COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 41 (2017) 2026–2052

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12463

Do Americans Have a Preference for Rule-Based Classification?

Gregory L. Murphy,[a] David A. Bosch,[a] ShinWoo Kim[b]

[a]*Department of Psychology, New York University*
[b]*Department of Industrial Psychology, Kwangwoon University*

Received 3 June 2016; received in revised form 3 October 2016; accepted 5 October 2016

Abstract

Six experiments investigated variables predicted to influence subjects' tendency to classify items by a single property (*rule-based* responding) instead of overall similarity, following the paradigm of Norenzayan et al. (2002, *Cognitive Science*), who found that European Americans tended to give more "logical" rule-based responses. However, in five experiments with Mechanical Turk subjects and undergraduates at an American university, we found a consistent preference for similarity-based responding. A sixth experiment with Korean undergraduates revealed an effect of instructions, also reported by Norenzayan et al., in which classification instructions led to majority rule-based responding but similarity instructions led to overall similarity grouping. Our American subjects showed no such difference and used similarity more overall. We conclude that Americans do not have a preference for rule responding in classification and discuss the differences between tasks that reliably show strong rule or unidimensional preferences (category construction and category learning) in contrast to this classification paradigm.

Keywords: Categorization; Classification; Concepts; Cultural differences; Rules

1. Introduction

One of the major advances in cognitive science was the discovery and eventual consensus that natural categories are not well-defined sets. The work of Eleanor Rosch and colleagues, in particular, demonstrated that people's classification is graded rather than all-or-none and that some items in a category are "better" members than others (Rosch, 1975; Rosch & Mervis, 1975; see Smith & Medin, 1981). Natural categories are generally thought to follow a *family resemblance* structure, in which items can vary continuously

Correspondence should be sent to Gregory L. Murphy, Department of Psychology, New York University, 6 Washington Place, 8th floor, New York, NY 10003. E-mail: gregory.murphy@nyu.edu

in how many of a category's common features they possess, resulting in a typicality gradient (Hampton, 1995). It is surprising, therefore, that laypeople seem to persist in the belief that categories can be given definitions and furthermore often hypothesize that categories in psychology experiments can be distinguished by a simple rule using a single stimulus dimension, even when they cannot (e.g., Brooks, Squire-Graydon, & Wood, 2007).

One example of this tendency is that people in category-learning experiments often first hypothesize that each category is associated with a single feature, for example, separating items into green and blue ones. Nosofsky, Palmeri, and McKinley's (1994) RULEX model assumes that this is people's first strategy in learning novel categories, and traditional rule-based learning experiments show that one-dimensional rules are easier to learn than those involving multiple dimensions (Bruner, Goodnow, & Austin, 1956; Shepard, Hovland, & Jenkins, 1961). Modern dual theories of category learning posit that one of the two category-learning modules is to hypothesize simple, verbalizable rules (reviewed by Maddox & Ashby, 2004).

In a different task, when people are asked to divide up novel items into the groups that they feel are "the best and most natural," they almost never form family resemblance categories (Medin, Wattenmaker, & Hampson, 1987). Instead, they pick a stimulus dimension on which items differ and then divide them up into two groups based on that dimension. Their desire to do this is strong enough to resist manipulations to discourage it, such as constructing the stimuli so that it isn't possible, using holistic stimuli, or using massive numbers of stimulus dimensions (Ahn & Medin, 1992; Regehr & Brooks, 1995). Ahn and Medin found that if no stimulus dimension could in fact divide up the stimuli, subjects would initially divide them based on a single dimension as much as possible and only then resort to using a different dimension or similarity. When stimulus dimensions are related by higher-level knowledge, subjects are more likely to form family resemblance categories, but this probably reflects use of a different single-dimensional rule, for example, put the underwater buildings into one pile and the space buildings into another (Spalding & Murphy, 1996; see also Kaplan & Murphy, 1999).

Where does this tendency to prefer single-dimensional rules come from? It probably has a number of sources. One possibility is that people expect experimenter-defined categories to be defined in this way, just as puzzles and reasoning problems often are. Indeed, in science classrooms, the emphasis on holding variables constant in order to find the single causal variable (Kuhn & Dean, 2005) might lead one to expect categories in a psychology lab to have a single dimension that explains them. Another factor is computational and attentional limitations. Shepard et al. (1961) interpreted the differences in learning their six category structures as reflecting the number of dimensions involved. Their rules with more dimensions were also more complex, however, requiring one to hypothesize more elaborate rules. As a result, not only did their higher-level rules require attention to more dimensions, the computation required to learn the correct rules was also greater.

There is another potential cause for the preference for rules, namely, a tendency to reason logically. Perhaps people think that a rule such as "all the things in Category A are

red," is a better basis for a category than "things in Category A tend to be red, have four legs, and meow." Probabilistic family-resemblance structures do not have the certainty and elegance of single-dimensional or simple multidimensional rules. This possibility was explored by Norenzayan, Smith, Kim, and Nisbett (2002, Experiment 2), who tested classification with a category structure borrowed from Kemler Nelson (1984). As part of an investigation of cross-cultural differences in cognition, they created groups of items with four features, for example flowers. One group might all possess a curved stem and the other group a straight stem (Fig. 1 shows our reconstruction of this design). However, the other three dimensions were more variable. In one group, three out of four flowers had a leaf, no circle in the "head," and rounded petals; in the other group, three out of four had no leaf, a circle in the head, and spiky petals. These dimensions therefore formed a family resemblance structure in which category members shared multiple typical properties. Test objects were constructed that matched the consistent feature of Group 1 on one dimension (curved stem), but matched the features of the family resemblance structure of Group 2 (no leaf, a circle, spiky petals). Another item matched the consistent feature of Group 2 and the three imperfectly predictive features of Group 1. Norenzayan et al. sought to discover whether subjects of Asian and European descent differed in their tendency to follow a strict rule or to use overall similarity, that is, sort the test item by the consistent feature or by family resemblance, respectively.

Subjects who were asked to classify the test items tended to follow the rule (60%–70% of the time), regardless of culture. However, the groups differed when they were asked to judge which group the test object was more similar to. Two Asian groups tended to choose the family resemblance match (55%–60% of the time), whereas the European Americans strongly preferred the rule match (almost 70% of the time). Given that the computational requirements of categorization and similarity judgments would be identical in the two cultures, the difference suggests a stronger preference for logical reasoning among their American subjects.

The reasoning behind Norenzayan et al.'s design was that subjects would feel a tension between two opposing determinants of category membership, which they called *formal* and *intuitive*. On the one hand, the fact that all of the items in a group have a consistent feature suggests that there is a rule that determines membership. On the other hand, multiple features that tend to go together—but not perfectly—suggest that the category is based on global similarity. Their design pitted these two opposing classification strategies against one another, strategies we refer to as *rule-based* or *family resemblance*. Their finding that European Americans preferred to rely on a one-dimensional rule ("they all have a curved stem") is consistent with the tendency reviewed above that subjects in many past studies, generally students attending American universities, seem to believe that categories can be defined by necessary and sufficient features.

