of training trials (25) than did the exception condition (11). So nonexception participants had opportunities to apply their representations to a wider variety of cases during training than did exception participants. This


Inference generalization trials after Exception feature inference training

Inference generalization trials after NonException feature inference training

Figure 7. Experiment 2 hierarchical clustering results for inference transfer generalization trials after exception or nonexception condition training. Distance is Euclidean (see text for details).

arguably facilitated the formation and strength of the rules in their representations, allowing stronger generalization performance in transfer and, correspondingly, less guessing than in exception participants. Label-absent transfer, quite novel anyway, would be expected to show this particularly strongly. The role of guessing in these data was more formally evaluated in the modeling described below, along with the hypothesis that feature inference training induces a set of rules representation reflecting inference task demands.

Modeling Analysis of the Results From Experiment 2

The models and procedures here were the same as for Experiment 1. However, the set of rules model uses opposite rules for the two conditions. Although the rules for the nonexception condition

Table 9
Average Feature Inference Testing Results for Trials on Which the Label Was Absent in Experiment 2

	Condition	
No. of A-prototype features	Exception	Nonexception
0	.424	.110
1	.460	.289
2	.510	.708
3	.531	.828

Note. All data are proportions of responses consistent with the prototype for Category A (A 1 1 1 1), collapsed across all testing trials with a given number of features consistent with the Category A prototype (on which the label was absent). No. = number.

are identical to those for the inference condition from Experiment 1 (see Table 5) and essentially equivalent to a prototype model for those data, the rules for the exception condition (see Table 7) are exactly opposite those rules and the category prototypes (see Table 4). Consequently, the set of rules model and the prototype model can be expected to be extremely different for the exception condition data while still similar for the nonexception condition.

Modeling Results for Averaged Data

The fits of the exemplar, prototype, and set of rules models to the averaged data from Experiment 2 are shown in Table 10, broken down by condition and by classification, label-present and label-absent inference transfer (with the average of the fits to individual participants in parentheses). Figures 8 and 9 plot the data versus the models' predictions.

Consistent with Experiment 1's inference condition, the prototype model does better than the exemplar model for both classification (0.065 RMSD vs. 0.154 RMSD) and label-present inference transfer (0.046 RMSD vs. 0.170 RMSD) after nonexception training. Further, the set of rules model is very similar to the prototype model for these data (0.065 RMSD vs. 0.065 RMSD for classification transfer and 0.053 RMSD vs. 0.046 RMSD for label-present inference transfer).

More important, the model fits to the transfer data following exception training are very different. Although the set of rules model reasonably accounts for the classification transfer (0.072 RMSD) and moderately accounts for label-present inference transfer (0.122 RMSD), the exemplar and prototype models provided extremely poor accounts of these data, as shown in Figure 8. Clearly, the exemplar and prototype models are accounting for essentially none of the variance in the classification and label-present inference data from the exception condition.

To deal with feature inference trials on which the label was absent, the set of rules model uses the mechanism of first classifying the instance and then using the implicitly invoked category label to predict the missing feature via the same set of rules used to account for the rest of the data. This enabled the set of rules model to do somewhat better than the prototype model (see Table 10) for both the nonexception condition (0.077 RMSD vs. 0.084 RMSD) and the exception condition (0.075 RMSD vs. 0.091 RMSD). In addition, the exemplar model was even worse than the prototype for both conditions.

Table 10
Model Fit Results for Averaged Data (and Individual Participant Data, Given in Parentheses) as Measured by Root-Mean-Square Deviation for Each Condition in Experiment 2

	Model		
Type of training	Exemplar	Prototype	Set of rules
Classification transfer			
Nonexception (n = 34)	0.154 (0.336)	0.065 (0.274)	0.065 (0.274)
Exception (n = 33)	0.192 (0.465)	0.192 (0.484)	0.072 (0.337)
Label-present inference transfer			
Nonexception (n = 34)	0.170 (0.307)	0.046 (0.224)	0.053 (0.238)
Exception (n = 33)	0.241 (0.478)	0.242 (0.499)	0.122 (0.398)
Label-absent inference transfer			
Nonexception (n = 34)	0.264 (0.481)	0.084 (0.312)	0.077 (0.307)
Exception (n = 33)	0.118 (0.468)	0.091 (0.422)	0.075 (0.416)

Note. Numbers in parentheses are averages of the fits to the individual participant data.

Modeling Results for Individual Participant Data

As for Experiment 1, the model fits to individual participant data here generally correspond to the fits of the group averaged data. The numbers in parentheses in Table 10 are the average of the models' fits to each participant's data by condition and transfer type. The prototype model does significantly better than the exemplar model for the classification transfer following nonexception training, replicating the finding from Experiment 1, $t(33) = 2.724$, $p = .010$, or Wilcoxon $z = -2.281$ (n = 28 nonties), $p = .023$, and the set of rules model performed almost identically to the prototype model. The prototype and set of rules models also do better than the exemplar model for label-present inference transfer following nonexception training (but see the discussion of the definition of the Category B prototype in the modeling results for Experiment 1). More important, the set of rules model does significantly better than the exemplar model (and the prototype model) in accounting for classification transfer after exception training, $t(32) = 4.345$, $p < .001$, or Wilcoxon $z = -3.485$ (n = 33 nonties), $p < .001$. In addition, the set of rules model does significantly better than the exemplar model in accounting for label-present inference transfer after exception training, $t(32) = 2.936$, $p = .000$, or Wilcoxon $z = -2.591$ (n = 33 nonties), $p = .010$. Finally, the fits to label-absent inference transfer indicate that the prototype model accounted for these data significantly better than the exemplar model: exception training, $t(32) = 2.129$, $p = .041$, or Wilcoxon $z = -1.543$ (n = 32 nonties), $p = .125$; nonexception training, $t(33) = 6.086$, $p = .000$, or Wilcoxon $z = -4.544$ (n = 32 nonties), $p = .000$. The set of rules model did microscopically better than the prototype model.

