Memory & Cognition
2003, 31 (2), 199-214

Domain differences in the structure of artifactual and natural categories

ZACHARY ESTES
University of Georgia, Athens, Georgia

In three experiments, different methodologies, measures, and items were employed to address the question of whether, and to what extent, membership in a semantic category is all or none (i.e., absolute) or a matter of degree (i.e., graded). *Resemblance theory* claims that categorization is based on similarity, and because similarity is graded, category membership may also be graded. *Psychological essentialism* asserts that categorization is based on the presumption of the *category essence*. Because artifactual (e.g., FURNITURE) and natural (e.g., FRUIT) categories have different sorts of essences, artifacts and natural kinds may be categorized in qualitatively different manners. The results converged on the finding of a robust domain difference in category structure: Artifactual categories were more graded than natural categories. Furthermore, typicality reliably predicted absolute category membership, but failed to predict graded category membership. These results suggest that resemblance theory and psychological essentialism may provide a concerted account of representation and categorization across domains.

Concept representation is fundamental to cognition. Because concepts are the building blocks of cognition, the theory of concept representation that one endorses will largely determine one's theory of cognition (Fodor, 1998). The present investigation addresses the question of whether, and to what extent, semantic concepts such as FRUITS and FURNITURE have absolute or graded structure. In other words, is membership in a semantic category all or none (i.e., absolute), or is category membership a matter of degree (i.e., graded)? If category membership is absolute, then an object either is a member of a category or it is not, and all category members are equivalent in membership status. For example, a *tomato* is no less a full member of the category FRUIT than an *apple* is. If, on the other hand, category membership is graded, then an object may partially belong to the category, and category members may be nonequivalent (Mervis & Rosch, 1981; Rosch, 1978). For instance, a *rug* may be a partial member of the category FURNITURE, meaning that it is less of a member than, say, a *chair*.

This question of category structure has important implications for theories of concept representation, as is described below. Reported in what follows are three experiments in which different methodologies, measures, and items were employed in an attempt to provide convergent evidence of category structure. Specifically, the experiments tested for a difference between the structures of artifact categories (i.e., those occurring by human production or intention—e.g., FURNITURE) and those of natural categories (i.e., those occurring independently of human production or intention—e.g., FRUIT).[1] I begin by briefly reviewing two current theories of concept representation—namely, resemblance theory and psychological essentialism.

This research was supported in part by a Faculty Research Grant from the University of Georgia Research Foundation. I am indebted to Sam Glucksberg, James Hampton, and Phil Johnson-Laird for guidance during the development of this research, and I thank Lara Jones, Michelle Verges, and three anonymous reviewers for insightful comments on its exposition. Correspondence may be addressed to Z. Estes, Department of Psychology, University of Georgia, Athens, GA 30602 (e-mail: estes@uga.edu).

Resemblance Theory and Psychological Essentialism

According to resemblance theory, categorization entails a comparison between the to-be-categorized object and the representation of the target category (i.e., a prototype or a set of exemplars). If the object in question is judged to be sufficiently similar to the category representation, then it is included in that category. Importantly, this claim has implications for category structure: Categorization is based on similarity. Similarity, of course, is a matter of degree. Therefore, category membership may also be a matter of degree.

In support of this resemblance-based model, Rosch and others demonstrated that the facility with which people learn (Mervis & Pani, 1980; Rosch, 1973; Rosch, Simpson, & Miller, 1976), categorize (Hampton, 1979, 1997; Rips, Shoben, & Smith, 1973; Rosch, 1973), and remember (Rosch, 1975; Rosch et al., 1976) objects is determined by the similarity between a given object and the category prototype. That is, typical category members (e.g., *robin* for the category BIRDS) are learned more readily, categorized more quickly, and recalled more reliably than atypical members (e.g., *penguin*). Rosch and her colleagues concluded from these typicality effects

that some members of semantic categories have greater membership status than do other, less typical members (but, for alternative interpretations, see Armstrong, Gleitman, & Gleitman, 1983; Bourne, 1982; Landau, 1982). Furthermore, an overall high positive correlation between typicality and categorization is observed in most semantic categories (see, e.g., Diesendruck & Gelman, 1999; Hampton, 1998; McCloskey & Glucksberg, 1978; Rosch & Mervis, 1975). Thus, resemblance theory is generally accurate in predicting categorization decisions, and a host of current prototype models (e.g., Hampton, 1995) and exemplar models (e.g., Estes, 1994; Heit, 2001; Lamberts, 1995; Nosofsky, 1986) has emerged to replace the early models of resemblance theory (e.g., Medin & Schaffer, 1978; Rosch, 1975).

Medin (1989) argued, however, that "it is perhaps only a modest exaggeration to say that similarity gets at the shadow rather than the substance of concepts" (p. 1474; see also Quine, 1969). In their seminal article, Murphy and Medin (1985) suggested that categorization is inference to the best explanation; whatever theory offers the best explanation of an object or event will be inferred to categorize that object or event. Psychological essentialism is one such theory of explanation-based categorization. Specifically, when we categorize an object, we infer that the object possesses the *essence* of the category. The essence, in turn, is believed to constrain and generate the features of the object (Gelman & Wellman, 1991; Keil, 1989; Medin, 1989; Medin & Ortony, 1989; Putnam, 1975; Rips, 1989). For instance, when we see a horselike animal with black and white stripes, we infer that it has the essence of a ZEBRA, and we further infer that it looks and acts like a zebra precisely because it has this ZEBRA essence. In other words, the inference that the animal has the ZEBRA essence provides the best explanation of why it is horselike with black and white stripes.[2]

To be sure, several different types of essences have been posited (Gelman & Hirschfeld, 1999; see also Marcus, 1971; Teller, 1975), and many different versions of psychological essentialism have been proposed (see, e.g., Strevens, 2000). Some researchers suggest that category essences are an all-or-none matter (e.g., Diesendruck & Gelman, 1999; Kalish, 1995, 2002), and, therefore, essentialist categorization must be absolute. Others argue that essences may be partially possessed (e.g., Gelman & Hirschfeld, 1999, p. 409), and, hence, essentialist categorization may be graded. Some versions of essentialism claim that natural kinds have essences, whereas artifacts do not (e.g., Atran, 1998, p. 551; Diesendruck & Gelman, 1999; Diesendruck, Gelman, & Lebowitz, 1998; Keil, 1989; Schwartz, 1978). Other versions of essentialism claim that both artifacts and natural kinds have essences (e.g., Bloom, 1996, 1998; Putnam, 1975; and perhaps Medin, 1989, p. 1477). The important point for the present purposes is that even if artifacts do have essences, they nonetheless have a qualitatively different type of essence from natural kinds (e.g., Bloom, 1996, 1998), and, hence, may be categorized in a qualitatively different manner (see, e.g., Gelman & Hirschfeld, 1999, especially pp. 422 and 429). Thus, the experiments reported below were not intended to differentiate empirically between these various models of psychological essentialism. Rather, these models can all account for a domain difference in categorization via a domain difference in essences. Resemblance theory, in contrast, has no a priori way to account for a domain difference in categorization without a concomitant domain difference in typicality.

Evidence of Category Structure

Early research on category structure (e.g., Rosch, 1975) confounded typicality and category membership; in fact, typicality was regarded as a measure of membership. Thus, it is now unclear whether the early research has any bearing whatsoever on conclusions about category structure (e.g., Armstrong et al., 1983). Since this realization, few studies have been conducted to directly investigate whether semantic categories have absolute or graded structure. In the first of these, Barr and Caplan (1987) used a category-membership scale ranging from nonmembership to full membership. Their logic was that if membership is absolute, then membership values at one or the other scalar endpoint should be chosen, on the basis of the belief that the instance either fully is a member or fully is not a member. But if an intermediate (i.e., non-endpoint) value is chosen, then that value is assumed to reflect the degree to which the instance is a member of the category. As an index of category gradedness, Barr and Caplan reported *partial membership ratings*, which were the proportions of responses not falling at either endpoint of the membership scale. That is, on their scale of 1–7, any response from 2 to 6 was scored as indicating partial membership. Although Barr and Caplan did not test for differences in category structure across domains, they did find a reliable amount of gradedness across various semantic categories.