More recent work on rule versus family resemblance strategies has suggested that designs such as Norenzayan et al.'s (2002) may not be ideal for identifying individuals' strategies (Wills, Inkster, & Milton, 2015; Wills, Milton, Longmore, Hester, & Robinson, 2013). For example, people may not notice that one property is found in all category members and so cannot choose between rule and similarity-based options. Alternatively,

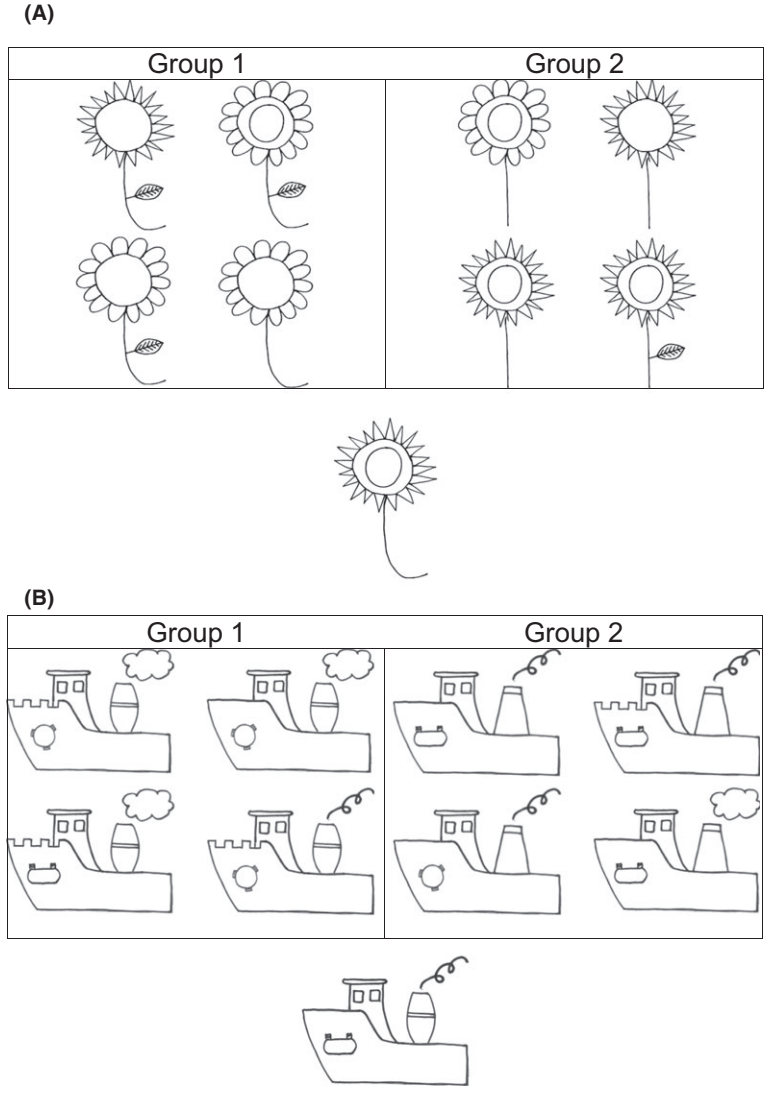

Fig. 1. Two displays illustrating the experimental design. The first was based on an item published by Norenzayan et al. (2002). In each case, the target at the bottom must be assigned to one of the two categories shown. The target matches one category perfectly on one dimension (e.g., the smokestack of the boat matches all boats in Group 1) and three dimensions of the other category imperfectly (the bow shape, life saver, and smoke type match Group 2) for three out of four items. For each category pair, another target item was constructed that was the opposite on all dimensions, thereby presenting the complementary test between a perfect rule versus matching multiple features.

they may use a single dimension but nonetheless make what appears to be a family resemblance choice. For example, if subjects believe that the petal shape is most important, they could attempt to match the target with the category that has the most identical

petal shapes. This would be one-dimensional responding but in the experimental design would result in a "family resemblance" match.

These are important methodological issues, which we return to in the General Discussion. As a result, the responses we call *rule-based* or *family resemblance* should be understood to be descriptive rather than indicating people's strategies, as one cannot tell from an individual response how the subject was matching. However, it is difficult to attribute consistent responding or findings of group differences to the possibilities Wills et al. mention. The greater use of rules by the European American group seems likely to reflect their preference for a perfectly predictive feature, which is different in every item, rather than preferences for curved stems and the like. Following a similar rationale, we hoped to investigate the factors that lead people to favor rule-based reasoning. In particular, we carried out a number of experiments in which we induced cognitive mindsets to see if we could increase or decrease the use of rules. A shift in group responding probably could not be explained by different preferences for a given stimulus dimension. In general, those manipulations had little effect, so we will not dwell on them. What was more significant was that the majority of our American subjects seemed not to favor rules after all, contrary to Norenzayan et al.'s (2002) result. This then casts a doubt on the preference for logic-based categories that explained their effect, which is a surprising finding given the number of phenomena in which people prefer to use single dimensions. Thus, the experiments have significance beyond their implications for Norenzayan et al.'s well-known finding.

In what follows, we first present our initial studies that attempted to influence rule use through a number of manipulations. We then explored possible explanations for the low rule use we observed. The final study was conducted in Korea to check for cultural differences. Our results replicated some aspects of Norenzayan et al.'s results, but not all. In particular, we did not find a strong bias toward rule use in Americans, nor less rule use in our one study with East Asian subjects.[1]

2. Experiments 1 and 2

The first two experiments explored two potential influences on rule use, using the general method of Norenzayan et al.'s (2002) Experiment 2. The first experiment varied construal level (CL), in which people consider the abstract, general properties of stimuli, or specific and concrete properties (Trope & Liberman, 2010). We did this through a standard abstraction manipulation developed by Fujita, Trope, Liberman, and Levin-Sagi (2006; see Burgoon, Henderson, & Markman, 2013). In this task, subjects provided either superordinates (*high CL*) or subordinates (*low CL*) of 40 category terms. For example, they might respond to the question, "Pasta is an example of what?" eliciting an answer like "food." This directs attention to the abstract properties of pasta and away from concrete properties that distinguish pasta from other foods. In contrast, "An example of pasta is what?" might elicit the answer "spaghetti," thereby emphasizing concrete, specific properties of pasta. Our hypothesis was that high construal level might lead to more

rule-based responding, on the assumption that rules are abstractions that ignore differences between category members.

Experiment 2 used a global-local manipulation. Focus on global properties of a group could direct attention to configurations, either at the level of the individual items or of groups, thereby encouraging family resemblance responding. Focus on local properties could lead to detection of specific features, facilitating detection of a defining rule based on a single consistent feature. We evoked these processing styles, using Navon's (1977) classic letter stimuli, in which small letter L's, for example, make up a large capital E. Subjects saw such Navon letters and provided as quickly as possible the local letter (L) or the global shape the letters made up (E). Such a task requires one to both perceive the stimulus at one level and to ignore its value at the other level.

Using this manipulation, Macrae and Lewis (2002) found that global processing improved performance in a face recognition task while local processing impaired it. Although that does not map directly onto rule versus similarity grouping, given that face recognition is often believed to be holistic, it is possible that attention to global configurations could encourage holistic processing, which in this case would presumably increase family resemblance sorting. Pre-training with the Navon task has also been demonstrated to influence strategy use in associative learning. The predictive value of a stimulus with multiple cues can be evaluated additively by attending to and summing the outcomes of each individual cue (elemental approach) or by viewing the configuration holistically and discriminating it from other configurations (configural approach) (Williams, Sagness, & McPhee, 1994). Byrom and Murphy (2014) found that use of a configural or an elemental approach was facilitated by global or local pre-training, respectively. These findings suggest that global processing may have a strong influence on individual differences in approaches to predictive learning (although our family resemblance categories were based on summing cues rather than configural properties). Similarly, Förster (2009) reported that global versus local perceptual styles led to an emphasis on similarities or differences, respectively. Perhaps a search for differences may be best satisfied by a single feature that is found in every single member of one category but in no member of the other. However, this paper has been retracted ("Relations between perceptual and conceptual scope... ," 2016), so further evidence for the influence of global and local processing on similarity and categorization seems called for.

In the test phase of both experiments, subjects saw a number of displays modeled after those of Norenzayan et al. As in that experiment, they classified one item for each display first and in a second block classified the display's "opposite" test item. For example, a test item that matched the rule of Group 1 but was similar to Group 2 would be followed in the second block by one that matched the rule of Group 2 and was similar to Group 1. Each subject received a score of how often he or she gave a rule or family resemblance response. Because of the high level of rule-based responding by all Norenzayan et al.'s groups in the classification task, we used similarity instructions, to avoid ceiling effects that would obscure differences between the conditions.

