The Dissertation Committee for Jonathan R. Rein certifies that this is the approved version of the following dissertation:

Reevaluating the Determinants of Category-based Induction

Committee:

_____
Arthur Markman, Supervisor

_____
Zenzi Griffin

_____
Cristine Legare

_____
Jeffrey Loewenstein

_____
Todd Maddox

Reevaluating the Determinants of Category-based Induction

by

Jonathan R. Rein, B.A.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2010

Reevaluating the Determinants of Category-based Induction

Jonathan R. Rein, Ph.D.

The University of Texas at Austin, 2010

Supervisor: Arthur Markman

What makes one more or less likely to project a novel property from an item to that item's broader category? Research on category-based induction has documented a consistent typicality effect: typical exemplars promote stronger inferences than atypical exemplars. This work has been largely confined to categories whose central tendencies are the most typical members of the category. Experiments 1 and 2, using natural and artificial categories, showed that central tendencies have greatest induction strength even for categories that are best represented by ideal exemplars. Experiments 3-7 investigate the role of familiarity in induction. Experiments 3 and 4 directly contrast statistical averageness against familiarity through category learning procedures. Experiment 5 creates this contrast through frequency differences across stimuli. Experiments 6 and 7 investigate how the familiarity advantage found in Experiments 3-5 can be modified through fluency manipulations, independent of actual experience. Taken together, these studies suggest that category-based induction is driven largely by a familiarity heuristic.

Table of Contents

1. Introduction

To navigate the ever-changing world effectively, people must generalize from limited experience to a potentially infinite set of related but not-yet-experienced objects and events. One way to use this experience efficiently is to make inferences about entire categories, not just individual items. For instance, consider a homeowner who notices that the sparrows that live in her trees have been eating more of the new and improved birdseed that she purchased. She could reasonably infer not only that these specific sparrows will continue to prefer the new seed, but also that all sparrows and possibly all birds would as well. This kind of predictive inference is a primary function of category knowledge (Markman & B. H. Ross, 2003; Osherson, E. E. Smith, Wilkie, López, & Shafir, 1990).

There are many factors that influence the strength of category-based property induction. These include the number of items known to have the property, the variability of the category one is generalizing to, the kind of information made salient by the property, and the causal relationships between categories and properties (Heit, 2000). In Experiments 1 and 2 of this dissertation, I focused on the impact of how the item one is generalizing from relates to its broader superordinate category. In particular, I examined how an item's typicality—how good an example it is of its category—affects induction. In the example above, the homeowner may generalize from sparrows to all birds because sparrows are highly representative of the bird category. In contrast, if she observes

vultures enjoying the seed instead of sparrows, she will be much less likely to generalize to all birds because vultures are considered a fairly poor example of the bird category.

Many studies have documented a robust typicality effect in category-based induction. However, this work has been largely confined to categories whose central tendencies are also the best examples of the category. Experiments 1 and 2 extend this work to role-governed categories, which are best represented by ideal members (Goldwater, Stilwell, & Markman, 2009). In these studies, induction strength was greatest for the central tendency exemplars, regardless of whether the central tendency or the ideal was rated more typical. These results suggest that the so-called "typicality" effect is a special case of a more universal central tendency effect in category-based induction.

Why do central tendency exemplars universally have this advantage in induction? What might this advantage tell us about the processes that support induction? One possibility is that central tendencies are special because of their unique position at the statistical center of the category distribution. This is the account favored by all extant computational models of category-based induction and is consistent with a large body of evidence suggesting that induction is essentially statistical reasoning. Another related possibility is that central tendencies are special because they are the most familiar category members. On this account, induction can also be thought of as a heuristic process. Experiments 3 through 7 provide evidence for this latter view.

Experiment 3 directly contrasted familiar versus central tendency stimuli using a category discrimination learning procedure. Specifically, familiar prototype stimuli of

individual subordinate categories were contrasted against central tendency stimuli that contained features of the entire superordinate category. Experiment 4 created a similar contrast between familiar extreme values and unfamiliar but central values following continuous-dimension category learning. Experiment 5 created a similar contrast using differential frequency of stimuli within a single category. In all of these studies, participants judged the familiar stimuli to have greater induction strength than the central tendency stimuli. Experiment 6 fostered the familiarity advantage in induction solely by manipulating processing fluency to create the *feeling* of familiarity, independent of actual experience. Experiment 7 showed that this fluency effect is dependent on the reliability of the fluency cue as a marker of familiarity All of these demonstrations of familiarity advantages will indicate that category-based induction is governed largely by a simple familiarity heuristic.

Chapter II: The Source of the Typicality Effect in Induction

Rips (1975) was the first to document the typicality effect in category-based induction. In his study, participants were asked to imagine an island that contained a small set of bird or mammal species. They were told that a particular species had a contagious disease and were asked to estimate the percentage of each other species that would share the disease. Estimates were significantly higher when the given instance was a typical category member, as measured by its distance from the superordinate category label in a multidimensional space. For example, when told that sparrows had the disease, people estimated that 32% of geese also had the disease. In contrast, they judged that only 17% of geese would have the disease when they were told that eagles had it. Sparrows and eagles are approximately equal in their similarity to geese, but sparrows had greater inductive strength because they are considered better examples of birds overall.

Another classic examination of category-based induction is Osherson et al.'s (1990) work on argument evaluation. These authors found that typicality has a substantial effect on judgments of inductive arguments. For example, participants were asked to choose which of the following is a stronger argument:

(1)

Robins have a higher potassium concentration in

their blood than humans.

Therefore, all birds have a higher potassium concentration in

their blood than humans.

(2)

Penguins have a higher potassium concentration in

their blood than humans.

Therefore, all birds have a higher potassium concentration in

their blood than humans.

Over ninety percent of people choose the robin argument.

This increase in induction strength for better category examples is a robust phenomenon. The typicality effect has been documented in multiple tasks, like those just discussed; with multiple domains of category knowledge (Rothbart & Lewis, 1988); and for multiple kinds of subject populations, including children (López, Gelman, Gutheil, & E. E. Smith, 1992; Rhodes, Brickman, & Gelman, 2008), indigenous Mayans in Guatemala (López, Atran, Coley, Medin, & E. E. Smith, 1997), and Alzheimer's disease patients (E. E. Smith, Rhee, Dennis, & Grossman, 2001). While varied in several respects, these studies have one thing in common. They have all focused on a particular class of categories: what Markman and Stilwell (2001) refer to as "feature-based" categories.

Feature-based categories are those, like *bird*, that are represented by their properties or dimension values. There have been several proposals for exactly how feature-based categories are represented, such as summary representations of average or characteristic values (Posner & Keele, 1968; Rosch & Mervis, 1975), the set of unique property combinations for each exemplar in the category (Medin & Shaffer, 1978; Nosofsky, 1986), and networks of causally related features (Rehder & Kim, 2006; Sloman, Love, & Ahn, 1998). For each of these, the category label denotes the set of features that are associated with that category. In other words, feature-based category representations contain information that is subordinate to, or part of, the category.

For these categories, an exemplar is considered to be a good or typical member of the category to the extent that it has features that are characteristic of the category (Rosch & Mervis, 1975). These exemplars are closest to the central tendency of all category members. For example, swallows have many of the characteristic features of birds: they have wings, fly, build nests, have a common shape and size, etc. They also do not have any features that are highly uncommon of birds, as penguins do. These central tendency exemplars are privileged in many ways beyond being considered the best examples of their category and promoting the strongest induction. For example, these category members are classified faster (Rips, Shoben, & E. E. Smith, 1973) and with more certainty (Dale, Kehoe, & Spivey, 2007), brought to mind more readily (Mervis, Catlin, & Rosch, 1976), and learned more quickly (Rosch, Simpson, & Miller, 1976) than are less central exemplars.

Although feature-based categories have been the primary object of study in psychological research on concepts, there are other types of categories that have different determinants of typicality. Role-governed categories, for instance, are qualitatively different from feature-based categories (Gentner & Kurtz, 2005; Markman & Stilwell, 2001). A role-governed category label refers to the role that an object plays in a larger relational system that connects multiple categories. Consider the relational system depicted in Figure 1.

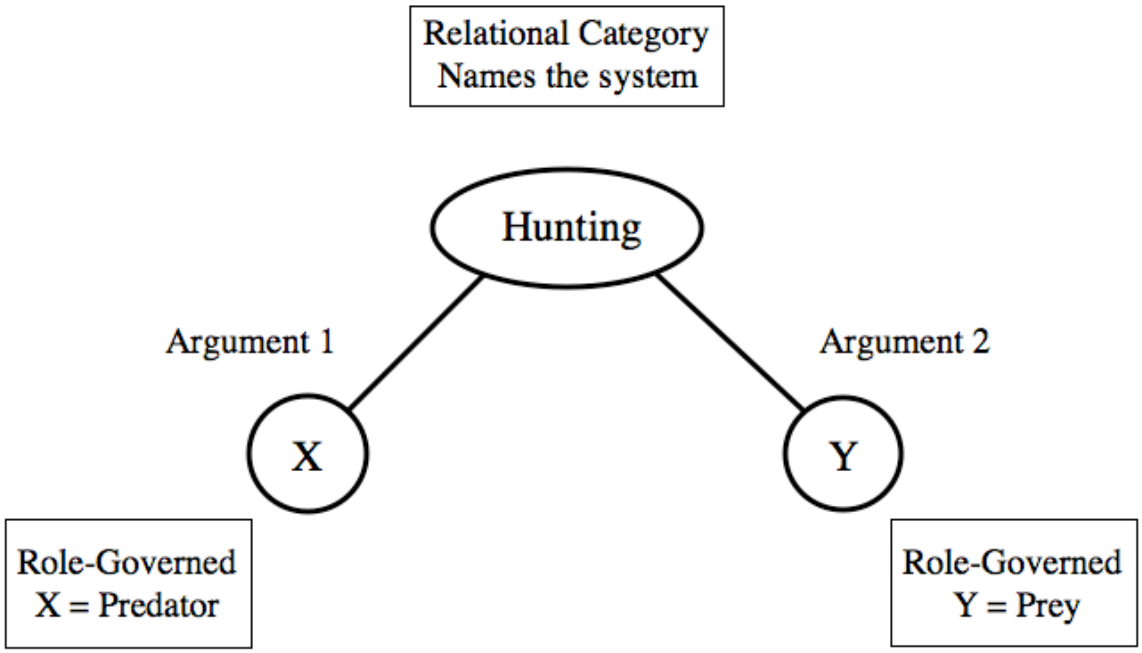


Figure 1. Relational structure of *hunting* category. Arguments in the system are role-governed categories.

The relational category *hunting* names the entire relational structure of objects, properties, and their interactions. This kind of structured representation is representative of verbs in general (Gentner, 1982). Role-governed categories are arguments in this relational structure. In order for an object to qualify as a *predator*, it must play the first

role in the relation "X hunts Y to satisfy a need." X is a free variable; anything that plays this role is a predator, regardless of its feature composition. Lions and spiders are both predators, despite having drastically different features, because they both play this role.

While there are some features that many predators share, these features are not relevant for an object's membership in the *predator* category. Furthermore, the common features of role-governed categories are generally relational in nature. Barr and Caplan (1987) distinguished between two kinds of features: intrinsic and extrinsic. Intrinsic features are those inherent to an object, such as "has a head" for the category *hammer*. Extrinsic features are those that refer to the object's relations to other objects, such as "pounds nails". Goldwater et al. (2009) found that the majority of the features of role-governed categories have this explicitly relational character.

Though they have been underrepresented in the research literature, role-governed categories are abundant. Of the 100 most common nouns, approximately half denote categories of roles or relations (Gentner & Kurtz, 2005). Classes of role-governed categories that have been studied include verb-specific thematic roles (McRae, Ferretti, & Amyote, 1997), event categories (Klein, 1998) abstract coherent categories (Rehder & B. H. Ross, 2001), ad hoc categories (Barsalou, 1983), and goal-derived categories (Barsalou, 1985).

Considering the differences between feature-based and role-governed categories, one would expect differences in the nature of the best examples of these categories. Role-governed categories are not represented as bundles of correlated common features, but as elements that fit into a relation or serve a function. Consequently, a good example

of a role-governed category is not an exemplar near the central tendency of the category, but rather an exemplar that best fits its relation or serves its function. Consider the category *diet food*. The central tendency exemplar of a diet food is one with an average amount of calories and a mediocre taste. In contrast, the best example of a diet food is one with zero calories and a great taste, even though this particular exemplar may not even exist. This kind of ideal exemplar is considered the most typical[1] for a variety of relational and goal-relevant categories (Barsalou, 1985; Borkenau, 1990; Davis & Love; Goldstone, 1996; Goldstone, Steyvers, & Rogosky, 2003; Goldwater, Stilwell, & Markman, 2008; Levering & Kurtz, 2006; Lynch, Coley, & Medin, 2000).

This dissociation in typicality between feature-based and role-governed categories raises an interesting question. Is the typicality effect in category-based induction that has been documented extensively with feature-based categories a result of the enhanced inductive strength of typicality *per se* or proximity to the central tendency? Studies of

[1] "Typicality" has generally been treated in the literature as synonymous with "representativeness". Since Rosch's early work, both of these constructs have traditionally been assessed by goodness-of-example ratings. As a result, researchers describe ideal examples as most typical when they are rated as the best example of their category, even though they are not typical in the "average" sense. Following these conventions, "typicality", "representativeness", and "example goodness" are used interchangeably while emphasizing that the third is best thought of as an operationalization of the first two.

feature-based categories alone cannot resolve this question because for these categories, an exemplar's representativeness and its proximity to the central tendency are highly, if not perfectly, correlated. Fortunately, these are decoupled in role-governed categories for which the ideal, not the central tendency is most typical.

There are two straightforward possibilities. If there is truly a typicality effect in category-based induction, then central tendency exemplars will promote stronger induction than ideal exemplars for feature-based categories, while ideal exemplars will promote stronger induction than central tendency exemplars for role-governed categories. In contrast, if the "typicality" effect is actually a central tendency effect, central exemplars will promote stronger induction for both kinds of categories, regardless of whether they are more typical.

In Experiments 1 and 2, I evaluated the relative contribution of typicality and central tendency to category-based induction. In Experiment 1, participants made judgments about example goodness and induction strength for ideal and central tendency exemplars of natural feature-based and role-governed categories. In Experiment 2, the learning context of artificial categories was manipulated to foster feature-based or role-governed representation, again measuring how idealness and central tendency influence example goodness and induction strength. These experiments were conducted to determine whether the typicality effect is truly a function of typicality or central tendency

Experiment 1

Goldwater et al. (2008) documented several differences between common feature-based and role-governed categories. In one study, participants chose either an ideal or

central tendency exemplar as the best example to explain the category to someone who was unfamiliar with the category. As expected, people chose significantly more ideals to explain role-governed than feature-based categories. The present study is designed to replicate this dissociation in typicality judgments and to evaluate whether typicality or central tendency is the main locus of induction strength.

Method

Participants

46 University of Texas at Austin students participated for course credit or payment of $8.

Materials

All stimuli were adapted from Goldwater et al. (2008). There were 12 role-governed categories: *guest, job, game, predator, hobby, gift, drug, customer, home, author, friend, pet*; and 12 feature-based categories: *television, chair, cell phone, fridge, truck, beer, website, shoes, knife, table, bicycle, microwave*. These were all artifacts. While natural kinds—particularly plants and animals— are more often used in studies of category-based induction (e.g., Osherson et al. 1990), artifacts were used because they are a stricter control for role-governed categories than natural kinds. Artifacts have functions and therefore have intuitive and salient ideals.

For each category, there was an ideal and central tendency exemplar. Goldwater et al. obtained these exemplars by having separate groups of participants list the attributes of an ideal and average member of each category. Ideal exemplars were defined by the five most common attributes listed by the ideal group but not the central tendency group,

and likewise for the central tendency exemplars. There were 48 total exemplars. See Appendix A for the entire set of categories and exemplars.

Procedure

Participants were tested at individual computers. Each participant made example goodness and induction strength judgments for all 48 exemplars. Trials were blocked by judgment task. For the following task descriptions, I will use the category *predator* as an example.

The following instructions preceded the example goodness block:

> In this section of the study, we are interested in how people think of individual items as representative of the entire category to which the items belong. You will be presented with a series of items, one at a time. Each item has a list of five attributes. For each item, you will judge how good of an example of its category that item is. You will answer by selecting a point on a scale with your mouse. Although you'll see multiple members of each category, try to make each of your judgments independently from the rest.

On each trial of the exemplar goodness task, there was an instruction at the top of the screen to "Imagine a predator has the following properties." The five properties of the exemplar appeared below. For the ideal predator exemplar, these properties were "Smart", "Strong", "Agile", "Cunning", and "Camouflaged". Beneath the exemplar properties was the question "How good of an example of authors is this?" There was an eleven-point scale at the bottom of the screen with the lowest point labeled "poor

example", the midpoint labeled "moderate example", and the highest point labeled "excellent example". Participants clicked on a point on the scale with the mouse.

The following instructions preceded the induction strength block:

> In this section of the study, we are interested in how people generalize from an individual item to the entire category to which the item belongs. You will be presented with a series of items, one at a time. Each item has a list of five attributes. For each item, you will be asked to imagine that the item has some additional attribute and to estimate the percentage of the rest of the category that also has that attribute. You will answer by selecting a point on a scale with your mouse. Although you'll see multiple members of each category, try to make each of your judgments independently from the rest.

The induction strength task was very similar to the example goodness task. On each trial, participants were given an instruction of the form "Imagine that a predator with the following properties also has property X". "Property X" is an example of what is referred to in the induction literature as a *blank predicate*, in the sense that it has no meaningful content for the participant. This removes any role of background causal knowledge. The five properties of the exemplar were listed, followed by the question "What percentage of other predators have property X?" There was an eleven-point scale at the bottom of the screen with the lowest point labeled "0%", the midpoint labeled "50%", and the highest point labeled "100%".

Order of block presentation was counterbalanced. Exemplar order within a block was randomized, with two restrictions. First, the initial 24 trials in a block were exemplars of all 24 categories, as were the latter 24 trials. Second, within each of these sets, 12 of the exemplars were ideals and 12 were central tendencies. Before each block, participants were instructed to make each judgment independent from the other judgments.

Results

The mean rating was calculated for each category type (feature-based vs. role-governed), exemplar type (central tendency vs. ideal), and judgment task (goodness vs. induction), collapsing across participants and items for separate analyses. Table 1 contains the mean ratings for each category. Figure 2 shows the mean goodness ratings collapsed across participants. Figure 3 shows the mean induction ratings, collapsed across participants. Because the main question of interest is whether the relationship between category type and exemplar type is the same or different across tasks, I analyzed each task separately with 2×2 repeated-measures analyses of variance (ANOVA) on participant and item data. For the interested reader, results from the full 2×2×2 ANOVAs are in Appendix B. For each effect discussed below, the result of the participant analysis is first, followed by the item analysis. In order to generalize to the broader population of participants and items, treating both as random effects simultaneously, min F' was also calculated for all effects (Clark, 1973).

Table 1
Example Goodness and Induction Strength Ratings by Category

		Goodness		Induction	
		Central	Ideal	Central	Ideal
Role-governed	author	6.4	7.2	6.0	5.0
	customer	4.3	6.5	5.0	4.7
	drug	6.1	4.9	5.4	4.2
	friend	8.2	8.3	6.0	5.7
	game	7.6	6.7	6.8	5.5
	gift	7.4	6.3	6.6	4.8
	guest	6.3	7.1	6.1	5.6
	hobby	5.7	7.1	5.7	5.9
	home	7.4	6.4	7.0	4.4
	job	5.1	6.8	6.1	4.4
	pet	7.3	7.8	6.8	5.8
	predator	6.9	7.7	6.4	6.3
	Mean	6.6	6.9	6.1	5.2
Feature-based	beer	8.0	4.4	7.3	4.3
	bicycle	8.5	5.6	7.4	5.3
	cell phone	7.5	5.9	6.8	5.1
	chair	7.6	6.4	5.9	5.3
	knife	7.6	6.9	6.4	5.8
	microwave	7.6	6.7	6.7	5.4
	refrigerator	7.8	7.0	7.0	5.8
	shoes	6.7	6.0	6.0	4.6
	table	7.1	6.5	5.8	5.7
	television	7.6	6.2	7.2	4.9
	truck	7.4	6.4	6.4	5.1
	website	7.5	7.2	7.3	5.5
	Mean	7.6	6.3	6.7	5.2

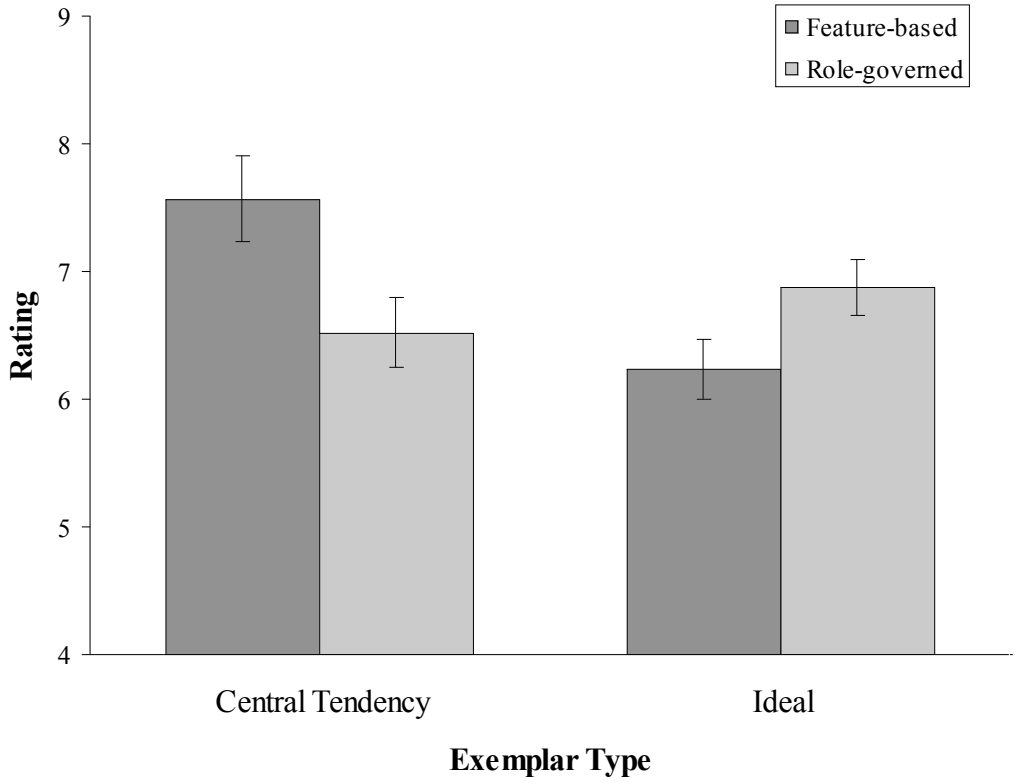


Figure 2. Experiment 1 example goodness judgments collapsed across participants. Error bars indicate ± one SEM.

