Confidence and gradedness in semantic categorization: Definitely somewhat artifactual, maybe absolutely natural

ZACHARY ESTES
University of Georgia, Athens, Georgia

Artifacts tend to be categorized in a graded (i.e., continuous) manner, whereas natural categorization tends to be absolute (i.e., discrete). This domain-specific categorization is assumed to reflect a domain difference in representation. However, another tenable but untested explanation is that graded categorization arises from uncertainty, which is greater in artifact categories than in natural categories. Confidence ratings were used as an index of certainty in two experiments that tested whether confidence in category judgments can explain the apparent gradedness of those categories. Both experiments revealed that artifact categories were more graded and were judged with greater confidence than were natural categories. Confidence and gradedness were negatively correlated within both domains. Thus, confidence did indeed predict gradedness *within* the artifact and natural domains but failed to predict the difference in gradedness *between* those domains. There is more to gradedness than just uncertainty.

Any given category may be either absolute (i.e., discrete) or graded (i.e., continuous). To illustrate, most people believe that the category of BIRDS is absolute, which is to say that all members are thought to belong in the category to the same extent. A *penguin* may be an atypical bird but is a BIRD nonetheless. In contrast, most people believe that the category of FURNITURE is graded, with different objects thought to belong in the category to different degrees. A *rug* can "sort of" be FURNITURE. As these examples illustrate, categorization tends to be graded for artifacts and absolute for natural kinds (Diesendruck & Gelman, 1999; Estes, 2003).[1]

This issue of category structure is important because it constrains models of categorization and representation. Specifically, graded category structure suggests a similarity-based model of categorization, with degrees of category membership explained by degrees of similarity. And conversely, absolute category structure implies an essentialist model of categorization, whereby every member fully possesses the category essence. Thus, the category representation is presumed to explain the occurrence or absence of graded category judgments (Hampton, 1998; Kalish, 1995; Rosch, 1975), with differential grad-

edness suggesting diverse representations (Diesendruck & Gelman, 1999; Estes, 2003).

However, an *uncertainty hypothesis* (Johnson-Laird, 1983; Rey, 1983) stands as another tenable but untested explanation of graded categorization. That is, uncertainty about a category judgment may evoke graded membership responses, regardless of whether the target category is absolute or graded. Intuitively, if one is uncertain whether a *tomato* is a FRUIT, one may respond that it is "sort of" a FRUIT, even though one believes that FRUIT is an absolute category. In other words, uncertainty may be confounded with gradedness. The domain difference in gradedness, then, may be attributable to a domain difference in uncertainty: If gradedness arises from uncertainty and if this uncertainty is greater in artifact categories than in natural categories, artifact categories will be more graded than natural categories. Crucially, if this explanation of gradedness were correct, it would undermine the previous research on category structure, as well as its implications for models of categorization and representation. With confidence ratings as an index of certainty, the experiments reported below tested whether confidence in category judgments could indeed explain the apparent gradedness of those categories.

Gradedness

The standard measure of category gradedness was developed by Barr and Caplan (1987), who elicited category judgments on a scale that ranged from nonmembership to full membership. As a measure of gradedness, Barr and Caplan reported *partial membership* scores, which were the proportions of responses not falling on either endpoint of the membership scale. That is, on their scale ranging

from 1 to 7, any response from 2 to 6 was scored as *partial membership*. Their logic was that if membership is absolute, membership values at one or the other scalar endpoint should be chosen, indicating a belief that the instance either fully is a member or fully is not a member. But if intermediate values are chosen, that value is assumed to reflect the degree to which the instance is a member of the category.

As a direct test of domain differences in category gradedness, Kalish (1995) presented 10 artifact items and 10 natural items for categorization, using the scalar method of Barr and Caplan (1987). Kalish (1995) found no difference in the gradedness of artifact and natural categories. Subsequently, however, Diesendruck and Gelman (1999) conducted a similar experiment, but with a larger and more controlled set of items. Contrary to Kalish (1995), Diesendruck and Gelman found that the artifact categories elicited more partial membership ratings than the natural categories did. This result clearly revealed a domain difference in category gradedness.