Overall, it must be acknowledged that the fits to the individual participant data are not as good as the corresponding fits from Experiment 1 (cf. Table 6 with Table 10), presumably reflecting a larger amount of guessing in Experiment 2, as discussed above. Still, the prototype model and the set of rules model do better than the exemplar model for the nonexception condition, and the set of rules model does better than the prototype model for the exception condition. So, only the set of rules model accounts for all of the inference training conditions.

Further Constraining the Set of Rules Model

The exemplar and prototype models failed badly on the exception condition data, so these models are not considered further. But although the set of rules model performed quite well on each transfer data subset separately—classification, label-present inference, and label-absent inference—it is also important to test whether it can account for all of the transfer data from a given training condition simultaneously. In addition, there was evidence

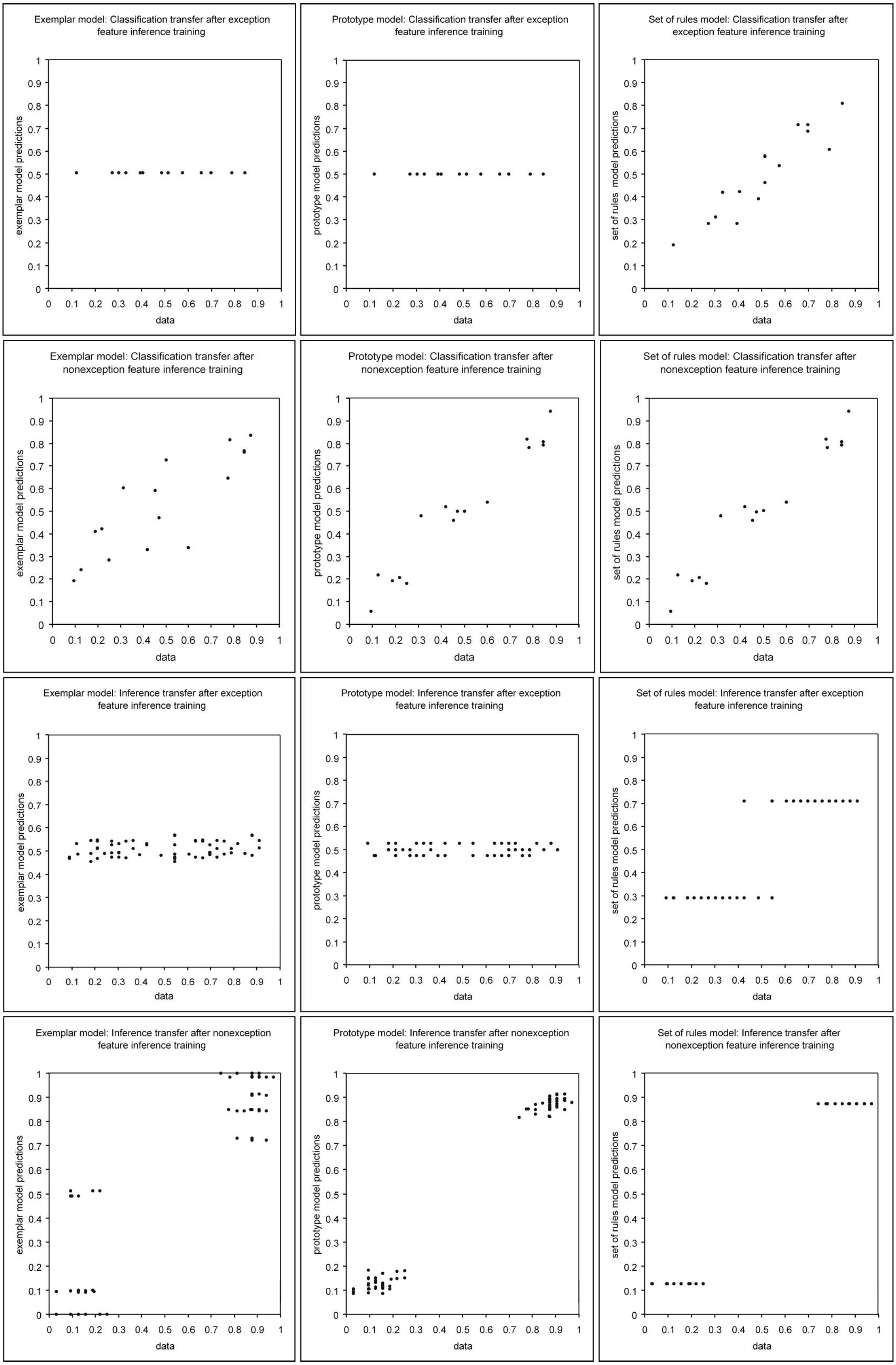


Figure 8. Predictions of the exemplar, prototype, and set of rules models plotted against the averaged data for Experiment 2 from classification and label-present transfer trials by training condition. Each data point represents the proportion of responses consistent with the Category A prototype.