Analyzing goodness ratings alone, there was no significant main effect of category type, $F(1,45) = 2.93$, *ns*; $F(1,22) = .52$, *ns*; *min* $F'(1,30) = .44$, *ns*. There was no main effect of exemplar type by participants, $F(1,45) = 2.86$, *ns*. This effect did reach significance by items, $F(1,22) = 5.71$, $p = .026$, $\eta^2 = .07$; though it was not significant across participants and items, *min* $F'(1,67) = 1.91$, *ns*. There was, however, a significant interaction between the category type and exemplar type, $F(1,45) = 18.47$, $p < .001$, $\eta^2 = .10$; $F(1,22) = 16.94$, $p < .001$, $\eta^2 = .20$; *min* $F'(1,57) = 8.84$, $p = .004$. Central tendency exemplars were considered better examples of feature-based categories ($M = 7.6$) than role-governed categories ($M = 6.5$), $F(1,45) = 19.64$, $p < .001$, $\eta^2 = .44$; $F(1,22) = 8.62$, p

= .008, η^2 = .28; *min F'*(1,42) = 5.99, p = .02. In contrast, ideal exemplars were considered better examples of role-governed than feature-based categories, $F(1,45)$ = 8.37, p = .006, η^2 = .19; though this difference was only marginally significant by items, $F(1,22)$ = 3.59, p = .071, η^2 = .14; *min F'*(1,41) = 2.51, *ns*.

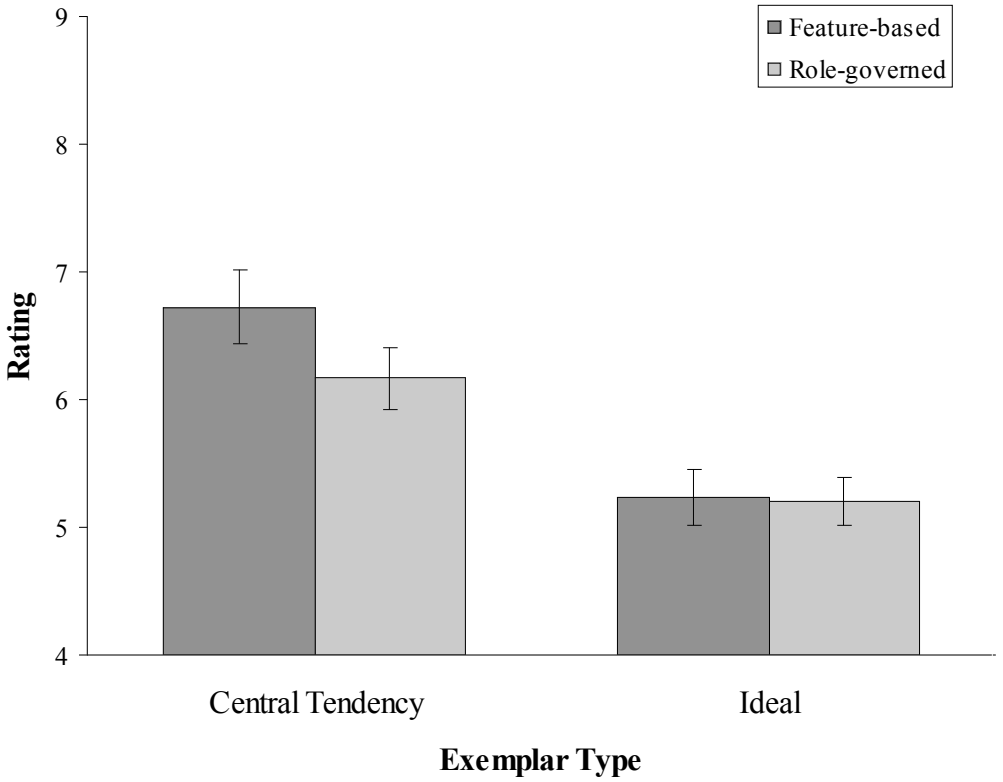


Figure 3. Experiment 1 induction strength judgments collapsed across participants. Error bars indicate ± 1 SEM.

The induction rating behavior was much different. There was a main effect of category type, $F(1,45)$ = 5.39, p = .025, η^2 = .02; though this did not reach significance by items, $F(1,22)$ = 2.26, *ns*; *min F'*(1,41) = 1.59, *ns*. There was also a main effect of exemplar type, $F(1,45)$ = 35.22, $p < .001$, η^2 = .27; $F(1,22)$ = 68.27, $p < .001$, η^2 = .50;

min F'(1,67) = 23.23, $p < .001$. However, there was no significant interaction, $F(1,45) = 2.67$, *ns*; $F(1,22) = 3.12$, *ns*; *min F'*(1,62) = 1.44, *ns*. For feature-based categories, central tendency exemplars were considered stronger bases for induction than ideals, $F(1,45) = 21.83$, $p < .001$, $\eta^2 = .48$; $F(1,22) = 46.40$, $p < .001$, $\eta^2 = .68$; *min F'*(1,67) = 14.84, $p < .001$. For role-governed categories, central tendency exemplars ($M = 6.2$) also had higher induction ratings than ideals ($M = 5.2$), $F(1,45) = 26.18$, $p < .001$, $\eta^2 = .58$; $F(1,22) = 12.38$, $p = .002$, $\eta^2 = .36$; *min F'*(1,43) = 8.41, $p = .006$.

Overall, there was a positive correlation between goodness ratings and induction ratings for individual items, $r(48) = .68$. It is informative to break this correlation down by category type. There was a significant correlation for role-governed categories, $r(24) = .47$. There was also a significant correlation for feature-based categories $r(24) = .89$, $p < .001$; *min F'*(1,103) = 6.52, $p = .01$. As expected, the correlation between goodness and induction ratings was stronger for feature-based than role-governed categories, $z = 2.95$, $p = .003$.

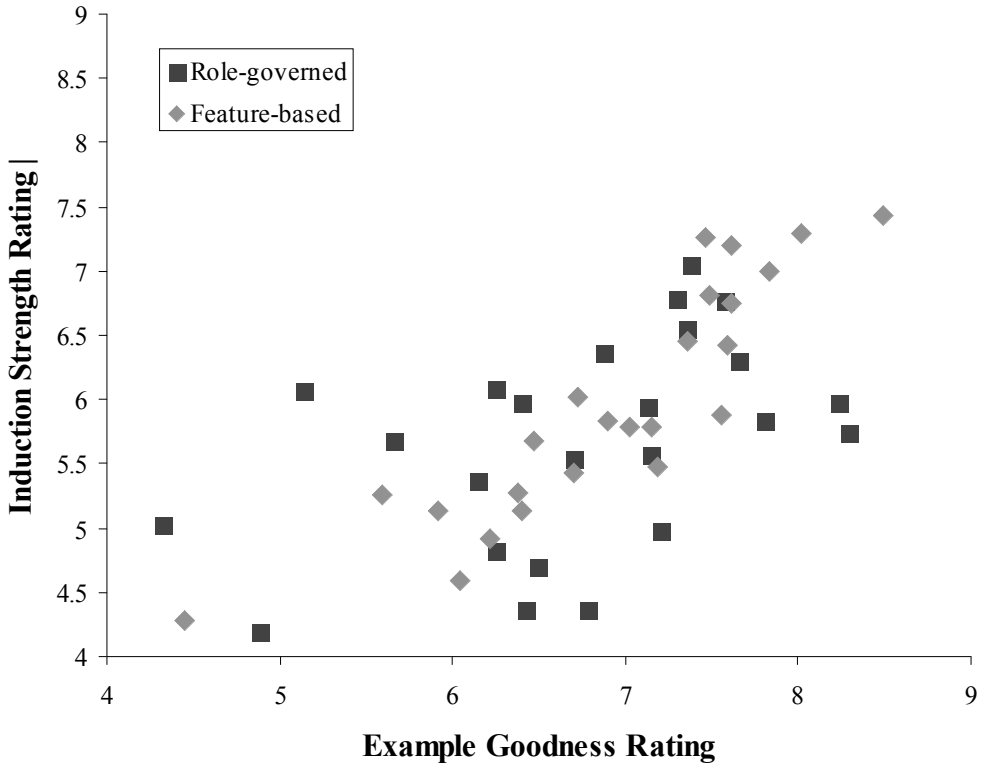


Figure 4. Scatterplot of example goodness and induction strength ratings.

Discussion

Participants' judgments of goodness of example for the different exemplar types mirrored previous work. Ideal exemplars were considered to be much better examples of role-governed categories than feature-based categories. In contrast, central tendency exemplars were considered better examples of feature-based than role-governed categories.

A different pattern was observed for induction strength. Participants rated central tendency exemplars as promoting stronger induction for both feature-based and role-governed categories. For feature-based categories, this replicates the canonical typicality effect. For role-governed categories, however, the most representative exemplars do not

promote the strongest induction. This suggests that representativeness *per se* is not in fact responsible for the typicality effect.

Although not central to the main points of this dissertation, it is interesting to note that induction ratings were lower for role-governed categories than feature-based categories overall. This may be because categories that are represented primarily with relational properties exhibit less within-category similarity (Barr & Caplan, 1987; Gentner & Kurtz, 2005). Role-governed category members all share an identical role, but they typically do not all share a bundle of correlated attributes like feature-based category members do. This within-category coherence has been shown to influence inductive strength (Patalano, Chin-Parker, & B. H. Ross, 2006). Care should be taken drawing conclusions from this finding, however, because it was not statistically reliable in the items analysis.

Experiment 1 provides preliminary evidence that category-based induction is actually influenced principally by exemplars' proximity to their categories' central tendency. However, this study used natural categories and therefore leaves the causality underlying this effect ambiguous. Role-governed categories differ in several ways from feature-based categories (Goldwater et al., 2008), and any one of these might have overridden the true typicality effect. Experiment 2 addresses this ambiguity by using an artificial category learning task.

Experiment 2

In this study, I adopted a method used by Barsalou (1985) to foster feature-based or role-governed representation of identical category information simply by manipulating

the learning context. Participants learned to classify two categories of teachers, with each teacher exemplar defined by five spare-time activities in which they engage. Each category had a defining activity. For one category, all of the exemplars had the activity *reads the newspaper*, while all of the members of the other category had the activity *jogs*. Individual exemplars varied in how often they engaged in this defining activity: *daily*, *weekly*, or *monthly*. In addition to this defining activity, each of the categories had three characteristic activities that predicted membership in that category but not the other. For instance, one category's members would tend to have the activities *invests in stocks*, *writes poetry*, and *goes to the movies*, while members of the other category would generally have *plays chess*, *renovates houses*, and *cooks Mediterranean food*. Individual exemplars varied in how many characteristic activities they had: 3, 2, or 1. Critically, one group of participants learned these categories with role-relevant labels: "current events teachers" and "physical education teachers", who read the newspaper and jog, respectively. Another group of participants learned to distinguish the same exemplars under role-irrelevant labels: "Q programming language teachers" and "P programming language teachers".

Unlike the natural categories in Experiment 1, in which idealness and central tendency were mutually exclusive, this design manipulates both independently. Barsalou (1985) found that each has an influence on judgments of exemplar goodness, depending on label relevance. When participants learned the categories with role-relevant labels, exemplars' frequency of performing the defining activity determined how good of an example it was. Teachers who jogged daily were considered better examples of physical

education teachers than those who jogged monthly. In other words, ideal exemplars that fit their role best had the highest typicality. For these role-relevant learners, an individual exemplar's number of characteristic features had little to no effect.

The relative influence of ideals and central tendency was reversed for role-irrelevant learners. An exemplar's frequency of the defining activity had no effect on category goodness judgments. Jogging has no obvious impact on how well Q programming language teachers fill their functional role, so jogging daily does not make one any better of a programming teacher than jogging monthly. In contrast, the number of characteristic activities an exemplar had did influence goodness judgments. Without functional information, participants relied on statistical similarity. The programming teachers were essentially feature-based categories. As has been documented extensively, the most typical exemplars for these categories are those closest to the central tendency.

Experiment 2 replicates Barsalou's (1985) work on typicality, extending it to category-based induction. As in Experiment 1, if an exemplar's representativeness of its category is truly the determinant of induction, then induction strength ratings should map onto example goodness ratings. If, however, exemplars' proximity to the category central tendency is the determinant of induction, then exemplars with more characteristic features should promote stronger induction, even if they are not considered most representative of their category.

This experiment also allows us to test anti-ideals—category members that fit their role very poorly. A physical education teacher who only jogs monthly, for example, is an anti-ideal. In Experiment 1, there were only ideal and central tendency exemplars, so the

inductive potential of truly awful examples of role-governed categories could not be assessed. It is possible that these are the exemplars that promote the strongest induction. Consider the following two arguments:

(3)

Poodles can bite through barbed wire.

Therefore, German shepherds can bite through barbed wire.

(4)

Dobermans can bite through barbed wire.

Therefore, German shepherds can bite through barbed wire.

Despite the greater similarity between German shepherds and Dobermans, people consider argument 3 stronger (E. E. Smith, Shafir, & Osherson, 1993). This reasoning is intuitively straightforward. Poodles appear to very poor fits for the role of barbed wire biter. If a dog like that can bite through barbed wire, then a German shepherd must be able to. It is possible that category-based induction of blank predicates for role-governed categories is similar to this kind of functional, causal reasoning about feature-based categories.

Method

Participants

Sixty-five University of Texas at Austin students participated for course credit or payment of $8. 32 were in the Role-Irrelevant condition, 33 in the Role-Relevant condition.

Materials

Each category consisted of 9 exemplars. For each level of the defining activity (*daily*, *weekly*, and *monthly*) there was an exemplar with each level of characteristic activities (3, 2, and 1). The 3 characteristic activities were equally diagnostic of category membership, with each appearing in 6 of the 9 exemplars. Every exemplar had 1-3 filler activities, yielding 5 activities per exemplar. There were 18 such filler activities, with each activity appearing in a single member of each category. For example, one category might have the defining activity *reads the newspaper* and the characteristic activities *plays chess*, *renovates houses*, and *cooks Mediterranean food*. In this category, a high-defining/high-central tendency exemplar would have the following properties: *reads the newspaper daily, plays chess, renovates houses, cooks Mediterranean food,* and *goes fishing*. This last property is a filler activity that is true of only this exemplar and one other exemplar in the contrast category. A low-defining/low-central tendency exemplar would have the following properties: *reads the newspaper monthly, renovates houses, plays the drums, rides a motorcycle,* and *goes to yard sales*. This exemplar has 3 filler activities.

Procedure

Participants were tested at individual computers. They were told to imagine that they worked for a personnel agency, where they were learning to use information about individuals' spare-time activities to predict the course that they would be best suited to teach. In the Role-Irrelevant condition, they were told that they would be distinguishing between Q programming language teachers and P programming language teachers. In

the Role-Relevant condition, they were told that they would be distinguishing between current events teachers and physical education teachers.

Prior to the classification phase, participants were familiarized with the exemplars from each category. During this familiarization phase, a category label appeared at the top of the screen, with the five activities describing all nine exemplars presented below. Participants studied the exemplars for a minimum of two minutes, then viewed all of the exemplars from the other category. Order of category presentation was randomized.

On each trial of the classification phase, an instruction appeared at the top of the screen, telling participants to "Choose the better teacher of [*course*];" with each course label appearing in half of the trials. An exemplar from each category appeared below this instruction, one on the left and one on the right. Categories' position was randomized across trials. Participants pressed the "Z" or "/" button on the keyboard to select the left or right exemplar. After each classification choice, "Correct" or "Incorrect" appeared at the bottom of the screen, the correct exemplar was highlighted by a white box, and a pleasant tone or harsh buzzer sound played. Participants studied the correct exemplar for eight seconds, then were presented with a single activity and were asked if it had been true of the correct exemplar. On half of the trials this activity was true of the correct exemplar, on the other half it was true of the incorrect exemplar. Because participants could correctly classify all of the exemplars using only the defining activity, the memory probe and the familiarization phase were included to promote complete knowledge of the categories, including the characteristic activities. There were fifty-four total classification trials, three each for all eighteen exemplars.

Following classification, participants provided example goodness and induction strength judgments for all eighteen exemplars. As in Experiment 1, these judgments were blocked, with block order counterbalanced across participants. Exemplar presentation within blocks was randomized. The presentation of each judgment trial was identical to Experiment 1, except induction judgments referred to "activity X" instead of "property X".

Results

The mean of participants' ratings across exemplars for each level of idealness on the defining activity (*daily*, *weekly*, and *monthly*), characteristic activities (3, 2, and 1), judgment task (goodness vs. induction), and label relevance (irrelevant vs. relevant) was calculated. Figures 5 and 6 show the means across participants for goodness and induction ratings, respectively. As in Experiment 1, I analyzed the data for each task in each labeling condition separately with 3×3 repeated-measures analyses of variance (ANOVA). The full 4-way ANOVA table can be found in Appendix B.

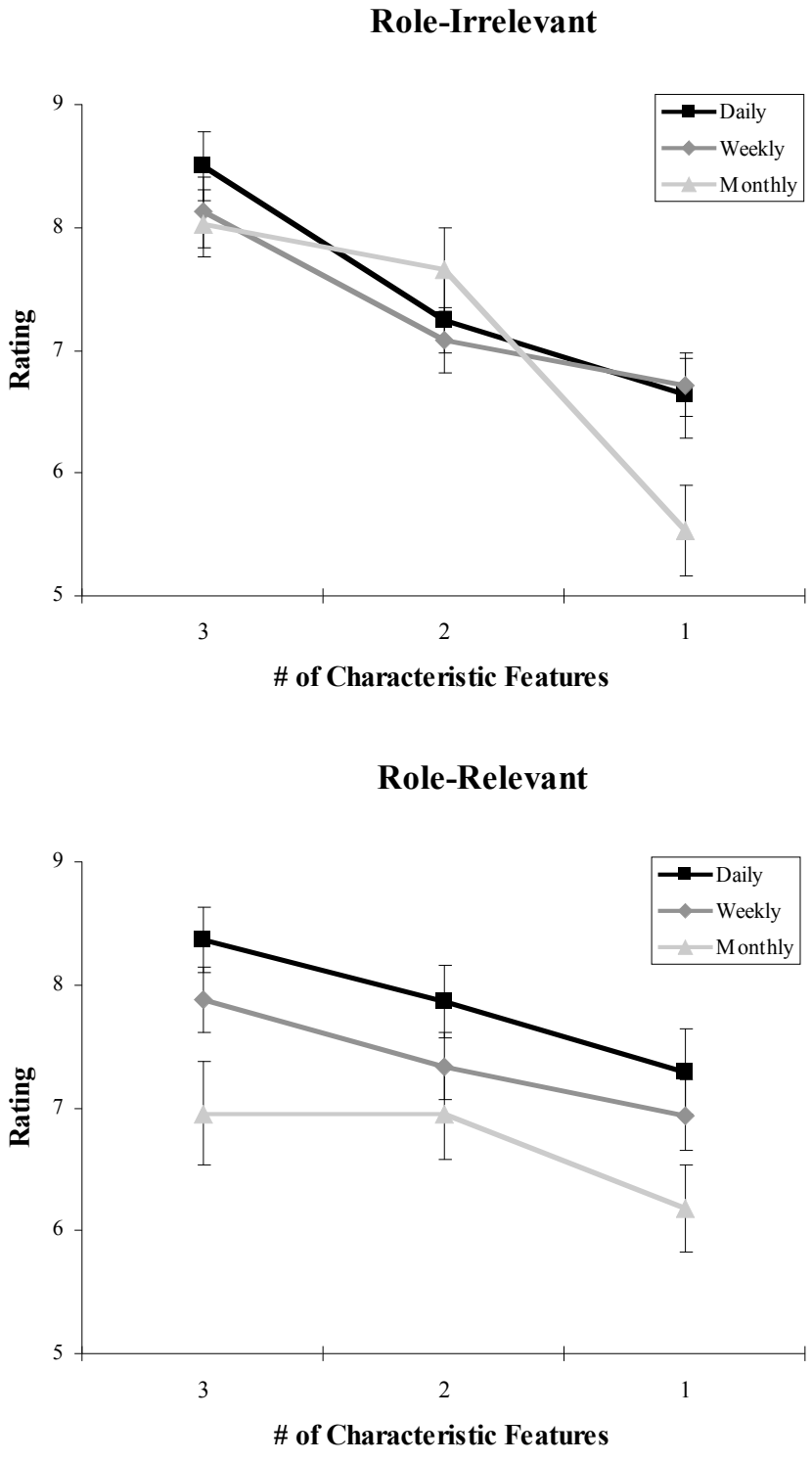


Figure 5. Experiment 2 example goodness judgments. Error bars indicate ± 1 SEM.

For example goodness judgments following role-irrelevant learning, more characteristic activities—i.e. greater central tendency—led to higher ratings, $F(2,62) = 32.74$, $p < .001$, $\eta^2 = .25$; this was a significant linear trend, $F(1,31) = 40.81$, $p < .001$, $\eta^2 = .25$. There was no effect of ideal level overall, $F(2,62) = 1.76$, *ns*. There was, however, an interaction between central tendency level and ideal level $F(4,124) = 4.57$, $p = .002$, $\eta^2 = .04$. This effect appears to be due to the exceptionally low ratings for exemplars with only one characteristic activity and infrequent performance of the defining activity.

Example goodness judgments following role-relevant learning were much different. As in the role-irrelevant condition, ratings increased as central tendency increased, $F(2,64) = 11.84$, $p < .001$, $\eta^2 = .06$; this was a significant linear trend $F(1,31) = 15.8$, $p < .001$. However, this effect was substantially smaller than in the role-irrelevant condition, as evidenced by the relative effect sizes as well as a significant 2-way interaction between central tendency and label relevance in the context of a 3-way (central tendency × ideal level × label relevance) ANOVA on goodness judgments alone, $F(2,126) = 5.02$, $p = .008$, $\eta^2 = .04$. There was also an effect of ideal level, with more ideal exemplars rated as better examples, $F(2,64) = 6.09$, $p = .004$, $\eta^2 = .09$; this was a significant linear trend $F(1,31) = 6.65$, $p = .015$, $\eta^2 = .09$. There was no significant interaction between these variables, $F(4,128) = .75$, *ns*. This increased influence of ideals on example goodness judgment following role-relevant learning is a replication of Barsalou's (1985) findings.

