TWO KINDS OF FEATURE? A TEST OF TWO THEORIES OF TYPICALITY EFFECTS IN NATURAL LANGUAGE CATEGORIES

Robin A. Barr & Leslie J. Caplan,
Department of Psychological Science, Ball State University, Muncie, IN 47306

The authors thank the teachers and students of Burris Laboratory School, Muncie, Indiana for their cooperation. Experiment 1 was supported by a Ball State University research grant.

In two experiments, we tested predictions of a model which states that natural language categories are represented primarily by intrinsic features (features true of an exemplar in isolation) or by extrinsic features (features true of an exemplar interacting with some other entity. We hypothesized that categories whose intensions consisted primarily of extrinsic features would have "fuzzier" extensions than those whose intensions consisted primarily of intrinsic features. This hypothesis was supported by the results of Experiment 1. In Experiment 2, we demonstrated that the features which represent a category are equally descriptive of it, regardless of whether the representation is primarily intrinsic or extrinsic - a finding inconsistent with a probabilistic theory's account of the results of Experiment 1. We argue that investigation of the properties of different kinds of feature is an appropriate focus for future research on natural-language categorization.

Typicality effects in studies of natural language category representation have frequently been interpreted as evidence that categories are "fuzzy" (e.g., Rosch, 1973, 1975). Recently, however, other investigators have questioned whether the existence of typicality effects does, in fact, imply that categories are necessarily fuzzy (e.g., Armstrong, Gleitman, & Gleitman, 1983; Barsalou, 1982; Osherson & Smith, 1981).

Our purpose here is to show that some typicality effects are caused by a particular kind of category intension, rather than by the fuzziness of the category per se. We previously (Barr and Caplan, note 1) have distinguished between intrinsic and extrinsic features in describing the intensions of categories. Intrinsic features are those which are true of a category member considered in isolation. For example, an intrinsic feature of "dog" is "has fur". Extrinsic features are true only when the category member interacts with some other entity. For example, an extrinsic feature of "clothing" is "covers people".

In that earlier paper we found that most of the features chosen by subjects to define artifactual categories (e.g. "tools" and "furniture") were extrinsic. Conversely, most of the features chosen to define naturally-occurring categories (e.g. "fruit" and "trees") were intrinsic. We

also found that subjects agreed with each other just as much in choosing defining features for artifactual categories as in choosing features for naturally-occurring categories. However, they agreed less when generating exemplars of artifactual categories than when generating exemplars of naturally-occurring categories. In other words, although the intensions of the two types of category were equally clear, the extensions of artifactual categories were "fuzzier" than those of naturally-occurring categories.

We argued that extrinsic features cause fuzzy extensions while intrinsic features result in clearer extensions. Because intrinsic features are always true of a category member, it makes little sense to apply a qualifier when deciding whether an exemplar possesses a particular feature (e.g., "Sparrows are sometimes hatched from eggs"). Either the exemplar possesses the feature, or it does not. On the other hand, it does make sense to apply qualifiers when making decisions about extrinsic features (e.g., "A sled is sometimes used for transportation"). Borderline category members will be those which can be considered a member only with the addition of strong qualifiers (e.g. "Rubber bands are SOMETIMES used to hurt people"). Therefore, categories which are represented primarily by extrinsic features should have fuzzier extensions than those represented primarily by intrinsic features.

Intrinsically represented categories will, however, still yield some effect of typicality on membership judgments. Several different factors will contribute to the size of a typicality effect. For example, as some investigators (e.g., Armstrong, Gleitman & Gleitman, 1983; Osherson & Smith, 1981) have suggested, subjects may well use an "identification function" in typicality experiments. However, because intrinsic features are true of exemplars without qualification, the effect of typicality should be smaller in intrinsically represented than in extrinsically represented categories.

We argued in Barr and Caplan (note 1) that, in practice, many categories will be represented by a mixture of extrinsic and intrinsic features. Accordingly, the relative importance of extrinsic and intrinsic features to the representation of the category will determine how fuzzy the extension becomes. Our data implied that the artifactual categories we used in these experiments were represented more by extrinsic features than were the naturally-occurring categories that we used. Therefore, in Experiment 1, we tested the prediction that the rated typicality of an item will yield larger effects on membership judgments for artifactual categories than on membership judgments for naturally-occurring categories.