In the first direct test of domain differences in category structure, Kalish (1995) used a measure of category gradedness similar to that developed by Barr and Caplan (1987). In addition to artifact and natural categories, Kalish (1995) also included a graded category (e.g., *B+* as a GOOD EXAM SCORE) and an absolute category (e.g., a *$2 bill* as U.S. CURRENCY). As validation of the measure of category structure, he demonstrated that the graded category did indeed show more gradedness than the absolute category. However, Kalish (1995) found no difference in the gradedness of artifact and natural categories. More interestingly, he also found that neither the artifact categories nor the natural categories differed significantly from the graded category, though they did differ from the absolute category. He therefore concluded that categories in both domains were graded.

In a more recent investigation of semantic category structure, however, Diesendruck and Gelman (1999) found that artifact categories were more graded than natural ones. They also found that the correlation between

typicality and membership ratings was significantly higher for artifact items than for natural items, suggesting that membership in artifact categories was a function of typicality, whereas membership in natural categories was not (or, at least, was so to a lesser degree). Their findings clearly revealed a domain difference in people's categorization behavior, thus contradicting the results of Kalish (1995).

Kalish (2002) investigated people's *beliefs* about category structure—that is, whether people believe various categories to be graded or absolute. Using nine artifact items (e.g., *chair*) and eight natural items (e.g., *chicken*), Kalish (2002) asked participants several questions about classifying those items, such as whether the classification of the item should be resolved by scientific investigation or by legislation. Critically, a preference for scientific investigation is thought to indicate an essentialist belief, because it implies that the object has one true, discoverable essence. Kalish (2002) found that the preference for classification by scientific investigation was above chance for natural items, but not for artifacts. Conversely, the preference for classification by legislation was greater than chance for artifacts, but not for natural items. In fact, over all the measures used in the study, essentialist beliefs were reliably stronger for natural items than for artifactual items. Thus, Kalish (2002) found a domain difference in beliefs about category structure (see also Kalish, 1995; Malt, 1990). But he also found that, within domains, there was significant variability in beliefs about category structure. That is, the participants appeared to believe that category structure varies along a continuum from absolute to graded. The participants believed that some artifact categories (e.g., MRI SCANNER) are as absolute as some natural categories (e.g., SARDINE). Presumably, this variability in *beliefs* about category structure within domains should produce variability in actual categorization *behavior* as well. Thus, the issue of stimulus sampling will be important for any investigation of category structure; the use of different stimuli by different researchers may be responsible for their differential patterns of results.

On the whole, then, the recent research provides only sparse and inconclusive evidence of a domain difference in category structure. Below, I report three experiments that employed different methodologies, measures, and items in an attempt to provide convergent evidence of category structure. In Experiment 1, a strictly controlled set of items was used to investigate the relationship between typicality and category structure across domains. Experiment 2 was a methodological replication of Experiment 1, but with a novel set of items. Finally, in Experiment 3 a novel method for testing category structure was introduced. If resemblance theory explains category structure, then typicality should predict categorization in both domains. If essentialist beliefs explain category structure, then the aforementioned domain difference in essentialist beliefs should yield a domain difference in category structure.

EXPERIMENT 1

Previous investigations of category structure have employed a methodology in which items are rated on a category-membership scale (i.e., Barr & Caplan, 1987; Diesendruck & Gelman, 1999; Kalish, 1995). The rationale for this measure is that endpoint responses indicate absolute judgments. Responses between these endpoints, conversely, are assumed to indicate degrees of membership in the category, and therefore provide a measure of the gradedness of the category. This scalar method, however, may produce an inflated measure of gradedness: By random chance, 71% of responses on the standard 7-point membership scale would count as "graded." In order to reduce this bias toward "graded" responses, in the present experiment a three-alternative forced-choice methodology was used. The response options were "nonmember," "partial member," and "full member." By this methodology, the experimental demand for "graded" responses is only 33% (i.e., the partial-membership choice). It was explained to the participants that choosing partial membership meant that the item belonged in the target category not fully, but only to a degree. Thus, graded membership was measured explicitly.

Three types of items were included in the experiment. *Definite member* items (e.g., *apple*) were clearly members of the category (e.g., FRUIT), and *definite nonmember* items (e.g., *spinach*) were clearly not members of the category. *Borderline* items (e.g., *tomato*) were neither clearly in nor clearly out of the category (i.e., they were on the category borderline). The borderline items were critical in this study. Because it was unclear whether they belonged in the category, they were the most likely candidates to receive "graded" membership responses.

Items from artifactual and natural categories were tested. One graded category (i.e., GOOD EXAM SCORE) and one absolute category (i.e., U.S. CURRENCY) were also included for purposes of comparison. The graded and absolute categories were selected because GOOD EXAM SCORE is a typical example of a category that allows degrees of membership, whereas U.S. CURRENCY is typical of categories that have all-or-none membership (Kalish, 1995). If artifact categories are in fact graded, then they should resemble the graded comparison category, and if natural categories are in fact absolute, then they should resemble the absolute comparison category.

Resemblance theory claims that typicality ratings will predict category-membership judgments. In Experiment 1, two implications of this claim were tested. First, items that are equivalent in typicality should have equal membership status in their respective categories. In the present experiment, artifact and natural items were matched for mean typicality ratings. Therefore, resemblance theory predicts no difference between the artifactual and natural categories; any domain difference in category structure observed in the present experiment would not be predicted by typicality. Second, typicality should be positively correlated with category member-

ship. To test this prediction, the category membership of each item was plotted as a function of its typicality. If this prediction were to be supported, then items from both domains should show significant (i.e., nonzero) positive slopes, such that increases in typicality are correlated with increases in category membership (see, e.g., Hampton, 1998; McCloskey & Glucksberg, 1978). Alternatively, if a difference in slopes between domains were to be obtained, then psychological essentialism would receive support (Diesendruck & Gelman, 1999).

Method

Participants. Twenty undergraduates at the University of Georgia participated for partial course credit.

Materials and Procedure. Ten categories were used: 4 artifactual categories (FURNITURE, TOOLS, VEHICLES, WEAPONS), 4 natural categories (BIRDS, FRUITS, TREES, VEGETABLES), 1 graded category (GOOD EXAM SCORE), and 1 absolute category (U.S. CURRENCY). Each category consisted of 15 items: 5 definite members, 5 borderline items, and 5 definite nonmembers. Items in the graded and absolute comparison categories were based on those used by Kalish (1995). Artifact and natural items were taken from Barr and Caplan (1987). Each definite nonmember had a mean membership rating between 1.00 and 3.00 in Barr and Caplan's norms, in which the scale ranged from 1 (*clear nonmember*) to 7 (*clear member*). The mean membership ratings of borderline items were between 3.01 and 5.00 in Barr and Caplan, and those of definite members were between 5.01 and 7.00 in the same study.

Items in the artifact and natural categories were matched for typicality. Mean typicality ratings from Barr and Caplan (1987) were entered into a 2 (domain: artifact, natural) × 3 (item type: member, borderline, nonmember) analysis of variance (ANOVA). The main effect of item type was significant [$F(2,114) = 1{,}110.03$, $MS_e = 0.21$, $p < .001$], whereas neither the main effect of domain nor the domain × item type interaction was reliable (both ps > .40). Thus, the three item types differed in typicality. Importantly, though, items in the artifact and natural categories did not differ in typicality. This is important, because resemblance theory claims that categorization is a function of typicality. So, if the present experiment were to reveal a domain difference in category gradedness, resemblance theory would have difficulty explaining the result.