2.1. Method

2.1.1. Subjects

Given that we were not investigating cultural differences, we did not carefully sample subjects according to their ethnic group. In Experiments 1 and 2, all subjects were Amazon Mechanical Turk (*MTurk*) workers with American addresses. Therefore, the large majority of them are likely to be of European American ancestry. In all experiments, we required subjects to be at least 18 years old and to be fluent in English. In Experiment 1, there were 36 subjects in the low construal level group and 40 in the high construal level group, after discarding subjects who did not follow instructions in the first phase. In Experiment 2, there were 42 and 40 subjects in the global and local groups, as well as 38 subjects in a control group that received no global-local induction. Subjects were randomly assigned to conditions in each experiment. In most experiments with MTurk subjects, we asked about ethnic identity at the end of the experiment. We report these data below.

2.1.2. Materials

We constructed 10 category pairs along the lines of Norenzayan et al.'s (2002), including one based on their flower example depicted in the article. In general, however, we avoided presence versus absence of features as in the flower display (e.g., the leaf was present on some flowers but not others), because the absence of a property does not provide a strong perceptual basis on which to group items. As described above, each category had one property that was found in every object, and the three other properties appeared in three out of the four members. Two test (*target*) items were constructed for each display, which had the consistent property of one category and the family resemblance properties of the other category. This resulted in 20 test displays (10 category pairs $\times$ 2 target items). Fig. 1 shows examples of two displays, one based on the Norenzayan et al. flower example and one novel display.

2.1.3. Procedure

In Experiment 1, subjects were told that they would first engage in a "mind-clearing exercise" before proceeding to the main experiment. They were told either that they would be asked to name superordinate or subordinate category labels for common items, in the high or low CL conditions, respectively. The low CL subjects read that if they were given, "An example of singer is what?" they might answer "Taylor Swift." The high CL subjects read that if they were given "Singer is an example of what?" they might answer "an artist." Subjects completed 40 such items in the induction phase. Their answers were later examined, and anyone who consistently reversed the tasks or who gave obviously bogus responses was omitted from the analysis.

In Experiment 2, subjects were instructed to name either the "overall letter" (global) or the "smaller letter" (local). They were shown an example of a Navon letter, which was used to explain their task, for example, "In this picture, 'F' is the overall letter, so you would press the F key." Subjects completed 24 trials in the global or local conditions, while control subjects proceeded directly to the main task.

At test, everyone received the same displays. The two groups appeared at the top of the screen, labeled Group A and Group B. The target object appeared centered underneath them. We modeled our instructions on those of Norenzayan et al. (2002):

> In this phase, two groups of objects will be shown, labeled A and B. Below these groups, you'll see a single object and you'll be asked to decide which group the target object is most similar to by pressing the A key for group A or the B key for group B.

No further information was given about how people should make their decisions. The items appeared in two blocks, with random ordering of the displays within blocks, and there was no feedback.

2.2. Results and discussion

We calculated the proportion of family resemblance responses across the 20 trials for each subject. The means and SEs of each group are shown in Table 1. (Note that rule responding is the complement of these figures.) There was essentially no difference between the low and high construal level inductions in Experiment 1, with both responding close to 66% family resemblance choices, $t(74) = .33$, $p > .50$, $d = .08$. The three groups of Experiment 2 were also not significantly different, $F(2, 117) = .67$, $p > .50$, $\eta^2 = .01$. The means varied between 63% and 67% family resemblance choices.

The failure to find any effect of the manipulations may be important for studies of construal level and global versus local processing, but the biggest surprise to us was the failure to find a preference for rule-based processing. Norenzayan et al. (2002, p. 66) reported 69% rule-based responses for similarity judgments, whereas we got almost the opposite result—about 65% family resemblance classification. An examination of the 10 category pairs revealed that eight of them showed strong family resemblance preferences, with the other two having a moderate or strong rule-based response. Those differences were consistent across the experiments; we address item differences in later experiments. Furthermore, the flower categories modeled after the example in Norenzayan et al.'s article yielded 76% and 72% family resemblance responses in Experiments 1 and 2.

Table 1
Percentage family resemblance responses (and *SE*s) in Experiments 1–3, by condition

	Variable		
Experiment 1	Low CL	High CL	
	65 (3)	66 (2)	
Experiment 2	Global	Local	Control
	63 (3)	67 (3)	64 (3)
Experiment 3	"Belongs"	"Similar"	
	57 (4)	63 (2)	

Note. CL = construal level.

It is possible that our mental set inductions somehow changed people's attention or strategies. However, the control group in Experiment 2 received no induction at all, yet also gave 64% family resemblance responses.

We asked our MTurk subjects to identify their races (Asian, Black, Native American, Pacific Islander, White, or refuse to answer) at the end of the experiment. There were few self-identified Asian subjects—five in Experiment 1, four in Experiment 2—not enough to warrant a cultural comparison. However, we note that their levels of family resemblance sorting were 65% and 58% in Experiments 1 and 2, similar to the overall means and by no means higher than average.

3. Experiment 3

Norenzayan et al. (2002) reported that all three of their populations used rules when classifying items; it was only when the question asked was about similarity that Asian groups then used family resemblance the majority of the time. This suggests that these different questions can induce different strategies. They did not find such an effect in their European Americans, but that group was already giving high rates of rule responding in the similarity condition. Given that our population was giving predominantly family-resemblance responses, the instructions might have a greater chance of shifting to a rule basis with them.

3.1. Method

Eighty MTurk workers served in this experiment. It was identical to the previous studies with two changes. First, there was no induction of any kind. Second, half the subjects were asked to choose the group that the target item was most similar to, as before, and half were asked which group the target item "belongs to," which is the language used by Norenzayan et al. (2002). These terms were used both in the general instructions and in the question on the screen of each trial.

3.2. Results

There was a slight difference between the two tasks, as shown in Table 1, but it was not significant. When asked about similarity, 63% of subjects chose family resemblance responses (virtually identical to the equivalent control group of Experiment 2), and when asked about belonging to a category, 57% of the responses were family resemblance choices, $t(78) = 1.21$, $p > .20$, $d = .27$. We would not make a strong claim that there is no difference between the two tasks based on this null result. However, it is clear that asking about classification did not engender a predominant rule-based response, as it did for Asian subjects in Norenzayan et al.'s experiment, who had almost a 30% increase in rule responding relative to similarity. (Their European Americans had a strong rule

response regardless of task.) We return to the issue of instructions in Experiment 6, which tested East Asian subjects.

There were six self-identified Asian subjects in the sample, and their mean family resemblance sorting was 56%, certainly no higher than the overall group.

4. Experiment 4

In the basic paradigm, people simply make their choices, without accounting for them in any way. Perhaps this method encourages subjects to be rather casual about their choices. After all, if they can pick either answer and not have to explain or justify it, perhaps once they notice that the target object usually matches a group on a salient feature, they pick that group. This would often lead to FR responses. (We consider this possibility at greater length in the General Discussion.)

Smith and Sloman (1994) examined similarity and rule-based categorization using a very different method with natural categories (e.g., coins and pizzas). They did not present groups of objects, but rather described a somewhat ambiguous object that was similar to two categories but that violated a rule associated with one of them. For example, a 3-inch-wide round object is closer to a quarter in size than it is to a pizza, but it violates the rule that quarters have a fixed, smaller size. Following the results of Rips (1989), they expected people to say that the object was more similar to a quarter but was actually a pizza, as similarity judgments would be influenced by relative closeness, but categorization judgments would be influenced by the rule (fixed size). In fact, they initially found no such difference, as subjects were about evenly split in their categorizations. In a second study, they asked subjects to talk aloud while making their judgments (as Rips did) and found that the categorization results now followed the rule 67% of the time. Such classifications were especially common for people who mentioned the rule in their verbal protocols. Other studies have shown at least a slight increase in accuracy in other reasoning tasks when justifications are required (Rehder, 2014).