Estes (2003) measured gradedness in a more direct and explicit manner. In a three-alternative forced choice task, participants indicated whether each item was a nonmember, a partial member, or a full member of the target category. The participants were informed that partial membership meant that the item belonged in the category, but not to the same extent as some other items. With this more explicit measure, artifact categories were again more likely than natural categories to receive partial membership judgments. Another experiment introduced a novel paradigm for testing category gradedness: Category judgments were elicited for two simultaneously presented items. For instance, *tomato* was presented alongside *apple* for the target category FRUITS. In this paradigm, a judgment that both items belong in the target category, but to different degrees, is prima facie evidence of gradedness. In corroboration of the previous experiment, this novel method demonstrated that artifact categories were more likely than natural categories to exhibit graded membership. Whereas membership in a natural category tended to be an absolute matter, membership in an artifact category tended to be a matter of degree.

Confidence and Gradedness

It is crucial to distinguish uncertainty from gradedness in semantic categorization. Considering the truth value of such categorical propositions as *A lampshade is a piece of furniture*, Johnson-Laird (1983) advised, "The fact that the evaluation of a proposition may vary over a continuum from absolute conviction to absolute rejection creates an atmosphere conducive to thinking that truth values may similarly vary, but *degree of conviction should not be confused with truth value*" (p. 199, italics added). That is, Johnson-Laird emphasized the possible confounding of certainty and gradedness: What appears to be graded category structure might really be graded certainty about absolute structure. Indeed, Rey (1983) charged that measures of graded categorization are, in

fact, guilty of this confound. For instance, a partial membership response to *lampshade* for the target category FURNITURE could mean either (1) that a *lampshade* is a partial member of the FURNITURE category or (2) that the uncertainty of whether a *lampshade* is FURNITURE has been confounded with the membership judgment itself. Researchers have assumed the former possibility without empirically addressing the latter. If the latter possibility were correct, semantic categories would not really have graded structure after all.

Although neither Johnson-Laird (1983) nor Rey (1983) directly addressed the possible differences between artifacts and natural kinds in certainty and gradedness, their argument can be logically extended to yield the following hypothesis: The domain difference in category gradedness might be attributable to a domain difference in the certainty of category membership. To see how, suppose that artifact categories were judged with less certainty than natural categories were. This domain difference in certainty would, according to the argument, cause more graded category judgments for artifacts than for natural kinds.

If uncertainty causes graded category judgments, one would expect an inverse relation between confidence (i.e., certainty) and gradedness; lower confidence should produce higher gradedness. In terms of artifact and natural categorization, this uncertainty hypothesis has two implications: (1) Confidence and gradedness ought to be negatively correlated within artifact and natural categories, and (2) artifact categories, which exhibit more gradedness than do natural categories, should also elicit less confidence than do natural categories. So in summary, the uncertainty hypothesis predicts a negative relation of confidence and gradedness both within and between domains. Surprisingly, this prediction remains untested. Therefore, category membership judgments of artifact and natural objects, as well as confidence ratings for those judgments, were collected in two experiments.

EXPERIMENT 1

Items were systematically sampled from four artifact and four natural categories. Within each category, there were five *borderline* items that were neither clearly in nor clearly out of the target category. Although clear

Table 1
Borderline Items, Experiment 1

Artifact
Furniture: clock, mantel, picnic table, refrigerator, shelves
Tools: computer, funnel, gun, paint, shaver
Vehicles: gondola, horse, roller skates, tricycle, wheelchair
Weapons: chair, drugs, fingernails, fork, gas