In fact, due to the importance of this control, and in light of the possibility of differences in populations (i.e., Ball State University students in the mid-1980s vs. University of Georgia students nearly 2 decades later), 15 undergraduates at the University of Georgia participated in a replication of Barr and Caplan's (1987) typicality-rating procedure. Only the items described above were included in the replication. The procedure was modeled after that of Barr and Caplan, with the participants rating the typicality of each item on a scale of 1 (*very poor example*) to 7 (*very good example*). Mean typicality ratings were submitted to one set of analyses in which the participants were treated as a random variable (F_p and t_p) and to another set of analyses in which items were treated as random (F_i and t_i). The present norming study precisely replicated the pattern of results found by Barr and Caplan. The main effect of item type was significant [$F_p(2,28) = 410.57$, $MS_e = 0.49$, $p < .001$ and $F_i(2,114) = 549.70$, $MS_e = 0.48$, $p < .001$]. Mean typicality ratings of member items ($M = 6.57$, $SE = 0.16$) were higher than those of borderline items ($M = 3.24$, $SE = 0.24$), which, in turn, were higher than those of nonmember items ($M = 1.50$, $SE = 0.12$). Most importantly, the typicality ratings of borderline items in the artifact ($M = 3.23$, $SE = 0.27$) and natural ($M = 3.24$, $SE = 0.25$) categories were virtually identical [$t_p(14) = 0.08$, $p = .94$ and $t_i(38) = 0.02$, $p = .98$]. Again, neither the main effect of domain (both ps > .95) nor the item type by domain interaction (both ps > .35) was significant in either analysis. Therefore, any domain differences in category gradedness obtained in Experiment 1 would not be attributable to differences in typicality across domains.

In Appendix A, a complete list of the stimuli used in the experiment is provided. Item order was randomized within the experimental list, but was constant across participants. For each of the 150 items, the participants were instructed to select the response option corresponding to that item's membership in the given category. The three options, which were included directly below each *item*–CATEGORY pair, were labeled "nonmember," "partial member," and "full member." The instructions included a description of the meaning of partial membership. The relevant part of the instructions, in which *billiards* was used as an example, reads as follows:

> If you believe that billiards is not a sport, then you should check the "nonmember" box. Or if you think that billiards is only somewhat a member of the category, then you should check the "partial member" box. But if you believe that it's just as much a member of the category as any other sport, then you should indicate that it's completely a member by checking the "full member" box . . . . Partial membership means that the item does belong in the category, but not to the same extent as some other items.

Results and Discussion

The dependent measure of interest in this study was gradedness, which was defined as the proportion of "partial member" responses. The proportion of such "graded" responses for each item is presented in Appendix A. The critical items were the borderline items. If "graded" responses were to occur, they should occur with the borderline items, because those items were judged as neither clearly in nor clearly out of the category in Barr and Caplan's (1987) norms. One set of analyses tested for domain differences in category gradedness, whereas another set of analyses examined more closely the relationship between typicality and category membership. These analyses are reported separately below.

Category gradedness. In Figure 1, the mean proportions of "partial member" responses to borderline items in the graded, artifact, natural, and absolute categories are displayed. The graded (i.e., GOOD EXAM SCORE) and absolute (i.e., U.S. CURRENCY) categories were included in the experiment for two purposes. First, they allowed a validity check of the gradedness measure. If the forced-choice methodology used in the present experiment provided a valid measure of gradedness, then the borderline items in the graded category should elicit a high proportion of "graded" responses, whereas the borderline items in the absolute category should elicit a low proportion of "graded" responses. This result was in fact obtained (see Figure 1): The mean gradedness of the borderline items in the graded category was near ceiling ($M = .87$, $SE = 0.07$), whereas the mean gradedness of the borderline items in the absolute category was near floor ($M = .11$, $SE = 0.04$). The measure of gradedness was thus validated.

The second purpose of the graded and absolute categories was the qualitative comparison with the artifact and natural categories. If artifact categories are in fact graded, then they should resemble the graded compari-

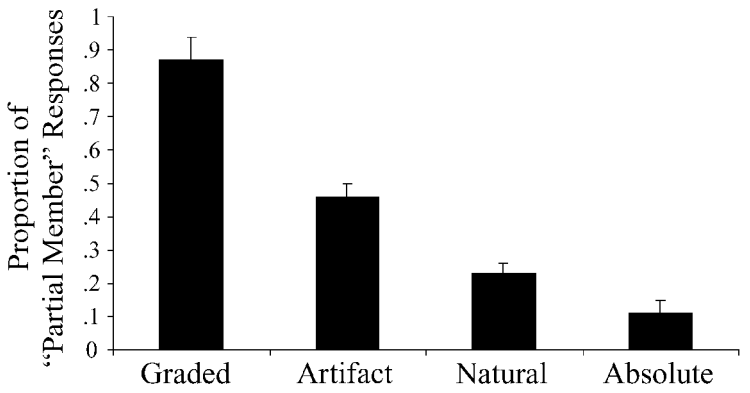

Figure 1. Proportions of "partial member" responses to borderline items of graded, artifact, natural, and absolute categories in Experiment 1. Error bars represent 1 standard error of the mean.

son category. If natural categories are in fact absolute, then they should resemble the absolute comparison category. However, because there was only one graded and one absolute comparison category (each with five arbitrarily selected borderline items), whereas there were four artifact and four natural categories (each with five systematically controlled borderline items), the graded and absolute comparison categories were not included in the quantitative analyses.

A one-way ANOVA indicated that there were no reliable differences in the gradedness of the borderline items across the four artifact categories. A separate ANOVA also found no reliable differences in the borderline items across the four natural categories. These results indicate that the various categories within each domain were uniform in their gradedness—that is, none of the natural categories was more or less graded than any other. Nor were any of the artifact categories reliably more or less graded than any of the other artifact categories. Therefore, all further analyses were collapsed across categories within domains (i.e., artifact, natural). The data of one participant, who did not follow instructions, were excluded from all analyses. In one set of analyses, participants was treated as a random variable, whereas in another items was treated as random. The participant analyses (t_p) were paired samples, whereas the item analyses (t_i) were independent samples.

Critically, borderline items of artifact categories (M = .46, SE = 0.04) were twice as likely to receive "graded" responses as were borderline items of natural categories (M = .23, SE = 0.03). This result clearly demonstrates a reliable domain difference in category structure [$t_p(18) = 5.00$, $p < .001$ and $t_i(38) = 5.26$, $p < .001$]. As is evident in Figure 1, the artifact categories were more graded than the natural categories. However, it is also important to note that the artifact categories were not as graded as the graded comparison category, nor were the natural categories as absolute as the absolute comparison category. Therefore, although it is true that the artifact categories were more graded than the natural categories, the natural categories were not absolute; they were in fact graded, though to a lesser degree than the artifact categories.

Turning, finally, to the definite-member and definite-nonmember items, recall that these items were selected specifically because the participants in Barr and Caplan's (1987) study overwhelmingly agreed that the member items belonged in the category and that the nonmember items did not. Thus, few "graded" responses were expected to occur with these items. Instead, the member items should have received a high proportion of "full member" responses, and the nonmember items should have received a high proportion of "nonmember" responses. Indeed, the proportion of "full member" responses to the member items was at ceiling and did not differ between artifact and natural categories (M = .98, SE = 0.01 in both cases). However, the proportion of "nonmember" responses to the nonmember items differed between artifact (M = .78, SE = 0.03) and natural (M = .88, SE = 0.03) categories. This difference was significant in the participant analysis [$t_p(18) = 2.38$, $p = .03$], but not in the item analysis [$t_i(38) = 1.56$, $p = .13$]. In other words, the participants were somewhat reluctant to exclude the nonmember items from the category absolutely, particularly when the category was artifactual. Rather, the proportion of "partial member" responses was higher for artifact nonmembers (M = .19, SE = 0.03) than for natural nonmembers [M = .10, SE = 0.02; $t_p(18) = 2.63$, $p = .02$ and $t_i(38) = 1.71$, $p = .10$]. This hint of a domain difference with the nonmember items, though unanticipated, is consistent with the result of the borderline items.

The domain difference in category structure found here is also consistent with the recent results of Diesendruck and Gelman (1999), but is inconsistent with the data reported by Kalish (1995). One potential explanation of why the present results differ from those obtained by Kalish (1995) is that his sample of items may not have been representative of the population as a whole (Diesendruck & Gelman, 1999). Kalish (1995) used 20 items across five

categories, and the items were not controlled for typicality. In the present experiment, on the other hand, 40 borderline items were used across eight semantic categories, and the items were systematically selected and matched for typicality ratings across domains.

Typicality and category membership. For a closer examination of the relationship between typicality and category membership, mean proportions of category-membership judgments (obtained in the experiment proper) were plotted as a function of mean typicality ratings (obtained in pretesting), and linear regressions were used to test whether the slopes differed from zero. Only the borderline items in the artifact and natural categories were analyzed.[3] In the present experiment, "full member" and "partial member" responses both indicated membership in the target category. Thus, the proportions of "full member" responses were plotted as a function of mean typicality ratings in one analysis (β_{full}), and those of "partial member" responses were plotted in a separate analysis ($\beta_{partial}$). In both analyses, slopes were calculated independently for artifact items and natural items, and those slopes were statistically compared with one another.