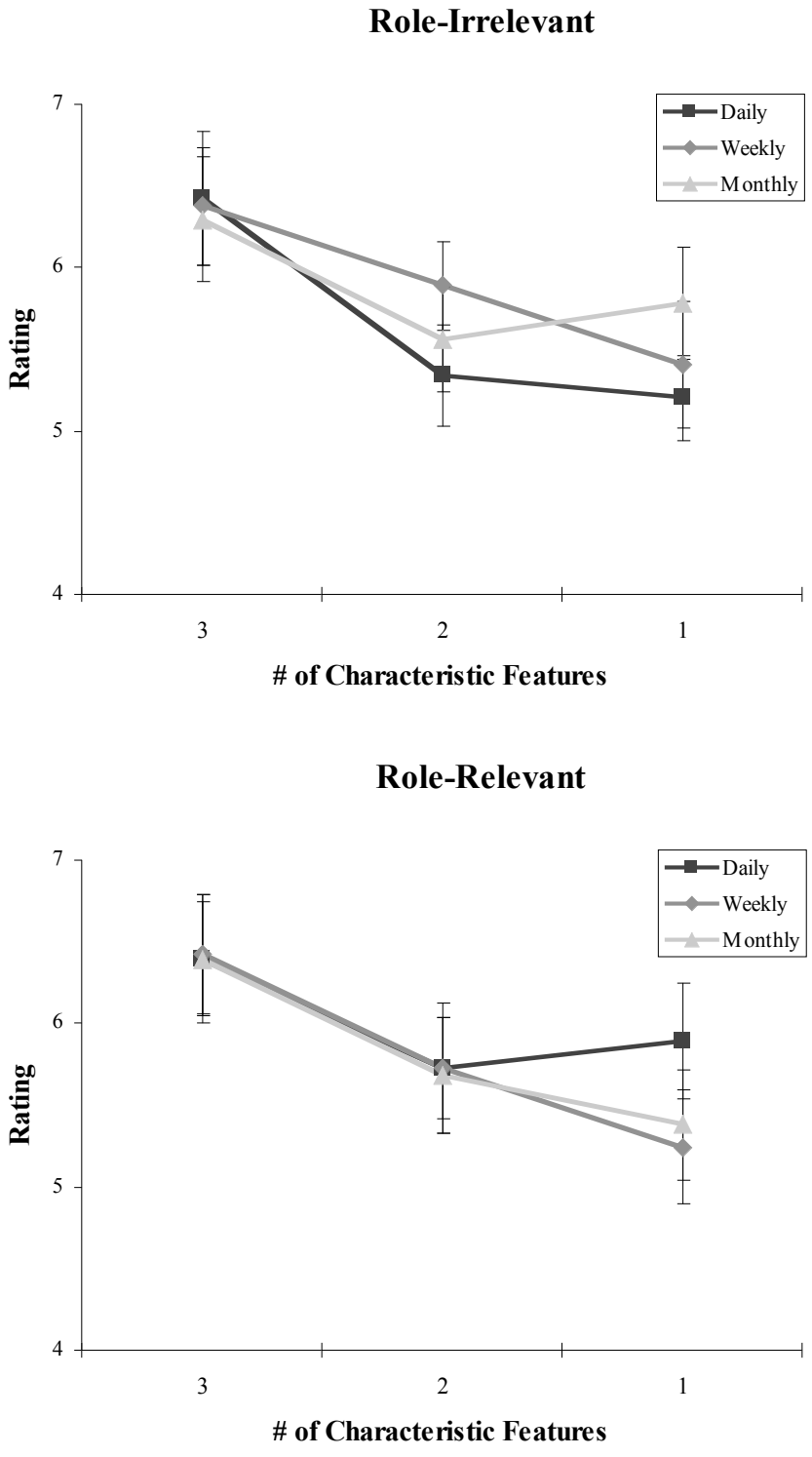


Figure 6. Experiment 2 induction strength judgments. Error bars indicate ± 1 SEM.

Induction strength judgments following role-irrelevant learning were similar to example goodness judgments. Greater central tendency led to higher induction strength ratings, $F(2,62) = 6.40$, $p = .003$, $\eta^2 = .07$; this was a significant linear trend, $F(1,32) = 7.16$, $p = .012$, $\eta^2 = .06$. There was no effect of ideal level, $F(2,62) = 1.15$, *ns*; nor was there any interaction between the two, $F(4,124) = .91$, *ns*.

In contrast, induction strength judgments following role-relevant learning did not correspond to example goodness ratings. As in the role-irrelevant condition, increased central tendency led to increased induction strength, $F(2,64) = 9.94$, $p < .001$, $\eta^2 = .11$; this was a significant linear trend, $F(1,32) = 13.15$, $p = .001$, $\eta^2 = .10$. There was no effect of ideal level, $F(2,64) = 1.01$, *ns*. Although the low central tendency, high ideal data point appears somewhat higher, there was no statistically significant interaction between central tendency and ideal level, $F(4,128) = 1.22$, *ns*.

Discussion

In the role-irrelevant condition, there was no obvious relationship between the categories' functions and their constituent activities. Consequently, participants represented the teacher groups as feature-based categories described by the distribution of activities. As with other feature-based categories, central tendency exemplars were considered the best examples. The level of the defining activity had little effect on example goodness because it had no bearing on the statistics of the category, nor could it be construed as a role-relevant ideal dimension.

Judgments of induction strength mirrored example goodness in the role-irrelevant condition. Proximity to central tendency increased induction strength, while idealness

had no effect on these judgments. This is another replication of the traditional typicality effect observed in feature-based categories.

In the role-relevant condition, the categories' functions and the relationship between those functions and the defining activities were salient. As a result, the teacher groups were represented as role-governed categories. As with other role-governed categories, ideal exemplars were considered the best examples. The level of central tendency also had an effect on example goodness. This is not surprising, considering the design's emphasis—through the familiarization phase and memory probes—on learning all category information. Nevertheless, this effect was smaller than the effect of ideals in this condition and substantially smaller than the effect of central tendency in the role-irrelevant condition.

In contrast to example goodness judgments, participants only considered central tendency during judgments of induction strength in the role-relevant condition. There was no effect of idealness on induction, either for ideals or anti-ideals. The influence of central tendency information on induction for role-relevant categories was remarkably similar to role-irrelevant categories, despite the divergence in example goodness judgments.

It is possible that there is a true typicality effect in category-based induction for feature-based categories and a central tendency effect for role-governed categories. However, a more parsimonious explanation of these findings is that central tendency exemplars are inductively privileged for all category types. The typicality effect for feature-based categories appears to be a special case of this more general phenomenon.

This universality in category-based induction also places boundary conditions on the feature-based/role-governed dissociation. Although these two kinds of categories differ in many ways (Goldwater et al., 2008; Markman & Stilwell, 2001), there is at least one cognitive process in which they behave identically.

Chapter III: The Role of Familiarity in Induction

While Experiments 1 and 2 have shown that central tendency exemplars are privileged in induction, they have not shown why they are privileged. There are a number of ways in which these exemplars are unique compared to all other exemplars, including ideals. In order to go beyond the central tendency effect, it is important to clarify which of these factors influences induction strength. In particular, Experiments 3 through 5 will contrast two of these unique aspects: statistical averageness and familiarity. This is intended to shed light on the *process* underlying category-based induction more broadly. Over the next several pages, I will discuss the evidence and justification for two theoretical positions about category-based induction: that it is a form of statistical reasoning and that it is the product of a familiarity-based heuristic.

The most obviously unique characteristic about central tendency exemplars is that they are the central tendency; they lie at the statistical center of the category distribution. As a result, any kind of similarity-based generalization gradient from the central tendency will extend to more of the other category members, assuming those members cluster toward the center of the distribution. Consider the simplified example in Figure 7. This represents the category *birds* in a two-dimensional similarity space, with each kind of bird represented as a point in that space. A generalization gradient centered on an exemplar near the center of this space, like *robin* covers many of the other exemplars. In contrast, an equally broad gradient from an exemplar in a more extreme portion of the space, like *penguin* covers fewer exemplars.

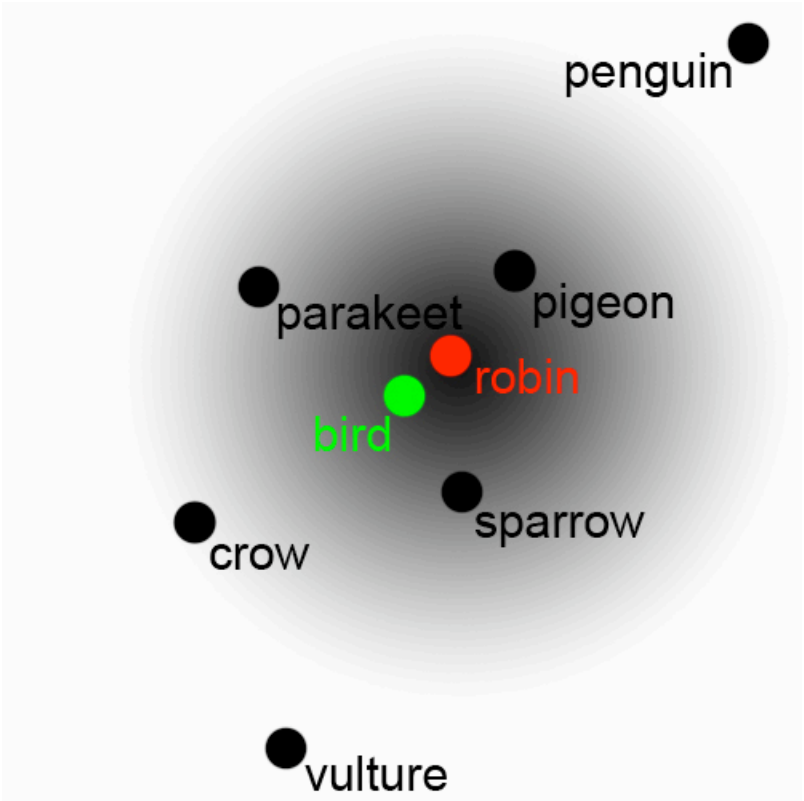


Figure 7. Generalization gradient from robin to other bird members.

Several computational models of category-based induction capture this statistical advantage of the central tendency. This is precisely because many of these models were designed to accommodate induction behavior for natural kind categories, whose central tendencies are most typical. For example, under Osherson et al.'s (1990) model, the strength of an induction from the category *robins* to the category *birds* is a function of *robins'* "coverage" of *birds*. *Robin* is on average very close to other categories in the multidimensional space of *birds* (i.e. more similar to other birds). In other words, *robins* covers the space of *birds* more broadly than *penguins*. Sloman's (1993) model embodies this idea of coverage through exemplars' and categories' feature overlap in a

connectionist architecture. Again, central tendencies have greater induction strength as premises because they have the characteristic features of the category and therefore have substantial overlap with the category's features. Bayesian models of induction (Blok, Medin, & Osherson, 2007; Heit, 1998; Kemp & Tenenbaum, 2009) can also account for central tendency advantages, though they could also account for ideal advantages, depending on the nature of the hypothesis space.

All of these models are consistent with the experiments reported here because they all perform some kind of rough statistical calculation in which the central tendency has a special status. The central tendency lies at the statistical center of the category distribution and is therefore consistent with the most hypotheses about the property's distribution within the category. Critically, this statistical relationship holds for feature-based *and* role-governed categories.

In addition to these statistical models of category-based induction, there are also a number of behavioral phenomena that support the idea that induction is akin to statistical reasoning. In accordance with a basic principle of inferential statistics, people's confidence in the proposition that all category members have some property increases as the number of members known to have the property increases. This intuitive notion is illustrated in what Osherson et al. (1990) dubbed the "premise monotonicity" effect. When asked to choose the better of the following arguments,

(5)

Hawks have sesamoid bones.

Sparrows have sesamoid bones.

Eagles have sesamoid bones.

Therefore, all birds have sesamoid bones.

(6)

Sparrows have sesamoid bones.

Eagles have sesamoid bones.

Therefore, all birds have sesamoid bones.

participants overwhelmingly selected Argument 5 as stronger.

In another parallel to inferential statistics, people also take the variability of the population to which they are generalizing into account. For example, when told of a single sample of the substance "floridium" that conducts electricity, participants readily inferred that all instances of floridium conduct electricity. In contrast, when told of a single member of the Barratos tribe who was obese, participants were less likely to infer that all Barratos were obese (Nisbett, Krantz, Jepson, & Kunda, 1983). Presumably, this more limited generalization is due to the assumption of greater variability of humans (at least as related to obesity) than material compounds. Relatedly, the size of the category also affects people's willingness to generalize. Osherson et al. (1990) labeled this phenomenon "conclusion specificity". When asked to choose the better of the following arguments,

(7)

Bluejays require Vitamin K for the liver to function.

Falcons require Vitamin K for the liver to function.

Therefore, all birds require Vitamin K for the liver to function.

(8)

Bluejays require Vitamin K for the liver to function.

Falcons require Vitamin K for the liver to function.

Therefore, all animals require Vitamin K for the liver to function.

participants overwhelmingly selected Argument 7 as stronger.

People's category-based inductions also often reflect the statistical predictions of their underlying causal models of categories and events. For example, Rehder and Hastie (2004) taught participants about fictional categories like Lake Victoria Shrimp that had four characteristic properties. In some conditions, these properties were presented with causal models explaining how they were related. In the common-cause schema, one of the characteristic properties was identified as the cause of the other three. In the control condition, the properties were presented without any explicit causal model. During an induction phase, participants estimated the percentage of the category that had a novel property, provided that some exemplar had that property. When the exemplar was missing one of the "effect" properties, participants were as likely to extend the novel property as they were from the identical exemplar in the control condition, in which no causality was provided. In contrast, when the "cause" property was missing, there was less generalization compared to the equivalent exemplar in the control and the missing-

effect exemplar. The novel property in this case is less consistent with the underlying causal model and therefore less consistent with the likely distribution of the property that would result from that model. Rehder and his colleagues have found similar reliance on statistical causal information in inference for unknown feature values and classification behavior (Rehder, 2003; Rehder & Burnett, 2005; Rehder & S. W. Kim, 2006).

Shafto, Kemp, Bonawitz, Coley, and Tenenbaum (2008) have shown that people's use of causal models is sensitive to the applicability of the particular model. When participants performed induction about the presence of a disease in animal categories, they based their generalization on food web relationships. When told that some animal had the disease, their likelihood of inferring that another animal had that disease increased as the number of links separating the animals in the food chain decreased. Induction also increased if the second animal was in a predator (as opposed to prey) relation to the first animal or one of the first animal's predators. In contrast, when participants performed induction about the presence of a gene in animals, inference likelihood was based on taxonomic distance between animals. Both of these effects were well-characterized by (distinct) Bayesian models. This and the Rehder and Hastie study are just two of several examples of the causal relations between features and categories influencing category-based induction (Hadjichristidis, Sloman, Stevenson, & Over, 2004; Lassaline, 1996). This use of the most statistically likely causal model in inductive reasoning is particularly prevalent in experts' category-based induction (Bailenson et al., 2002; Proffitt et al., 2000).

Between computational models and behavioral evidence, there are good reasons to believe that category-based inductive reasoning *is* statistical reasoning. However, there are also a number of behavioral studies indicating that reasoners do not appropriately employ statistical principles. Malt, Ross, and Murphy (1995) presented participants with stories of ambiguous categorization. For instance, in one story, a person coming up the driveway was described as probably a realtor, but possibly a burglar. In another version the same person was described as probably a realtor, but possibly a cable repairman. Participants infer that the person will ring the doorbell with equal probability in both cases, apparently uninfluenced by the burglar possibility in the first category. A Bayesian agent would perform this induction by estimating the probability that each category had the property (i.e. each hypothesis), weighted by the likelihood of each category being the correct classification. Across several studies, people regularly make their inductions based on single categories instead of using the statistically appropriate strategy (Murphy & B. H. Ross, 2005; B. H. Ross & Murphy, 1996), though they will take the alternative category into account when explicitly provided numerical probabilities (Murphy & B. H. Ross, 1999).

One phenomenon of non-normative induction in argument evaluation is the inclusion fallacy (Osherson et al. 1990). When asked to choose the better of the following arguments,

(9)

Robins secrete uric acid crystals.

Therefore, all birds secrete uric acid crystals.

(10)

Robins secrete uric acid crystals.

Therefore, ostriches secrete uric acid crystals.

participants selected Argument 9 as stronger, even though the probability of the conclusion in Argument 10 cannot logically be lower. Similar non-normative errors occur in evaluation of deductively valid arguments (Sloman, 1998). Another error is a reversal of the premise monotonicity effect discussed above. When asked to choose the better of the following arguments,

(11)

Crows secrete uric acid crystals.

Peacocks secrete uric acid crystals.

Therefore, all birds secrete uric acid crystals.

(12)

Crows secrete uric acid crystals.

Peacocks secrete uric acid crystals.

Rabbits secrete uric acid crystals.

Therefore, all birds secrete uric acid crystals.

participants selected Argument 11 as stronger, despite having fewer evidence premises (Osherson et al. 1990). Similarly, the preference for narrower conclusions in induction can also be reversed. When asked to choose the better of the following arguments,

(13)

Chickens have Property X12.

Therefore, cows and pigs have Property X12.

(14)

Chickens have Property X12.

Therefore, cows have property X12.

participants selected Argument 13 as stronger, even though the probability of the conclusion in Argument 14 cannot logically be lower, as in the inclusion fallacy (Medin, Coley, Storms, & Hayes, 2003).

So, there is substantial evidence that induction is statistical reasoning, but also substantial evidence that it fails to meet normative standards. As a result, the evidence is mixed for the idea that central tendency exemplars are privileged in induction because of their special status as the statistical center of the category distribution. What else is special about central tendency exemplars that may account for their advantage in induction? One possibility is that the central tendency has the most inductive strength because it is the most familiar member of the category. As noted earlier, a food with zero calories and a great taste is an excellent example of a diet food, but we are much more familiar with less ideal examples. This was likely the case for the natural categories in Experiment 1. In Experiment 2, the exemplars were presented equally often during

training, but central tendency exemplars contained properties that occurred more often, making them the most familiar items in that task.

It is difficult to disambiguate familiarity and central tendency, particularly in natural categories like those generally used in induction studies. Adults, even college freshmen, have a great deal of direct or indirect experience with categories like *birds*, *predators*, and *shoes*. They have sampled the category space sufficiently enough to render the true central tendency most familiar, whether it be a particular exemplar or a combination of characteristic feature values. In contrast, children have much more limited experience. As a result, they are most familiar with a narrow range of exemplars, often not the central tendency. With time and experience, their induction preferences switch from these more familiar exemplars to the central tendency, becoming more like adults (Carey, 1985; Inagaki, 1990; N. Ross, Medin, Coley, & Atran, 2003).

When comparing category-based induction with heuristic reasoning, one readily thinks of the representativeness heuristic (Kahneman & Tversky, 1972; Shafir, E. E. Smith, & Osherson, 1990). This comparison is not very apt, however, in light of the dissociation between representativeness and induction in Experiments 1 and 2. But what about the availability heuristic (Tversky & Kahneman, 1973)? Other researchers have discussed how different kinds of properties in an induction task make different domains of background knowledge more (Shafto, Coley, & Vitkin, 2007). The remaining experiments explore a different role for availability in induction. Specifically, they address the possibility that items are assigned greater inductive strength to the extent that they are thought of easily.

In the next set of experiments, central tendency and familiarity were directly contrasted. While these are perfectly correlated in natural categories (and particularly natural kinds), it is possible to dissociate them in an artificial category setting. The relative advantage for induction between central tendency and familiar exemplars can in turn be interpreted as evidence for the relative importance of statistical reasoning or heuristics in that task setting.

It is important to note that familiarity-based heuristics and statistical reasoning should not necessarily be construed as mutually exclusive. There are many circumstances, particularly experts' induction, where the role of statistical reasoning is unassailable. There is also a certain amount of overlap between the two. As I have discussed, central tendency is generally a reliable marker of familiarity. I deliberately separate them here to provide an existence proof for the role of familiarity in induction. This is intended to counter the tacit assumption—principally implied by mathematical models—that inductive reasoning is *purely* statistical in nature. I am not, however, suggesting the opposite.

In Experiments 3 and 4, familiarity was fostered through a category discrimination learning task. This classification process created greater familiarity for prototype (Experiment 3) and extreme-valued (Experiment 4) exemplars. A fuller description of these familiarity advantages is provided in the introductions of the respective experiments. Following learning, participants were told that each category was actually a single cluster of a larger, multi-cluster category. In Experiment 5, participants learned a single multi-clustered category through observational learning.

Following the learning phase, participants in each experiment performed example goodness and induction strength judgments for familiar and central tendency exemplars. In this context, central tendency exemplars are those that lie between the clusters (i.e. the previously-learned categories) in the category's dimensional space. As a result, central tendency and familiarity were mutually exclusive in these designs.

While the primary motivation for these experiments is to clarify the role of statistics and heuristics, the design also presents an opportunity to test the limits of the diversity effect in category-based induction. This effect was first illustrated by Osherson et al. (1990). When asked to choose the better of the following arguments,

(15)

Hippopotamuses have a higher sodium concentration in their blood than humans.

Hamsters have a higher sodium concentration in their blood than humans.

Therefore, all mammals have a higher sodium concentration in their blood than humans.

(16)

Hippopotamuses have a higher sodium concentration in their blood than humans.

Rhinoceroses have a higher sodium concentration in their blood than humans.

Therefore, all mammals have a higher sodium concentration in their blood than humans.

participants selected Argument 11 as stronger. This phenomenon actually runs counter to the influence of typicality, as *rhinoceros* is considered a better example of *mammals* than *hamster*. It is, however, consistent with the principle of coverage embodied in all of the formal models of induction discussed.

The diversity effect is remarkably robust. In addition to college students, the effect is exhibited by children as young as 8 years (Lopez et al. 1992) and by some domain experts, such as tree taxonomists (Proffitt et al. 2000). People also value diversity in a range of tasks, including evaluating arguments, seeking evidence in testing arguments (López, 1995), and diagnostic causal reasoning (N. S. Kim & Keil, 2003). People can also apply the diversity principle flexibly across different dimensions of similarity (Heit & Feeney, 2005).