Natural
Birds: bat, Big Bird, duck-billed platypus, sandcrane, thunderbird
Fruits: avocado, coconut, cucumber, rhubarb, tomato
Trees: hemlock, juniper, lilac, sage, sassafras
Vegetables: cloves, gourd, hominy, pumpkin, rice

member and clear *nonmember* items were included, the borderline items (see Table 1) were most critical for the present purposes, since they are, a priori, most likely to exhibit category gradedness (Diesendruck & Gelman, 1999; Estes, 2003; McCloskey & Glucksberg, 1978). The participants judged the category membership of each item and their confidence in that membership judgment. For the category judgments, a main effect of domain (artifact or natural) was expected. In order to replicate prior research, artifact categories should elicit more graded membership responses than natural categories do (Diesendruck & Gelman, 1999; Estes, 2003). For the confidence ratings, the uncertainty hypothesis predicts a main effect of domain, so that artifact categories should be judged with less confidence than natural categories are. This would reveal a negative relation of confidence and gradedness between domains. Furthermore, the uncertainty hypothesis predicts that confidence and gradedness should be negatively correlated within domains as well.

Method

Twenty undergraduates at the University of Georgia participated for partial course credit. The materials were the same as those used by Estes (2003, Experiment 1) and included four artifact categories and four natural categories. Each category consisted of five members, five borderline items, and five nonmembers, all sampled from Barr and Caplan (1987). Each nonmember had a mean membership rating between 1.00 and 3.00 in Barr and Caplan's norms, where the scale ranged from 1 (*clear nonmember*) to 7 (*clear member*).

The mean ratings of borderline items were between 3.01 and 5.00, and the mean ratings of members were between 5.01 and 7.00. Items in the artifact and the natural categories were also matched for typicality (see Estes, 2003).

Each item was presented with its target category—for instance, "item: tomato, category: fruits." Below each item were a membership scale, which ranged from 0 (*not at all a member*) to 10 (*completely a member*), and a confidence scale, which also ranged from 0 (*not at all confident*) to 10 (*completely confident*). The participants were instructed to judge the membership of each item and their confidence in that membership judgment ("circle a number on the scale that corresponds to how confident you are in your membership judgment"). Item order was random within the experimental list but was constant across participants.

Results

Artifact categories were more graded than natural categories. Artifact categories were also judged with greater confidence than were natural categories. Despite this positive relation of confidence and gradedness *between* domains, confidence and gradedness were negatively correlated *within* both domains. Gradedness and confidence of the borderline items are illustrated in Figure 1.

Gradedness. In the interest of brevity, statistical analyses focus on the critical borderline items; results from the member and nonmember items are summarized in note 2. The proportion of nonendpoint responses (i.e., any rating from 1 to 9) served as the measure of category gradedness. These proportions were submitted to two sets of analyses, one with participants (t_p) and another
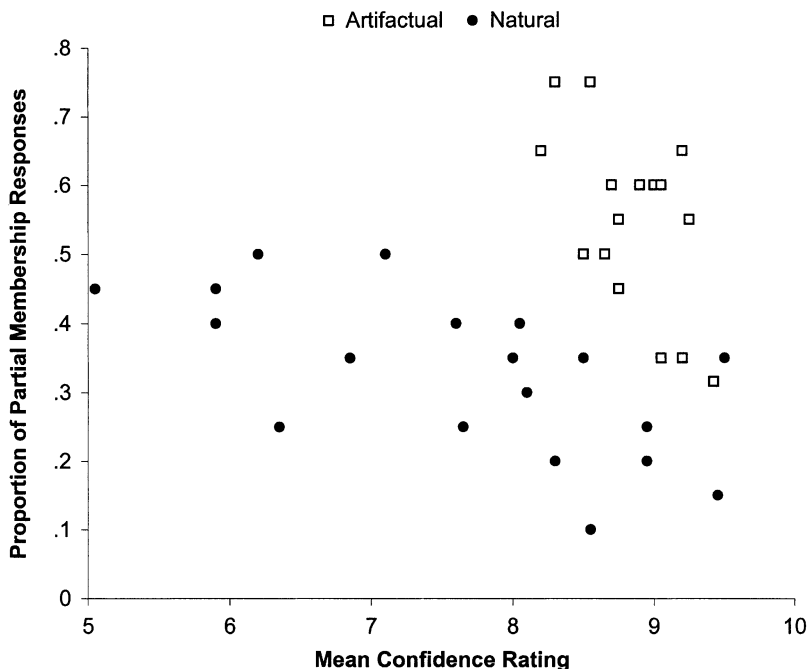