It also follows that the extensions of intrinsically represented categories should be easier to learn than those of extrinsically represented categories.To learn intrinsically represented categories, children need only attend to the exemplar itself when learning the category label. However, to learn extrinsically represented categories, children need to attend both to the exemplar and to the context in which the exemplar occurs. Atypical exemplars of extrinsically represented categories will rarely be encountered in contexts consistent with the category's features In the first experiment, we therefore tested the hypothesis that school-age children's extensions of artifactual categories would be less adult-like than their extensions of naturally-occurring categories. In particular, we predicted that children's category-membership judgments will be least adult-like for atypical, artifactual exemplars.

Experiment 1

The artifactual categories we chose were "furniture" and "clothing". In our previous paper, we asked three independent judges to rate the five most popular features defining a category (as chosen by our subjects) as either intrinsic or extrinsic. According to these judgments, three of the top five features for "furniture" were extrinsic, and four of the top five features for "clothing" were extrinsic. The remaining features for both categories were judged to be intrinsic. The naturally-occurring categories used in this experiment were "birds" and "fruit". According to our independent judges, four of the top five features for "birds" were judged intrinsic, and three of the top five features for "fruit" were judged intrinsic. The remaining features for both "fruit" and "birds" were judged to be extrinsic.

We asked subjects in this experiment to decide whether exemplars which differed in category goodness-of-example ratings according to Rosch (1975) were members of categories. We predicted that the differences in RT and "no" judgments for "true" sentences between naturally-occurring and artifactual categories would be greater for atypical than for typical instances of the categories.

Method

Subjects. Subjects were 28 college students, some of whom received course credit for participation, and 33 school-age children. Seventeen of the children were from combined second- and third-grade classes (mean age = 8 years, 7 months), thirteeen of them were from the fourth grade (mean age = 10 years, 2 months), and nine were sixth graders (mean age = 11 years, 9 months). All of the children were students at Burris School, a public laboratory school run by Ball State University which serves grades kindergarten through twelve. The adult volunteers were all undergraduates at the same university.

Stimuli. The stimuli for the practice trials were constructed from two sets of three exemplars from the categories "animals" and "tools". They were presented in sentences of the form "A hammer is a tool". Each exemplar was presented twice, once in a true sentence, and once in a false sentence (i.e., a sentence using the exemplar from one of the two categories, paired with the other superordinate), yielding a total of twelve sentences.

Stimuli for the main part of the experiment were constructed from sets of ten exemplars drawn from each of the naturally-occurring categories "fruit" and "bird" and from each of the artifactual categories "furniture" and "clothing". As in the practice trials, each stimulus was a sentence of the form "Apples are fruit". For each category, two exemplars were selected at each of five different levels of goodness-of-example (Rosch, 1975). As far as possible, mean typicality ratings for each level were equated across the different categories. However, as only one "bird" ("bat") has a mean goodness-of-example rating above 5.00 in the Rosch data, it was not possible to completely control for differences in these numerical ratings. In addition, the mean number of letters in exemplar words from all categories was equated as much as possible across typicality levels.

As in the practice trials, each of the exemplars used in the experimental trials was presented once in a true sentence, and once in a false sentence. To construct the false sentences, bird exemplars were paired with the superordinate "fruit", furniture exemplars were paired with the superordinate "birds", fruit exemplars were paired with the superordinate "clothes", and

clothing exemplars were paired with the superordinate "furniture". The resulting 80 sentences (40 true and 40 false sentences) were divided into two trial blocks of 40 sentences each. Each trial block included 20 true sentences, one at each level of typicality for each of the four categories. The remaining 20 false sentences were divided equally among the four categories used.

Apparatus. Stimuli were presented using a three-channel tachistoscope. Reaction times were recorded by a voice-activated relay and timer system.

Procedure. Each subject was tested individually. At the beginning of each session, children were tested for their ability to read the exemplars and superordinates to be used in practice and experimental trials, by reading these words aloud when they were presented on flashcards. If a child was unable to read a word, the child was corrected and that word was later re-presented. Words read correctly were not re-presented. This procedure was repeated until the child had read each word aloud correctly once. This procedure was not used with adult subjects.

Subjects were instructed that they were going to see a series of sentences, and that their task was to decide whether each sentence was true or false. If they decided the sentence was false, they were to say "false" aloud. If they decided it was true, they were to say "true" aloud. All subjects were instructed to respond as quickly and accurately as possible. They were also instructed to fixate a fixation point at the beginning of each trial.

Practice trials were then presented, followed by the experimental trials. Each sentence was preceded by a fixation point for 1.5 seconds, and each sentence was presented for 4 seconds. The order in which the two trial blocks were presented was counterbalanced. Reaction times were recorded from the onset of the stimulus sentence, for experimental trials only.