Critically, all of these studies have embodied diversity in two or more premises. In the coming studies, central tendency exemplars embody diversity within a single item that is located between two clusters in the category space. As with *hamsters* in the earlier example, these exemplars should not be considered typical because they do not have the correlated features or standard dimension values of either cluster, but they do contain features or values more representative of *both* clusters than any exemplar from within a cluster. If the central tendency exemplars are judged to have greater inductive strength, this would establish a single-item diversity effect and extend validation of the coverage principle.

Experiment 3

In this experiment, participants began by learning to correctly classify text descriptions of imaginary "animals" into one of four categories. There were four dimensions on which category members varied: body color, largest internal organ, primary sensory system, and protective head feature. Each category had a single feature on each dimension that was characteristic of the category. There were four members for each category presented during training, which conformed to a prototype-plus-exception, or Type IV (Shepard, Hovland, & Jenkins, 1961), category structure. So, if the characteristic values of Category A were "red", "heart", "sight", and "horns", each individual Category A exemplar would have three of these features plus a feature on the remaining dimension that was characteristic of another category.

This initial learning phase was followed by a second learning phase, in which the four original categories were collapsed into two larger categories or "species", each with two of the original categories as clusters or "subtypes". Following this second phase, participants made example goodness and induction strength judgments about two classes of exemplar. One of these was the set of prototypes of the original four categories. These were exemplars that had all four of their categories' characteristic features. The other class was the set of mixture exemplars, which had two features from one cluster within a category and two features from the other cluster. Although the prototypes' particular feature combinations constituted novel exemplars, they were expected to be highly familiar. Prototype stimuli are generally classified more accurately than other novel category exemplars (Knowlton & Squire, 1993; Posner & Keele, 1968) and judged

to be more familiar than novel and, often, old exemplars (Bruce, Doyle, Dench, & Burton, 1991; Homa, Goldhardt, Burruel-Homa, & J. C. Smith, 1993; Solso & McCarthy, 1981). I will use the term "familiar" to refer to a psychological metric, distinct from actual experience, throughout this paper. This distinction will be especially important in Experiments 6 and 7.

The mixture exemplars also constituted novel exemplars, though these were expected to be non-familiar because they violated feature-feature correlations established during the learning phase. They were, however, the central tendency exemplars of the category. These exemplars lie in the center of the category distribution (i.e. between the two component clusters) at the minimal distance between all of the previously-learned exemplars. If people perform induction purely via statistical reasoning, these central tendency exemplars should be preferred over the familiar but extreme prototypes. If familiarity has greater influence in this task, the opposite should be true.

Method

Participants

36 University of Texas at Austin students participated for course credit or payment of $8.

Materials

There were four dimensions: body color, largest internal organ, primary sensory system, and protective head feature. An individual exemplar could have one of four features on each of these dimensions. These features were "red", "green", "white", and "black" for body color; "heart", "lungs", "stomach", and "liver" for largest internal

organ; “sight”, “sound”, “smell”, and “taste” for primary sensory system; and “horns”, “spike”, “crest”, and “antlers” for protective head feature. These features were displayed as text on a computer screen.

Table 2 depicts the category structure of the training stimuli in abstract notation. Each of the four original categories (“subtypes”) had a prototype-plus-exception structure. For example, the modal value on each dimension for Category A was 0, and every exemplar had this value on three of the four dimensions. Each category had a distinct modal prototype value. The exception value for each exemplar was one of the modal values from the contrasting combined category (“species”). For example, all of the exception values of category A exemplars were 2’s and 3’s, the modal values of the Category X subtypes, Categories C and D. There were 16 total training exemplars; four in each original category and eight in each combined category. The particular features associated with each category were assigned randomly for each participant.

Table 2
Experiment 3 Training Stimuli

Categories		**Dimensions**			
		1	**2**	**3**	**4**
X	A	0	0	0	2
		0	0	3	0
		0	2	0	0
		3	0	0	0
	B	1	1	1	3
		1	1	2	1
		1	3	1	1
		2	1	1	1
Y	C	2	2	2	0
		2	2	1	2
		2	0	2	2
		1	2	2	2
	D	3	3	3	1
		3	3	0	3
		3	1	3	3
		0	3	3	3

Table 3 depicts the category structure of the rating stimuli for the induction strength and example goodness judgments. Half of the rating stimuli were prototype exemplars. These had the modal value of their respective original category on all four dimensions. For example, the prototype exemplar for Category A was composed entirely of 0's. The other half of the rating stimuli were mixture exemplars. These had the modal value of one cluster on two dimensions and the modal value of the major category's other cluster on the remaining two dimensions. There were 8 total rating stimuli.

Table 3
Experiment 3 Example Goodness and Induction Strength Rating Stimuli

Categories	**Dimensions**			
	1	**2**	**3**	**4**
X	0	0	0	0
	1	1	1	1
	0	0	1	1
	1	1	0	0
Y	2	2	2	2
	3	3	3	3
	2	2	3	3
	3	3	2	2

The rating stimuli were constructed so that prototype and mixture exemplars were as comparable as possible. Both were novel in the sense that their exact feature combinations did not appear during training. Both were composed entirely of feature values that were characteristic of the category during training. They were equally similar to their category's training exemplars, regardless of whether similarity is computed with an additive (e.g. Tversky, 1977), multiplicative (e.g. Medin & Shaffer, 1978), or exponential (e.g. Nosofsky, 1984) function.

Procedure

Participants were tested at individual computers. At the beginning of the experiment, participants were given the following instructions:

> Imagine that you are a scientist exploring a remote island where you find new kinds of animals that have never been documented. The animals mainly differ in four ways: their body color, their largest internal organ, their primary sensory

system, and their protective head feature. Each animal can be classified into one of four categories according to these features. Your task is to classify the animals into one of the four categories. For the sake of simplicity, let's call these categories A, B, C, and D. All four features will be important for classifying the animals, so it will be important to pay attention to all of them.

You will see a series of animal descriptions, presented one at a time. Each description will include all four of its features. To classify the animal, press the keyboard button corresponding to its category (A, B, C, or D). At first you will just be guessing, but you should soon figure out which features go together for each category.

On each trial, the dimension labels and features appeared in text at the top of the screen, and the prompt "Which Category?" appeared at the bottom of the screen. Feedback appeared after participants made a classification response, informing them whether they were correct or incorrect and stating the correct category label. Positive and negative feedback appeared in blue- and red-colored font, respectively. This feedback remained on the screen for 2 seconds, followed by a blank screen intertrial interval for 500 ms. Trials were organized into blocks of 16 trials, with each stimulus appearing once per block. Stimulus order was randomized within blocks. If participants met a learning criterion of 14 correct trials in a block, beginning with the fifth block, they immediately moved to the second training phase. All participants moved on after 12 blocks, regardless of performance.

Before the second training phase, participants were given the following instructions:

> After further testing, you have learned that the four animal categories are actually two larger categories. Categories A and B are subtypes of a single species (category X), and likewise for categories C and D (category Y). Your next task is to classify the animals into one of the two larger categories.
>
> You will see a series of animal descriptions, presented one at a time. Each description will include all four of its features. To classify the animal, press the keyboard button corresponding to its category (X or Y). You might need some practice, but this should be fairly simple. You are simply combining the categories that you have already learned into two new species categories.

This task was identical to the first training phase, except for the button responses and feedback labels. Again, there was a minimum of 5 and maximum of 12 blocks, dependent on reaching the 14-trial accuracy criterion.

The example goodness and induction strength tasks were identical to those in Experiments 1 and 2, with one exception. Instead of "property X", participants were told that an exemplar had "neuropeptide [AA00]", with a unique combination of two letters and two numbers on each trial. Again, these tasks were blocked, and block order was randomized across participants. Stimulus order was randomized within each block.

Results

Participants performed reasonably well in the first training phase, considering the difficulty of a 4-category task. Mean accuracy across this entire first training phase was 56.7%, which was reliably greater than the 25% expected for chance responding, $t(35) = 17.39$, $p < .001$, $d = 2.90$. The mean number of blocks in this phase, including participants who met the criterion and those who required the whole phase, was 10.5. Mean accuracy for participants' final block in the first training phase was 74.1%, which was also reliably greater than chance, $t(35) = 15.35$, $p < .001$, $d = 2.56$.

Although there was some drop-off in performance between the last block of the first training phase and the subsequent 2-category training phase, participants nonetheless did rather well. Mean accuracy across the entire second training phase was 66.1%, which was reliably greater than the 50% expected for chance responding, $t(35) = 9.21$, $p < .001$, $d = 1.54$. The mean number of blocks in this phase was 9.3. Mean accuracy for participants' final block in the second training phase was 78.0%, which was also reliably greater than chance, $t(35) = 10.40$, $p < .001$, $d = 1.73$.

Each participant's mean example goodness and induction strength rating was calculated for prototype and mixture stimuli. Figure 8 depicts the means of the example goodness ratings for both stimulus types and Figure 9 depicts the mean of the difference between them. Overall, prototype stimuli ($M = 7.3$) were rated as significantly better examples of their categories than mixture stimuli ($M = 6.5$), $t(35) = 3.20$, $p = .003$, $d = .53$. Figure 10 depicts the means of the induction strength ratings and Figure 11 depicts the mean of the difference between them. As with example goodness, prototype stimuli

(M = 6.9) were rated as having significantly greater induction strength than mixture stimuli (M = 6.3), $t(35) = 2.81$, $p = .008$, $d = .47$. Overall, there was a significant correlation between example goodness and induction strength judgments, $r(72) = .79$, $p < .001$.

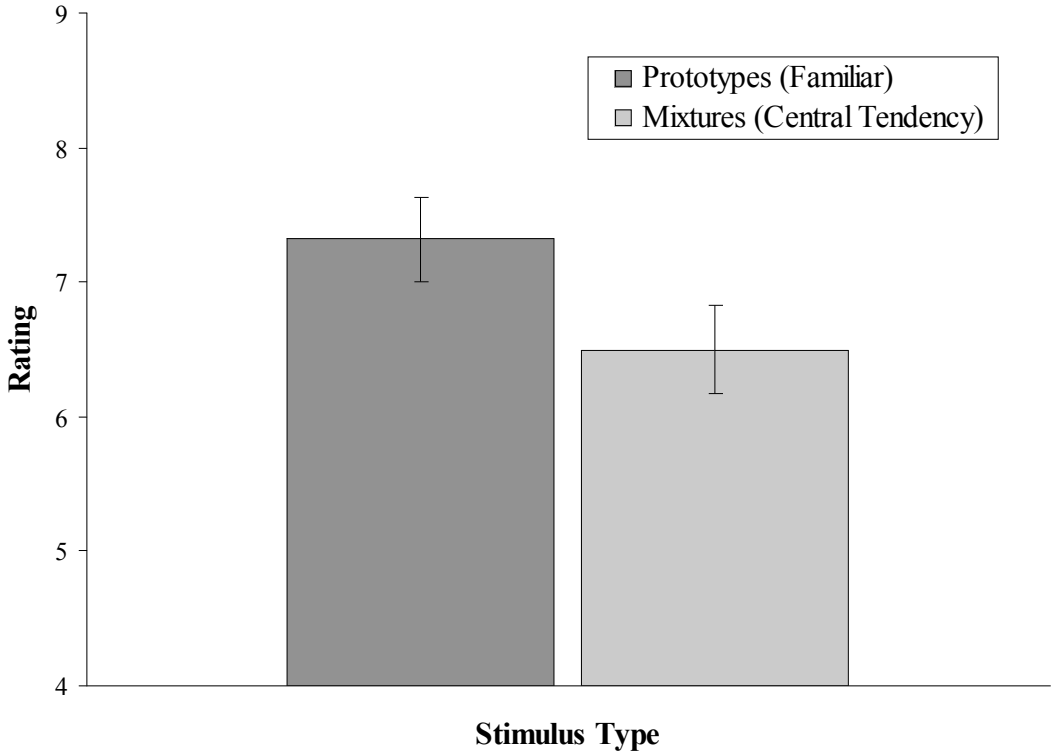


Figure 8. Experiment 3 example goodness ratings. Error bars indicate ± 1 SEM.

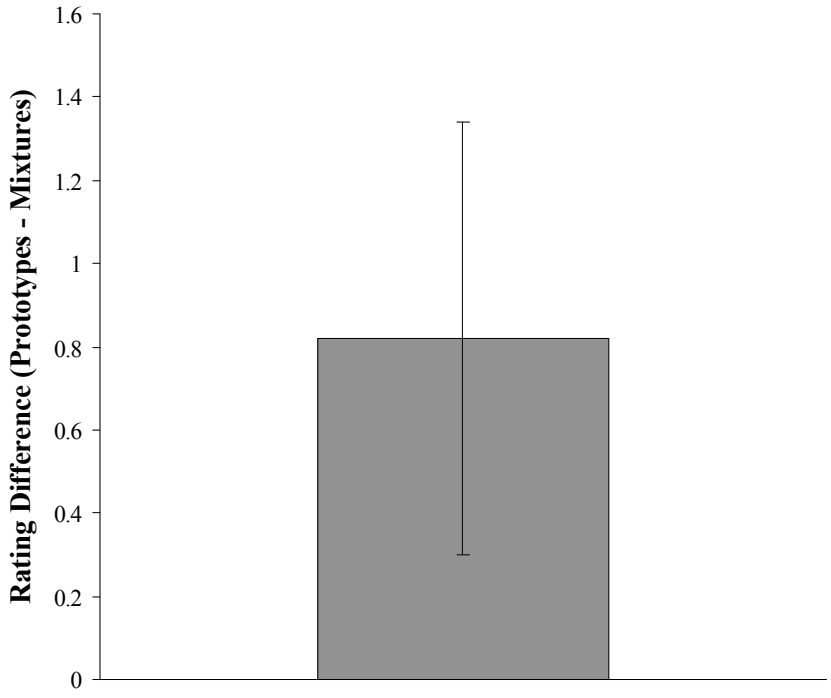


Figure 9. Experiment 3 example goodness mean difference score. Error bar indicates 95% confidence interval.

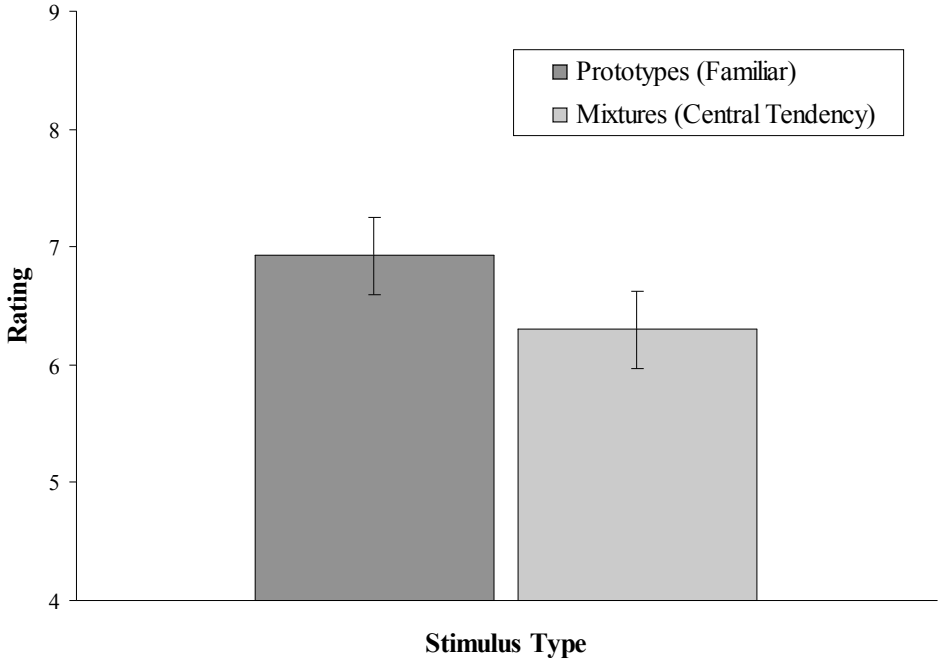


Figure 10. Experiment 3 induction strength ratings. Error bars indicate ± 1 SEM.

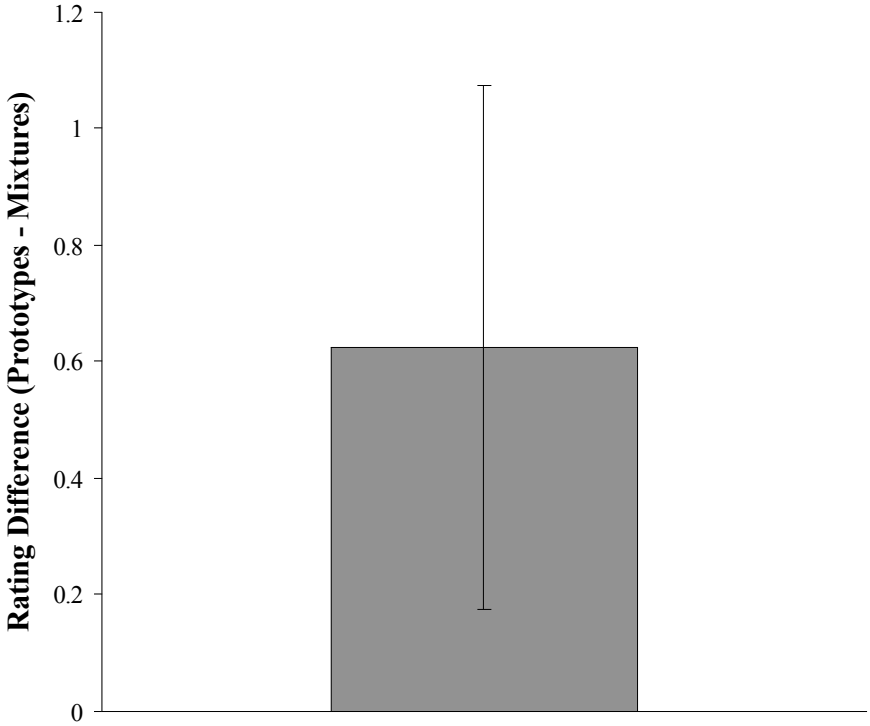


Figure 11. Experiment 3 induction strength mean difference score. Error bar indicates 95% confidence interval.

While it is important to analyze the data from all participants, some caution is warranted. With this kind of prototype-plus-exception category structure, many people do not learn correlations between dimension values, instead classifying stimuli according to simple unidimensional rules (Nosofsky, Palmeri, & McKinley, 1994). If participants in the present experiment used this kind of simplistic strategy, one could argue that the prototype stimuli were not actually more familiar, as these participants did not learn the correlational structure of the categories. To assuage this concern, additional analyses were performed on the set of participants that met the learning criterion in the final

training block, whose performance exceeded the maximum possible with a unidimensional rule. There were 20 such participants.

For this set of participants (hereafter "learners"), mean accuracy across the entire first training phase was 60.1%, which was reliably greater than the 25% expected for chance responding, $t(19) = 19.87$, $p < .001$, $d = 4.44$. The mean number of blocks in this phase, including participants who met the criterion and those who required the whole phase, was 9.7. Mean accuracy for learners' final block in the first training phase was 81.9%, which was also reliably greater than chance, $t(19) = 24.64$, $p < .001$, $d = 5.51$.

As with the more inclusive group, there was some drop-off in performance between the last block of the first training phase and the subsequent 2-category training phase. Mean accuracy across the entire second training phase was 71.9%, which was reliably greater than the 50% expected for chance responding, $t(19) = 12.54$, $p < .001$, $d = 2.8$. The mean number of blocks in this phase was 7.2. Mean accuracy for learners' final block in the second training phase was 90.3%, which was also reliably greater than chance, $t(19) = 47.69$, $p < .001$, $d = 10.66$.

Each learner's mean example goodness and induction strength rating was calculated for prototype and mixture stimuli. Figure 12 depicts the means of the example goodness ratings for both stimulus types and Figure 13 depicts the mean of the difference between them. Overall, prototype stimuli ($M = 8.0$) were rated as significantly better examples of their categories than mixture stimuli ($M = 7.0$), $t(19) = 2.35$, $p = .03$, $d = .53$. Figure 14 depicts the means of the induction strength ratings and Figure 15 depicts the mean of the difference between them. As with example goodness, prototype stimuli ($M =$

7.7) were rated as having significantly greater induction strength than mixture stimuli (M = 6.9), $t(19) = 2.84$, $p = .011$, $d = .63$. Overall, there was a significant correlation between example goodness and induction strength judgments, $r(40) = .79$, $p < .001$.

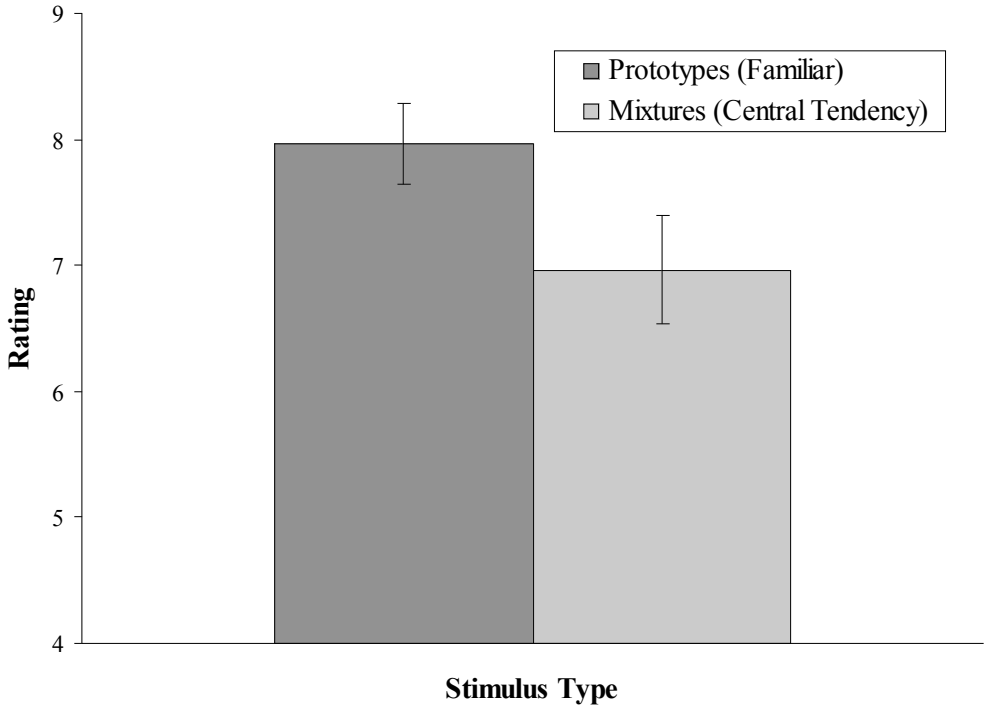


Figure 12. Experiment 3 learners' example goodness ratings. Error bars indicate ± 1 SEM.