Figure 1. Gradedness as a function of confidence, Experiment 1. Artifact categories were more graded and were judged with greater confidence than natural categories were. Despite this positive relation of confidence and gradedness between domains, confidence and gradedness were negatively correlated within each domain (artifact, $r = -.55$; natural, $r = -.64$).

with items (t_i) treated as random. Artifact items ($M =$ .56, $SE = .06$) were more likely to elicit graded membership judgments than were natural items [$M = .34$, $SE = .05$; $t_p(19) = 3.34$, $p < .01$, and $t_i(38) = 4.98$, $p < .001$]. This result replicates previous findings of a domain difference in category gradedness (Diesendruck & Gelman, 1999; Estes, 2003) and is illustrated as the tendency for the artifact items (i.e., white squares) to occur higher in Figure 1 than the natural items (i.e., black circles).

Confidence. Artifact items ($M = 8.77$, $SE = 0.16$) were judged with greater confidence than natural items were [$M = 7.50$, $SE = 0.24$; $t_p(19) = 5.19$, $p < .001$, and $t_i(38) = 3.93$, $p < .001$]. This result is contrary to the uncertainty hypothesis—the artifact categories, which were more graded, were also judged with greater confidence. That is, confidence and gradedness were positively related across domains. This domain difference in confidence is shown in Figure 1 as the tendency of the artifact items to occur to the right of the natural items.

Confidence and gradedness. Correlations between the proportions of graded category ratings and mean confidence ratings were calculated across the borderline items. Two outlying artifact items (i.e., *gun*–TOOL and *fork*–WEAPON) and one outlying natural item (i.e., *sassafras*–TREE) were excluded from the analysis on the basis that their means for either gradedness or confidence were two standard deviations beyond the group mean. (Removing these three items from analyses had no effect on the pattern of significant results reported above or below.) As is evident in Figure 1, confidence and gradedness were negatively correlated within the artifact domain ($r = -.55$, $n = 18$, $p = .02$), as well as within the natural domain ($r = -.64$, $n = 19$, $p < .01$). Thus, despite the positive relation of confidence and gradedness observed between the artifact and the natural domains, confidence and gradedness were negatively correlated within both of those domains.

Discussion

Experiment 1 revealed that artifact categories were more graded and were judged with greater confidence than were natural categories. Within both domains, however, confidence and gradedness were negatively correlated. Generally, then, confidence and gradedness were positively related across domains but were negatively related within domains. The uncertainty hypothesis can account for the within-domain variance in gradedness: The lower the confidence (i.e., certainty) in a category judgment, the more likely that judgment was to be graded. However, that relationship clearly did not hold between domains: Artifact judgments were both more graded and more confident than natural judgments. So the uncertainty hypothesis fails to account for the between-domain variance in gradedness.

As a first investigation of the relation between confidence and gradedness, the preceding experiment is informative. However, in one respect, the results are somewhat ambiguous. Given the scalar (0–10) method of eliciting

category judgments (cf. Barr & Caplan, 1987; Kalish, 1995), it is unclear whether the confidence ratings represent the participants' confidence in the *general* category membership judgment or their confidence in the *specific* scalar value chosen. To illustrate, a confidence rating of 5 could indicate marginal confidence that the item belongs in the category, or it could indicate marginal confidence that the participant's exact rating truly reflects his or her belief. The scalar method in Experiment 1 does not allow discrimination of these alternatives. This ambiguity was addressed in the following experiment.

EXPERIMENT 2

In the present experiment, a three-alternative forced choice method of category judgment was used. Response options were *nonmember*, *partial member*, and *full member* (Estes, 2003). Because the exact degree of category membership was not judged, this paradigm eliminated (or at least greatly reduced) the ambiguity inherent to the scalar membership ratings in Experiment 1. Experiment 2 also used a different set of borderline items. Kalish (2002) correctly pointed out that, in studies of category structure, the sampling of items can dramatically affect the results obtained. Thus, it was important to generalize the effects of Experiment 1 to another set of items (see Table 2). The question of present interest was whether the differential within- and between-domain relation of confidence and gradedness observed in Experiment 1 would replicate with the different method and items in Experiment 2.