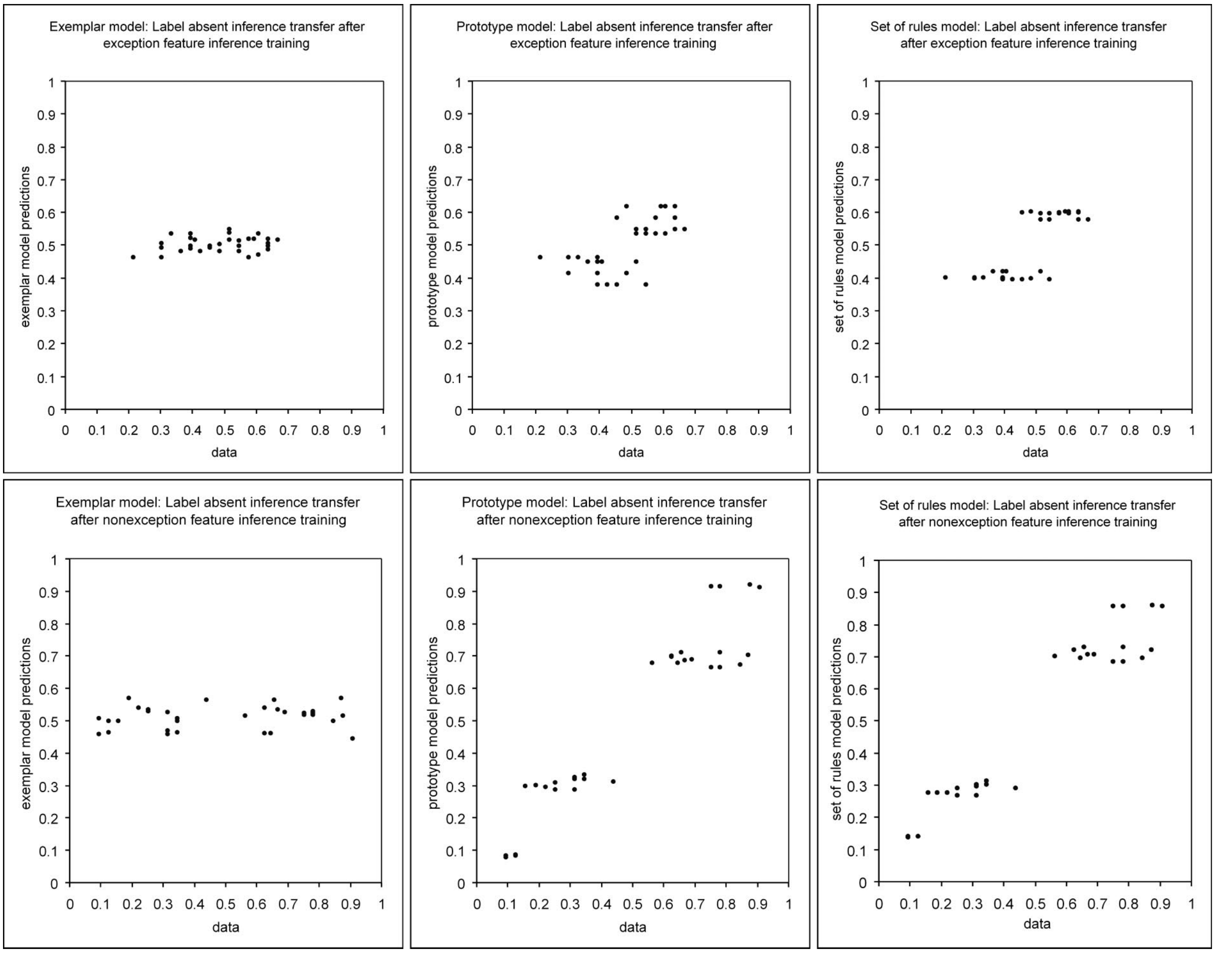


Figure 9. Predictions of the exemplar, prototype, and set of rules models plotted against the averaged data for Experiment 2 from only the label-absent transfer trials by training condition. Each data point represents the proportion of responses consistent with the Category A prototype.

of guessing, especially in the exception condition, so a guessing parameter was added to the set of rules model:

$$p(\text{Response}) = (1 - \text{Guessing}) * p(\text{Response}|i) + \text{Guessing}, \quad (9)$$

where $p(\text{Response}|i)$ is the model's response probability without guessing (see Equation 2 and/or Equation 5), and $p(\text{Response})$ is the model's response probability with guessing. The guessing parameter ranges from 0 to 1. Larger fitted values indicate a greater guessing component in the data.

The set of rules model was fit to all of the averaged transfer data for the exception training condition simultaneously, with a resulting fit of 0.109 RMSD, corresponding to 72.2% of the variance for 112 data points. The scatter plot of the model's predictions versus the data is shown in Figure 10. For comparison with the data subset fits in Table 10, the subset RMSDs for this simultaneous fit for classification, label-present inference transfer, and label-absent inference transfer, respectively, were 0.098, 0.123, and 0.083, corresponding to 74.0%, 74.4%, and 51.2% of the variance. This shows that constraining the model to account for all of exception condition data simultaneously resulted in very little reduction in its performance relative to the subset fits.