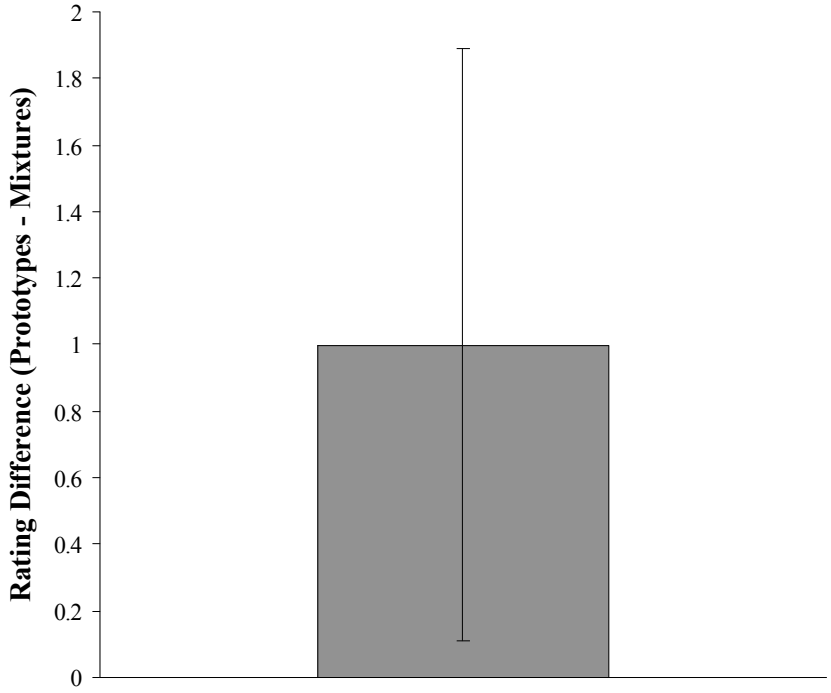


Figure 13. Experiment 3 learners' example goodness mean difference score. Error bar indicates 95% confidence interval.

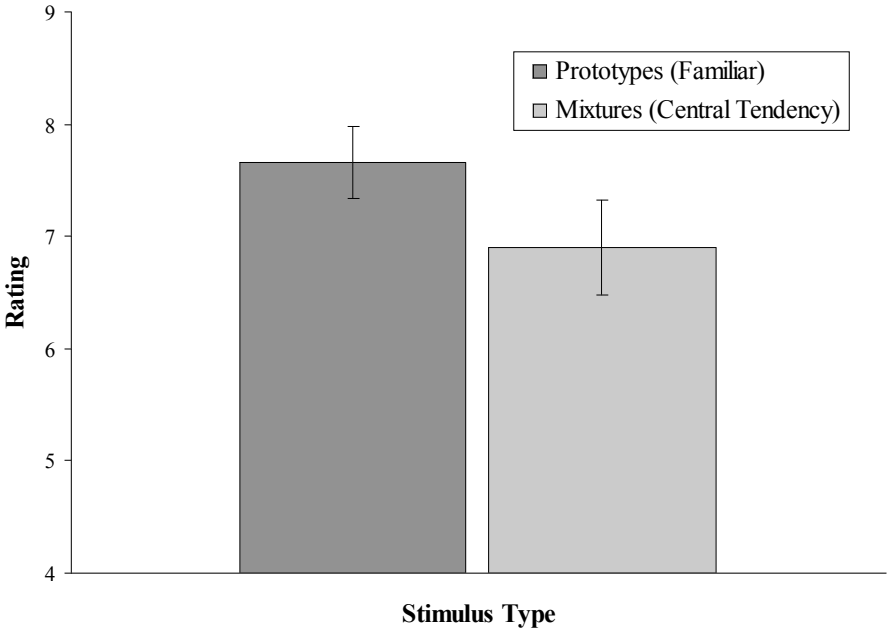


Figure 14. Experiment 3 learners' induction strength ratings. Error bars indicate ± 1 SEM.

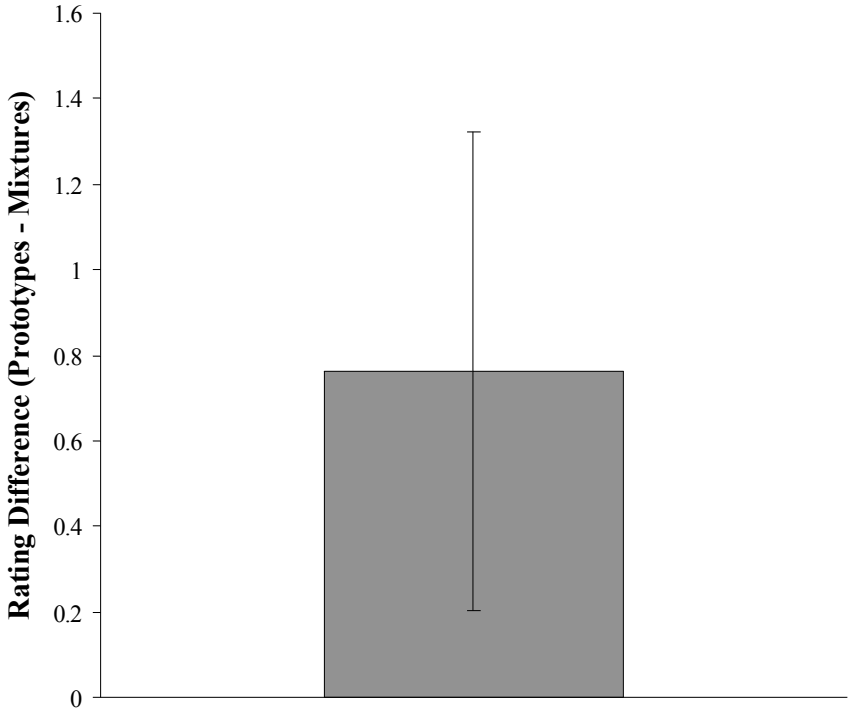


Figure 15. Experiment 3 learners' induction strength mean difference score. Error bar indicates 95% confidence interval.

Discussion

Consistent with findings throughout the categorization literature, familiar prototype exemplars were judged to be better examples of their category than mixture exemplars that had feature values from multiple category clusters. Induction strength judgments were also higher for the familiar stimuli than the central tendency stimuli. This latter finding suggests that there is no diversity effect when diversity is embodied in a single item representing values that fully cover the category space. Both effects held for the entire sample, as well as the more limited set of objectively-defined learners. Overall, these results are good preliminary evidence for the proposition that category-

based induction can be based mainly on a familiarity heuristic, and less on statistical reasoning.

However, some caution is warranted. One might argue that the mixture exemplars were not appropriate test cases for statistical reasoning. The multi-dimensional, discrete-valued category space does not have an "average" in the traditional sense. The central tendencies lie at the minimal distance between all other category exemplars, but there is no true center of the distribution. This kind of discrete-valued average is conducive with some featural models of induction (e.g. Sloman, 1993), but other models, and perhaps intuition, require a more continuous dimensional space with a central tendency that lies in the true center of that space.

Experiment 4

In this experiment, participants learned to distinguish between two categories that varied on a single continuous dimension. Following this kind of contrastive learning, exemplar familiarity increases with distance from the category boundary. Extreme-valued exemplars are classified more accurately (Steyvers et al. 2003) and participants recall them as being most common (Davis & Love, in press). As in Experiment 3, participants were told after training that the categories were actually subsets of a single more inclusive category. They then provided example goodness and induction strength judgments for exemplars covering the entire continuum, including familiar extreme exemplars and central tendency exemplars near the original category boundary.

Method

Participants

44 University of Texas at Austin students participated for course credit or payment of $8.

Materials

Stimuli were cartoon-like pictures of Martians. See Figure 16 for an example.

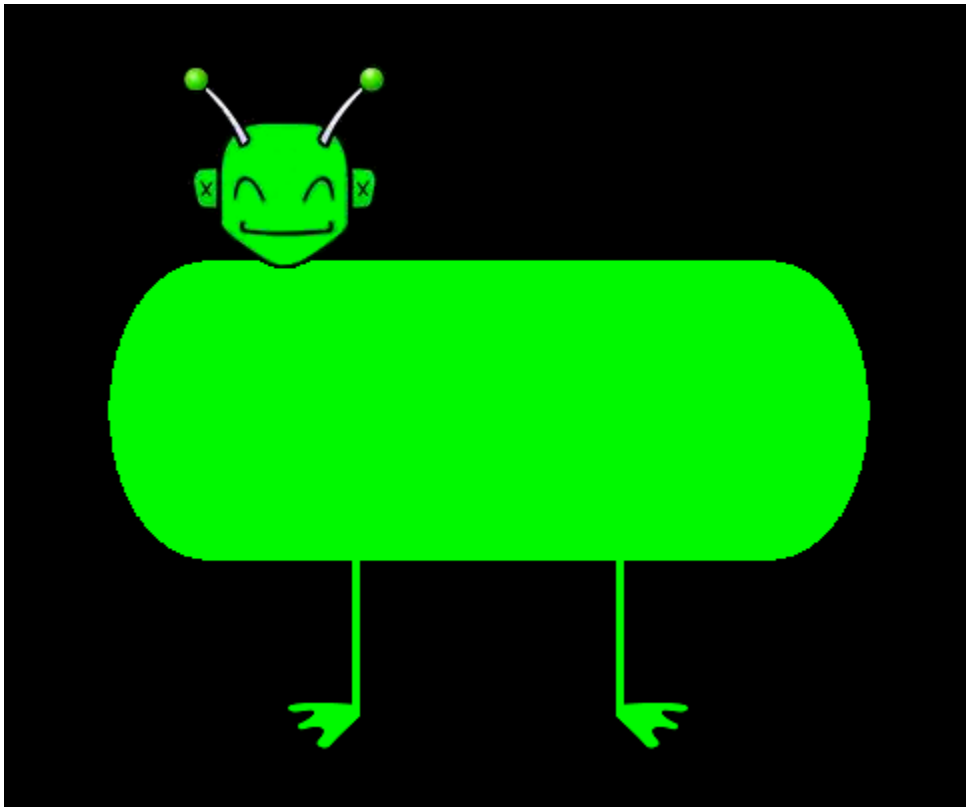


Figure 16. Example of Martian stimuli.

Stimuli were 10.0 cm in height (11.4° visual angle) and 13.2 cm in width (15.0°). They varied in the horizontal position of the head. Training stimuli position values were drawn randomly from two normal distributions, one with a mean 3.5 cm to the left of center, the other 3.5 cm to the right of center. Each distribution had a standard deviation of .5 cm

and individual stimuli were restricted to fall within 1 cm of the mean. The Gaussian curves in Figure 18 depict the distribution of training stimuli along the horizontal dimension. There were 19 rating stimuli with position values between 4.5 cm left of center and 4.5 cm right of center, in .5 cm increments. The rectangular tick marks in Figure 18 depict the position of these stimuli.

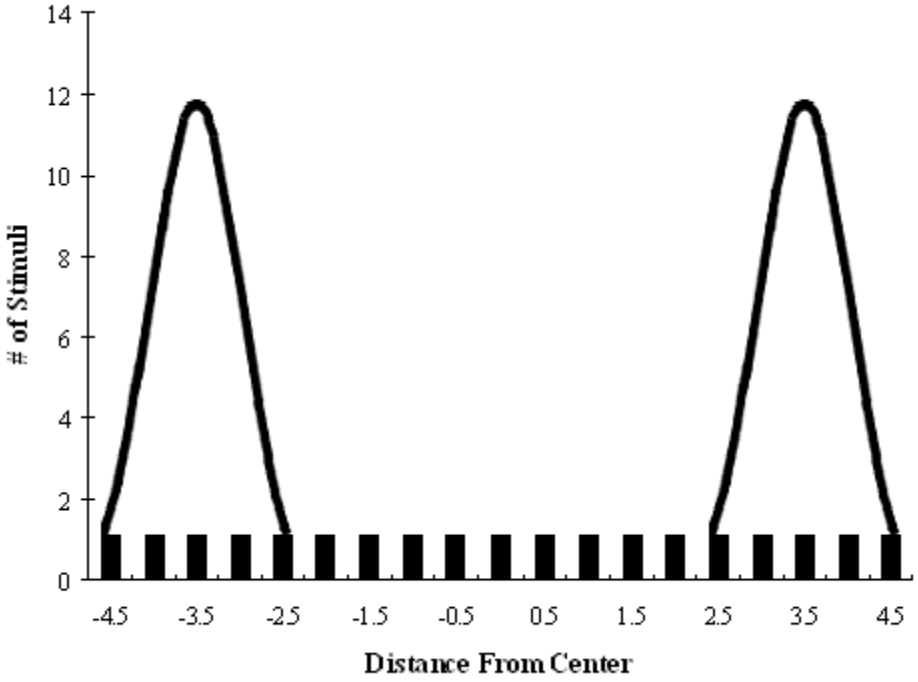


Figure 17. Experiment 4 stimuli

Procedure

Participants were tested at individual computers. At the beginning of the experiment, participants were given the following instructions:

> Imagine that you are an astronaut exploring Mars, where you discover alien lifeforms. You have noticed that the Martians can be separated into two

> categories based on the position of their heads on their bodies. Your task is to classify the Martians into one of the two categories. For the sake of simplicity, let's call these categories X and Y.
>
> You will see a series of Martians, presented one at a time. To classify a Martian, press the keyboard button corresponding to its category (X or Y). At first you will just be guessing, but you should soon figure out how to tell the categories apart.

Individual training trials were identical to Experiment 3, except for the stimuli. There were 120 total training trials. Each category response was correct on half of the trials. Following the training phase, participants made example goodness and induction strength judgments about all Martians, based on individual exemplars. These tasks were identical to those in Experiment 3.

Results

Due to the considerable distance between categories, participants performed extremely well in the training phase, with a mean accuracy of 97.1%.

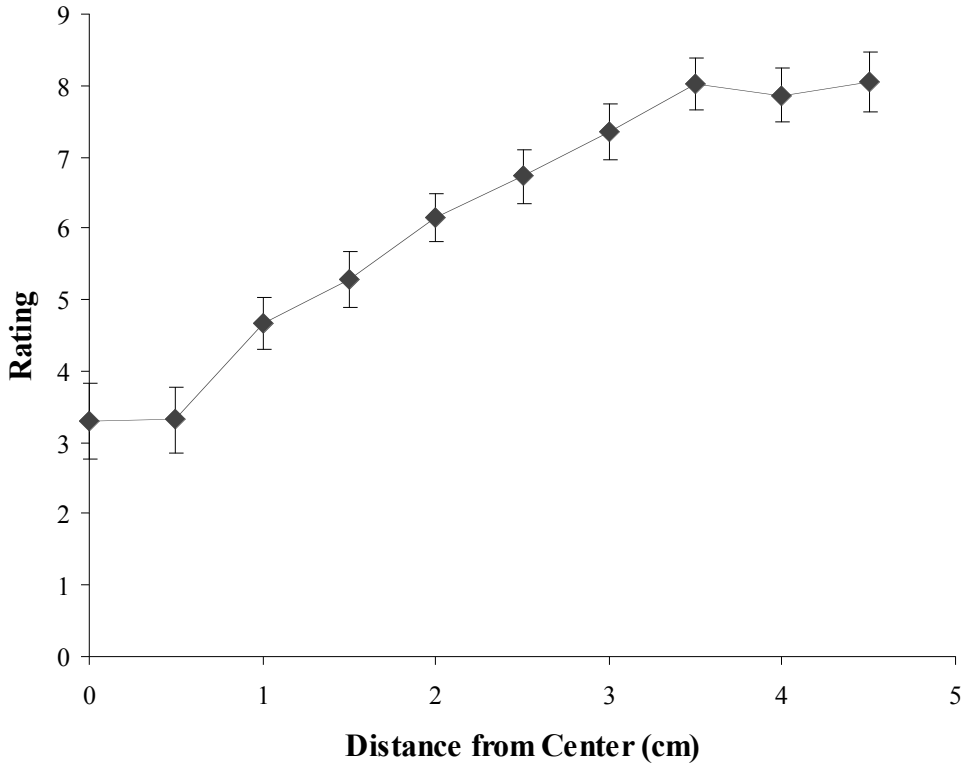


Figure 18. Experiment 4 example goodness ratings.

Figure 18 depicts example goodness ratings as a function of head position's distance from the center. Example goodness differed significantly by distance, $F(9,387) = 31.74$, $p < .001$, $\eta^2 = .42$. There was a significant linear trend, $F(1,43) = 43.72$, $p < .001$, $\eta^2 = .40$; a significant quadratic trend, $F(1,43) = 9.69$, $p = .003$, $\eta^2 = .01$; and a significant cubic trend, $F(1,43) = 9.04$, $p = .004$, $\eta^2 = .004$. The extreme value ($M = 8.0$) was rated as a significantly better example of Martians than the central value ($M = 3.3$), $t(43) = 6.04$, $p < .001$, $d = .91$.

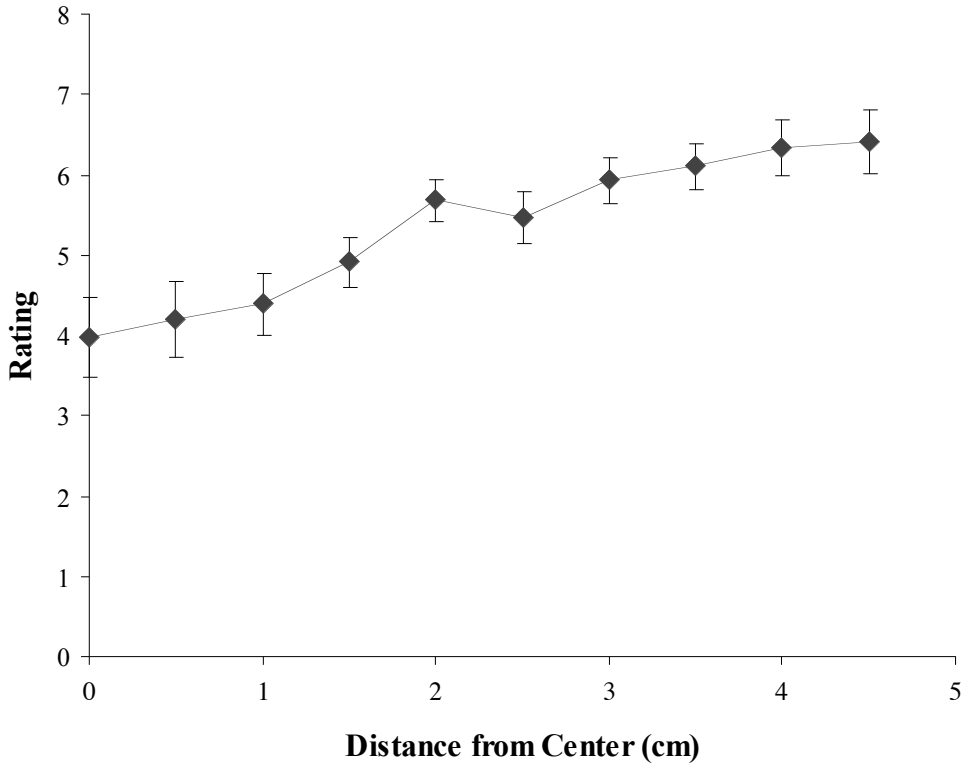


Figure 19. Experiment 4 induction strength ratings.

Figure 19 depicts induction strength ratings as a function of head position's distance from the center. Induction strength differed significantly by distance, $F(9,387) = 10.00$, $p < .001$, $\eta^2 = .19$. There was a significant linear trend, $F(1,43) = 15.92$, $p < .001$, $\eta^2 = .18$. The extreme value ($M = 6.4$) was rated as having significantly higher induction strength for all Martians than the central value ($M = 4.0$), $t(43) = 3.75$, $p = .001$, $d = .57$.

Discussion

As expected, extreme values of the Martian category were rated as the best examples of the category. These values were expected to be most familiar because they were near the original training exemplars and far from the category boundary. Also as

expected, the extreme familiar values were judged to have the highest induction strength for the Martian category. The unfamiliar central tendency exemplar, by contrast, was rated as the worst example of the category and the least inductively strong. Again, this represents another failure to find a single-item diversity effect.

As in Experiment 3, the familiar exemplars were preferred in induction over the central tendencies. The low induction strength of the central tendency exemplar is particularly notable in this study because of its salience. The central tendency in this case was literally the center of the dimensional space. The center of the body is arguably also the most natural place for a head to be positioned, even for a Martian. This study, using a very different category structure from Experiment 3, is further evidence that category-based induction can be driven more by a familiarity-based heuristic process than statistical reasoning.

Some caution is warranted before strong claims are made about the role of heuristics and statistics in category-based induction. The previous two experiments showed that familiar exemplars provide more inductive strength than central tendencies *in these task contexts*. One might argue that these are contrived tasks that artificially and unnaturally boosted the inductive strength of the familiar items and/or disrupted that of the central tendency exemplars. There are two prominent potential criticisms of Experiments 3 and 4 in this vein.

First, one might think that the relative strength of the two kinds of exemplars was an artifact of combining categories following a discrimination learning task. Under this argument, it is unnatural to form incoherent multi-cluster categories created by

combining simpler categories, as was done in the previous experiments tasks. These might be too different from the natural categories that would otherwise take advantage of induction by statistical reasoning. In other words, one might argue that multi-cluster categories are just too abnormal for these experiments to be generalizable to real-world induction.

In contrast to the artificial categories in Experiments 1 and 2, many natural categories have many features that are common across all category members. Natural kind categories are intuitively thought of as having a single cluster. Individual birds may vary in how well they fit the category *birds*. Nonetheless, they all share a basic correlational structure of attributes. In one of the first discussions of multi-cluster categories, Medin & Schaffer (1978) pointed out that the category *spoons* could be broken down into two sets. Some spoons are small and metal; others are large and wooden. Meanwhile, there are very few small wooden spoons or large metal spoons, at least in modern American culture. Despite this demarcation, however, all spoons at least have a common shape and a (roughly) common function. Even the category *animals*, which can be broken down into many clusters that are highly distinct (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), has many features common to all members, which can be the basis for broad inductive generalization. In contrast, a participant in the previous two experiments may have seen very little in common between clusters of the same category. In Experiment 3, there were very few values in common, all of which were actually modal values of contrast category clusters. In Experiment 4, the Martian clusters were decidedly far apart in the horizontal dimension. Of course, all exemplars

fell into the same category, animals or Martians. Furthermore, the animals of Experiment 3 shared four common dimensions, and the Martians of Experiment 4 shared a single dimension. Still, it is possible that unnatural clustering eliminated what would have otherwise been a preference for central tendencies.