Method

Thirty-nine undergraduates at the University of Georgia participated for partial course credit. Materials consisted of 39 artifact borderlines and 39 natural borderlines selected from Barr and Caplan (1987), Kalish (1995), and McCloskey and Glucksberg (1978). The items were the same as those used by Estes (2003, Experi-

Table 2
Items, Experiment 2

Artifact

Clothing: corduroy, handkerchief, headband, pocket
Furniture: clock, piano, picnic table, pillow, sofa, wheelchair
Kitchen utensils: broom, dishwasher, dustpan, garbage disposal, mop, refrigerator, stove
Ships: canoe, catamaran, gondola, houseboat, hovercraft, kayak, raft, rowboat, spacecraft
Toys: backgammon, bat, cards, drum, guitar, music box, musical instrument, racquet, string
Weapons: car, fork, rubber band, satellite

Natural

Animals: bacterium, fungus, poet, virus, yeast
Dogs: coyote, hyena, wolf
Fish: clam, crab, jellyfish, lobster, octopus, oyster, plankton, porpoise, sea anemone, seahorse, shrimp, sponge, squid, starfish
Flowers: fern, heather, hyacinth, ivy, philodendron, schefflera
Horses: donkey, mule, zebra
Insects: caterpillar, leech, praying mantis, scorpion, spider, worm
Mammals: bluejay, goose

ment 2). As in Experiment 1, borderline items sampled from Barr and Caplan were selected on the basis that they had a mean membership rating between 3.01 and 5.00 (on their 1–7 scale), with the restriction that the categories from Experiment 1 not be included here. Kalish (1995) used 10 artifacts and 10 natural kinds that were, according to his intuition, borderline items. All 20 of those were included here. Finally, McCloskey and Glucksberg elicited binary (yes/no) category membership judgments. Items about which their participants disagreed most—that is, items with 30%–50% non-modal responses—were selected for the present experiment. Sampling from these three independent sources is advantageous in that they all offer different criteria for determining borderline items, thereby reinforcing the generality of the obtained results.

Category membership judgments were measured via a three-alternative forced choice method, as described above. The participants were informed that *partial membership* meant that the item belonged in the category, but not to the same degree as some other items. These responses thus served as the measure of category gradedness. Below each item–category pair were three boxes corresponding to the three response options, and the participants were instructed to check one box for each item. The 0–10 confidence scale was presented directly below the categorization response options, as in Experiment 1. The participants were instructed to circle a number on the scale to indicate their degree of confidence in the category judgment. The items were presented in random order.

Results

Artifact categories were again more graded and were judged with greater confidence than natural categories were. The negative correlation of gradedness and confi-

dence within both domains was also replicated. Thus, in the present experiment, a disambiguated measure of gradedness was used to extend the results of Experiment 1 to a larger, potentially more representative sample of items. The results are presented in Figure 2.

Gradedness. The proportion of partial membership judgments was reliably higher for artifact categories ($M = .53$, $SE = .03$) than for natural categories [$M = .34$, $SE = .02$; $t_p(38) = 6.01$, $p < .001$, and $t_i(76) = 5.32$, $p < .001$].

Confidence. Confidence ratings were also reliably higher for artifact categories ($M = 7.68$, $SE = 0.16$) than for natural categories [$M = 7.20$, $SE = 0.22$; $t_p(38) = 2.86$, $p < .01$. and $t_i(76) = 2.28$, $p < .05$]. Between domains, then, confidence and gradedness were positively related.