Not surprisingly, the set of rules model did even better when fit to all of the averaged transfer data from the nonexception condition simultaneously: 0.063 RMSD, corresponding to 96.4% of the variance. The scatter plot of the model's predictions versus the data is shown in Figure 10. Again, for comparison with the data subset fits in Table 10, the subset RMSDs for this simultaneous fit for classification, label-present inference transfer, and label-absent inference transfer, respectively, were 0.065, 0.054, and 0.078, corresponding to 94.0%, 98.0%, and 91.4% of the variance. There was almost no reduction in the model's performance, despite the large added constraint of having to account for all 112 data points simultaneously with the same number of free parameters.

As expected, the guessing parameters for the set of rules model fits for the two training conditions were very different: 0.558 for the exception condition and 0.225 for the nonexception condition. Although it is important not to overinterpret these numbers in terms of their absolute values, they strongly suggest that, as anticipated, there was considerably more guessing in the exception condition.

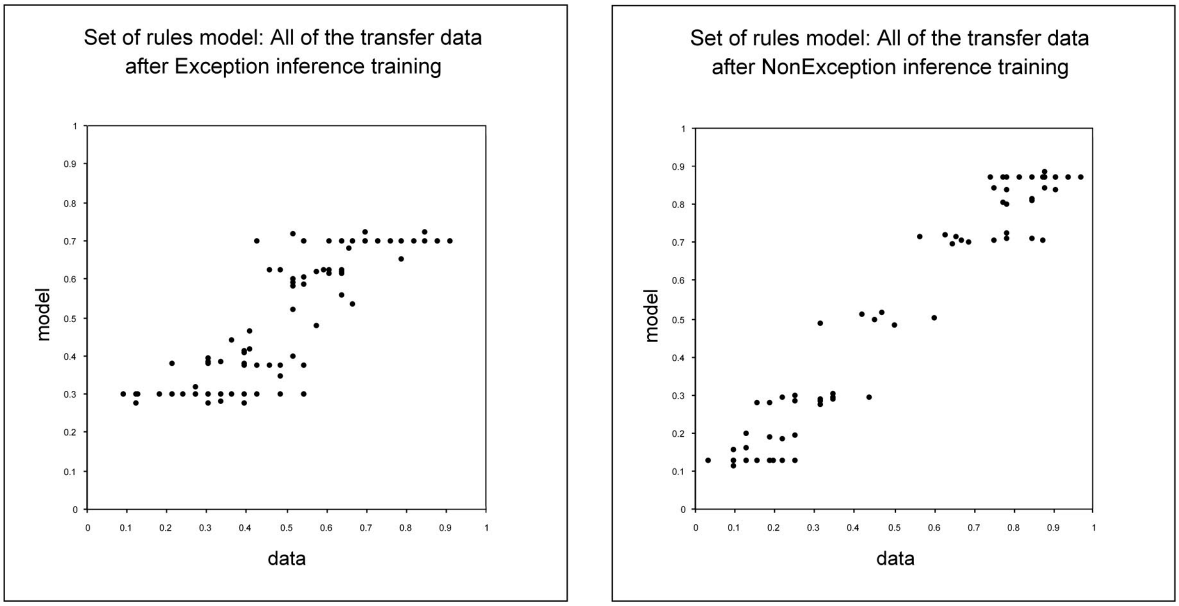


Figure 10. Predictions of the set of rules model with a guessing parameter to all of the transfer data by training condition from Experiment 2.

Adding Explanatory Mechanisms to the Set of Rules Model

The data and modeling analyses discussed so far suggest that there are both systematic differences in the nonexception and exception data and issues for the set of rules model and the exception data (though the exemplar and prototype models have far larger problems). The purpose of this section is to explore several additional mechanisms that allow the set of rules model to better account for all of the exception condition data simultaneously and, as such, provide both evidence for those mechanisms and a better understanding of the data.

Mechanisms in the elaborated set of rules model. The principles that these additional mechanisms embody are all specified in terms of moderating attention to the category label from testing trial to testing trial, depending on the feature and response properties of each trial. Formally, this corresponds to the attention weight for the category-label dimension in Equation 7, w_{label}, where the dimension k equals the label dimension, potentially taking on different values for different testing trials. However, as specified in Appendix B, the elaborated set of rules model allows different values of the attention weight for the category-label dimension without adding any free parameters beyond those that are already in the simple set of rules model, as specified previously.

Going down through the fits of the set of rules model in Table 10, the first noticeable difference between the exception and nonexception condition fits is for label-present inference transfer: 0.122 RMSD versus 0.053 RMSD. The set of rules model accounts for considerably less of the variance in the label-present inference data from after exception training (74.7%) than it did for the corresponding data following nonexception training (98.0%; see Figure 8). This result suggests that additional mechanisms not in the simple set of rules model were contributing to the variance in the label-present data following exception training. Two plausible candidates are (a) the feature values on the other dimensions in addition to the label dimension and (b) the large differences in the number of exception features on different dimensions. Both of these have been specified in terms of moderating attention to the category-label dimension (see Appendix B for the formal specification of the model).