Typicality ratings reliably predicted "full member" responses in both domains (see Figure 2). Linear regressions revealed highly significant positive slopes for the artifact [$\beta_{full} = .89$, $t(18) = 8.20$, $p < .001$] and natural [$\beta_{full} = .79$, $t(18) = 5.46$, $p < .001$] items. This outcome replicates the finding of Diesendruck and Gelman (1999), who also reported a significant relationship between typicality ratings and proportions of absolute category judgments in both domains. As is evident in Figure 2, the present relationship between typicality and category membership was strictly linear, and the difference between the slopes of the artifact and natural items did not approach significance [$t(36) = 0.40$, $p = .69$]. Thus, typicality predicted "full member" responses about equally well across domains.

Typicality ratings did not predict "partial member" judgments in either domain. Linear regressions revealed no significant relationship between typicality and partial membership for artifact items ($\beta_{partial} = .16$, n.s.) or for natural items ($\beta_{partial} = .09$, n.s.). There was also no significant difference between the corresponding slopes ($p = .46$). Typicality was a poor predictor of partial membership in both domains.[4]

Summary. The analyses above yielded an interesting and informative pattern of results. Typicality reliably predicted full category membership in both artifactual and natural domains. This finding is unequivocally supportive of resemblance theory, which predicts a positive correlation between typicality and category membership. However, typicality did not predict partial category membership in either domain. Resemblance theory fails to explain this lack of relationship between typicality and partial membership judgments. One possible explanation is that partial category membership is attributable not to the typicality of the item, but to the participants' essentialist beliefs about the category. If the participants did in fact hold essentialist beliefs about the target do-

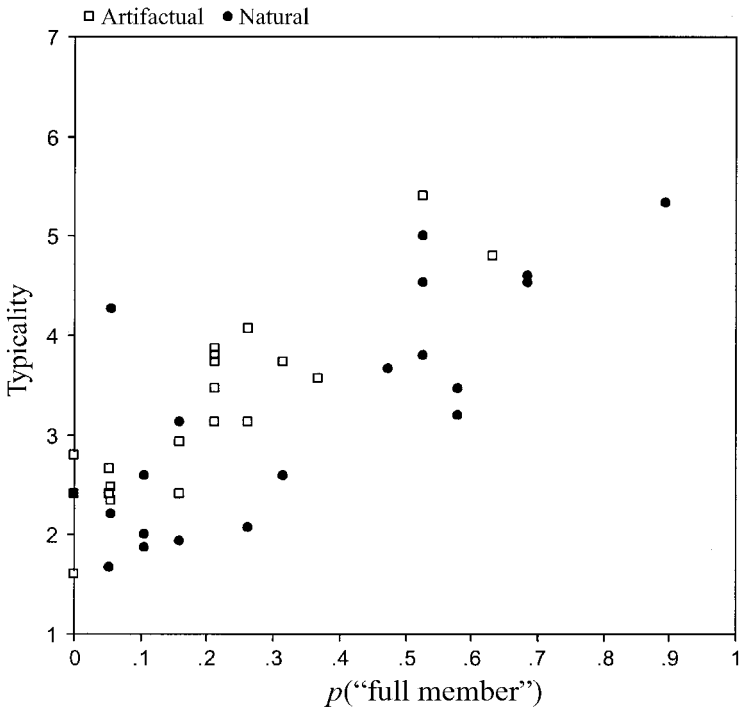

Figure 2. Proportions of "full member" responses to borderline items as a function of mean typicality ratings for artifactual and natural items in Experiment 1. Black squares indicate a black circle eclipsing a white square.

mains, then, according to previous research (e.g., Diesendruck & Gelman, 1999; Keil, 1989; Medin & Ortony, 1989; Rips, 1989), one might expect a systematic difference in the categorization of items in those domains. Indeed, artifact categories exhibited more graded membership responses than did natural categories. In summary, then, evidence for both resemblance theory and psychological essentialism was obtained in Experiment 1. Resemblance theory naturally accounts for the relationship between typicality and absolute category membership in both domains, but psychological essentialism may be necessary to explain category structure, particularly the domain difference in partial category membership.

EXPERIMENT 2

In Experiment 1, a domain difference in category structure was revealed. That experiment included 40 borderline items across eight categories (i.e., four artifactual and four natural). Elsewhere, however, Kalish (2002) has shown that some natural categories (e.g., SARDINE) are believed to be just as graded as some artifactual categories (e.g., MRI SCANNER). Given this variability in *beliefs* about category structure, one might expect actual categorization *behavior* to exhibit great variability as well. In light of this presumed variability in category gradedness, then, the generality of the results of Experiment 1 may be suspect.

Experiment 2, designed to test the generality of the results obtained in Experiment 1, was an exact procedural replication of the latter, but with a new set of items and categories. For a relatively conservative yet general test of the hypothesized domain difference in category structure, the items in Experiment 2 were selected from three independent sources of published category membership ratings. Those stimulus sources were McCloskey and Glucksberg (1978), Barr and Caplan (1987), and Kalish (1995). Sampling from the former two sources provided the benefit of using different operational definitions of a borderline item. Barr and Caplan used graded membership judgments, as was described above. McCloskey and Glucksberg used the proportion of nonmodal responses—that is, the amount of disagreement among participants about whether a given item is a member of the target category. McCloskey and Glucksberg (1978) and Barr and Caplan (1987) did not test for domain differences, whereas Kalish (1995) found no reliable domain difference in category gradedness. Therefore, if there was any bias at all, the items used in Experiment 2 were biased *against* a domain difference in category structure. If the results of Experiment 1 are replicated with these new items, then general statements about domain differences in category gradedness may be made with greater confidence.

Method

Participants. Fifty-eight University of Georgia undergraduates participated for partial course credit.

Materials and Procedure. The procedure of Experiment 2 was identical to that of Experiment 1. The materials consisted entirely of borderline items selected from McCloskey and Glucksberg (1978), Barr and Caplan (1987), and Kalish (1995). All of the items are presented in Appendix B, along with the source from which they were sampled. Kalish (1995) used 10 artifact items and 10 natural items. The present experiment included all 20 of these. The borderline items sampled from Barr and Caplan were selected according to the criteria described above in Experiment 1 (i.e., a mean membership rating of 3.01–5.00 in their study). Barr and Caplan's study included two artifact categories (i.e., CLOTHING, TOYS) and two natural categories (i.e., FLOWERS, MAMMALS) that were not used in Experiment 1 above. Therefore, items from these four categories were included in the present experiment. Borderline items sampled from McCloskey and Glucksberg were selected according to the criterion that they have a mean proportion of nonmodal responses of .30–.50 in the original study. That is, between 30% and 50% of the participants in that study disagreed with the modal categorization judgment for that item. Because no more than 50% of responses to any item can be nonmodal, the selected 30%–50% range represents the items for which there was the most disagreement. Two artifact categories (i.e., KITCHEN UTENSILS, SHIPS) and three natural categories (i.e., ANIMALS, FISH, INSECTS) were sampled from that study. All items from McCloskey and Glucksberg (1978) and Barr and Caplan (1987) that met these sampling criteria were pooled together. Then a few items were excluded, to equalize the number of artifactual items and natural items. This sampling procedure yielded 29 artifactual and 29 natural items from the two original studies. Thus, from the three stimulus sources, there was a total of 78 items (39 artifacts and 39 natural items). Item order was randomized, and the participants received the same instructions as did the participants in Experiment 1.

Results and Discussion

The full results of Experiment 2 are presented in Appendix B. As Kalish (2002) has suggested, there was substantial variability in category gradedness within domains. Appendix B shows clearly that certain individual artifact items were less graded than some individual natural items. Nonetheless, the *mean* gradedness was reliably greater for the artifactual than for the natural categories. That is, the artifactual categories ($M = .45$, $SE = 0.02$) were reliably more likely than the natural categories ($M = .35$, $SE = 0.02$) to receive partial membership judgments [$t_p(57) = 3.86$, $p < .001$ and $t_i(76) = 2.72$, $p < .01$]. Thus, Experiment 2, using a new set of 78 borderline items from six artifactual and seven natural categories, replicated the domain difference obtained in Experiment 1. Hence, the domain difference in category gradedness appears to be a fairly robust and general phenomenon, despite within-domain variability in category gradedness.