We borrowed this idea in an attempt to discover whether people would be more careful when justifying their categorizations and therefore provide more rule-based responses in the present paradigm. Furthermore, the requirement to give a reason in particular (as opposed to just talking aloud) may encourage people to seek a simple, clear basis for their judgment. Saying "All the flowers have a curved stem" is a compelling answer to why a test flower should go into Group 2. In contrast, saying that some—but not all—of the items in Group 1 have the same petals, some have the circle in the middle, etc., may not seem a very good reason. Of course, Norenzayan et al. (2002) did not ask for such justifications, but there may have been some other difference in procedure or subject population that made their subjects take the classification task more seriously.

Experiment 4 compared two groups, one of which provided a reason for its classification and one that simply made the classification, as in the other experiments. We used classification rather than similarity judgments as the task, because (a) there was little difference between the two in Experiment 3, and (b) classifications seem to require

justification more than similarity judgments do. To identify something as being in a category may imply that it shares an underlying property with other category members (Gelman, 2003), and the consistent feature could be a sign of that underlying property. Another difference from the earlier experiments was that we conducted this experiment in our lab at New York University. Perhaps the attention of an experimenter as well as the selection of university members would yield more rule-based responding. Norenzayan et al.'s study was conducted in the lab with University of Michigan students.

Finally, we decided to perform a more careful selection of the displays' defining features. Because Norenzayan et al. found such a strong tendency to rule responding in European Americans, and they did not report performing any pretests of the stimuli, we had assumed that people would notice any rule with readily visible features. However, our failure to find consistent rule-based responding brings this assumption into question. The two items that were consistently rule-based in Experiments 1 and 2 may have had defining features that were especially salient. Furthermore, if some of the family resemblance features in the other items were particularly salient, this could have discouraged people from choosing the rule-based response. For example, if curved stems are hardly noticeable, then people might choose a different dimension to focus on, which would result in a family resemblance response. Thus, we evaluated the salience of the features and constructed new displays for this experiment.

4.1. Method

4.1.1. Subjects

The subjects were 59 paid members of the NYU community. They were randomly assigned to the two groups, resulting in 31 subjects in the reason group and 28 in the simple group.

4.1.2. Materials

The 10 picture sets used in the previous experiment were modified for use in Experiment 4. We first obtained rating data on how salient each dimension of the pictures was. We did this by displaying the two groups of items for 6 s, during which time the subjects were asked to try to identify the differences between them. Then subjects rated each dimension on how much they had noticed it in comparing the two groups. The groups reappeared, with questions listed in alphabetical order, in the form "How salient are the antennae? How salient is the presence or absence of the leaf?" They were told that some differences might be very obvious and others would hardly be noticed, and that they should rate each dimension on a 1–7 scale indicating how obvious or salient the properties were as follows:

Twenty Mechanical Turk subjects performed these ratings. We then examined the mean rating for each dimension and chose a dimension to be defining that was neither the most nor the least salient. This way, we would not be biasing people toward or away from using a defining dimension by virtue of its salience. For example, for the animal category, the mean ratings were 6.1 (antennae), 5.5 (head), 5.2 (tail), and 5.0 (feet). We

selected the tail for the defining feature, as being neither the most obvious nor the least salient feature (though all features were rated as conspicuous in this case). Table 2 shows the ratings of the selected properties. In general, the defining feature was about as salient as the average of the other features.

With these defining features, new displays were constructed, following the design of the previous experiments. Two test items were again constructed for each display, each one matching the defining feature of one category and the typical features of the other category. These displays are available in the article's Supporting Information.

Recall that two items received consistently rule-based responses in Experiments 1 and 2, the robots and airplanes. We looked at the salience ratings to discover whether the rule dimension was particularly salient in those cases. For the airplane, that seemed to be the case, as the rule dimension was rated 6.1, and all the others were under 5.0. For the robots, the defining dimension, the antenna, was one of three about equally salient dimensions. Thus, perceptual salience likely influenced similarity choices for the airplane case, but some other factor was probably driving the rule-based responding for robots. Perhaps the fact that the antenna was on top made it popular as a rule dimension.

4.1.3. Procedure

The stimuli were displayed in a notebook whose pages subjects turned to reveal each pair of categories. They wrote their responses on a response sheet. The *simple* group was told that they would see two groups of objects labeled Group 1 and Group 2. Underneath was a test object, and they were to indicate whether it belonged to Group 1 or 2 on the answer sheet. The *reason* group received the same instructions, with the addition, "Then you will be asked to explain why you thought the object belonged to that group."

The response sheet for each group had a separate question for each display, in the form, "Which group does the robot belong to? Group 1 Group 2." They circled their response. The reason group then answered the question, "Why does it belong to that

Table 2
Mean rated salience of rule-based and other features, Experiment 4

Items	Defining Feature	Salience, Defining Feature	*M* Salience, Other Features
Airplane	Rudder	5.0	5.0
Animal	Tail pattern	5.2	5.6
Boat	Smokestack	5.6	5.2
Bug	Legs	5.8	5.1
Car	Wheels	5.0	5.2
Flower	Stem	5.4	5.9
House	Window	5.4	5.6
Robot	Foot	5.4	5.4
Teapot	Knob	5.2	5.3
Tree	Stump/ring on the trunk	5.2	5.4
Mean		5.3	5.4

group?" for each item. As before, the stimuli were blocked so that each set was questioned once with one test item in the first half and again with the other test item in the second half. The items were randomly ordered within each block. The experiment took 10–15 min to complete.

4.2. Results

As before, we coded whether subjects classified the item based on the defining feature or family resemblance, initially without reference to their reasons. There were 52% family resemblance responses in the simple group and 55% in the reason group, a small difference in the opposite direction from what had been predicted, $t(58) = -.34$, $p > .50$, $d = .09$. Perhaps more important is the fact that there is still no evidence that this American population relied primarily on the defining feature. The item means ranged from 48% to 61% family resemblance choices, suggesting that the stimulus pre-testing did even out some of the item differences found in the earlier experiments. Now no item had a strong rule-based preference.

We next examined the rationales given for classification by the reason group. The analysis revealed a subset of subjects who consistently claimed to use the intended defining feature. Of 31 subjects in the reason condition, only ten had 10 or more correct one-dimensional justifications (out of 20 questions); eight of those provided 19 or 20 such reasons. These eight clearly identified the defining features and consistently used them to classify the test items. But among the other 23 subjects, the correct defining feature was only cited 1.8 times out of the 20 trials on average (median = 0). Some gave what were clearly family resemblance explanations, by either stating that overall similarity was used (e.g., "very similar to group 1") or by mentioning multiple properties (e.g., "the horizontal lines, the ears, and the antennas/eyes" in a particularly clear case). Altogether, 13 of the subjects gave majority family-resemblance explanations (ignoring uninterpretable responses). Overall, then, there is a slight advantage for family resemblance responding in people's stated reasons.

As is not uncommon when subjects provide explanations of their choices, some of the responses seemed uninformative or peculiar. Some subjects claimed to use a one-dimensional strategy, but it was consistently not the actual defining dimension, so they actually chose the FR option most of the time. As warned by Wills et al. (2013), the response itself in such tasks may be ambiguous: People may choose the "family resemblance" option but are really focusing on one feature, just not the one involved in the rule. However, the very consistent nature of these responses, usually or always avoiding the actual defining feature, casts doubt on whether such people were really using a rule that did not happen to correspond to the actual rule,[2] or instead were making similarity-based judgments which they explained in a lazy manner, identifying only one of the family resemblance dimensions (perhaps the most important one). Self-report data must always be taken with a few grains of salt, but recall that the primary motivation for this experiment was not to rely on stated reasons to classify responses, but to discover whether asking for reasons would increase rule-based responding. It did not.

4.3. Discussion

Overall, the results were similar to those of the previous experiments. The defining feature was now in the middle range of stimulus salience, so the experiment was not biased towards or away from using it. People chose the category consistent with the defining feature less than half the time, and this proportion did not increase when they were asked to give reasons for their decision. Although the reasons given were not always coherent or even correct in describing the choice made, about a third of subjects claimed to have used the (actual) defining feature on a majority of trials.