As shown in Table 4, Dimension 2 had four exception features, Dimension 4 had three exception features, and Dimensions 1 and 3 both had only two exception features. As a consequence, participants were required to make Dimension 2 responses twice as often as Dimension 1 and 3 responses in exception training, whereas the dimensions were much closer to equal in this regard in nonexception training. Differential response strength is added to the model via greater or lesser attention to the category-label dimension, depending on whether the response dimension on a given transfer trial was a response dimension during training a lot or a little (see Appendix B for the formal specification of response strength in the model without added free parameters).

Given the clear centrality of the label dimension for the exception condition data—it accounted for the majority of the variance (74.7%)—it seems reasonable that the influences of the other feature dimensions might be moderated by the category-label dimension as well. That is, the more specific features occur with a given category label during training, the more attention the category label receives during testing to the degree that those same features still occur with it on a given transfer trial. Far from being a prototype model in disguise, this is exactly opposite the prototype model for the exception data. So, for example, attention to the category-label dimension would be high for the trial A 1 1 1 ?, a training instance, and low for A 2 2 2 ?, because these 2 features were unlikely to occur with the A label. For the exception condition data, both of these cases would moderate label attention in the application of the rule subcomponents A _ _ _ 2 and B _ _ _ 1, with high label attention for A 1 1 1 ?, hence strong prediction of a 2 feature, compared with A 2 2 2 ?, which predicts a 2 feature only weakly (see Appendix B for a specification of this mechanism without added free parameters).

In summary, although the elaborated set of rules model includes several additional mechanisms—response strength and feature-to-feature agreement in the context of specific category labels—no free parameters have been added beyond the attention and scaling parameters that are already part of the model. Further, these additional mechanisms have all been formalized via the simple process of moderating attention to the label dimension, which already serves as the basis for the rules in the simple set of rules model.

Modeling results for the elaborated set of rules model. The elaborated set of rules model was fit to all of the averaged transfer data for the exception training condition simultaneously, with a resulting fit of 0.083 RMSD, corresponding to 84.2% of the variance for 112 data points. In more detail, the set of rules model accounted for 77.0% of the variance for the classification data, 89.6% of the variance for the label-present inference data, and 48.2% of the variance for the label-absent inference data. For comparison, the fit of the simple set of rules model to all of the averaged transfer data reported above was 0.109 RMSD, corresponding to 72.2% of the variance. In particular, the elaborated set of rules model accounted for 89.6% − 74.4% = 15.2% more of the variance for the label-present inference data than did the simple set of rules model, suggesting that the principles formalized in these additional mechanisms are partly responsible for the exception condition data.

Further analysis of the four feature dimension attention parameters for the fits of the elaborated set of rules model above suggested that the model could be simplified in several ways that further emphasize the critical aspects of the data. In particular, the attention parameters for the two different conditions apparently reflect the relative degree of training for the hypothesized rules, which is not surprising given the response-weighting mechanism described above.

In the nonexception training condition, the number of trials within a training block involving a given response dimension (see Table 4) varied only from five for Dimension 2 to seven for Dimensions 1 and 3. If participants did, in fact, learn a set of label-to-feature rules, then the rules should have all been learned about equally well in that the amount of training consistent with each rule was roughly the same. In fact, the feature-dimension attention parameters in the fit of the elaborated set of rules model to the nonexception data—respectively, 0.875, 0.815, 0.866 and 0.886 RMSD—suggested that the feature attentions were all essentially the same and very close to 1. Consistent with this, fitting the model to these data with the attention parameters fixed at 1 and only the guessing and response-scaling parameters free to vary resulted in only a slight decline in the fit, to 0.064 RMSD from 0.063 RMSD.

In the exception training condition, however, the number of trials within a training block involving a given response dimension (see Table 4) varied considerably, from two for Dimensions 1 and 3 to four for Dimension 2. If participants did, in fact, learn a set of label-to-feature rules, then the rule for Dimension 2 responses should have been considerably overlearned fully twice as well as two of the other rules. In fact, the feature dimension attention parameters—respectively, 0.264, 1.662, 0.099, and 0.243 RMSD—suggested that the Dimension 2 parameter was much larger than the other dimensions. Consistent with this, fitting the model to these data with the attention parameters for Dimensions 1, 3, and 4 fixed at 0 resulted in only a moderate decline in the fit, to 0.094 RMSD from 0.083 RMSD.

Taken together, these fits with the attention parameters fixed at values consistent with the learning tasks strongly suggest that the set of rules model accounts for these data because it embodies reasonable theoretical principles. The set of rules model is not merely taking advantage of arbitrary parameter flexibility, which, together with its simplicity, makes it highly falsifiable.

In summary, the empirical and mathematical modeling results for Experiment 2 support the conclusion that feature inference training encourages the formation of a set of bidirectional rules based on the category labels that are consistent with the inference task demands. The results of this experiment do not support prototype or exemplar representation as a result of inference training, and they indicate that although prototype and set of rules representations can result in similar predictions (as in Experiment 1), they are fundamentally distinct.

General Discussion

In the experiments and modeling described here, we have evaluated the category representations resulting from learning a category structure by classification or feature inference. The data from Experiment 1 indicate that the category representations from classification and feature inference training are very different. These results are moderately consistent with the conclusion that classification training results in exemplar rather than prototype representation, as suggested by prior research. More important, the results of inference training are much more consistent with prototype than with exemplar representation, according to standard model fits.

The results of Experiment 1 replicate and extend several of the findings from Yamauchi and Markman’s (1998) Experiment 1 using a different category structure and a richer set of testing trials, including classification and inference generalization trials. In particular, we have shown that these data better differentiate exemplar and prototype representation after inference training (see Table 6) than did Yamauchi and Markman’s (1998) Experiment 1 (see Table 3).