EXPERIMENT 3

Experiments 1 and 2 demonstrated a robust domain difference in the gradedness of artifactual and natural categories. The purpose of Experiment 3 was to provide convergent evidence of this domain difference by introducing a novel paradigm for testing category structure. All previous investigations of category structure, including Experiments 1 and 2 above, have measured gradedness via a paradigm of individual presentation—that is, participants were presented a single item (e.g., *tomato*), and their task was to judge its membership in the category (e.g., FRUIT). Another possible paradigm for

testing category structure is the simultaneous presentation of two or more items. For instance, *tomato* might be presented alongside *banana* for the target category FRUIT. The logic of this paradigm is that if both items are judged to belong in the category, *but to different degrees*, then that category must have graded structure.

In the present experiment, the more carefully controlled items of Experiment 1 were used. Borderline items were presented simultaneously with definite members of the target category in a four-alternative forced-choice paradigm. The response options were "neither is a member," "one is a member, one is not a member," "unequal members," and "equal members." The proportion of "unequal members" responses served as the measure of gradedness. If the simultaneous presentation paradigm is to corroborate the results of the individual presentation paradigm of Experiments 1 and 2, then artifact categories should elicit more "unequal members" responses than should the natural categories.

Method

Participants. Forty Princeton University undergraduates participated for course credit.

Materials and Procedure. The materials of Experiment 1 were adapted for the present purposes. Each borderline item was assigned one member item and one nonmember item from its target category. For instance, the borderline item *fork* of the category WEAPONS was assigned the member item *handgun* and the nonmember item *hairspray*, whereas the borderline item *tomato* was assigned the member item *banana* and the nonmember item *onion* from the FRUITS category. This was done for every borderline item in each of the 10 categories, creating 50 triads.

Each of the 50 borderline items was included in all of the experimental lists. The lists were counterbalanced so that, in any given list, half of the borderline items in each category were presented with their assigned member items, whereas the other half were presented with their assigned nonmember items. Each borderline item was presented with its member item in one list and with its nonmember item in another list. The order in which the two items in any pair were presented (e.g., *tomato and onion* vs. *onion and tomato*) was also counterbalanced across lists. Hence, 4 lists were generated: 2 (item pair: borderline with member, borderline with nonmember) $\times$ 2 (item order: borderline before member/nonmember, borderline after member/nonmember).

Additionally, for each category, 1 pair consisting of two definite member items (e.g., both *pear* and *pineapple* are FRUITS) and 1 pair consisting of two definite nonmember items (e.g., neither a *carrot* nor an *onion* is a FRUIT) were included in the experiment. This was done in order to give the participants ample opportunity to respond "equal members" and "neither is a member." These 20 filler pairs were included in all the lists. Thus, each list consisted of 70 item pairs. No item appeared in any list more than once. Item order was random within lists but fixed across participants.

For the two items of each item pair, the participants were instructed to check one of four boxes labeled "equal members," "unequal members," "one is a member, one is a nonmember," and "neither is a member" of the target category. The target category label was presented with each item pair. It was clearly explained that "unequal members" meant that both items belonged in the category, but not to the same degree.

Results and Discussion

Gradedness was defined as the proportion of "unequal members" responses elicited by each item pair. "Unequal members" responses were expected to occur when a borderline item was presented simultaneously with a definite member of the category. These proportions are presented in Appendix A. An initial one-way ANOVA indicated that the mean gradedness of the borderline items, when presented with a definite member of the category, did not differ across the four artifact categories. A separate ANOVA found that the gradedness of the borderline items did not differ across the four natural categories either. Thus, as in Experiment 1, the various categories within each domain were again uniform in their gradedness. All further analyses were therefore collapsed across categories within domains.

Figure 3 shows the mean proportions of "unequal members" responses to borderline items when they were presented with member items. Borderline items in the graded comparison category (M = .62, SE = 0.07) were most likely to be judged as unequal members of the category. The probability of unequal membership judgments of borderline items in the absolute comparison category (M = .18, SE = 0.06) was unexpectedly high. Some participants judged the different monetary values as unequal members of the category. For instance, when *$2 bill* (a borderline item) was presented with *quarter ($.25) coin* (a member item), some participants judged that they were unequal members of the category U.S. CURRENCY. Both items in this example are in fact members of the category, but one is of greater monetary value than the other. In retrospect, it is obvious that this task would be confusing with such items, and it is more than likely that the relatively high proportion of "unequal members" responses in the absolute comparison category was due to this confusion.

The more critical comparison, however, was that between artifact and natural categories. Borderline items of artifact categories were reliably more likely to be judged as unequal members (M = .33, SE = 0.04) than were borderline items of natural categories [M = .07, SE = 0.01; $t_p(39) = 7.81$, $p < .001$ and $t_i(38) = 5.13$, $p < .001$]. This indicates a domain difference in category gradedness and replicates the finding of Experiments 1 and 2 with a novel paradigm. Note also that in this paradigm the categories exhibited less gradedness than they did in Experiments 1 and 2. In fact, the gradedness of the natural categories was essentially at floor. Again, the artifact categories appeared to be less graded than the graded comparison category. This point will be elaborated upon in the General Discussion.

The borderline items were presented with member items as well as with nonmember items. For instance, the borderline item *fork* was presented alongside the nonmember item *hairspray* for the category WEAPONS, and *tomato* was presented with *onion* for the category FRUITS. Borderline nonmember stimulus pairs of artifact categories were more likely to receive "unequal members" responses (M = .08, SE = 0.02) than those of natural categories [M = .01, SE = 0.01; $t_p(39) = 4.15$, $p < .001$ and $t_i(38) = 2.50$, $p = .02$]. This result provides yet another instance of the domain difference in category structure.

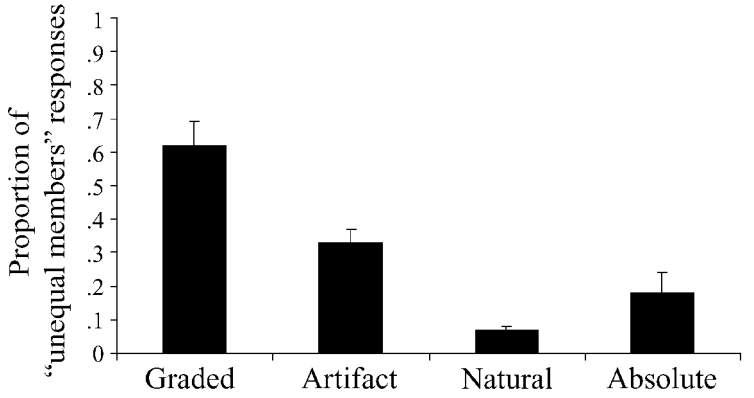

Figure 3. Proportions of "unequal members" responses to borderline items (when presented with definite member items) of graded, artifact, natural, and absolute categories in Experiment 3. Error bars represent 1 standard error of the mean.

GENERAL DISCUSSION

Evidence from a traditional test of category structure (Experiment 1), a traditional test with novel items (Experiment 2), and a novel test of category structure (Experiment 3) converged on the finding of a robust domain difference: Artifact categories were more graded than natural categories. In Experiment 1, artifact categories were twice as likely as natural categories to elicit "partial member" responses. For instance, a *rug* was more likely to be judged a graded member of the category FURNITURE than a *tomato* was to be judged a graded member of the category FRUIT. In addition, Experiment 1 showed that typicality reliably predicted full category membership, but failed to predict partial category membership. Experiment 2 extended the domain difference in category structure to a novel set of items and categories, thereby demonstrating the generality of the phenomenon. In Experiment 3, a novel paradigm for testing category structure was introduced. With the simultaneous presentation of two category members, a judgment that they differ in their degree of membership is prima facie evidence of gradedness. As corroboration of the first two experiments, Experiment 3 demonstrated that artifact categories were more likely than natural categories to exhibit unequal membership. In summary, then, natural categories were reliably less likely than artifact categories to receive "partial member" and "unequal members" responses across three experiments with different methodologies and items.