The mean level of family resemblance responses is about the same as that of the classification task in Experiment 3 (57%, compared to 52% and 55% here). Thus, we can say with greater certainty now that the failure to find rule-based responding the majority of the time in Experiments 1–3 was not due to uncontrolled differences in stimulus salience. We test this claim further in Experiment 5, which returned to the MTurk population, allowing comparison with the earlier experiments.

5. Experiment 5

The previous experiments all labeled the categories with neutral terms like *Group 1* or *Group 2*. These labels do not suggest the nature of the categories, or even suggest that there is a good reason for the items to be grouped together. Perhaps this label actively discouraged people from using defining features. After all, if the items are merely a group perhaps made only for convenience, there's no reason to expect that the group will have a consistent property binding the items together.

Research in categorization suggests that presence of a category name may indicate that there is an underlying, defining basis for categorization. That is, once you call an animal a *zork*, you have suggested that something is common to zorks, or else they would not all have that name (Gelman, 2003). Furthermore, people attribute deep properties to categories they know little about, such as elms or sturgeon, in large part because they assume that categories with names like this possess many features in common (Coley, Medin, & Atran, 1997).

We attempted to take advantage of this propensity by adding labels to the categories. However, the labels were not simple category names. Especially for artifacts, names could be construed merely as brands (e.g., are *Zork airplanes* a type of plane or a brand name like Cessna?), which would not necessarily indicate important differences between the groups. Thus, we also used functional or location labels (e.g., airplanes that carry cargo or passengers). We took care to avoid any names that would be associated with the objects' featural differences. Our hypothesis was that meaningful names might lead to the inference that the objects sharing that label must all have something in common, leading to rule-based classification.

5.1. Method

The design was very similar to that of the previous experiments. The main difference was that half of the subjects received categories with the Group 1 or 2 labels (as before) and half received meaningful labels. Those labels were as follows: airplanes carrying people/cargo, carnivorous/herbivorous animals, ferries/tugboats, fireflies/beetles, cars fueled by gasoline/electricity, flowers that grow in sunlight/shade, urban/rural houses, robots used in homes/factories, teapots for green/black tea, and forest/desert trees. In all other respects, the experiment followed previous procedure. The displays were the ones designed for Experiment 4, with rule features equated for salience.

The participants were 40 MTurk workers who were randomly assigned to the two groups. All subjects read the same classification instructions telling them that they would see two groups and would then choose which one of them the object below belonged to. Through an oversight, we did not request information on ethnic identity in this experiment.

5.2. Results and discussion

When there was no meaningful label, subjects made the family resemblance response 75% of the time (SD = 24.0). With a label, they did so 66% of the time (SD = 25.4). The difference was not significant, $t(38) = 1.22$, $p > .20$, $d = .39$. Although there was a hint of an increase in rule use with the meaningful labels, two-thirds of responses were still family resemblance choices. We do not strongly argue that there is no effect of names, and perhaps a search for more efficacious labels would strengthen the effect. However, the present results do indicate that the failure to find rule-based categorization is not due to the emptiness of the labels Group 1 and Group 2. The rate of family resemblance sorting by the meaningful label group here (66%) was essentially identical to those of Experiments 1 and 2—and more than in Experiment 3—none of which used meaningful labels (Table 1).

The results also show that the earlier results from MTurk subjects were not due to imbalances in the salience of the rule and family resemblance stimulus dimensions. The dimension underlying the rule was now in the middle of the salience distribution, following the construction of new displays in Experiment 4, and the present results, also from MTurk subjects, show no more rule use than in Experiments 1–3. Indeed, in the similar no-label condition, they show less.

6. Experiment 6

As explained earlier, we did not begin these studies to investigate cultural effects on cognition. Rather, we hoped to manipulate our American subjects' expected tendency to use rules. Thus, we did not carefully control our subjects' ethnic identity, or even assess it in all cases. The results of the previous experiments were based primarily on data from

European-American subjects, and when we had data on the few Asian American subjects, they did not differ noticeably.

Given our results, it seemed worth testing an Asian sample for comparison to Norenzayan et al.'s results. Since Americans were already sorting by family resemblance at moderate levels, we did not expect an Asian group to produce much more family resemblance sorting, but we were now able to address another of Norenzayan et al.'s findings. They asked some subjects to choose the group that the test object belonged to (classification) and others to choose the group that it was most similar to, finding that the first instruction led to rule responding, but the second to more family resemblance responding in their Asian groups. We did not find such a difference in Experiment 3, but that was with American subjects. Norenzayan et al. did not find a difference with European American subjects either, but their subjects were predominantly rule responders in both cases, whereas ours were majority family resemblance responders. Regardless of the difference in the overall rule responding in our studies, it is possible that East Asian subjects are more sensitive to these instructions than Americans are, so we tested two groups of Korean students at Kwangwoon University in Seoul. One group was asked to choose the more similar group and the other the group to which the object belonged.

We planned to compare these results to those of Experiment 3, which also compared instructions. Although this is technically a cross-experiment comparison, it is effectively the same as any cross-cultural study that tests different populations in different countries, with different experimenters, using different languages, etc. The important thing is that the materials and instructions are as identical as possible, as they were in this case.

6.1. Method

The overall method was very similar to that of the previous experiments. The categories and test items were the same as in Experiments 4 and 5, with the displays presented in a notebook. Each page asked which group the target object was similar to or belonged to; the answer was circled on a response sheet. The test items were presented in one order for half of the subjects and in the reverse order for the other half.

Forty-eight undergraduates at Kwangwoon University (Seoul, South Korea) participated for course credit (Norenzayan et al. had 53 East Asians in their sample). The materials and instructions were written and the experiment conducted in Korean. Subjects were randomly assigned in equal numbers to one of the two conditions (similarity, classification) and to one of the two category orders. They were first asked to read the instructions carefully. The experimenter then verbally repeated the task in their assigned conditions and told them to begin.

6.2. Results

There was a near-significant difference in the two instructional groups: Those asked to classify the objects chose the family resemblance option 45% of the time, whereas those asked to choose the most similar did so 62% of the time, $t(46) = 1.94$, $p < .06$, $d = .56$.

Although not reaching statistical significance, the result must be judged as a replication of Norenzayan et al.'s (2002) result. In their East Asian population (University of Michigan students), the classification instructions resulted in only about 30% family resemblance responding (read from their Fig. 5), whereas the similarity instructions yielded 59%. (Norenzayan et al. analyzed the two instruction groups separately, so we do not know whether that difference was statistically reliable. However, given the effect size and size of their error bars, it seems very likely that it was for their East Asian group.) This difference is considerably larger than ours, but the two results are clearly consistent. However, the overall percentage of rule responding is less in our sample than in theirs, consistent with our other results.

6.3. Discussion

Comparing our results to those of Norenzayan et al. (2002) is not straightforward. They found that all their groups used rule responding when the instructions were to classify, but the two Asian groups (especially the East Asians) switched to primarily family resemblance responding with similarity instructions. The European Americans used rule responding regardless. The simplest way to analyze that pattern would seem to be a two-way ANOVA with factors culture and instruction, which we report first. However, Norenzayan et al. did not report this analysis, but instead did separate ANOVAs comparing groups in the two instruction conditions.[3] We do a parallel analysis for the purposes of comparison.

Our Experiment 3 varied instructions with MTurk subjects in the same way as the present one, so we use that as our American sample, although recognizing that there are differences between the MTurk population and a college sample in addition to any cultural influence. A 2 × 2 ANOVA with variables culture and instructions revealed a main effect of instructions, $F(1, 124) = 6.16$, $p < .02$, partial $\eta^2 = .05$, as there was more family resemblance responding with similarity than with classification instructions. However, neither the effect of culture, $F(1, 124) = 2.10$, $p = .15$, nor the interaction, $F(1, 124) = 1.49$, $p = .22$, was reliable.