It is worth emphasizing that Experiment 1 provided data from a perceptual category learning task, which is better fit by a prototype model than by a standard exemplar model. This is particularly ironic given the historical importance of the 5–4 category structure in establishing the dominance of exemplar representation and this structure’s continuing importance (Johansen & Palmeri, 2002; Nosofsky, 2000; Nosofsky & Johansen, 2000; Smith & Minda, 2000).

The results of Experiment 2, however, suggest that the prototype model provides a reasonable account of inference condition data from Experiment 1, because of the inference task demands and because it is mimicking a set of rules model that integrates across bidirectional label-feature rules, one for each response dimension. This conclusion is particularly supported by results for the exception condition in Experiment 2, in which participants learned the category structure by feature inference on the prototype-incompatible features. Both the exemplar and prototype models predicted essentially none of the variance in these data, but a set of rules model provides a reasonable account. The set of rules model emphasizes the importance of rules based on the category labels. These rules can be used in the label-to-feature direction for feature inference and in the feature-to-label direction for classification. Further, this model emphasizes the centrality of the labels by being able to account for inference trials on which the label was absent by assuming that the label is implicitly invoked via the same set of rules. In summary, only the set of rules model can account for all of the inference learning results in these experiments.

As it would have been nice after Experiment 1 to conclude in general that feature inference training induces prototype representation, so would it have been nice to conclude after Experiment 2 that feature inference training in general induces representation that exactly matches our elaborated set of rules model. We believe that we have compellingly demonstrated that feature inference training does not in general result in a representation that is compatible with either standard exemplar or prototype models. This is an important conclusion given the prior success of these models. And to the degree to which exemplar and/or prototype representation has been shown to result from classification learning, classification and feature inference learning must result in different representations. However, just as the recent proliferation of mixed-representation models has shown the difficulty of reaching general conclusions about the representation that results from classification learning (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; Love et al., 2004), so too our results emphasize the greater difficulty of determining both what representation results from feature inference learning in general and whether it is qualitatively different from classification learning in particular.

Our data suggest that feature inference learning tasks result in category representation by a set of label-based rules, with the specific rules largely determined by the inference task demands. If participants learn prototype-compatible features, they generalize to new instances using a representation based largely on prototype-compatible features and the category labels. If participants learn prototype-incompatible features, they generalize using a representation based largely on prototype-incompatible features and the category labels. The exemplar model performed worse than the set of rules model on all of the inference learning conditions in these experiments. Still, it seems likely that there is some inference learning task that will induce participants to form a representation based on the configuration of features together with the label. For example, knowing that an entity is an animal is not by itself sufficient to accurately predict whether or not it is a vegetarian, but knowing it is an animal *and* has canine teeth allows a much more accurate prediction. This kind of task would necessitate the inclusion of more complex rules in our set of rules model, but it still seems conceptually compatible.

Relationship of These Results to Prior Research

We do not attempt here to summarize all of the recent research on the feature inference paradigm and its relationship with the standard classification paradigm. A clear summary can already be found in Markman and Ross (2003), which, in particular, cogently argues that the dominance of the classification paradigm has resulted in the partial neglect of other aspects of categorization in terms of both acquisition and use. However, we do discuss the relationship of our research to some of the recent feature inference research in the context of the hypothesis (Markman & Ross, 2003; Yamauchi, Love, & Markman, 2002; Yamauchi & Markman, 1998) that classification and feature inference category learning in general result in fundamentally different types of category representations. Specifically, the hypothesis is that classification learning induces category representations that reflect a focus on what differentiates the categories, whereas feature inference learning induces representations that focus on the internal structure of each category while deemphasizing what differentiates the categories. It is not our intent to dismiss this intuitively compelling hypothesis and its associated research, particularly given that it has been our working hypothesis. Our conclusion is not that there are no systematic differences between classification and feature inference learning. Rather, more work is needed to clarify potential differences in the representations resulting from these two kinds of learning.

The results of our Experiment 2 are not supportive of the hypothesis that feature inference training results in a representation that emphasizes the internal category structure, in that two different types of feature inference training resulted in extremely different data sets—one of which is strongly compatible with prototype representation and the other of which is very incompatible. Although, in a weaker sense, the internal category structure can be said to be emphasized because of the clear importance of the category labels, this is not the same as emphasizing typical features in preference to diagnostic features that differentiate the categories.

The main empirical result of Yamauchi et al. (2002) was that participants found feature inference on a linearly inseparable category structure extremely difficult, and many did not even learn the task: "This finding reflects that the prototype of a nonlinearly separable category does not provide a good summary of the category members" (p. 585). This conclusion was made in the context of the following idea: "Learning categories by making predictive inferences focuses learners on an abstract summary of each category (e.g., the prototype)" (Yamauchi et al., 2002, p. 585). Although Experiment 2 has led us to question that this result demonstrates prototypes as the default representation for feature inference learning, it is worth pointing out that our set of rules model is conceptually compatible. Like the prototype model, the set of rules model, as formulated, does not allow accurate feature inference performance on a linearly inseparable category structure. Following the same logic as Yamauchi et al., this can be taken as evidence for the model, because participants did, in fact, have difficulty learning this structure.