Of course, it must be cautioned that the results from this sample of items may not necessarily generalize to all other categories, especially given the variability in category structure suggested by Kalish (2002). In the present experiments, cross-domain variability was manifest as a domain difference in category gradedness, and substantial within-domain variability was observed in Experiment 2. This pattern of results supports Kalish's (2002) suggestion of a continuum of category gradedness, with some categories having more or less graded structure than others. As Kalish (2002) emphasized, this variability in category structure implies that stimulus sampling is critical in studies of category structure.

Two points favor the generalizability of the present results. First, the careful control of the materials used in Experiments 1 and 3 should be emphasized. Because resemblance theory claims that category membership is a function of typicality, its advocates might claim that the domain differences in category structure could be attributable to concomitant domain differences in typicality. But this was not the case. Items were matched for typicality across domains, and typicality failed to predict "partial member" responses within domains. Second, the domain difference in category structure was obtained with two different sets of stimuli.

Although it is true that the artifact categories were more graded than the natural categories, the artifact categories were not as graded as the graded comparison category, nor were the natural categories absolute. Although the proportion of "graded" responses to the natural categories was near floor (i.e., .08) in Experiment 3, in Experiment 1 that proportion was .23 and in Experiment 2 it was .35. So it would not be correct to say that the natural categories were absolute. Rather, they were graded, though less so than the artifact categories. This result corroborates other studies of category structure in demonstrating a *relative* difference in the gradedness of artifactual and natural categories (Diesendruck & Gelman, 1999), but failing to consistently demonstrate *absolute* structure in either type of category (Diesendruck & Gelman, 1999; Kalish, 1995).

Although resemblance theory and psychological essentialism both account for some of the present data, neither theory alone can explain all of them. Resemblance theory, with its prediction of typicality-based categorization, naturally accounts for the gradedness exhibited by the categories in both domains. If category membership is a function of typicality and typicality is a matter of degree, then category membership may be a matter of

degree. In support of this prediction, typicality reliably predicted full membership judgments in both domains. However, resemblance theory failed to account for within-domain gradedness: Typicality did not predict "partial member" responses in either domain. Nor did resemblance theory account for cross-domain gradedness: Artifactual and natural items were matched for typicality, yet they differed in gradedness. Ironically, then, although resemblance theory specifically arose from a need to account for graded category structure (see, e.g., Rosch, 1975), it did account for absolute categorization but failed to account for graded categorization.

The relative difference in the gradedness of artifactual and natural categories may be explained by psychological essentialism. According to this theory, natural categories differ from artifactual categories either in that (1) natural categories have essentialist structure, but artifactual categories do not (Atran, 1998; Diesendruck & Gelman, 1999; Diesendruck et al., 1998; Keil, 1989; Schwartz, 1978) or in that (2) natural categories and artifactual categories have qualitatively different types of essences (Bloom, 1996, 1998; Gelman & Hirschfeld, 1999). In either case, this domain difference in beliefs about essences (see, e.g., Kalish, 1995, 2002; Malt, 1990) may explain the domain difference in category structure. This suggestion is supported by the lack of a relationship between typicality and partial category membership, for, if typicality doesn't explain graded membership, then what does? Essentialist beliefs present themselves as one viable explanation. Of course, the present experiments do not establish any causal relationship between essentialist beliefs and category structure. That remains for future investigation.

Psychological essentialism clearly accounts for the *relative* difference in category gradedness, but its explanation of the lack of *absolute* categorization in either domain is less clear. In particular, some versions of psychological essentialism claim that essences are all or none and cannot be partially possessed (see Diesendruck & Gelman, 1999; Kalish, 1995, 2002). Because essences cannot be partially possessed, category membership must be absolute. This clearly was not the case in the present experiments; categories in both domains exhibited significant gradedness. As a consequence, absolute versions of essentialism may be rejected. Other versions of essentialism, though, do allow gradedness. Gelman and Hirschfeld (1999), for instance, argued that both artifactual and natural categories can admit degrees of membership by possessing degrees of the category essence: "[S]ubjects may believe that a certain inner quality or process of inheritance is needed in order for an animal to be a horse, but that in the real world different instances possess that quality or participate in that process to various degrees . . . ." (p. 409). Therefore, at least nominally, some essentialist models can account for the gradedness exhibited by categories in both domains. Notice, however, that this account is essentially descriptive, offering no basis on which to predict whether a given category will have an absolute or a graded structure.

The final result that psychological essentialism must address is the high positive correlation between mean typicality ratings and the likelihood of full membership judgments. Essentialism provides no direct explanation of this finding. It does provide an indirect explanation, though. Recall that, according to the essentialist theory, category judgments entail inference of the category essence that best explains the object's observable features. In the case of natural kinds, that essence is thought to consist of deep, biological features (see, e.g., Barton & Komatsu, 1989; Gelman & Wellman, 1991; Hirschfeld & Gelman, 1994; Keil, 1989; Medin & Atran, 1999; Rips, 1989). In the case of artifact kinds, the essence is thought to be the object's intended function (see, e.g., Bloom, 1996, 1998; Johnson-Laird, 1983; Keil, 1989; Kemler Nelson, Frankenfield, Morris, & Blair, 2000; Rips, 1989). But, importantly, Gelman and Medin point out that "essences are typically not known, almost always unobservable, and may not exist. So, the essence itself cannot usually serve as the basis of how people categorize items" (1993, p. 163). This raises the following question: On what basis are category essences inferred? Advocates (e.g., Bloom, 1998; Gelman & Bloom, 2000; Medin, 1989; Medin & Ortony, 1989; Putnam, 1975) and opponents (e.g., Malt & Johnson, 1992, 1998) of the theory agree that similarity is ordinarily the basis of essential inferences. For instance, if an animal is horselike with black and white stripes, one may infer that it has the essence of a ZEBRA, not of a HORSE, because it resembles zebras, not horses. That is, because the category essence constrains and generates objects' observable features, and because all members of a category possess the same category essence, the members of that category will tend to resemble one another (Kemler Nelson et al., 2000). Consequently, resemblance is a useful (though fallible) heuristic for essentialist categorization (Medin, 1989).

The preceding explanation implies that, in many cases, essentialism will require similarity-based categorization. Indeed, several theorists have recently argued that the distinction between similarity-based (i.e., resemblance theory) and theory-based (e.g., psychological essentialism) categorization is a false dichotomy (e.g., Hahn & Ramscar, 2001; Heit, 2001; see also Hahn & Chater, 1998). Hahn and Ramscar express this position clearly:

> [S]imilarity is instrumental in explaining what categories to form in the first place and how it is that we subsequently assign items to these categories . . . . This is not to say that similarity is the *only* thing that plays a role here. . . . it seems more than likely that extant "theoretical" knowledge can refine originally similarity-based (categories), and that background knowledge can influence individual classification decisions. (p. 269)

In fact, the category essence itself may be conceived of as a specialized, essential sort of similarity—a view that is favored even by essentialist philosophers. For instance, John Locke argued that the basis of a category's essence is some microstructural similarity, or some microscopic property common to all members of that cate-

gory (Dupré, 1981), and Putnam (1983) suggested that an essence might best be characterized as a specialized resemblance (see also Quine, 1969). In a general sense, possession of the category essence is a similarity shared by all members of the category. For example, whatever the essence of a ZEBRA may be, all zebras are similar in that they have that essence.

Resemblance theory and psychological essentialism, then, may not be mutually exclusive. Rather, the two theories may provide a concerted account of representation and categorization across domains (see, e.g., Hahn & Ramscar, 2001; Heit, 2001). Essentialism may be viewed as a constraint on similarity-based categorization, in that essentialist beliefs and other theoretical knowledge influence which features are selected for comparison with the category representation and how those features are weighted in the similarity computation (cf. Heit, 1998; Lamberts, 1994).

REFERENCES

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

Atran, S. (1998). Folk biology and the anthropology of science. *Behavioral & Brain Sciences*, **21**, 547-611.

Barr, R. A., & Caplan, L. J. (1987). Category representations and their implications for category structure. *Memory & Cognition*, **15**, 397-418.

Barton, M. E., & Komatsu, L. K. (1989). Defining features of natural kinds and artifacts. *Journal of Psycholinguistic Research*, **18**, 433-447.