The closest comparison we can make to Norenzayan et al.'s analysis is to compare the two ethnic groups within each instruction condition. (We recognize that the interaction of culture and instructions was not reliable, but given that Norenzayan et al. did not test that interaction, we perform the comparisons that they did.) With classification instructions, the two groups were marginally different, $t(62) = 1.67$, $p = .10$, $d = .43$, but the Americans gave *more* family resemblance answers than the Koreans did (57% vs. 45%), contrary to Norenzayan et al., who reported no effect of culture for this question. With similarity instructions, the two groups gave nearly identical responses (62% vs. 63%), $t(62) = 0.19$, $p > .80$, whereas Norenzayan et al. found a culture effect here.

Although this seems to be very different from Norenzayan et al.'s results, the form of the interaction is actually the same in their and our results: Koreans give more family resemblance answers for the similarity question than for classification, whereas Americans were less sensitive to the questions. What differs is the effect of culture, as Koreans

gave overall more rule-based answers than our Americans did, and so the (marginal) difference was found in classification instructions rather than similarity instructions. Note that the amount of rule responding by Koreans in the classification group is greater than that of any condition in our previous five experiments with Americans (i.e., the 45% family resemblance mean is lower than that of any other condition, which were all over 50%; e.g., see Table 1). We will address this as part of the broader comparison of the two studies in the General Discussion.

7. General discussion

The individual experiments generally took the form of testing whether a particular variable would influence the use of rules versus family resemblance in similarity or classification. Only one of those attempts had a significant effect; the main empirical finding is that our subjects seldom chose the category based on a rule, as shown by family resemblance responses around two-thirds of the time.

Many forms of categorization or reasoning are subject to strong strategic differences, such that some people consistently give one kind of response and others consistently give a different one. For example, people often show strong preferences for taxonomic or thematic categorization (near 90% preference for one or the other) when making forced choices analogous to the method used here (Lin & Murphy, 2001; Simmons & Estes, 2008). Perhaps the same is true of the present contrast. To answer this, we constructed a histogram based on the combined responses of subjects in Experiments 1, 2, 3, and 5, ignoring the different conditions (which had little effect). These were all American MTurk subjects and so were generally comparable.

As Fig. 2 shows, there is little sign of strong individual differences. The mode is the 60–79 bin, which is also the mean of most of the conditions. Most of the remaining subjects are in the adjacent bins. Only a small number of subjects consistently used rule-based responding. Twelve subjects used rules greater than 80% of the time, whereas 58 subjects made family resemblance responses 80% or more of the time. So, unlike the

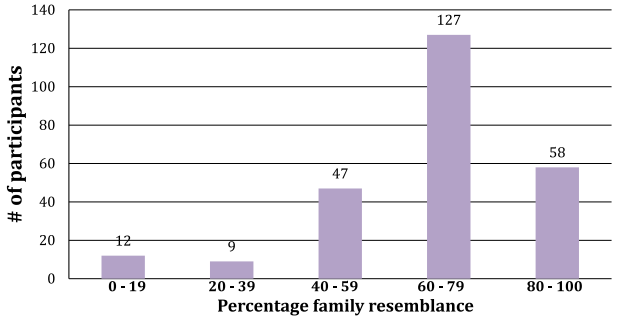

Fig. 2. Histogram of American MTurk subjects' mean percentage of family resemblance choices, Experiments 1, 2, 3, and 5.

taxonomic-thematic preference, here the vast majority of subjects have the same conceptual preference. The mean family resemblance sorting over 253 subjects was 64.4%, *SE* = 1.3. However, as we have emphasized, it must be remembered that "family resemblance" refers to a kind of response in this task. Not all of these responses are true use of a family resemblance strategy in which the test item is matched to a category by multiple features. But by the same token, the complementary distribution of "rule" responses also does not include only people who identified a criterial attribute and used it. In both cases, people may have chosen an attribute arbitrarily, which just happened to coincide with one or the other choice. However, as we will argue below, it seems unlikely that such an arbitrary choice happened much of the time (and see Experiment 4). In any case, it is clear that rule-based responding is in the small minority.

The manipulations we tested generally seemed to be ineffective. Although some may have had a weak effect not detectable with our sample sizes, none of them resulted in a preference for rule-based responding. The most noticeable difference among Americans was not due to a manipulation at all but to a difference between subject populations, which we discuss in the next section.

7.1. Comparison to Norenzayan et al. (2002)

Clearly, our results are different from those reported by Norenzayan et al. (2002). Our original goal was not to test or replicate their findings but rather to investigate the role of other manipulations on rule use. Our comments on the differences between their and our results are therefore necessarily speculative.

It is useful to break down the Norenzayan results into separate findings for purposes of comparison. First, they found a large amount of rule-based responding, especially in the European American population. We clearly did not replicate that result, as shown in Fig. 2. Second, they found that Asian but not American subjects changed their responses as a function of instructions, classification versus similarity. Surprisingly, we did replicate that result, even though our American subjects' overall responses were not similar to theirs. That is, MTurk subjects showed little difference between the two tasks (Experiment 3), whereas the Kwangwoon undergraduates showed a 17% decrease in rule-based responding with similarity instructions ($p < .06$). Although Norenzayan et al.'s effect was much larger for East Asians (about 30%), our effect was marginally significant and numerically larger than any other within-experiment comparison we made. Finally, under similarity conditions, Norenzayan et al. found a large difference between European American and East Asian subjects. We did not replicate that result, related to the first point noted above. Our Korean undergraduates made family resemblance choices 62% of the time under similarity instructions, very close to the overall average for our American subjects (Fig. 2). Indeed, if there is any point of difference between the Asian and American subjects, it is the Koreans' higher rate of rule-based responding under classification instructions (55%), which is greater than we found in any experimental condition with American subjects. That seems clearly contradictory to Norenzayan et al.'s conclusion.

Norenzayan et al. (2002) present a number of different paradigms to support their claim of cultural differences, and so their conclusions do not rest on this particular result. But as regards this paradigm, even if Asian cultures do prefer similarity-based processing compared to European cultures, it will be difficult to find this if Americans are using similarity two-thirds of the time in this task. Thus, one conclusion we would offer is that researchers investigating cultural differences in cognition might be better served by choosing a different task.

The one replicated result raises its own puzzles, however. Why should East Asian subjects show a difference between similarity and categorization instructions, while American subjects do not? In Norenzayan et al.'s results, the cultural difference made some sense: The Americans are always logical, but the East Asian subjects are more open to similarity-based reasoning. However, that story is less compelling when the Americans are not that logical, as in our results. Possibly there is some linguistic difference in the way these terms are understood, such that classification or "belongs" has an implication of following a rule in Korean and other Asian languages. Norenzayan et al. tested their subjects in English, but there could be a hold-over in how their East Asian subjects in particular understood terms in what was a second language for many of them.

The question remains, though, why their results with Americans are so different from ours. One contributing factor may be the use of university students versus the larger MTurk population. In our one experiment using New York University students, Experiment 4, we did find a slightly greater use of rules, with family-resemblance scores of 51% and 57%, instead of around 65% for MTurk subjects. Norenzayan et al.'s subjects were all University of Michigan students. As mentioned earlier, research in classification suggests that use of rules or taxonomic categories is related to education and even educational testing context, where many tasks require one to identify the single controlling variable or property (Kuhn & Dean, 2005; Murphy, 2002; Sharp, Cole, & Lave, 1979). However, the differences between our university and MTurk subjects does not seem sufficient to explain the high level of rule responding Norenzayan et al. report; and the Korean university students also produced strong family-resemblance responses under similarity instructions.

It would be tempting to attribute Norenzayan et al.'s (2002) results from their European American students to a statistical fluke. However, given that it happened in both instructional conditions, that is two flukes, which seems too many. In the end, we are thrown back on the possibility that there is some difference between Norenzayan et al.'s stimuli or procedure and ours that accounts for the differences. In the one item we shared, the flower stimuli, however, we found consistent family resemblance use.