A related possibility is that, due to the discrimination task, participants approached the induction task in some unintended way. For instance, it is possible that some participants restricted their generalization to an exemplar's original category, not the broader category. In the context of these experiments, this would be cluster-based, not category-based induction. Under this strategy, induction is strongest for the central tendency of a *cluster*. If this was at all common, then the phenomenon that I have characterized as a familiarity preference is actually a central tendency preference, albeit a limited central tendency. This particular strategy does seem somewhat unlikely, however. If participants were restricting generalization to clusters, their probability estimates should have never exceeded 50%, which would be the maximum proportion of category members in that cluster. This was not the case; induction from familiar exemplars exceeded 50% and therefore covered at least some of both clusters in the broader category. Still, even if people were not performing induction on individual clusters, the larger point remains that the discrimination learning task could have fostered some kind of alternative strategy. Any strong claims about the relative effect of familiarity and central tendency on induction require a further demonstration that does not contrast them through this kind of task.

The second prominent criticism, related to the nature of the learning task, is that the design of these tasks was biased against the central tendency exemplars. Under the category structures used, participants could reasonably infer that these exemplars were extremely atypical category members at best, and completely unviable members at worst. In Experiment 3, the mixture exemplars were unfamiliar because they contained feature combinations that violated the previously learned correlations. Due to these violations, participants may have judged the mixture exemplars to not actually be category members and therefore not viable bases for induction, despite their central tendency. The same could have been true in Experiment 4. There, the central exemplars were fairly far outside the range of familiar values. Again, participants could have reasonably assumed that such a central exemplar simply could not be a Martian and so could not be generalized from. This strong claim seems unwarranted because the induction strength ratings for these stimuli substantially exceeded 0%, indicating that people did see them as viable category members. That said, the weaker claim that these stimuli were viewed as *less* viable is still problematic. Penguins may be atypical birds, far from the *bird* central tendency, but no one would suggest that they are less viable members of the category than robins. To be able to argue that familiar stimuli have greater strength than central tendencies when the two are in conflict, it is necessary that they are both considered acceptable category members.

Experiment 5

The purpose of this experiment was to address the potential problems in Experiments 3 and 4. The stimuli were similar to those in Experiment 4, but instead of

forming a combined category after a discrimination learning phase, participants were presented with a single category to learn. This initial learning phase exposed participants to a continuous range of values, including the central tendency. The central tendency exemplar was therefore unambiguously viable because it was presented during training. Instead of assuming familiarity differences indirectly based on category learning phenomena, this experiment used a straightforward manipulation of familiarity: frequency of exposure. The central tendency, although presented during training, was presented less often than more extreme values. As a result, these extreme values were more familiar.

Method

Participants

62 University of Texas at Austin students participated for course credit or payment of $8.

Materials

Stimuli were the same Martian figures from Experiment 4, though exact values differed. Exemplars appeared in blocks of 82 trials. The central exemplar appeared in 2 trials, exemplars .5 cm to the left and right appeared 4 times each, exemplars 1 cm to the left and right appeared 6 times each, and so on. Every .5 cm increment corresponded to an increase of 2 exemplars on each side, up to 12 presentations for exemplars 2.5 cm to the left and right of center. There were 3 blocks, for a total of 246 trials. Exemplar presentation was randomized within blocks. The gray bars in Figure 19 depict the

relative frequency of the training values. As in Figure 18, the rectangular tick marks indicate the position of the rating stimuli.

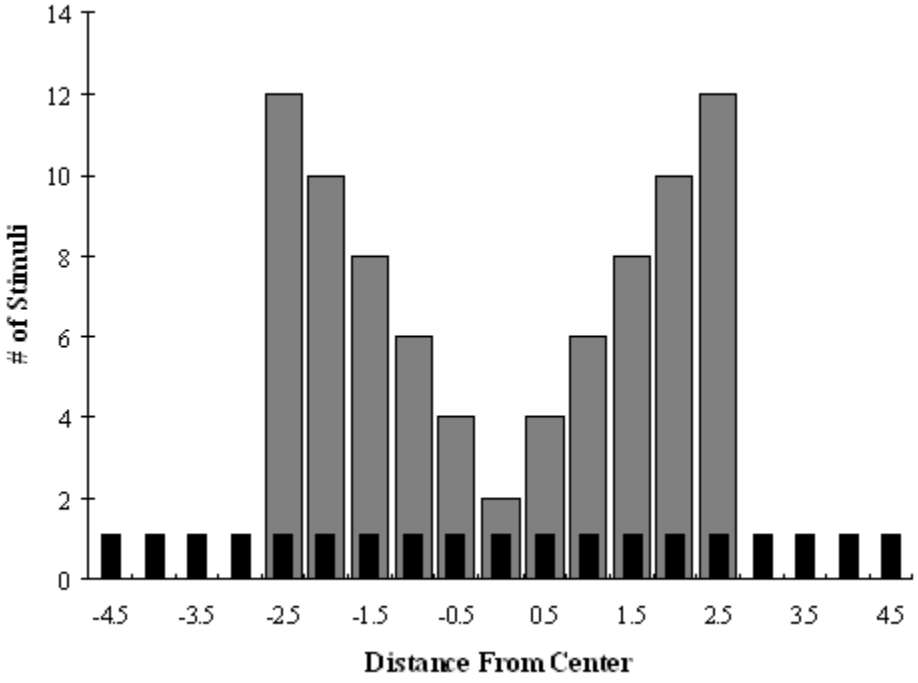


Figure 20. Experiment 5 Stimuli.

Procedure

On each trial, a Martian stimulus appeared in the center of the screen. After 1.5 seconds, the cursor appeared in the center of the screen. After participants clicked on the head of the stimulus, there was a 500 ms intertrial interval with a blank screen, followed by the next trial. This head-clicking procedure forced participants to attend to the relevant dimension on every trial.

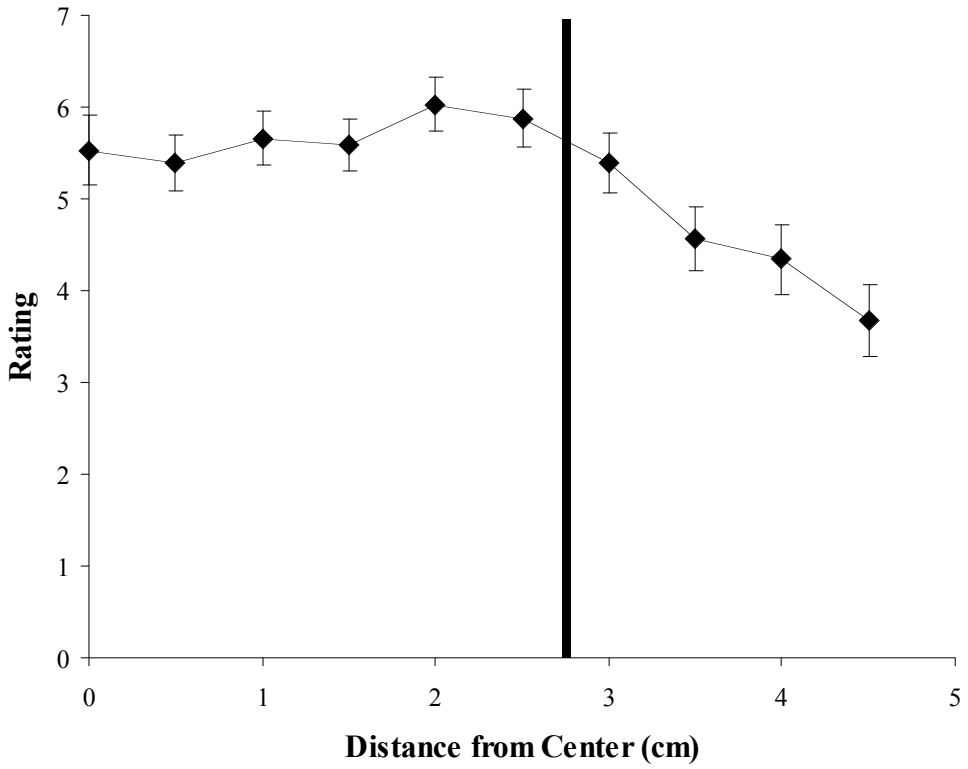


Figure 21. Experiment 5 example goodness ratings. Error bars indicate ± 1 SEM. The black line separates the range of stimuli that were presented during training (left) from the novel region (right).

Figure 21 depicts example goodness ratings as a function of head position's distance from the center. Example goodness differed significantly by distance, $F(9,549) = 8.89$, $p < .001$, $\eta^2 = .13$. However, there was no effect of distance for the 0 – 2.5 cm stimuli that appeared with differential frequency during training, $F(5,305) = 1.25$, *ns*. The most frequent and extreme value (2.5 cm, $M = 5.9$) was not rated as a significantly better example of Martians than the central value ($M = 5.5$), $t(61) = .77$, *ns*. The most

highly rated value (2.0 cm, $M = 6.0$) was also not significantly greater than the central value, $t(61) = 1.17$, *ns*.

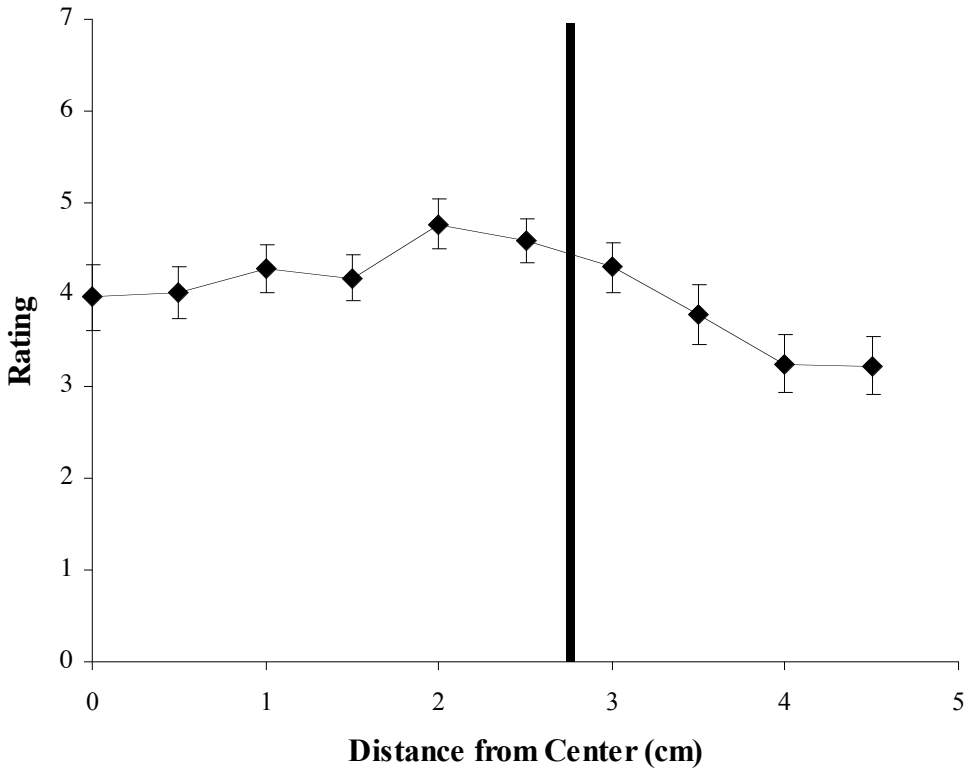


Figure 22. Experiment 5 induction strength ratings. Error bars indicate ± 1 SEM.

Figure 22 depicts induction strength ratings as a function of head position's distance from the center. Induction strength differed significantly by distance, $F(9,549) = 7.915$, $p < .001$, $\eta^2 = .11$. Furthermore, unlike the example goodness ratings, stimuli that appeared with greater frequency during training were rated more highly, $F(5,305) = 3.40$, $p = .005$, $\eta^2 = .05$. There was a significant linear trend, $F(1,61) = 4.87$, $p = .031$, $\eta^2 = .04$. The most frequent and extreme value (2.5 cm, $M = 4.6$) was rated marginally higher than the central value ($M = 4.0$), $t(61) = 1.82$, $p = .073$, $d = .23$. The most highly rated value

(2.0 cm, $M = 4.8$) was significantly greater than the central value, $t(61) = 2.282$, $p = .026$, $d = .30$.

Discussion

Frequency of training presentation did not have any appreciative effect on example goodness ratings. One possible explanation for this is that stimuli within the training range were all presented, albeit with different frequency. As such, they may have been considered equally viable, typical category members. Another possibility is that frequency and central tendency were countervailing forces that balanced each other out in typicality ratings.

In contrast to example goodness, induction strength was influenced by frequency. Frequently-presented familiar stimuli rated higher than the less familiar central tendency stimuli. Also as expected, the extreme familiar values were judged to have the highest induction strength for the Martian category. This is yet another case in which people did not exhibit a single-item diversity effect. Similarly to Experiments 1 and 2, this study is also a case of typicality and induction behavior dissociating.

Chapter IV: The Role of Fluency in Induction

Experiments 3-5 showed that central tendency exemplars are not universally preferred in category-based induction. I have argued that this is due to an influence of a familiarity-based heuristic. However, a proponent of the pure statistical reasoning view could argue that the previous studies only showed that participants did not rely on *global* statistics. The fact that people did not place much inductive value on the summary statistic of central tendency does not entail that they were not using more local statistics. Participants could have been basing their induction on the statistics of sub-regions of the category space. This could account for the influence of frequency in Experiments 4 and 5, and the advantage for the prototype—the central tendency of a cluster—in Experiment 3.

One should note that an appeal to local statistics does not make things any better for any of the extant computational models of category-based induction. As noted earlier, maximum inductive strength for central tendency exemplars is an architectural requirement of many of these models. Bayesian models of induction have been adapted to account for ideal advantages, but ideals are only one extreme in the category space. These models, as presently constructed, cannot account for the symmetrical extreme advantage in Experiments 4. Existing models cannot capture local statistical effects because they base induction on the totality of exemplars in a category, much like exemplar models of classification (e.g. Kruschke, 1992; Nosofsky, 1984). However, there are classification models that can account for local statistical effects, whether by

sampling a limited set of exemplars probabilistically (Nosofsky & Palmeri, 1997) or by dividing the category space into discrete clusters (e.g. Love, Medin, & Gureckis, 2004). An induction model that shared these properties could account for the previous data in a purely statistical manner. To make the case that induction—in these tasks and more broadly—can be driven by familiarity heuristics and not just statistical reasoning, it is necessary to document an effect of familiarity independent of any global or local statistical advantage.

Another way to evaluate the role of familiarity in inductive reasoning is to manipulate the *feeling* of familiarity while holding objective experience constant. Any resulting change in induction behavior would necessarily be attributable to a metacognitive heuristic because statistical properties would be unchanged. The remaining experiments test this heuristic process in induction by manipulating processing fluency. These studies draw on the large body of research showing that processing fluency can be a marker of previous experience and also occur in the absence of actual experience.

In a classic study, Jacoby & Dallas (1981) presented participants with a list of study words. Afterward, participants performed a speeded reading task, reading word presented individually as fast as possible. Some of these words had appeared in the study phase, while others were novel. Overall, old words were read more quickly—i.e. processed more fluently—than novel words. This advantage in fluency only occurred when the words were presented in the same sensory modality at test as they were during study, indicating that this phenomenon is more perceptual than conceptual.

Familiarity increases fluency and, presumably because of this relationship, fluency can create the illusion of familiarity. Johnston, Dark, and Jacoby (1985) had participants read a series of words, followed by another set that contained old and new words. Each word in the test phase was initially presented in a visually degraded form, and clarified progressively until the participant read the word. The participant then judged whether the word had been presented in the first phase or not. Interestingly, old/new classification was strongly predicted by response time. In fact, when they presented participants with pronounceable letter strings instead of common English words, speed and accuracy was more associated with incorrectly classifying new strings as old than incorrectly classifying old strings as new.

Familiarity can also be induced experimentally by manipulating fluency. Exposing people to a list of unfamiliar names increases the likelihood that they will later judge those "people" to be famous (Jacoby, Kelley, Brown, & Jasechko, 1989). Increasing the visual noise on a word decreases the likelihood that that word will be judged old in a recognition memory test, regardless of whether the word truly is old or new (Whittlesea, Jacoby, & Girard, 1990). This latter finding is particularly compelling because the manipulation of visual clarity impacts familiarity despite being truly independent of prior experience.

These fluency manipulations can affect more than familiarity judgments. They can indirectly influence other judgments for which subjective ease is a relevant cue, particularly when that ease can be attributed to familiarity. Visual clarity manipulations alone have been shown to affect judgments of statements' truth (Reber & Schwarz,

1999), confidence in task ability (Alter, Oppenheimer, Epley, & Eyre, 2007), preference formation (Novemsky, Dhar, Schwarz, & Simonson, 2007), and spatial distance (Alter & Oppenheimer, 2008); for reviews of the numerous forms of fluency manipulation and their effects on judgment, see (Alter & Oppenheimer, 2009; Schwarz, 2004). Most relevant for the present discussion, visual clarity has been shown to increase typicality ratings. Oppenheimer and Frank (2008) found that exemplars were rated as better examples of their category when they were presented in a legible font than when presented in an illegible font.

From this brief review, it should be clear that there are at least two distinct kinds of processing fluency manipulations. Memory-based manipulations increase the processing fluency of a stimulus at some point in time by presenting that stimulus at some earlier point. Stimulus-based manipulations increase (or, more often, decrease) processing fluency by changing features of the stimulus itself. It is informative to gauge people's response to both of these.

Experiments 6 and 7 assess the influence of fluency on typicality and induction ratings. Experiment 6 used the "instant fame" prior exposure paradigm to induce familiarity of exemplar labels by presenting some stimuli in an earlier, unrelated task. Experiment 7 used the font legibility paradigm to manipulate fluency by presenting stimuli in easy- or hard-to-read fonts.

Experiment 6

In this study, participants made judgments about the typicality and induction strength of individual exemplars for their broader category, as in earlier studies. Half of

these exemplar labels were presented earlier in an unrelated memory recall task. This was effectively a priming manipulation, making the semantic and orthographic representations of these terms more active. The other half did not appear in the earlier task and were therefore expected to be less active and fluent during the judgment tasks.

I also manipulated two other factors that could plausibly modulate the familiarity effect. One of these was the item's example goodness, as rated under normal circumstances. Some theorists argue that the difference between expected and actual fluency, not some absolute fluency measure, is what influences judgment (Whittlesea & Williams, 2000). If that is the case, then the fluency manipulation should have a relatively greater effect on poor exemples, as they should have a lower expected fluency. The other factor was whether exemplars were organized into their broader categories during the memory task or presented randomly. Organizing items was intended to steer participants into thinking of each word not just as a word to be remembered, but as a member of a category. Previous research (Barsalou, 1985) has shown that an item's frequency of instantiation as a category member—rather than its overall familiarity—is what predicts example goodness ratings.

Method

Participants

104 University of Texas at Austin students participated for course credit or payment of $8. 52 participants viewed category-organized stimuli during the memory task and 52 viewed with randomly organized stimuli.

Materials

4 previously-studied (Barsalou, 1985; McCloskey & Glucksberg, 1978) superordinate categories were used for the judgment tasks: fruits, vegetables, birds, and insects. For each superordinate, 12 exemplar categories were selected for rating. 6 of these had example goodness ratings higher than the median rating for their category in the previous norming studies. The other 6 had lower=than-median example goodness. For the 3 categories that came from Barsalou's norms, there was an additional restriction that high example goodness items also have higher-than-median central tendency values, as measured by pairwise similarity ratings across the category, and likewise for low example goodness items. When more than 12 items satisfied these constraints, selections were made according to experimenter discretion.

8 superordinate categories were used for the memory task: the 4 judgment categories and 4 filler categories, also taken from the norming studies. A full list of superordinate and exemplar categories can be found in Appendix C.

Procedure

For both tasks, participants were tested at individual computers.

Memory Recall. Participants were told that they were performing a memory study. On each trial, they were presented with 12 words. Words appeared one at a time in the center of the screen for 2 seconds, followed immediately by the next word. After the last word, 12 text boxes appeared. Participants typed in as many words as they could remember. They were instructed not to guess. After 2 minutes, if the participant had not

pressed the Enter key to indicate that they had finished, the screen flashed red and the next trial began. There were 8 total trials, one for each superordinate category.

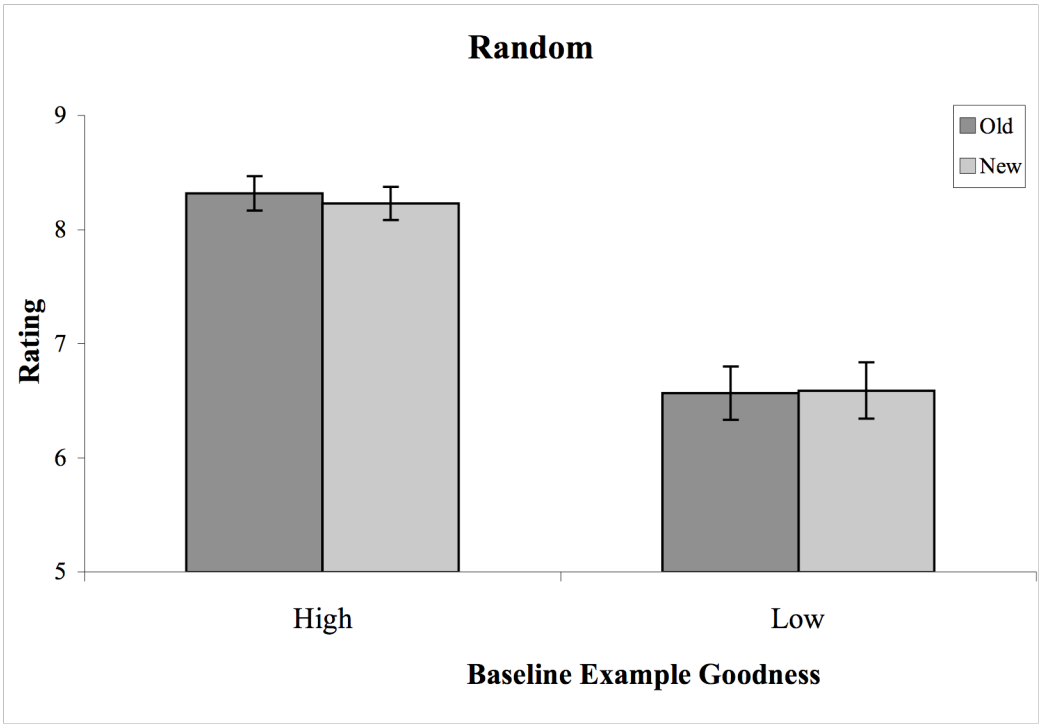
For participants in the Organized condition, each word in a particular trial was a member of the same superordinate category. This category organization was not explicitly highlighted in any way. For participants in the Random condition, the words were randomly distributed across trials.