Our results are consistent with the idea that it is difficult to draw general conclusions about whether feature inference learning results in representations that are systematically different from those that result from classification learning, because the representation resulting from feature inference learning seems to be heavily influenced by task demands. In this context, we have a more fundamental concern about the results of Yamauchi et al. (2002). They found that a linearly *inseparable* category structure was significantly harder to learn by feature inference than by classification, whereas Yamauchi and Markman (1998) found that a linearly *separable* category structure was significantly easier to learn by classification than by feature inference—hence the prototype representation conclusion. However, the task demands for the two inference conditions were different. Yamauchi et al. required participants to learn to infer both prototype-compatible and prototype-incompatible features, whereas Yamauchi and Markman (1998) only required participants to learn prototype-compatible features. Hence, Yamauchi et al.'s inference task required the participants to learn approximately *twice* as many different responses as Yamauchi and Markman's (1998) inference task. So it is not surprising that the one task was harder than the other relative to the corresponding categorization learning tasks, without regard to whether the category structure was linearly separable or inseparable. It is plausible that learning linearly inseparable category structures by feature inference should be harder than learning linearly separable ones, if for no other reason than the highly salient category label is an extremely valid predictor in the first but, by itself, a poor predictor in the second. However, further research is needed to clearly establish this.

Another task demand concern occurs in the context of the claim (Chin-Parker & Ross, 2004; Markman & Ross, 2003; Yamauchi & Markman, 1998) that classification and feature inference on the family resemblance structure in Table 1 are equivalent: If the category label is treated as just another feature, then an inference trial requires producing a response given four features, just as does a classification trial. But beyond this, the classification and inference tasks differ in at least three fundamental ways that are not inherent to the definitions of classification and inference in their most general senses (which, for inference, would be learning about a category structure by learning to infer some of the features of its instances): (a) In the classification condition, participants were trained to classify all of the instances in the category structure, but in the inference condition, participants were only trained to infer the prototype-compatible features, not all features of all instances. (b) Correspondingly, although there were only two different responses in the classification condition (the two category labels), there were eight different responses to keep track of across all of the learning trials in the inference condition, four different features for each of the two category prototypes. (c) A single, highly salient and perfectly valid dimension was present in the inference task—the category-label dimension—but not during the classification task. Again, this is not inherent to the definitions of classification and inference, because the category labels need not be perfectly predictive for learning the task (e.g., Yamauchi et al., 2002). Further, the relative salience and validity of the category-label dimension can be partially manipulated by, for example, using a category structure with another highly salient and valid dimension. These potential classification and inference task asymmetries do not all apply to or explain every published result in this paradigm, but neither have they been systematically specified and explored.

The results of Chin-Parker and Ross (2004) are another example in which the potential task asymmetries between classification and feature inference training may have played a role. Chin-Parker and Ross trained participants by either feature inference only on the prototype-compatible features or classification on a category structure in which the prototypes of the two categories overlapped on some of the features dimensions but not on others. It was possible to independently manipulate diagnosticity and prototypicality in the testing items, because the features on which the prototypes did not overlap were diagnostic of the categories, but the overlapping features were prototypical, not diagnostic. The key results were that inference learners showed greater sensitivity to prototypical features that were nondiagnostic than did classification learners, and correspondingly, classification learners showed greater sensitivity to diagnostic features than to nondiagnostic features. These results are consistent with the hypothesis that classification learning emphasizes diagnosticity, whereas inference learning emphasizes central tendency, but task demand differences may also have had an influence. If participants are only trained on the prototype-compatible features in inference training, it seems likely that they should be particularly sensitive during the testing phase to prototype-compatible features, regardless of their diagnosticity, because the task demands it. Had participants been trained on only a subset of the prototype-compatible features—for example, only the diagnostic ones or on both the prototype-compatible and prototype-incompatible features—it is possible that the results would have been considerably weaker. More important, for accuracy to be high, this particular inference task *required* participants to learn the prototype-compatible features, regardless of whether these were diagnostic, whereas the classification task *allowed* high accuracy even if the participants only attended to the diagnostic features. It is possible that these results can be explained by participants attending to as little as possible to sufficiently perform well in the task. It is not clear that these results indicate a general property of feature inference learning.

Consistent with our set of rules model, the importance of the category label for determining responding during inference transfer is supported by Yamauchi and Markman (2000), though they used a design in which all of the instances of the categories were displayed at once rather than being learned over a series of trials by classification or inference. They presented evidence that feature inference at test was strongly determined by the category label, even when the majority of the other features had greater similarity to the contrasting category. Using a similar methodology, Yamauchi (2003) concluded that inference at test was mediated by a rule-based decision process, which presumably emphasizes the category labels. Developmental work by Gelman and Markman (1986, 1987) also supports the supremacy of category labels, even in the context of contradictory similarity information, though Loose and Mareschal (1999) suggested that this conclusion may apply strongly only to relatively unhomogeneous categories and that both the label and feature similarity play a role.

Overall, it is intuitively plausible in the context of real-world category learning that category labels, when available, play a crucial role, because they indicate that a categorical structure relates some of the observed features in the environment. In other words, category labels may serve as a marker that label-based rules can be found and exploited.

References

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., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science, 5,* 144–151.

Blair, M., & Homa, D. (2003). As easy to memorize as they are to classify: The 5–4 categories and the category advantage. *Memory & Cognition, 31,* 1293–1301.