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, **60**, 1-29.

Bloom, P. (1998). Theories of artifact categorization. *Cognition*, **66**, 87-93.

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

Diesendruck, G., & Gelman, S. A. (1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. *Psychonomic Bulletin & Review*, **6**, 338-346.

Diesendruck, G., Gelman, S. A., & Lebowitz, K. (1998). Conceptual and linguistic biases in children's word learning. *Developmental Psychology*, **34**, 823-839.

Dupré, J. (1981). Natural kinds and biological taxa. *Philosophical Review*, **90**, 66-90.

Estes, W. K. (1994). *Classification and cognition.* New York: Oxford University Press.

Fodor, J. A. (1998). *Concepts.* New York: Oxford University Press.

Gelman, S. A., & Bloom, P. (2000). Young children are sensitive to how an object was created when deciding what to name it. *Cognition*, **76**, 91-103.

Gelman, S. A., & Hirschfeld, L. A. (1999). How biological is essentialism? In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 403-446). Cambridge, MA: MIT Press.

Gelman, S. A., & Medin, D. L. (1993). What's so essential about essentialism? A different perspective on the interaction of perception, language, and conceptual knowledge. *Cognitive Development*, **8**, 157-167.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition*, **38**, 213-244.

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition*, **65**, 197-203.

Hahn, U., & Ramscar, M. (2001). Mere similarity? In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 257-272). New York: Oxford University Press.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning & Verbal Behavior*, **18**, 441-461.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory & Language*, **34**, 686-708.

Hampton, J. A. (1997). Associative and similarity-based processes in categorization decisions. *Memory & Cognition*, **25**, 625-640.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition*, **65**, 137-165.

Heit, E. (1998). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 712-731.

Heit, E. (2001). Background knowledge and models of categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 155-178). New York: Oxford University Press.

Hirschfeld, L. A., & Gelman, S. A. (1994). *Mapping the mind: Domain specificity in cognition and culture.* Cambridge: Cambridge University Press.

Johnson-Laird, P. N. (1983). What is meaning? In P. N. Johnson-Laird, *Mental models* (pp. 182-204). Cambridge, MA: Harvard University Press.

Kalish, C. W. (1995). Essentialism and graded membership in animal and artifact categories. *Memory & Cognition*, **23**, 335-353.

Kalish, C. W. (2002). Essentialist to some degree: Beliefs about the structure of natural kind categories. *Memory & Cognition*, **30**, 340-352.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kemler Nelson, D. G., Frankenfield, A., Morris, C., & Blair, E. (2000). Young children's use of functional information to categorize artifacts: Three factors that matter. *Cognition*, **77**, 133-168.

Lamberts, K. (1994). Flexible tuning of similarity in exemplar-based categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1003-1021.

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

Landau, B. (1982). Will the real grandmother please stand up? The psychological reality of dual meaning representations. *Journal of Psycholinguistic Research*, **11**, 47-62.

Locke, J. (1989). *An essay concerning human understanding* (P. H. Nidditch, Ed.). Oxford: Oxford University Press, Clarendon Press. (Original work published 1690)

Malt, B. C. (1990). Features and beliefs in the mental representation of categories. *Journal of Memory & Language*, **29**, 289-315.

Malt, B. C., & Johnson, E. C. (1992). Do artifact concepts have cores? *Journal of Memory & Language*, **31**, 195-217.

Malt, B. C., & Johnson, E. C. (1998). Artifact category membership and the intentional-historical theory. *Cognition*, **66**, 79-85.

Marcus, R. B. (1971). Essential attribution. *Journal of Philosophy*, **68**, 187-202.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well-defined or fuzzy sets? *Memory & Cognition*, **6**, 462-472.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, **44**, 1469-1481.

Medin, D. L., & Atran, S. (1999). *Folkbiology.* Cambridge, MA: MIT Press.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179-195). Cambridge: Cambridge University Press.

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

Mervis, C. B., & Pani, J. R. (1980). Acquisition of basic object categories. *Cognitive Psychology*, **12**, 496-522.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, **32**, 89-115.

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

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

Putnam, H. (1975). The meaning of "meaning." In H. Putnam, *Mind, language, and reality: Philosophical papers* (Vol. 2, pp. 215-217). Cambridge: Cambridge University Press.

Putnam, H. (1983). Possibility and necessity. In H. Putnam, *Realism and reason: Philosophical papers* (Vol. 3, pp. 46-68). Cambridge: Cambridge University Press.

QUINE, W. V. (1969). Natural kinds. In W. V. Quine, *Ontological relativity and other essays* (pp. 114-138). New York: Columbia University Press.
RIPS, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). Cambridge: Cambridge University Press.
RIPS, L. J., SHOBEN, E. J., & SMITH, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning & Verbal Behavior*, **12**, 1-20.
ROSCH, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111-144). New York: Academic Press.
ROSCH, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, **104**, 192-232.
ROSCH, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale, NJ: Erlbaum.
ROSCH, E., & MERVIS, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.
ROSCH, E., SIMPSON, C., & MILLER, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception & Performance*, **2**, 491-502.
SCHWARTZ, S. P. (1978). Putnam on artifacts. *Philosophical Review*, **87**, 566-574.
STREVENS, M. (2000). The essentialist aspect of naive theories. *Cognition*, **74**, 149-175.
TELLER, P. (1975). Essential properties: Some problems and conjectures. *Journal of Philosophy*, **72**, 233-248.

NOTES

1. Over the last several decades, there has been considerable philosophical debate about exactly what an *artifact* is and what a *natural kind* is (see, e.g., Schwartz, 1978). I do not presume to settle this debate in the present research. I use the terms *artifactual* and *natural* in a general, inclusive sense, as is defined in the text. Although most previous theoretical claims have been made about natural kinds (e.g., BIRDS) specifically, note that my definition of *natural category* also includes some categories that may not technically be natural kinds (e.g., TREES). However, the present methodology would allow any difference between these two subtypes of natural categories to be evident.

2. Note that essentialist categorization does not necessarily require knowledge of what exactly the essence *is*. Not knowing the essence of a ZEBRA does not preclude the categorizer from inferring that there is some such essence, and that this essence causes its appearance. Note also that psychological essentialism is explicitly *not* a theory of real essences (Locke, 1690/1989). Psychological essentialism is a theory of the interaction between belief and behavior; it concerns how essentialist beliefs affect the categorization of and reasoning about objects.

3. When all three item types (i.e., definite members, borderline items, and definite nonmembers) were included in regression analyses, there was a highly significant relationship between mean typicality ratings and proportions of full-membership judgments. But this is unremarkable. We know from the item-selection process that the definite member items were highly typical of and highly likely to belong in the target category, and that the definite nonmember items were atypical of and unlikely to belong in the target category. More interesting are the analyses of the borderline items.

4. As an alternative hypothesis, resemblance theory might predict that partial membership judgments are an inverted U function of typicality. That is, for highly atypical items, nonmembership judgments are expected, and, therefore, few partial membership judgments should occur. For highly typical items, full membership judgments are expected, so again, few partial membership judgments should occur. However, for items of intermediate typicality, partial membership judgments are expected. Thus, the proportion of partial membership judgments should begin near floor with atypical items, should rise to a peak at intermediate levels of typicality, and should then fall back toward floor with high typicality, thereby creating an inverted U curve. To investigate this hypothesis, the artifactual and natural items were tested separately for fit to a quadratic function. Neither fit approached significance (both $ps > .30$).