Judgment tasks. The exemplar goodness and induction strength tasks were structured as in previous experiments. In the induction strength task, the blank predicate was again "Property X".

Results

For simplicity, I present the results separately for each judgment task and method of category organization during the memory task (Random vs. Organized) as 2 × 2 ANOVAs with category novelty in the judgment phase (Old vs. New) and baseline exemplar goodness (High vs. Low) as factors. As in Experiment 1, analyses were performed by participant and item. Results from both analyses are presented below, as well as the min F' statistic. Only graphs from participant analyses appear; item data was broadly similar.

Exemplar goodness rating data appears in Figure 23. The same data appears in Figure 24, collapsing across baseline exemplar goodness to focus on the Old-New effect.



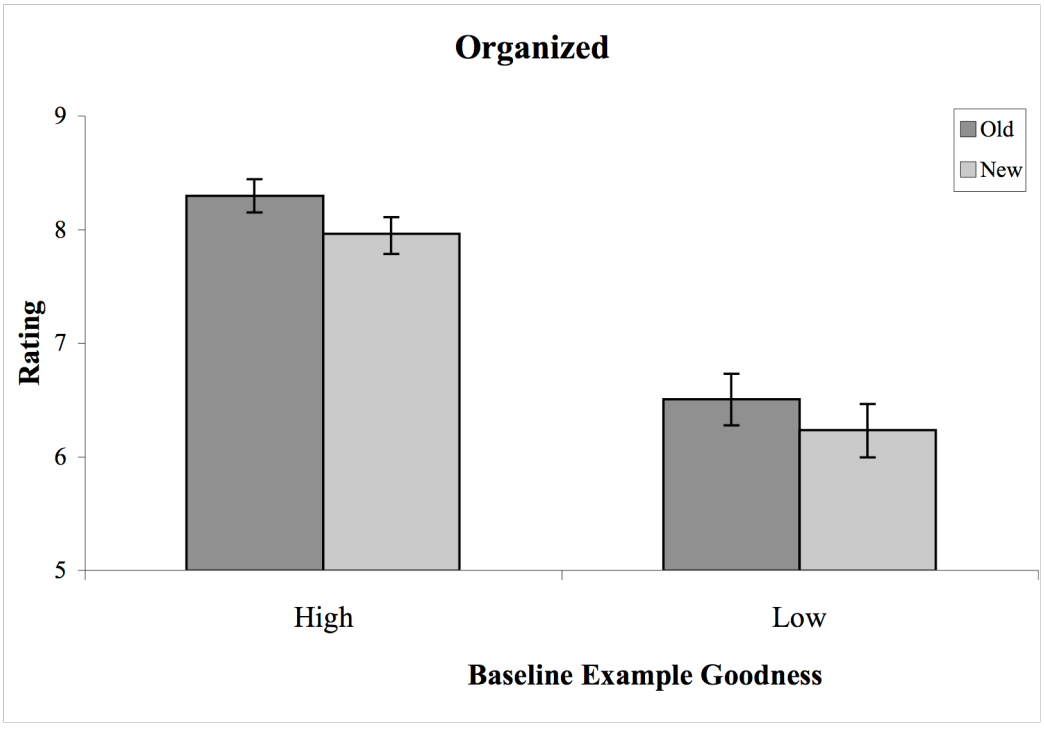


Figure 23. Experiment 6 example goodness ratings. Error bars indicate ± 1 SEM.

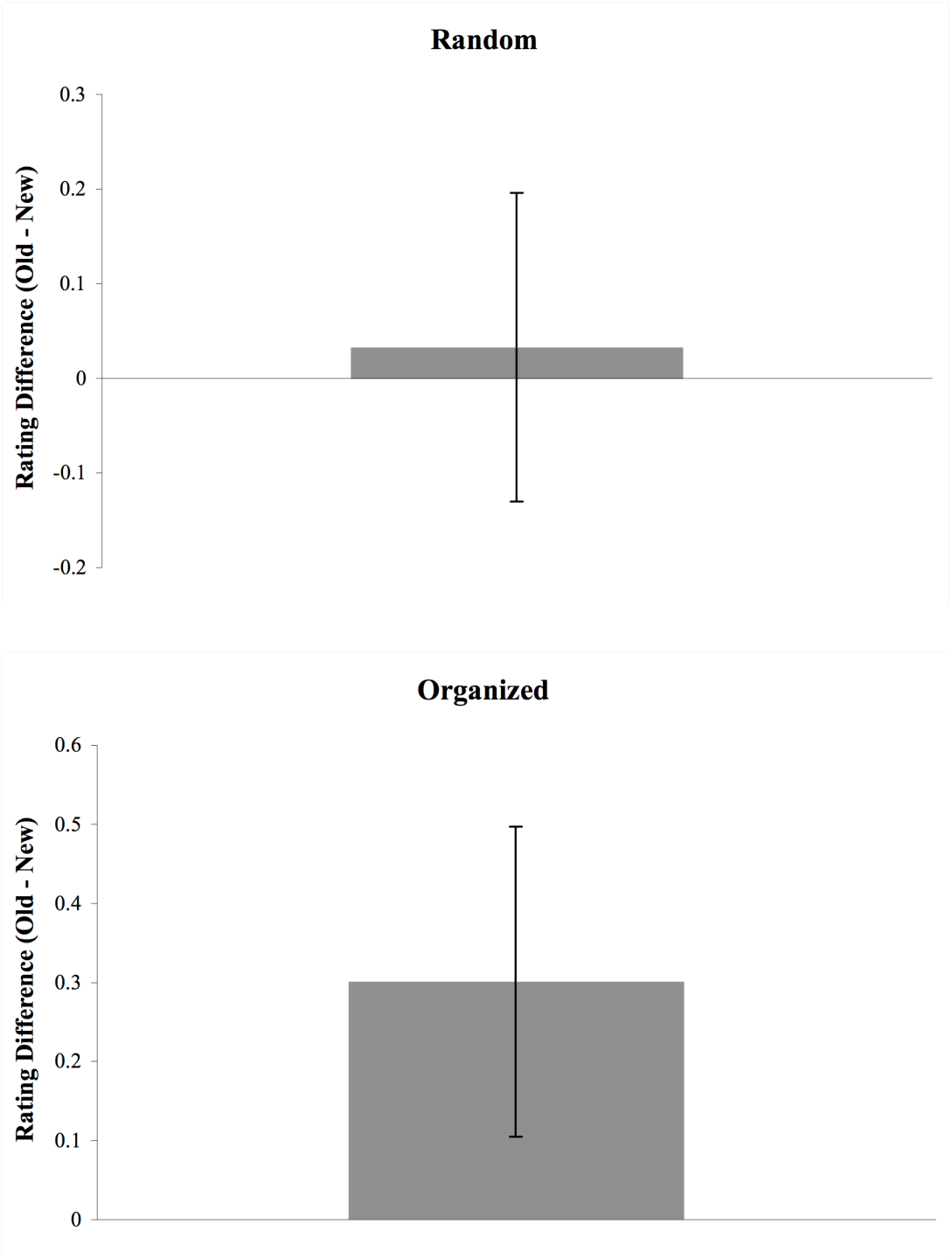
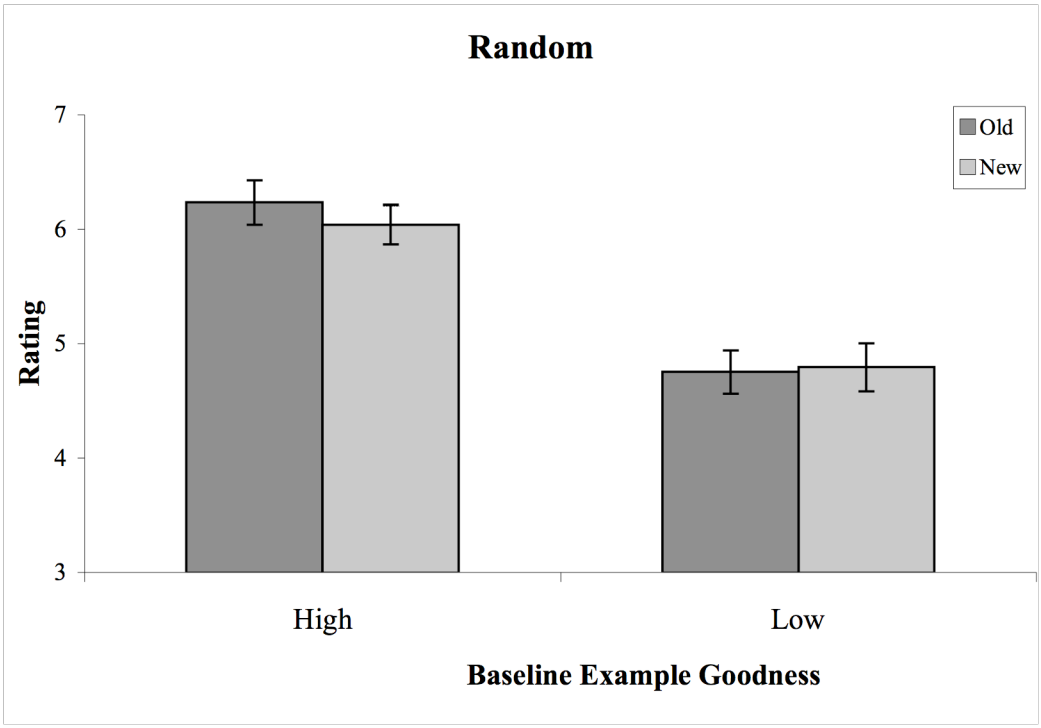


Figure 24. Experiment 6 example goodness ratings, collapsed across baseline example goodness. Error bars indicate 95% confidence intervals.

In the Random condition, there was no effect of stimulus novelty across participants, $F(1,51) = .16$, *ns*; nor was there an effect across items $F(1,46) = .08$, *ns*; *min* $F'(1,85) = .05$, *ns*. There was a strong effect of baseline exemplar goodness, $F(1,51) = 149.54$, $p < .001$, $\eta^2 = .26$; $F(1,46) = 27.52$, $p < .001$, $\eta^2 = .35$; *min* $F'(1,63) = 23.2$, $p < .001$. There was no significant interaction, $F(1,51) = .32$, *ns*; $F(1,46) = .04$, *ns*; *min* $F'(1,57) = .03$, *ns*.

In contrast, there was an effect of stimulus novelty in the Organized condition, $F(1,51) = 9.50$, $p = .003$, $\eta^2 = .01$; $F(1,46) = 8.86$, $p = .005$, $\eta^2 = .01$; *min* $F'(1,96) = 4.58$, $p = .035$. There was also again a strong effect of baseline exemplar goodness, $F(1,51) = 139.07$, $p < .001$, $\eta^2 = .27$; $F(1,46) = 27.57$, $p < .001$, $\eta^2 = .35$; *min* $F'(1,64) = 23.02$, $p < .001$. There was no significant interaction, $F(1,51) = .14$, *ns*; $F(1,46) \approx .00$, *ns*; *min* $F'(1,47) \approx .00$, *ns*.

Results for the induction strength task, which appear in Figures 25 and 26, were broadly similar.



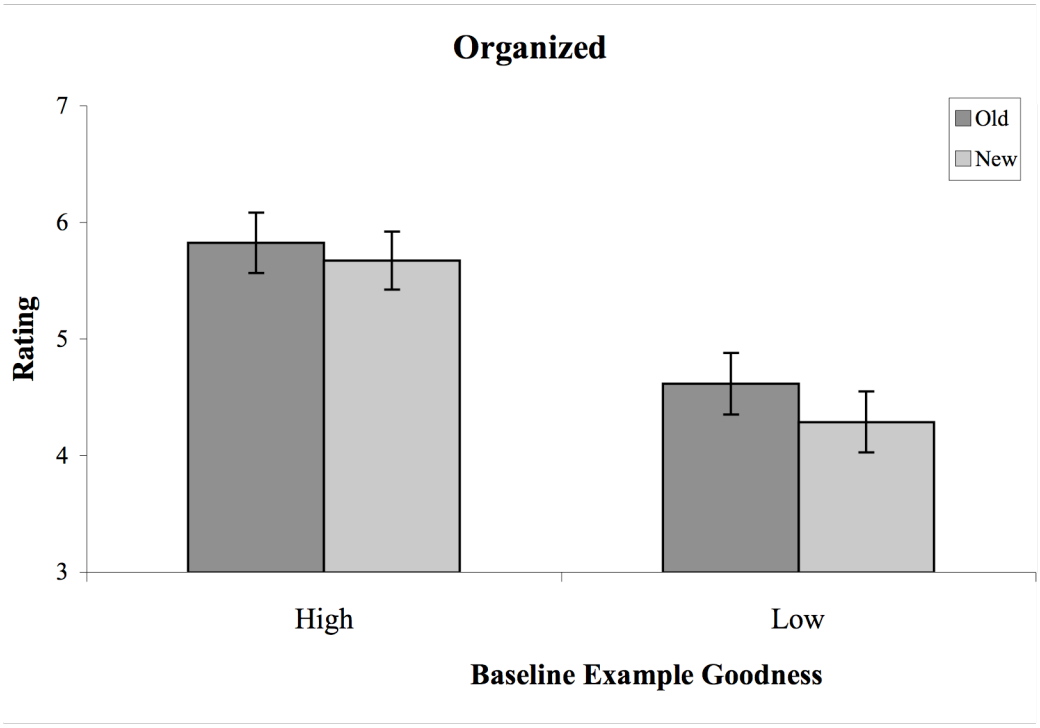


Figure 25. Experiment 6 induction strength ratings. Error bars indicate ± 1 SEM.

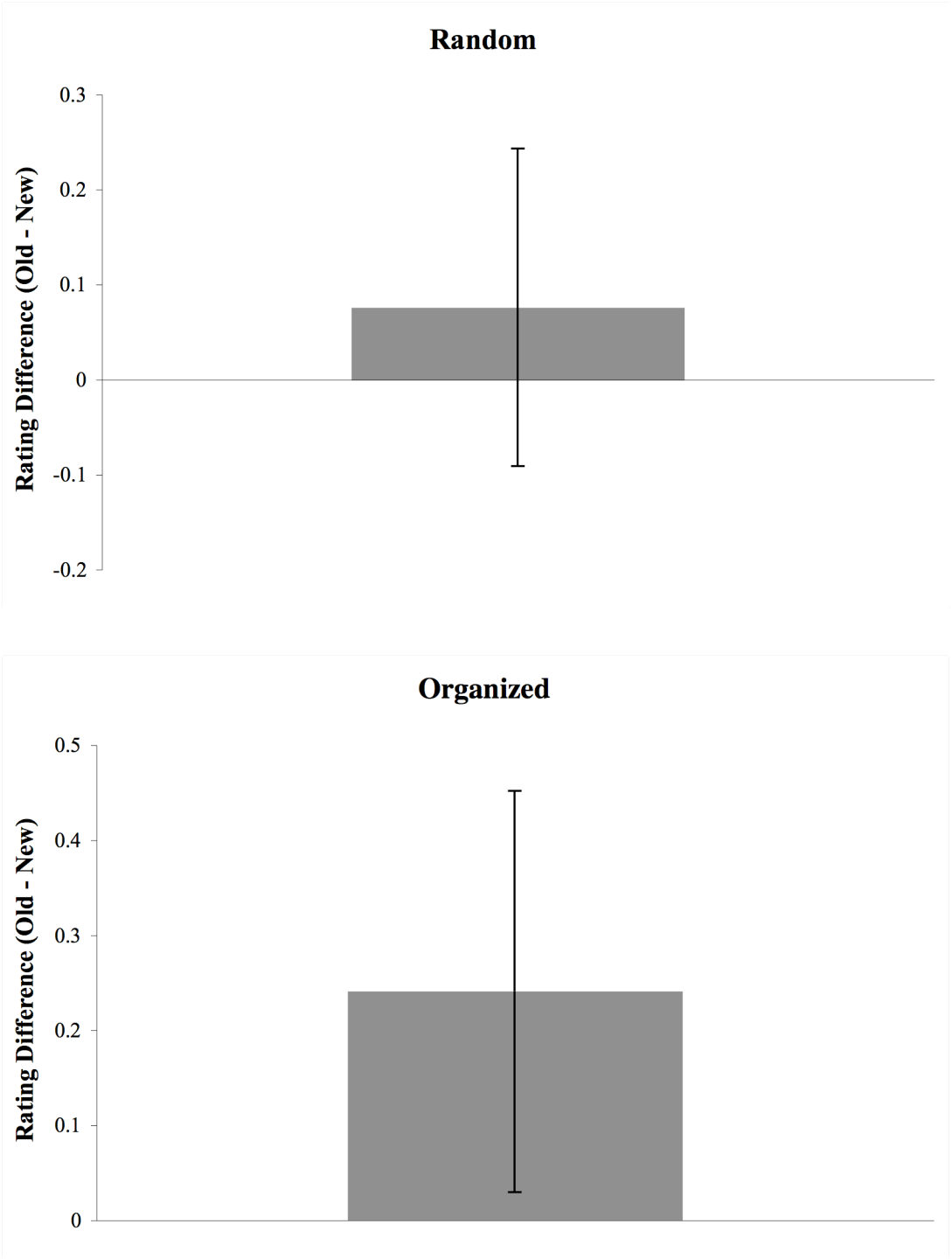


Figure 26. Experiment 6 induction strength ratings, collapsed across baseline example goodness. Error bars indicate 95% confidence intervals.

In the Random condition, there was no effect of stimulus novelty, $F(1,51) = .84$, *ns*; $F(1,46) = 2.72$, *ns*; *min* $F'(1,79) = .64$, *ns*. There was again a strong effect of baseline exemplar goodness, $F(1,51) = 110.24$, $p < .001$, $\eta^2 = .20$; $F(1,46) = 28.04$, $p < .001$, $\eta^2 = .35$; *min* $F'(1,68) = 22.35$, $p < .001$. There was no significant interaction, $F(1,51) = 2.16$, *ns*; $F(1,46) = .83$, *ns*; *min* $F'(1,78) = .60$, *ns*.

In contrast, there was an effect of stimulus novelty in the Organized condition, $F(1,51) = 5.26$, $p = .026$, $\eta^2 = .004$; $F(1,46) = 8.74$, $p = .005$, $\eta^2 = .02$; *min* $F'(1,93) = 3.28$, $p = .07$. There was again a strong effect of baseline exemplar goodness, $F(1,51) = 84.39$, $p < .001$, $\eta^2 = .11$; $F(1,46) = 31.63$, $p < .001$, $\eta^2 = .37$; *min* $F'(1,77) = 23.01$, $p < .001$. There was no significant interaction, $F(1,51) = 1.26$, *ns*; $F(1,46) = 1.46$, *ns*; *min* $F'(1,78) = .68$, *ns*.

Discussion

Not surprisingly, there was a strong typicality effect on example goodness and induction strength, replicating numerous prior studies. As one would expect, this effect occurred regardless of whether and how the items were presented during the earlier memory task. There was also a less robust but significant effect of prior exposure on both example goodness and induction strength ratings. Furthermore, this effect occurred when items were organized by category during the memory task, but there was no advantage for old stimuli when they were presented randomly during the earlier phase.

The advantage for old stimuli indicates that processing fluency influences typicality and induction judgment. The fact that the advantage obtained following

organized but not random presentation indicates that thinking of the words *as category members* was crucial to the subsequent fluency effect. Merely seeing the words beforehand had no discernible effect. In other words, the fluency effect was dependent on priming at a conceptual level.

The difference between old and new stimuli was numerically greater for low example goodness items, but the interaction between baseline goodness and prior exposure did not reach significance. One therefore cannot infer anything positive from this data about the role of relative fluency, though the effect may have simply been too small to pick up statistically. That said, the overwhelming typicality effect for old and new stimuli could be reasonably attributed to the primacy of absolute fluency. I will return to this point following Experiment 7.

Experiment 7

In this study, the fluency of category exemplars was modulated independently of actual familiarity by manipulating the perceptual fluency of stimuli. Exemplars were presented in a fluent, easily legible font for some participants and in a disfluent, less legible font for others.

Method

Participants

130 University of Texas at Austin students participated for course credit or payment of $8. 65 participants viewed fluently-presented stimuli during the judgment tasks and 65 viewed disfluently-presented stimuli.

Categories and exemplars were identical to those in Experiment 1.

Procedure

The procedure was identical to Experiment 1, with one exception. All text presented in the Fluent condition, including the directions, appeared in Times New Roman font (sample font). In the Disfluent condition, all text appeared in Vladimir Script font (sample font).

Results

As in Experiment 1, analyses were performed by participant and item for both judgment tasks. Results from both analyses are presented below, as well as the min F' statistic. Only graphs from participant analyses appear; item data was broadly similar.

Exemplar goodness rating data appears in Figure 27. The same data appears in Figure 28, collapsing across exemplar and category types to focus on the Fluent-Disfluent effect.