Confidence and gradedness. Within domains, according to the uncertainty hypothesis, confidence and gradedness ought to be negatively correlated. Outliers were removed if they were beyond two standard deviations from the group means for either gradedness or confidence, as in Experiment 1. This led to the exclusion of three outlying artifact items (i.e., *sofa*–FURNITURE, *catamaran*–SHIP, and *gondola*–SHIP) and three outlying natural items (i.e., *schefflera*–FLOWER, *philodendron*–FLOWER, and *hyena*–DOG). Removal of these six items had no effect on the overall patterns of significance reported above or below. Confidence and gradedness were indeed negatively
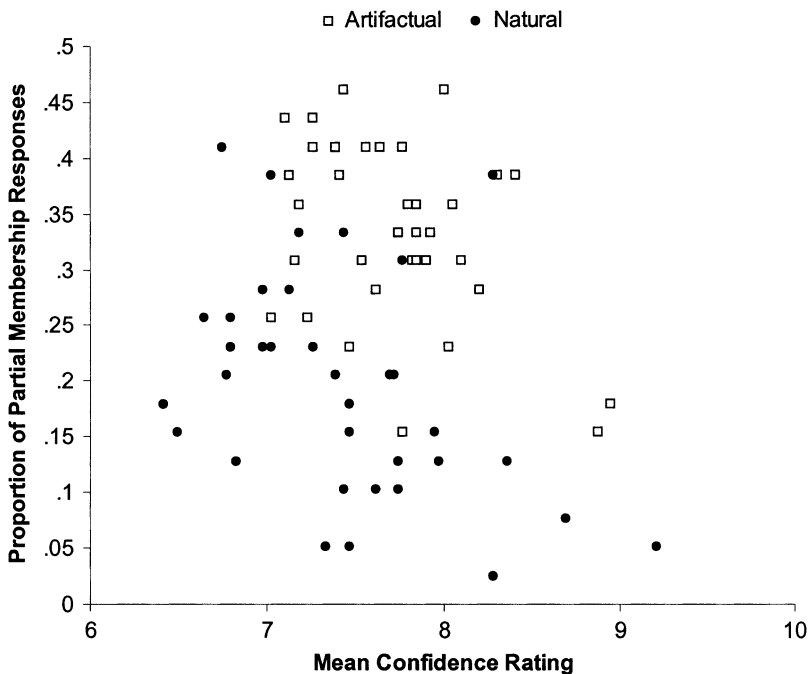

Figure 2. Gradedness as a function of confidence, Experiment 2. Confidence and gradedness were again positively related between domains, with confidence and gradedness both higher in artifact categories than in natural categories. However, confidence and gradedness were negatively correlated within each domain (artifact, $r = -.39$; natural, $r = -.42$).

correlated within both the artifact domain ($r = -.39$, $n = 36$, $p = .02$) and the natural domain ($r = -.42$, $n = 36$, $p = .01$). These correlations are illustrated in Figure 2.

Discussion

With a novel set of items and a modified measure of category gradedness, Experiment 2 corroborated the results in Experiment 1. Artifact categories were more graded and were judged with greater confidence than were natural categories. However, confidence and gradedness were negatively correlated within both of those domains. The uncertainty hypothesis can explain the variance in gradedness within the artifact and natural domains but fails to explain the variance in gradedness between domains. The results appear to reflect the participants' confidence in their general category judgments (Experiment 2), rather than their specific scalar rating (Experiment 1).

GENERAL DISCUSSION

Prior evidence of graded category structure (e.g., Barr & Caplan, 1987; Diesendruck & Gelman, 1999; Estes, 2003; Hampton, 1998; Kalish, 1995; McCloskey & Glucksberg, 1978; Rosch, 1975) rested on the assumption that graded category judgments truly indicate graded category structure. However, an *uncertainty hypothesis* (Johnson-Laird, 1983; Rey, 1983) suggested that graded category judgments might really just reflect graded certainty in those category judgments. If this hypothesis were correct—if gradedness could be explained as confounded uncertainty—then much prior research on categorization and representation would be rendered equivocal.