APPENDIX A
Stimuli and Their Gradedness in Experiments 1 and 3

Category	Item Type	Item	Gradedness: Experiment 1	Gradedness: Experiment 3
		Artifactual		
GOOD EXAM SCORE				
	Definite nonmember	77%	.26	
		78%	.42	
		79%	.53	
		80%	.89	
		81%	.84	
	Borderline	82%	.89	.45
		83%	.89	.45
		84%	.89	.65
		85%	.79	.90
		86%	.89	.80
	Definite member	87%	.84	
		88%	.74	
		89%	.68	
		90%	.32	
		91%	.26	
FURNITURE				
	Definite nonmember	closet	.00	
		telephone	.32	
		bicycle	.00	
		ceiling	.00	
		sugar bowl	.16	

APPENDIX A (Continued)

Category	Item Type	Item	Gradedness	
			Experiment 1	Experiment 3
	Borderline	picnic table	.26	.05
		shelves	.37	.25
		refrigerator	.26	.35
		mantel	37	.30
		clock	.37	.50
	Definite member	sofa	.00	
		coffee table	.00	
		loveseat	.00	
		bed	.00	
		desk	.00	
TOOLS				
	Definite nonmember	calendar	.47	
		camera	.37	
		hamster	.00	
		umbrella	.53	
		Band-Aid	.44	
	Borderline	paint	.37	.35
		funnel	.53	.05
		computer	.53	.45
		gun	.61	.30
		shaver	.58	.05
	Definite member	drill	.00	
		hammer	.00	
		screwdriver	.00	
		socket wrench	.05	
		wire cutters	.05	
VEHICLES				
	Definite nonmember	husky	.21	
		lawnmower	.47	
		bus driver	.00	
		carton	.00	
		newspaper	.00	
	Borderline	gondola	.21	.20
		tricycle	.58	.25
		wheelchair	.63	.40
		horse	.50	.50
		roller skates	.58	.40
	Definite member	bus	.00	
		car	.00	
		truck	.00	
		van	.00	
		taxi	.05	
WEAPONS				
	Definite nonmember	hairspray	.32	
		pillow	.32	
		marshmallow	.00	
		paper bag	.11	
		rain	.05	
	Borderline	drugs	.47	.35
		chair	.26	.40
		gas	.42	.20
		fingernails	.68	.80
		fork	.63	.45
	Definite member	rifle	.05	
		hydrogen bomb	.00	
		handgun	.00	
		fire bomb	.00	
		torpedo	.11	
BIRDS				
	Definite nonmember	flying squirrel	.05	
		airplane	.05	
		helicopter	.05	
		housefly	.00	
		butterfly	.11	

APPENDIX A (Continued)

Category	Item Type	Item	Gradedness	
			Experiment 1	Experiment 3
	Borderline	bat	.11	.05
		thunderbird	.22	.05
		sandcrane	.21	.00
		duck-billed platypus	.11	.10
		Big Bird	.58	.60
	Definite member	canary	.00	
		cardinal	.00	
		blackbird	.00	
		bluebird	.00	
		mockingbird	.00	
FRUITS				
	Definite nonmember	carrot	.05	
		onion	.00	
		potato	.00	
		rose	.00	
		spinach	.00	
	Borderline	avocado	.16	.10
		coconut	.37	.05
		tomato	.05	.20
		cucumber	.21	.05
		rhubarb	.26	.00
	Definite member	apple	.00	
		pear	.00	
		plum	.00	
		banana	.00	
		pineapple	.05	
TREES				
	Definite nonmember	leaf	.42	
		twig	.42	
		daffodil	.11	
		grass	.05	
		corn plant	.32	
	Borderline	hemlock	.11	.00
		lilac	.26	.00
		sage	.32	.00
		sassafras	.21	.00
		juniper	.26	.00
	Definite member	maple	.00	
		oak	.00	
		pine	.00	
		cedar	.00	
		hickory	.00	
VEGETABLES				
	Definite nonmember	oatmeal	.11	
		wheat	.11	
		pine needle	.00	
		egg	.00	
		macaroni	.11	
	Borderline	pumpkin	.26	.05
		rice	.05	.05
		gourd	.11	.00
		hominy	.39	.15
		cloves	.37	.00
	Definite member	broccoli	.00	
		cauliflower	.00	
		green bean	.00	
		radish	.05	
		turnip	.05	
U.S. CURRENCY				
	Definite nonmember	$12 bill	.00	
		75-cent ($.75) coin	.11	
		chicken	.00	
		Monopoly money	.05	
		poker chips	.05	

APPENDIX A (Continued)

Category	Item Type	Item	Gradedness	
			Experiment 1	Experiment 3
	Borderline	$1,250 bill	.05	.15
		$2 bill	.32	.15
		$50 bill	.00	.25
		I.O.U.	.11	.20
		silver dollar coin	.05	.15
	Definite member	$1 bill	.00	
		$100 bill	.00	
		dime	.00	
		penny	.00	
		quarter	.00	

Note—Gradedness is the proportion of "partial member" responses in Experiment 1 and "unequal members" responses (when presented with a definite member) in Experiment 3.

APPENDIX B
Stimuli and Their Gradedness, Experiment 2

Item	Category	Source	Gradedness
	Artifactual		
sofa	FURNITURE	Kalish	.03
picnic table	FURNITURE	Kalish	.24
mop	KITCHEN UTENSILS	M&G	.29
broom	KITCHEN UTENSILS	M&G	.29
corduroy	CLOTHING	B&C	.31
dustpan	KITCHEN UTENSILS	M&G	.31
stove	KITCHEN UTENSILS	M&G	.31
music box	TOYS	B&C	.33
dishwasher	KITCHEN UTENSILS	M&G	.38
refrigerator	KITCHEN UTENSILS	M&G	.38
bat	TOYS	B&C	.41
wheelchair	FURNITURE	Kalish	.41
spacecraft	SHIPS	M&G	.41
garbage disposal	KITCHEN UTENSILS	M&G	.41
catamaran	SHIPS	M&G	.43
cards	TOYS	B&C	.45
backgammon	TOYS	B&C	.45
houseboat	SHIPS	M&G	.45
hovercraft	SHIPS	M&G	.45
piano	FURNITURE	Kalish	.47
car	WEAPONS	Kalish	.48
raft	SHIPS	M&G	.48
drum	TOYS	B&C	.50
satellite	WEAPONS	Kalish	.50
clock	FURNITURE	Kalish	.50
gondola	SHIPS	M&G	.50
racquet	TOYS	B&C	.52
pillow	FURNITURE	Kalish	.52
musical instrument	TOYS	B&C	.53
pocket	CLOTHING	B&C	.53
kayak	SHIPS	M&G	.54
string	TOYS	B&C	.55
headband	CLOTHING	B&C	.55
rowboat	SHIPS	M&G	.57
rubber band	WEAPONS	Kalish	.59
fork	WEAPONS	Kalish	.59
handkerchief	CLOTHING	B&C	.60
canoe	SHIPS	M&G	.60
guitar	TOYS	B&C	.62

(*Continued on next page*)

APPENDIX B (Continued)

Item	Category	Source	Gradedness
	Natural		
praying mantis	INSECTS	Kalish	.05
bluejay	MAMMALS	B&C	.07
goose	MAMMALS	B&C	.09
poet	ANIMALS	M&G	.12
hyacinth	FLOWERS	B&C	.12
spider	INSECTS	Kalish	.14
worm	INSECTS	M&G	.16
fungus	ANIMALS	M&G	.18
philodendron	FLOWERS	B&C	.19
scorpion	INSECTS	Kalish	.19
leech	INSECTS	M&G	.21
porpoise	FISH	M&G	.21
caterpillar	INSECTS	Kalish	.22
virus	ANIMALS	M&G	.22
yeast	ANIMALS	M&G	.24
bacterium	ANIMALS	M&G	.26
heather	FLOWERS	B&C	.29
jellyfish	FISH	M&G	.29
schefflera	FLOWERS	B&C	.33
sponge	FISH	M&G	.38
shrimp	FISH	M&G	.38
plankton	FISH	M&G	.40
fern	FLOWERS	B&C	.41
lobster	FISH	M&G	.43
starfish	FISH	M&G	.43
octopus	FISH	M&G	.43
squid	FISH	M&G	.45
sea anemone	FISH	M&G	.47
crab	FISH	M&G	.48
ivy	FLOWERS	B&C	.52
seahorse	FISH	M&G	.57
clam	FISH	M&G	.57
coyote	DOG	Kalish	.59
oyster	FISH	M&G	.59
zebra	HORSE	Kalish	.60
hyena	DOG	Kalish	.60
mule	HORSE	Kalish	.62
wolf	DOG	Kalish	.64
donkey	HORSE	Kalish	.66

Note—"B&C" indicates that the item was sampled from Barr and Caplan (1987), "Kalish," from Kalish (1995), and "M&G," from McCloskey and Glucksberg (1978). Gradedness was the proportion of "partial member" responses.

(Manuscript received March 26, 2002;
revision accepted for publication December 4, 2002.)