Chin-Parker, S., & Ross, B. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 216–226.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107–140.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Gelman, S., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23,* 183–209.

Gelman, S., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development, 58,* 1532–1541.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45,* 482–553.

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

Kruschke, J. K., Johansen, M. K., & Blair, N. J. (1999). *Exemplar model account of inference learning: Commentary on Yamauchi and Markman (1998)* [Unpublished manuscript]. (Available from http://www.indiana.edu/~kruschke/yamauchicomment.html)

Loose, J., & Mareschal, D. (1999). Inductive reasoning revisited: Children's reliance on category labels and appearance. In Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (pp. 320–325). London: Erlbaum.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics, 61,* 354–375.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53,* 49–70.

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

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 3,* 333–352.

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 607–625.

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

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355–368.

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

Myung, I. J. (1997, August). *The importance of complexity in model selection.* Methods for Model Selection Symposium, Indiana University Bloomington.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–109.

Nosofsky, R. M. (1992). Exemplars, prototypes and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 149–167). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M. (1998). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance, 24,* 322–339.

Nosofsky, R. M. (2000). Exemplar representation without generalization? Comment on Smith and Minda's (2000) "Thirty Categorization Results in Search of a Model." *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1735–1743.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22,* 352–369.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7,* 375–402.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. *Psychology of Learning and Motivation, 28,* 207–250.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 221–233.

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

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 924–940.

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

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 3–27.

Vandierendonck, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin & Review, 2,* 442–459.

Yamauchi, T. (2003). Dual processes in the acquisition of categorical concepts. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 1259–1264). Boston: Cognitive Science Society.

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, 28,* 585–593.

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

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 776–795.

(*Appendixes follow*)

Appendix A

An Alternative Prototype Model

The prototype model can be reformulated to reflect the ambiguity on Dimension 2 for Category 2 by including two prototypes for that category—namely, B 2 1 2 2 and B 2 2 2 2—and then, for symmetry, including two instances of the Category A prototype—namely, A 1 1 1 1 and A 1 1 1 1. For this alternative prototype model, the superiority of the prototype model's fit (0.144 root-mean-square deviation [RMSD]) over the exemplar model's fit (0.140 RMSD) for averaged data completely disappears, as it does for the average of the fits to individual participants (0.249 prototype RMSD and 0.238 exemplar RMSD).

The alternative prototype model is not intended to be an appropriate model for these results, because clearly it is not a standard prototype model, and in any event, it fit the data poorly. Rather, the results for the alternative prototype model are intended to emphasize that the standard prototype model does well on the inference data from after inference training compared with the exemplar model, because the traditionally defined prototype for Category B somewhat arbitrarily matches what participants were trained on in inference training.

Appendix B

Formal Specification of the Elaborated Set of Rules Model

The elaborated set of rules model includes two additional mechanisms beyond the simple set of rules model—response strength weighting and feature-to-feature agreement—both of which are intuitively motivated in the main text. Both of these mechanisms are formalized below in terms of greater or less attention to the category-label dimension, depending on the response dimension and other features present on a given transfer trial.

Specifically, the elaborated set of rules model is identical to the simple set of rules model, except that it replaces the free parameter attention weight for the label dimension in Equation 7, w_{label}, when the dimension k equals the label dimension, with w^{*}_{label} as specified by this equation:

$$w^{*}_{\text{label}} = w_{\text{Label}} + w_{\text{FAgree}} + w_{\text{RespDm\_k}}. \tag{B1}$$

w_{Label} is the original label attention free parameter as used by the exemplar, prototype, and set of rules models. It is still needed as a free parameter, because w_{FAgree} and $w_{\text{RespDm\_k}}$ are not free parameters but, rather, are constrained by the free parameters that are already used by the exemplar, prototype, and simple set of rules models.

w_{FAgree} is a measure of the degree to which the features occurring with a given label on a testing trial also occurred with that label during training. It is specified as the sum of feature weights for those features that tended to occur with a given label during training—that is, the prototypical features for that label. Because there are four feature dimensions, there are four possible feature agreement weights $w_{\text{FAgree\_k}}$, but to avoid the proliferation of free parameters, these are respectively set to $1 - w_{\text{k}}$, where the w_{k}s are the dimensional attention weights used in Equations 1, 4, and 7 by the exemplar, prototype, and simple set of rules models. So, overall,

$$w_{FAgree} = \sum_{\substack{k \neq respdm \\ k = protof}} f_p(1 - w_k) \tag{B2}$$

where $f_p = 1$ if the feature is from the same prototype as the instance's label, and $f_p = 0$ otherwise.

Because the attention parameters, the w_ks, were not restricted to be less than 1, the feature agreement score and, correspondingly, the overall adjusted dimensional attention, w_{LabelDm}, had the potential to be negative. To prevent this, the adjusted label attention was forced to be greater than or equal to 0, the natural constrain on the other attention parameters.

In summary, the moderated attention to the category-label dimension, $w^{*}{}_{\text{label}}$, tends to be relatively high when the features on a given transfer trial tended to occur with the category label present on the trial during training and when the response dimension occurred with high relative frequency during training. Otherwise, $w^{*}{}_{\text{label}}$ tends to be lower. Note that this model is not a prototype model and can behave very differently, as discussed in the main text.

Received May 2, 2003
Revision received June 14, 2005
Accepted June 15, 2005 ■