The preceding experiments are empirically and theoretically informative of this critical relation between confidence and gradedness in semantic categorization. Empirically, the experiments provided evidence that confidence does indeed predict gradedness *within* the artifact and natural domains but fails to predict the difference in gradedness *between* those domains. Theoretically, the uncertainty hypothesis partially explained category gradedness (i.e., within domains) but failed to explain the domain difference in gradedness (i.e., between domains). Thus, to some extent, participants do appear to misattribute their uncertainty as category gradedness. But there is more to gradedness than just uncertainty.

The finding of a domain difference in confidence, although novel in itself, is consonant with prior research. Several studies collectively have demonstrated that disputes about natural categories are thought to be definitively resolvable (Kalish, 1995; Malt, 1990) into technically defined classes (Hampton, 1998) via scientific investigation (Kalish, 2002) by experts (Malt, 1990). Artifact categories, on the other hand, are legitimately the subject of disagreement (Kalish, 1995; Malt, 1990), are less likely to fall under the purview of experts (Malt, 1990), and are more likely to be legislated than discovered (Kalish, 2002), presumably because they lack technical definition (Hamp-

ton, 1998). In other words, natural categories are believed to be objective matters of fact, whereas artifact categories are believed to be subjective matters of opinion.

Now, given this apparent domain difference in perceived objectivity, research on judgments of fact and opinion become relevant. Wagner (1984) found that confidence in opinion statements (e.g., "Strict gun control would be too great an infringement of individual freedom") was greater than confidence in factual statements (e.g., "Imposing strict gun control will reduce the crime rate more than 15%"). So to the extent that artifact category membership is considered a matter of opinion, whereas natural category membership is considered a matter of fact, it follows that artifact (opinion) categories will be judged with greater confidence than natural (factual) categories. That is, confidence and gradedness ought to be positively related, as the experiments showed. This explanation makes intuitive sense: Subjective (artifact) category judgments cannot be incorrect and are, therefore, rendered with confidence. But objective (natural) category judgments may be incorrect and thus engender less confidence. Put another way, people have more confidence in their subjective opinions about artifacts than in their objective knowledge of natural categories. This leads to the present finding that artifact judgments are resolutely graded, whereas natural judgments are hesitantly absolute.

REFERENCES

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

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.

Estes, Z. (2003). Domain differences in the structure of artifactual and natural categories. *Memory & Cognition*, **31**, 199-214.

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

Johnson-Laird, P. L. (1983). *Mental models*. 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.

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

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

Rey, G. (1983). Concepts and stereotypes. *Cognition*, **15**, 237-262.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, **104**, 192-232.

Wagner, W. (1984). Social comparison of opinions: Similarity, ability, and the value–fact distinction. *Journal of Psychology*, **117**, 197-202.

NOTES

1. By *artifact* and *natural kind*, I simply mean to denote, respectively, those categories that result from human production (e.g., KITCHEN UTENSILS) and those that occur independent of humankind (e.g., ANIMALS).

2. Member items did not differ in gradedness between artifact ($M = .02$, $SE = .01$) and natural ($M = .04$, $SE = .02$) categories (both $ps >$

.15). But there was a difference in confidence between artifact ($M = 9.97$, $SE = 0.01$) and natural ($M = 9.79$, $SE = 0.06$) domains [$t_p(19) = 2.83$, $p = .01$, and $t_i(38) = 2.56$, $p = .01$]. For these member items, confidence and gradedness were negatively correlated within the artifact ($r = -.77$, $n = 20$, $p < .001$) and the natural ($r = -.77$, $n = 20$, $p < .001$) domains. Nonmember items differed in gradedness between artifact ($M = .34$, $SE = .04$) and natural ($M = .16$, $SE = .03$) categories [$t_p(19) = 5.68$, $p < .001$, and $t_i(38) = 2.70$, $p = .01$]. But confidence ratings did not differ between artifact ($M = 8.97$, $SE = 0.02$) and natural ($M = 9.00$, $SE = 0.18$) categories (both $ps > .80$). Confidence and gradedness were negatively correlated in the artifact ($r = -.89$, $n = 20$, $p < .001$) and the natural ($r = -.76$, $n = 20$, $p < .001$) domains.