We are familiar with only one other study that attempted to replicate Norenzayan et al.'s (2002) study. Klein et al. (2009) tested undergraduates in Japan, Korea, Taiwan, and the United States on a variety of tasks, including the Norenzayan et al. classification task used here. (It is not clear whether they constructed their own stimuli or had access to Norenzayan et al.'s.) They found no difference across groups using the similarity instructions, $F(2, 181) = 2.01$. The mean numbers of family resemblance responses (out of 20) were 10.2, 10.7, 12.0, and 11.5 for their Japanese, Korean, Taiwanese, and

American subjects, respectively. If anything, the American subjects were on the high side of family resemblance responding. Their absolute proportion, 57.5%, is about the same as what we found with NYU students. Klein et al. also report finding no group differences under classification instructions but do not provide details. They did find reaction-time differences in which American subjects responded more slowly than the other groups under similarity instructions. However, this cannot be interpreted as indicating different strategies in grouping judgments, absent any differences in actual choices. Thus, we take the Klein et al. result as being consistent with our failure to find rule-based responding as a dominant strategy.

7.2. Explanations of unidimensional preference

One consequence of the difference between our and Norenzayan et al.'s (2002) results is that it now seems less likely that Americans are expressing a preference for logical or formal solutions in this task. At least, if they have such a tendency, it was overcome by the other aspects of the task, such as the difficulty of identifying the defining feature. Therefore, we now reconsider other tasks that have shown a preference for unidimensional responding in this light. Should they be interpreted as reflecting a preference for rules?

We noted in the Introduction that there are many demonstrations that people prefer or find it easier to process categories in terms of a single dimension. This may make our consistent failure to find a rule-based preference in the present task seem surprising. However, there are important differences between tasks that very likely explain the different results. The most striking contrast involves the category construction task originated by Medin et al. (1987). When given a set of items to sort into groups, people very reliably form groups (most often two) based on a single dimension, even if the stimuli are drawn from two categories with strong family resemblance structures (Ahn & Medin, 1992; Lassaline & Murphy, 1996; Regehr & Brooks, 1995). For example, a set of fictional bugs that differ in head shape, body pattern, tail length, and number of legs could have been sorted into two groups in which all those dimensions are mostly—but not perfectly—consistent. Instead, *all* subjects divided them into groups that shared the same head shape, say, ignoring the other dimensions (Medin et al., 1987). Increasing the number of dimensions to six did not change people's strategy, only the number of dimensions that they ignored when making unidimensional sorts.

Why is this extremely strong unidimensional preference not found in the present study, in which people could categorize test items by using a perfect rule? One important difference is that in the category construction task, people scan unsorted stimuli and identify which dimension they think is most important. Then they can use that dimension to divide up the items. In the present task, the choice of the rule and family resemblance features is made by the experimenter. As a result, subjects are not scanning dimensions to see which ones they wish to use but are comparing the existing categories to see how they differ. When making that comparison, they are likely to notice multiple differences between the categories. One category of houses mostly has tall chimneys, while the other

mostly has short chimneys; one has mostly two doors, and the other mostly one door; and so on. Even if one dimension is perfectly divided in the two categories, subjects can still see that there are other differences, and they may then wish to use all the dimensions they have identified as different. In category construction, because the items are not already separated into groups, those almost-consistent features in each category are not evident. When a single dimension has been identified that can evenly separate the items, people can then form categories by focusing on only this one dimension, which is clearly easier than identifying the relations of multiple imperfect dimensions.

This analysis is supported by the fact that when subjects process the relations among the features prior to category construction, they are much more likely to form family resemblance categories (Lassaline & Murphy, 1996; Spalding & Murphy, 1996). Changing the task to a match-to-sample format increases family resemblance categories as well (Regehr & Brooks, 1995). If the unidimensional preference in the normal sorting task arose from a desire to have a logical basis for categorization, it is difficult to see why these manipulations should have much effect, as the logical response is equally possible in those conditions. Furthermore, Wills et al. (2013) found that when subjects were asked to be particularly slow and careful in their classification using the Regehr and Brooks matching task, they increased their family resemblance sorting, not their use of a single dimension (which declined 25%). Apparently, subjects did not believe that unidimensional responding was the "careful" answer.

Another task revealing unidimensional preference is category learning. The classic work by Bruner et al. (1956) found that it was easier to learn categories based on one feature (e.g., green things) than those based on two features (e.g., green and square; green or two). Shepard et al. (1961) found that their Type I categories based on one dimension (e.g., large vs. small) were learned faster than any other types, which involved multiple dimensions. This result has been replicated dozens of times, including with an MTurk population (Crump, McDonnell, & Gureckis, 2003). This is not surprising, given that the more complex categories are computationally more difficult to specify (Feldman, 2000). The Shepard et al. Type II categories had rules like "black and triangle or white and square." Virtually any theory of learning would predict that this would be harder to learn than a category based only on a single color. It is not necessary to propose a preference for unidimensional or logical categories to explain this.

The Norenzayan et al. task is not a learning task, however. The display is presented with the items already divided up into categories. There is no memory or learning requirement; subjects can look as long as they like and compare the categories at their leisure. For example, they could compare each feature of the test item successively to the two categories. Since they do not know in advance which feature will be used in the rule, they are very likely to notice that the item has more matches to one category than to the other. In contrast, during category learning, it is difficult or impossible to keep all the prior trials and their correct answers in memory (e.g., two recent members of Category A had tall chimneys, but one had a short chimney; all three had two doors; two of three members of Category B had round windows, etc.). Thus, hypothesis testing based on small numbers of features seem be more likely to be successful than trying to learn all

the available features. However, it should be noted that when people have enough trials in a learning experiment, they usually do not learn only one dimension but acquire more information about each category, even if it does not improve their classification performance (Bott, Hoffman, & Murphy, 2007; see also Hoffman & Murphy, 2006). Learners do not focus on a single highly salient dimension of a category to the detriment of the other features (Murphy & Dunsmoor, unpublished data). This may reflect an underlying belief that most categories are based on multiple dimensions, which should be learned (Bott et al., 2007). Thus, our finding that most Americans use a family resemblance strategy to classify the test items is consistent with the long-term learning of probabilistic categories in which multiple features are learned.

Wills et al. (2013, p. 304) make a very important observation about the use of simple rules in categorization tasks. They point out that it is very easy to *implement* a single-dimensional rule. If you know that members of category A are green and those of category B are red, it is trivial to classify a new object into one or the other. On the other hand, *discovering* a single-dimensional rule is often difficult. If the stimuli have many dimensions, one must rely on memory to test different dimensions and remember which ones have been ruled out. If other dimensions are correlated with category membership, they can mislead learners into thinking that they are the basis of a rule. In our task, subjects presumably look at multiple features at the beginning, especially because most of them are not perfect predictors. Noticing that one dimension perfectly separates the two categories actually takes a certain amount of attention and memory; noticing that other dimensions are predictive but *not* perfect is also a process that is likely not error-free. Given that, identifying one dimension as perfect and the others as not perfect may not be easy. Subjects might be very satisfied to respond after they have noticed that the test item matches most members of category B on two features that they've checked, without realizing that there is a different perfectly predictive feature. The result is that for many subjects, the two options may not represent the choice between logic and similarity, as either the defining feature or the family resemblance structure may not always be perceived.

Recall that Smith and Sloman (1994) found that talking aloud increased people's use of rules compared to simply making the judgments without any explanation. However, our Experiment 4 did not find any effect of justifying one's choices. We suspect that the reason is related to Wills et al.'s observations. Smith and Sloman tested natural categories, where people were already very familiar with the rules. That is, everyone knows that quarters have a fixed size. Therefore, a manipulation that causes subjects to think about such considerations would increase rule use. In our task, people do not know that there is a rule or what it is without carefully comparing the categories. Since giving a justification would not activate an already known rule in our case, it had much less of an effect. Again, implementing a known rule (as in Smith & Sloman's experiment) is easier than discovering a new one.