Results

Despite pretraining in the reading of stimulus words, some children were still unable to read tachistoscopically presented stimulus sentences. Therefore, data from four of the second- and third graders and from one fourth grader were not included in the following analyses.

We tested the following predictions: 1) that the number of "no" responses would be greater for artifactual than for naturally-occurring categories, especially for atypical exemplars, 2) that subjects would be slower to verify atypical artifactual exemplars than to verify atypical naturally-occurring exemplars, 3) that children's judgments would differ from adults' most for atypical artifactual category exemplars, and 4) that the difference in RT between children and adults would be greatest for atypical artifactual exemplars. A summary of the data used in the analyses reported below is presented in Table 1.

"No" judgment analyses. A three-way mixed design analysis of variance was conducted on the number of "no" judgments made for "true" sentences, with category type (artifactual vs. naturally-occurring) and typicality level (levels one through five) as within-subject variables, and age (children vs. adults) as the between-subject variable. For both this analysis, and the one presented below, children were not divided into age groups because previous analyses had failed to reveal any interaction between age and the other variables of interest when the data from children were considered alone.

More "no" judgments were made for artifactual categories than for

naturally-occurring categories, $F(1, 57) = 49.70$, $p < .0001$. These judgments also increased as typicality decreased, $F(4, 228) = 71.84$, $p < .0001$. However, an interaction between category type and typicality level was also obtained, $F(4, 228) = 29.74$, $p < .0001$. Subjects made more "no" judgments for atypical artifactual items than for atypical naturally-occurring items, although the number of "no" judgments was similar for typical artifactual and naturally-occurring categories. All remaining main effects and interactions failed to reach significance at the $p < .05$ level.

The predicted three-way interaction between age, category type, and typicality level was clearly not significant in the above analysis. However, our prediction originally involved atypical items. Therefore, in a second analysis, we investigated the effects of age, category type, and typicality level only for the three least typical conditions. In this analysis, the three-way interaction was significant, $F(2, 118) = 3.11$, $p < .0485$. Children's and adults' judgments were similar for naturally-occurring categories. For artifactual categories, children made more "no" judgments than adults for medium typicality items, and fewer "no" judgments than adults for the most atypical items. As in the previous analysis, there were significant effects of category type ($F(1, 59) = 63.14$, $p < .0001$), and of typicality level ($F(2, 118) = 64.58$, $p < .0001$), and an interaction between category type and typicality ($F(2, 118) = 42.40$, $p < .0001$).

Reaction time analyses. A three-way mixed design analysis of variance was conducted on individual subjects' mean reaction times, using the same design as that of the first analysis reported above. Subjects took longer to verify statements about artifactual categories than naturally-occuring categories, $F(1, 43) = 25.48$, $p < .0001$. Reaction times also increased as typicality decreased, $F(4, 172) = 17.40$, $p < .0001$. Children showed a greater increase in RT for atypical items than did adults, as reflected in an interaction between age and typicality, $F(4, 172) = 2.92$, $p = .0229$. The difference in RT between naturally-occurring and artifactual categories was marginally greater for children than for adults, $F(1, 43) = 3.63$, $p = .0634$. The predicted interaction between category type and typicality level was not obtained. Children's RT's were longer than those of adults, $F(1, 43) = 67.59$, $p < .0001$. All remaining effects failed to reach significance at the $p < .05$ level.

Finally, as in analyses of "no" judgments, we investigated the effects of age, category type, and typicality level on RT, using only the three least typical conditions. Once again, the effects of category type ($F(1, 43) = 8.45$, $p = .0058$, and age ($F(1, 43 = 61.26$, $p < .0001$) were significant. The main effect of typicality was no longer significant at the $p < .05$ level. All remaining effects failed to reach significance at the $p < .05$ level.

Discussion

The results of this study are consistent with those of our earlier work (Barr & Caplan, note 1). Once again, there was a clear difference between the the nature of the extensions of artifactual and naturally-occurring categories. Reaction times and the number of "no" judgments were higher for artifactual than for naturally-occurring categories. In addition, artifactual categories demonstrated a more pronounced effect of typicality in "no" judgments than did naturally-occurring categories. Finally, children's RT's

and judgments suggest that they have more difficulty learning the extensions of artifactual categories than of naturally-occurring categories.

Our original predictions were similar to the results obtained. However, we had expected to find an interaction between category type and typicality level in the RT data, which we did not find. Instead, judgments for artifactual categories were slower than those for naturally-occurring categories at all levels of typicality.

How might this result be explained? We had expected this interaction because of the nature of extrinsic features. Presumably, atypical artifactual items would have forced subjects to engage in a lengthy search for contexts appropriate to the category's features. This search would increase the reaction times for atypical artifactual exemplars. However, the difference between artifactual and naturally-occurring categories was the same, regardless of the exemplars being considered. This suggests that the difference between the two types of category is directly related to the features associated with the category label, not to those of the exemplar.

When a category is represented by intrinsic features, it usually also shares many extrinsic features with other members of the category. For example, if a category member has wings and feathers, it is very likely to fly. On the other hand, when a category is extrinsically represented, it is less likely to share intrinsic features with other members of the category. For example, if a category member is used to enhance house decor, it can take many shapes, sizes, etc. Therefore, when a subject tries to generate the features of an intrinsically represented category, either the intrinsic or extrinsic features will be useful in judging whether an item belongs to a category. However, when he tries to generate the features of an extrinsically represented category, any intrinsic features he may generate will not be useful. He must instead search for extrinsic features. Because this process will take time, and is independent of the typicality level of the exemplar involved, category membership judgments will be longer for artifactual than for naturally-occurring categories.

Although this explanation can easily account for our results, there is an alternative explanation. Some theories of category representation imply that features are only probabilistically associated with a category (e.g. Rosch, 1975). Membership is determined by some kind of weighted sum of these probabilistic features. Presumably, borderline instances have a lower overall sum than more typical instances. One might hypothesize that the features of artifactual categories are associated to the category with a lower probability than are those of naturally-occurring categories. This, in turn, would yield results similar to those we obtained. In the next experiment, therefore, we tested between these alternative explanations of our results.

Experiment 2

In this experiment we presented to subjects a list of features which were previously selected as defining of a category (Barr & Caplan, note 1). The subjects' task was to name the category described by the features.

If the probabilistic account of the results of Experiment 1 is correct, subjects should be less likely to name the artifactual categories described by the features than they are to name the natural categories. After all, according to the probabilistic account, the features of the artifactual categories we used should have a lower overall association to the category

label than the features of the natural categories that we used.

On the other hand, the intrinsic/extrinsic account explains the results of Experiment 1 by pointing to the kinds of features associated with the category, rather than to their relative probabilities. This account, therefore, predicts no difference in subjects' ability to name the artifactual and natural categories described by the features.

Method

Subjects. The subjects were 44 undergraduate volunteers at Ball State University. Some of the volunteers received course credit for participation.

Stimuli. The stimuli were sets of ten features written on pages of a small booklet, one feature to a page. The features were chosen from sets of features selected as defining by subjects in a previous study (Barr and Caplan, note 1; Experiment 1). The features used were the ten most popular features for each category in the earlier experiment. Seven artifactual categories (furniture, clothing, tools, weapons, vehicles, toys and sports) and seven naturally-occurring categories (fruits, birds, vegetables, trees, mammals, flowers and metals) were used. The order in which features were presented to subjects was determined by a Latin square.

Procedure. Subjects first completed the consent form. They were then told that they would receive 14 booklets, each containing 10 features. The features in each booklet described one particular category. They were instructed to go through the booklets, page by page. On each page they were to write down their idea of what was being described by the features. Subjects were encouraged to guess if they had no particular idea as to the category and to write down a response on each page before turning to the next page. They were permitted to look back over features on earlier pages, but they were not permitted to look forward beyond the current page. Finally they were encouraged to think of the broadest category (i.e., the highest level) which would fit the features. Each subject received the 14 booklets in a different random order.

Results

Because of space limitations, the results from only the four categories used in Experiment 1 will be reported. These were furniture, clothing, fruits and birds. In the results below we counted as a correct response either (1) the category label, or (2) the label of an exemplar of the category. For example, if a subject was working with the features of "furniture", his response would be scored as correct if he had responded with an exemplar such as "bed"

All 44 subjects correctly identified "birds" and 41 out of 44 subjects correctly identified "fruit". Among the artifactual categories, all 44 subjects correctly identified "clothing". Thirty-seven out of 44 correctly identified "furniture".

Subjects required a mean of 1.40 features before first identifying "bird", a mean of 3.24 features before first identifying the category "fruit", a mean of 4.43 features before identifying the category "furniture" and a mean of 1.67 features before identifying "clothing".

It was also possible to calculate which features proved to be most helpful to subjects in identifying the categories. In the summary below, we counted a feature as helpful if it allowed at least one-third of the subjects to identify the category correctly for the first time. The category "birds" had six such features (four were intrinsic and two were extrinsic). "Fruit"

had three such features (all three were intrinsic). "Clothing" had six such features (two were intrinsic and three were extrinsic). "Furniture" had just two such features (both were extrinsic).