There is another possible strategy in this task, which is to pick one dimension arbitrarily and base the decision on that match. For example, looking at the target house, one might focus on its two doors and decide to pick the group that also has two doors most of the time. This would be single-dimensional responding but with an imperfect rule.

(Indeed, we question whether using an arbitrarily chosen, imperfect feature should be called "rule use.") About three out of four times, the chosen dimension would be a family resemblance dimension, leading to about 75% family resemblance answers. That number is a bit higher than most of our experiment means, but it is possible that this kind of guessing could explain the overall high levels of family resemblance scores we found. However, the reasons subjects provided in Experiment 4 were generally not consistent with this strategy. Only three of the 31 subjects gave mostly rule-based explanations of which a minority were the correct dimension (i.e., 4 out of 11,[4] 3 of 19, and 7 of 15 one-dimensional answers). In contrast, there were 8 subjects who gave a correct one-dimensional explanation 19 or 20 times, and 16 who said that they were following the correct rule 0 or 1 time out of 20. Neither of these patterns (exhibited by 24 out of 31 subjects) is consistent with this guessing strategy. So, a few people may have been arbitrarily choosing a single dimension to classify the target object, but that strategy does not seem very frequent.

7.3. Conclusion

Our data suggest that the findings of unidimensional responding or performance advantages in many conceptual tasks do not reflect a preference for rules. Instead, many of those findings are explained by task analyses that reveal that family-resemblance (or multidimensional) responding requires more memory or computation. In the present task, identifying the single dimension that perfectly separates two categories is not trivial, and the majority of subjects do not correctly identify and use it. Whether they realize that there is a rule that separates items but choose to ignore it is not entirely clear, but it does not seem that the task engenders a rule-based strategy in American subjects as a whole. American subjects may indeed attempt to be logical, as Norenzayan et al. (2002) propose, but the "logical" response in this task may require greater attention and computation than the similarity response. US subjects appear very willing to use similarity or a non-universal feature to determine category membership under such conditions. In order to make a stronger test of such logical tendencies to compare different groups, it would be useful to construct a task in which the logical rule takes the same amount of computation as the similarity computation. Then any preference or cognitive style difference would have a greater opportunity to reveal itself. However, lacking such a demonstration, our results tend to contradict the generalization that American subjects have a preference for single-feature classification.

Notes

1. We refer to our subjects as *Americans*, because we did not always know the details of their ethnic identities. However, we refer to Norenzayan et al.'s group as *European Americans*, because they specifically selected for that group. As will be seen, we were able to identify Asian Americans in our sample in most cases. Finally,

like Norenzayan et al., we use *East Asian* to refer to people born in East Asia (i.e., not including Asian Americans). Norenzayan et al. identify that sample in their Experiment 2 as being University of Michigan international students from China or Korea.

2. Two subjects claimed to use a single feature on all 20 trials, but it was never the rule-based feature. The chance probability of 20 choices of a single feature that is never the rule-based dimension is only .003. This suggests that they were actually making family resemblance choices but only describing one of the relevant features. Another subject claimed to use a single feature on 10 trials, but only one was the rule-based feature.
3. A minor point is that Norenzayan et al. did not perform one-way ANOVAS comparing the rule responses of the three groups but instead two-way ANOVAS with the factors ethnic group and response type (rule vs. family resemblance). However, the two response types were not variables but dependent measures that were perfectly correlated (proportion rule = 1 − proportion family resemblance), so the interaction they report is equivalent to the main effect of either dependent variable, which is how we analyze our data.
4. Actually, only two of these answers were correct as a description of the subject's classification; the other two were reversed. This makes it unclear what the person was doing.

References

Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*, 81–121.

Bott, L., Hoffman, A. B., & Murphy, G. L. (2007). Blocking in category learning. *Journal of Experimental Psychology: General*, *136*, 685–699. doi:10.1037/0096-3445.136.4.685

Brooks, L. R., Squire-Graydon, R., & Wood, T. J. (2007). Diversion of attention in everyday concept learning: Identification in the service of use. *Memory & Cognition*, *35*, 1–14.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Burgoon, E. M., Henderson, M. D., & Markman, A. B. (2013). There are many ways to see the forest for the trees: A tour guide for abstraction. *Perspectives on Psychological Science*, *8*, 501–520. doi:10.1177/1745691613497964

Byrom, N. C., & Murphy, R. A. (2014). Sampling capacity underlies individual differences in human associative learning. *Journal of Experimental Psychology: Animal Learning and Cognition*, *40*, 133–143. doi:10.1037/xan0000012

Coley, J. D., Medin, D. L., & Atran, S. (1997). Does rank have its privilege? Inductive inferences within folkbiological taxonomies. *Cognition*, *64*, 73–112.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLos ONE*, *8*(3), e57410. doi:10.1371/journal.pone.0057410

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Förster, J. (2009). Relations between perceptual and conceptual scope: How global versus local processing fits a focus on similarity versus dissimilarity. *Journal of Experimental Psychology: General*, *138*, 88–111. [This paper has been retracted; see "Relations. . ." below.]

Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self-control. *Journal of Personality and Social Psychology*, *90*, 351–367. doi:10.1037/0022-3514.90.3.351

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford, UK: Oxford University Press.

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

Hoffman, A. B., & Murphy, G. L. (2006). Category complexity and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 301–315. doi:10.1037/0278-7393.32.3.301

Kaplan, A. S., & Murphy, G. L. (1999). The acquisition of category structure in unsupervised learning. *Memory & Cognition*, *27*, 699–712.

Kemler Nelson, K. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior*, *23*, 734–759.

Klein, H. A., Lin, M.-H., Radford, M., Masuda, T., Choi, I., Lien, Y., & Boff, R. (2009). Cultural differences in cognition: Rosetta phase I. *Psychological Reports*, *105*, 659–674. doi:10.2466/PR0.105.2.659-674

Kuhn, D., & Dean, D., Jr (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, *16*, 866–870. doi:10.1111/j.1467-9280.2005.01628.x

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

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, *130*, 3–28.

Macrae, C. N., & Lewis, H. L. (2002). Do I know you? Processing orientation and face recognition. *Psychological Science*, *13*, 194–196. doi:10.1111/1467-9280.00436

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-based systems of perceptual category learning. *Behavioural Processes*, *66*, 309–332. doi:10.1016/j.beproc.2004.03.011

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*, 242–279.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353–383.

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, *26*, 653–684.

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

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 347–363.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107. doi:10.1016/j.cogpsych.2014.02.002

"Relations between perceptual and conceptual scope: How global versus local processing fits a focus on similarity versus dissimilarity": Retraction of Förster (2009). (2016). *Journal of Experimental Psychology: General*, *145*, 265. doi:10.1037/a0040143

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.

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

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Sharp, D., Cole, M., & Lave, C. (1979). Education and cognitive development: The evidence from experimental research. *Monographs of the Society for Research in Child Development*, *44*, serial no. 148, nos. 1–2.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75* (13, Whole No. 517).

Simmons, S., & Estes, Z. (2008). Individual differences in the perception of similarity and difference. *Cognition*, *108*, 781–795. doi:10.1016/j.cognition.2008.07.003

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

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, *22*, 377–386.

Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 525–538.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*, 440–463. doi:10.1037/a0018963

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 694–709.

Wills, A. J., Inkster, A. B., & Milton, F. (2015). Combination or differentiation? Two theories of processing order in classification. *Cognitive Psychology*, *80*, 1–33. doi:10.1016/j.cogpsych.2015.04.002

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *Quarterly Journal of Experimental Psychology*, *66*, 299–318. doi:10.1080/17470218.2012.708349

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

SM S1: These are the final versions of the 10 displays, as described in the Materials section of Experiment 4. They are presented in two versions, differing in the test item.