Discussion

Clearly, nearly all subjects identified all four categories. There was little, if any, difference between the number of subjects who identified the naturally-occurring categories and the number who identified the artifactual categories. There were differences in the number of features needed before the subjects successfully identified the categories. But the differences did not appear related to the naturally-occurring versus artifactual distinction. Thus "birds" and "clothing" required relatively few features before they were identified, whereas "fruit" and "furniture" required more features.

On the other hand, there were dramatic differences in the kinds of feature which proved particularly helpful to subjects identifying the categories. The two artifactual categories were identified more with the aid of extrinsic features than of intrinsic features. The naturally-occurring categories, on the other hand, were identified more with the aid of intrinsic features.

The results, then support the "intrinsic/extrinsic" account of the results of Experiment 1 rather than the "probabilistic" account.

General Discussion

We have argued that "fuzzy" extensions are caused in part by the kind of features which form the intension of a category. Extrinsic features permit qualifiers to be applied when considering whether an exemplar is a member of a category (e.g., a pool table is OCCASIONALLY used inside houses). The extent to which qualifiers are necessary determines degree of membership in the category. The existence of degrees of membership, in turn, creates a fuzzy extension. Intrinsic features, on the other hand, do not permit qualifiers to be applied (e.g., an apple OFTEN contains vitamins?). Accordingly, when an intension is described entirely by intrinsic features, the extension is clear.

The results of the two experiments reported in this paper support our position. The two artifactual categories (furniture and clothing) are apparently represented largely by extrinsic features. The two naturally-occurring categories (fruit and birds) are represented largely by intrinsic features (see Experiment 2). The effect of typicality on membership judgments was very much more marked on the artifactual categories than on the naturally-occurring categories. (Experiment 1). Children also differed from adults more in their judgments of atypical artifactual items than in their judgments of atypical naturally-occurring items (Experiment 1). Presumably, the fuzzy extension created by extrinsic features is more difficult to learn than the clearer extension created by intrinsic features.

More traditional accounts of typicality effects (and other demonstrations of "fuzzy" natural categories) rely on the hypothesis that features are only probabilistically associated with a category to explain the demonstrated fuzziness. (e.g. Bourne, 1982; Rosch, 1975; also see Smith & Medin, 1981). Our account does not depend on this kind of uncertainty for its explanatory power. We believe that the features "covers people" and "is worn" are just as closely associated to the category "clothing" as the features "has feathers" and "has a beak" are to the category "bird". The results of Experiment 2 support this view, since subjects appeared almost as able to retrieve the artifactual

category names from a list of their features as they were to retrieve the naturally-occurring category names.

We suggest that a research focus on the properties of the features which describe a category's intension is more likely to elucidate categorization than earlier emphases on the uncertain nature of category boundaries.

Reference Note:

Barr, R. A., & Caplan, L. J. Some comparisons of the representations of two different classes of category Manuscript under review, Cognition.

References

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. Cognition, 13, 263-308.

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. Memory and Cognition, 10, 82-93.

Bourne, L. E. (1982). Typicality effects in logically defined categories. Memory & Cognition, 10, 3-9.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. Cognition, 9, 35-58.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), Cognitive development and the acquisition of language. New York: Academic Press.

Rosch, E. (1975). Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 104, 192-233.

Smith, E. E., & Medin, D. L. (1981). Categories and concepts. Cambridge, Mass.: Harvard University Press.

Table 1. Mean Response Times (sec) and Percentage of "No" Responses for "True" Sentences as a Function of Age, Category Type, and Typicality Level (all five typicality levels included; lower numbers indicate increased typicality).

ARTIFACTUAL CATEGORIES

		Typicality level				
		1	2	3	4	5
Adults	RT	3.118	3.127	3.219	3.267	3.318
	% "No"	0.0	8.9	9.5	34.5	66.4
Children	RT	4.264	4.438	4.792	4.579	4.773
	% "No"	3.5	7.8	21.5	32.1	57.9
Mean	RT	3.729	3.826	4.058	3.967	4.094
	% "No"	1.9	8.3	16.0	33.2	61.8

NATURALLY-OCCURRING CATEGORIES

		Typicality level				
		1	2	3	4	5
Adults	RT	2.958	3.011	3.148	3.256	3.207
	% "No"	1.8	0.9	5.9	24.4	17.5
Children	RT	4.133	4.116	4.544	4.564	4.451
	% "No"	4.5	4.5	8.3	21.0	19.4
Mean	RT	3.584	3.600	3.892	3.953	3.871
	% "No"	3.3	2.9	7.2	22.5	18.5