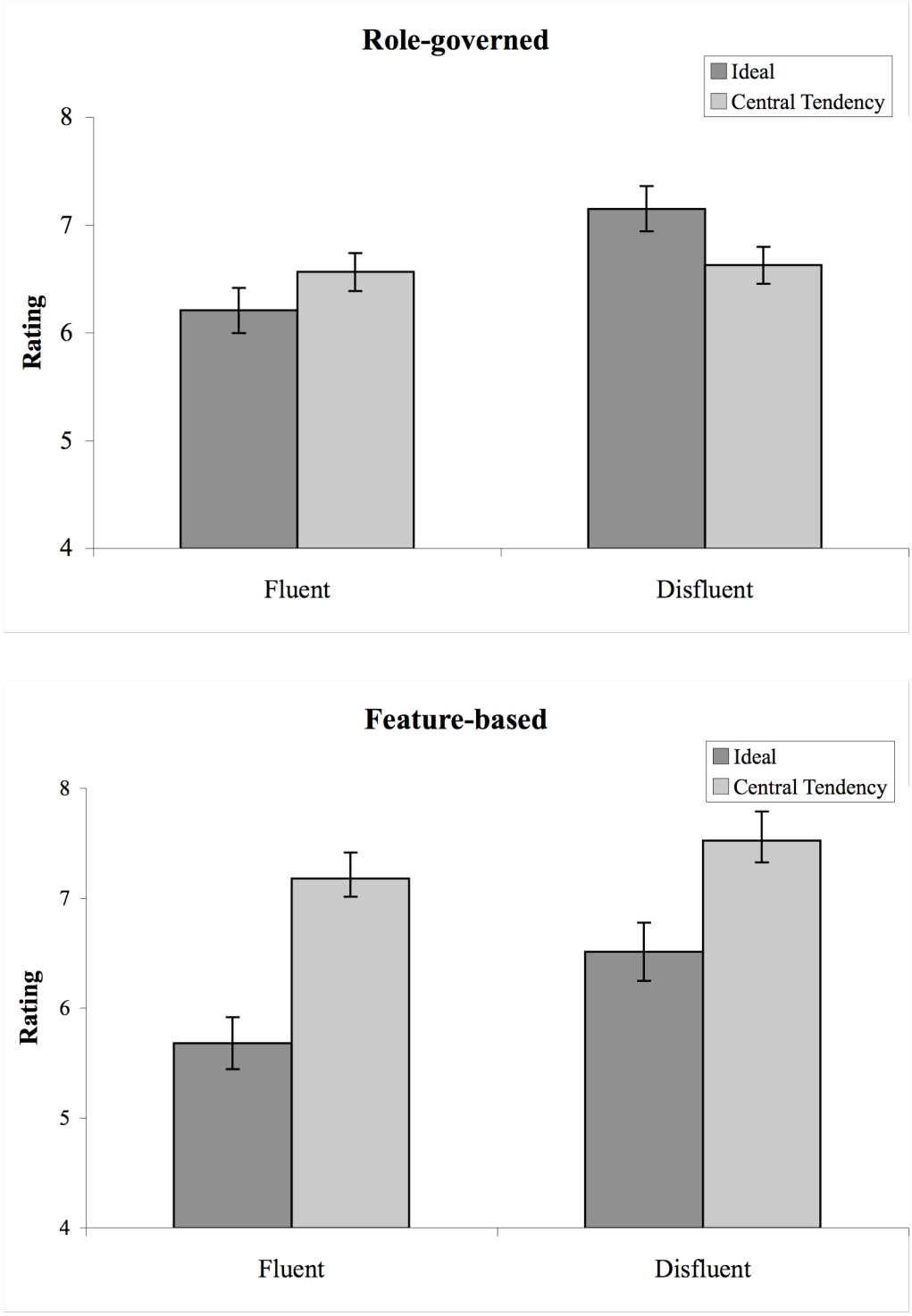


Figure 27. Experiment 7 example goodness ratings. Error bars indicate ± 1 SEM.

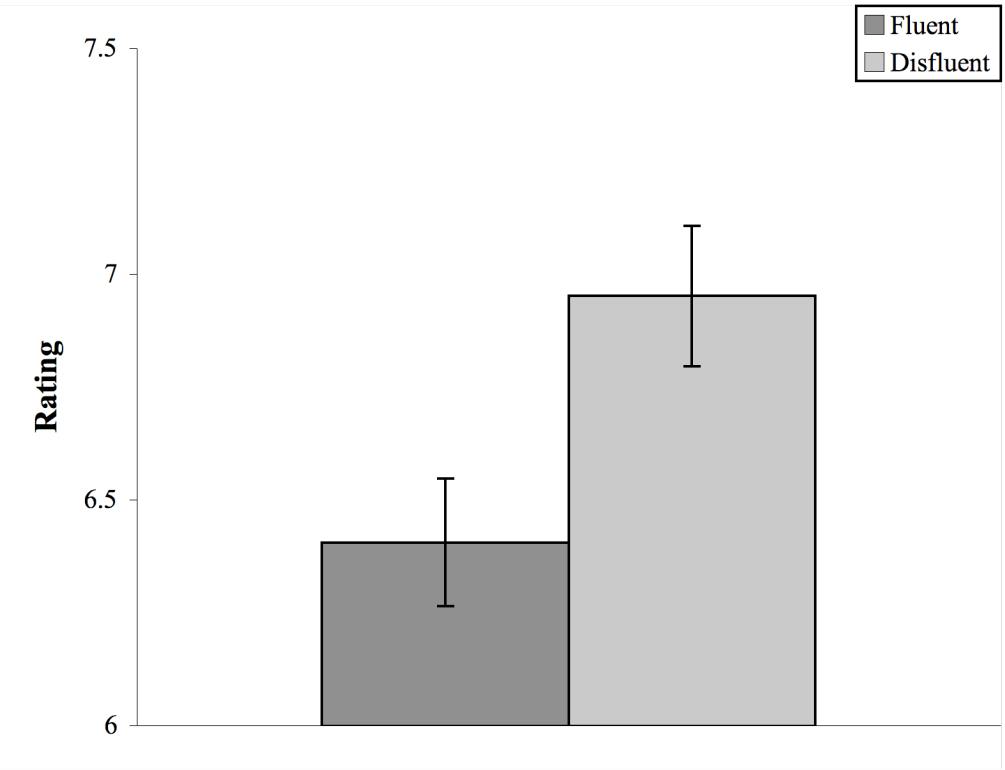


Figure 28. Experiment 7 example goodness ratings, collapsed across category and exemplar type. Error bars indicate ± 1 SEM.

As in Experiment 1, there was no significant main effect of category type, $F(1,128) = 1.70$, *ns*; $F(1,22) = .23$, *ns*; *min* $F'(1,28) = .20$, *ns*. There was a significant main effect of exemplar type by participants, $F(1,128) = 13.27$, $p < .001$, $\eta^2 = .03$; and by items, $F(1,22) = 6.66$, $p = .02$, $\eta^2 = .09$; *min* $F'(1,48) = 4.43$, $p = .04$. There was, as in Experiment 1, a significant interaction between the category type and exemplar type, $F(1,128) = 42.67$, $p < .001$, $\eta^2 = .04$; $F(1,22) = 8.36$, $p < .001$, $\eta^2 = .11$; *min* $F'(1,31) = 6.99$, $p = .01$. Critically for the present experiment, there was also a significant main effect of fluency across participants, $F(1,128) = 6.75$, $p = .01$, $\eta^2 = .02$; and items $F(1,22) = 40.17$, $p < .001$, $\eta^2 = .03$; *min* $F'(1,150) = 5.78$, $p = .02$. Interestingly, there was also a significant

interaction between exemplar type and fluency, $F(1,128) = 4.52$, $p = .035$, $\eta^2 = .01$; $F(1,22) = 22.89$, $p < .001$, $\eta^2 = .03$; *min* $F'(1,150) = 3.77$, $p = .05$; reflecting the relative increase in ratings for ideal exemplars in the Disfluent condition. No other interactions reached significance.

Induction strength rating data appears in Figure 29. The same data appears in Figure 30, collapsing across exemplar and category types to focus on the Fluent-Disfluent effect.

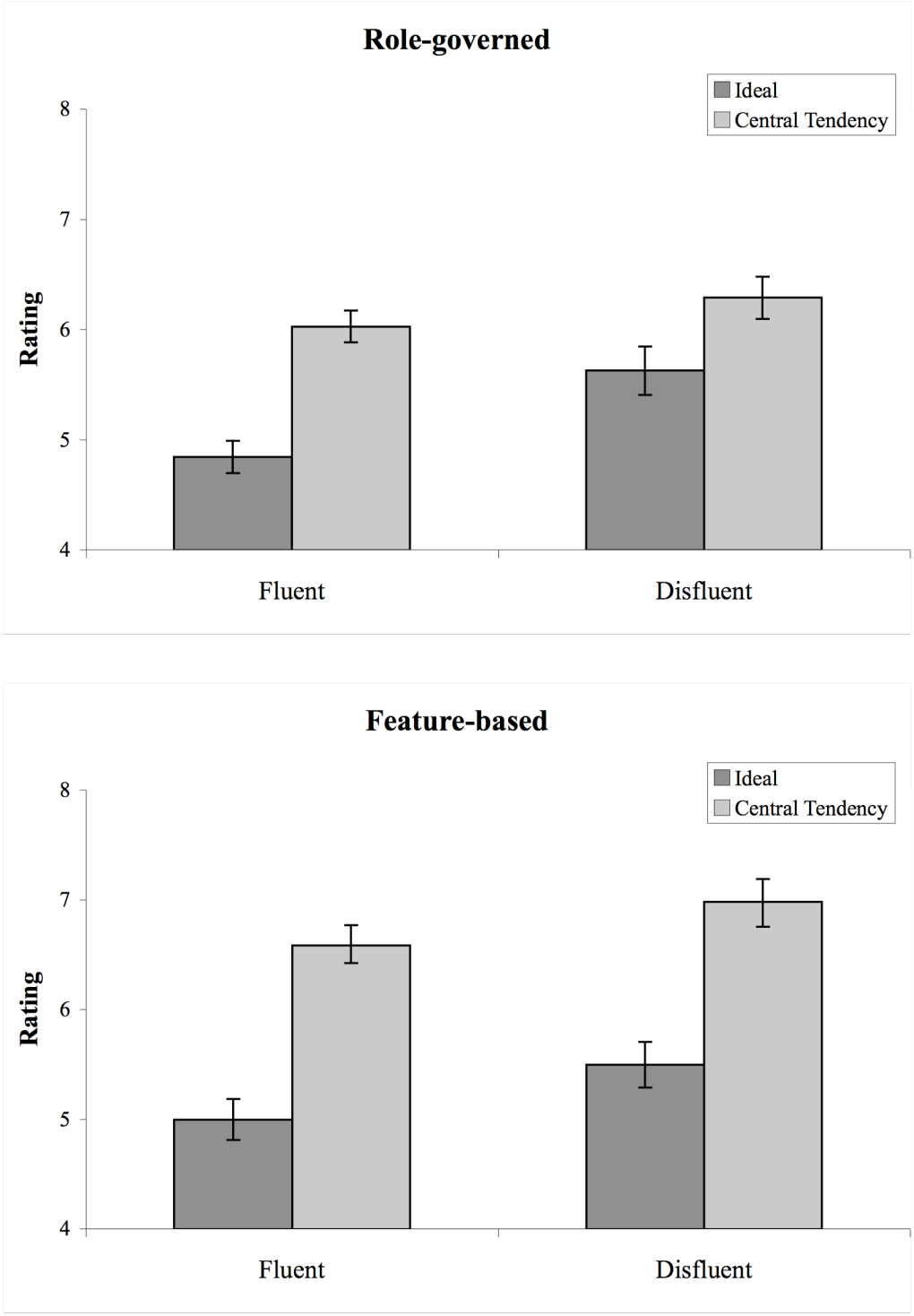


Figure 29. Experiment 7 induction strength ratings. Error bars indicate ± 1 SEM.

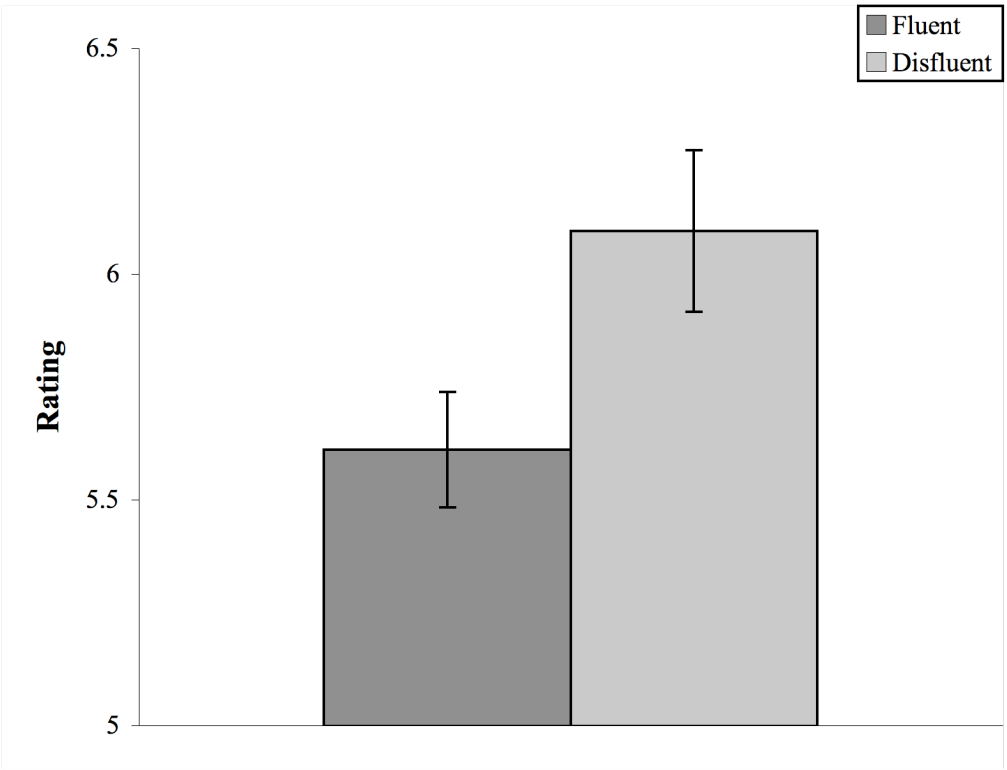


Figure 30. Experiment 7 induction strength ratings, collapsed across category and exemplar type. Error bars indicate ± 1 SEM.

As in Experiment 1, there was a main effect of category type by participants, $F(1,128) = 28.13$, $p < .001$, $\eta^2 = .01$; but not by items, $F(1,22) = 1.95$, *ns*; *min* $F'(1,25) = 1.82$, *ns*. There was also a main effect of exemplar type, $F(1,128) = 144.72$, $p < .001$, $\eta^2 = .14$; $F(1,22) = 40.65$, $p < .001$, $\eta^2 = .38$; *min* $F'(1,36) = 31.73$, $p < .001$. There was also a significant interaction between category and exemplar type by participants, $F(1,128) = 11.40$, $p = .001$, $\eta^2 = .01$; though not by items, $F(1,22) = 2.69$, *ns*; *min* $F'(1,33) = 2.18$, *ns*. As with exemplar goodness, there was a significant main effect of fluency, $F(1,128) = 4.86$, $p = .03$, $\eta^2 = .02$; $F(1,22) = 34.57$, $p = < .001$, $\eta^2 = .02$; *min* $F'(1,149) = 4.26$, $p = .04$. No other interactions reached significance.

Discussion

The results of Experiment 7 were broadly similar to those of Experiment 1. In the example goodness task, there was a substantial interaction between category type and exemplar type: central tendency exemplars were rated as much better examples than ideal exemplars for feature-based categories, but they were rated at similar levels for role-governed categories. For the induction strength task, the most significant pattern in the data was the sizable advantage of central tendency exemplars over ideals for both category types. Although there was a category type by exemplar type interaction, it was rather small by comparison.

As in Experiment 6, there was also a main effect of fluency. Interestingly, this effect was in the opposite direction. Unlike the advantage found in Experiment 6 for older fluent stimuli, participants gave higher ratings in both tasks for less legible disfluent stimuli in Experiment 7. This is likely due to "spontaneous discounting" of fluency information. Previous research has shown that when people are consciously aware that they are experiencing fluency from an irrelevant source, they ignore or even overcompensate for this unreliable cue (Alter & Oppenheimer, 2009; Bornstein & D'Agostino, 1994; Oppenheimer, 2004). In the present study, participants in the Disfluent condition overcorrected for the illegibility of the font, rating stimuli higher than they would have by basing judgment purely on semantic content, as in the Fluent condition.

The effect of fluency on example goodness ratings was particularly strong for ideal exemplars. As I argued above, these exemplars should less familiar than their

central tendency counterparts. This differential effect could suggest that relative fluency has a greater influence on judgment than absolute fluency. However, there was no statistically reliable trend for induction strength ratings. Given the similar ambiguity in Experiment 6, I will refrain from making any strong claims about the role of relative vs. absolute fluency.

Chapter V: Summary and Discussion

In Experiment 1, I explored the nature of the canonical typicality effect in category-based induction. Using a set of previously studied natural categories, I showed that the relative fit of central tendency and ideal exemplars as examples depends on whether their broader category is role-governed or feature based. In contrast, central tendency exemplars were rated as having a substantially greater induction strength for both kinds of categories.

In Experiment 2, I replicated these findings in an artificial category learning setting, in which central tendency and idealness were manipulated orthogonally and simultaneously. I also included negative ideal stimuli, which were exceptionally bad at fulfilling the functional role of the category. Results showed that distance from the central tendency determined example goodness following feature-based category learning, while both central tendency and idealness determined goodness following role-governed learning. As in Experiment 1, only central tendency influenced induction strength. I concluded that the typicality effect was not driven by typicality per se.

In Experiment 3, I began to address the nature of the central tendency effect. In particular, I addressed the possibility that familiarity could be a more powerful influence on induction than central tendency. After learning quaternary-valued, multi-clustered categories, I asked participants to rate two kinds of novel stimuli: prototypes of individual clusters (familiar) and mixtures of values across both clusters (central tendency). The familiar prototypes were rated as better examples and better sources of induction.

The nature of the central tendency exemplars in Experiment 3 did not necessarily match the lay understanding of central tendency. Experiment 4 addressed this by separating categories along a continuous-valued dimension. Following a category-learning phase similar to Experiment 3 participants again rated familiar and (now intuitive) central tendency stimuli. Familiar stimuli again received higher ratings.

In Experiment 5, I addressed potential concerns that the central tendency exemplars were not viable category members in Experiments 3 and 4, and that their lower ratings suffered from this unnatural non-viability. In this study, participants learned the category structure through an observational learning task in which the central tendency of the category was explicitly viable. While there was no strong effect on example goodness, participants again preferred familiar stimuli as induction sources over central tendencies. In addition to the familiarity effects, Experiments 3-5 also demonstrated that there is no single-item diversity effect. When a category was composed of two distinct clusters, the central tendency provided relatively little inductive value, despite offering maximal coverage of the category space.

Experiments 3-5 all induced familiarity through some statistical property, albeit not central tendency. Experiment 6 was designed to induce familiarity independently from actual experience. I used a perceptual fluency manipulation, by which a stimulus feels more familiar because it has been presented earlier. In this case, exposure to exemplar labels was manipulated while exposure to the exemplars themselves was unaffected. Results showed that fluency through prior exposure increased example goodness and induction strength ratings, but only when that prior exposure encouraged

participants to think of the exemplar labels *as category members* and not just to read them on a surface perceptual level.

Experiment 7 also employed a fluency manipulation, this time on a purely perceptual level. Some participants rated stimuli presented in a standard legible font while others viewed the same stimuli in an illegible font. Interestingly, participants in the illegible Disfluent condition rated the identical stimuli as having greater example goodness and induction strength.

It should be clear from these experiments that familiarity plays an important role in category-based induction. It should also be clear that this role is fairly complex. Some sources of familiarity clearly overlap with existing models of statistical category-based inductive reasoning. Other sources are clearly distinct from this statistical view. Some sources of fluency promote induction while others do not. Figure 31 is a schematic view of this complex relationship.

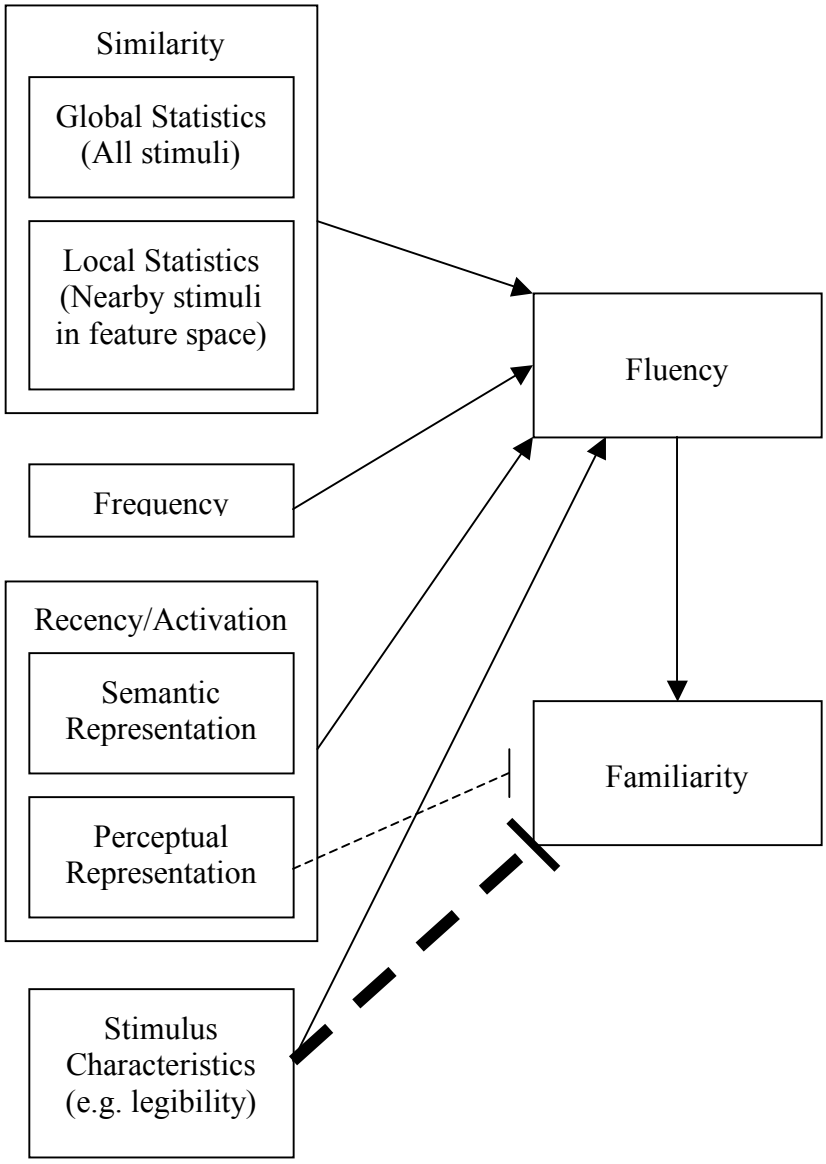


Figure 31. Sources of fluency and familiarity in category-based induction. Solid lines are excitatory; dashed lines are inhibitory. Weight of lines corresponds to relative influence.

Figure 31 depicts all of the inputs that were manipulated in the seven experiments and may or may not be an exhaustive list of the influences on familiarity more generally. Similarity refers to feature overlap or distance in some representational space. Global statistics, like central tendency, register an item to all exemplars in the category. This is

the kind of information that was most dominant in Experiments 1 and 2. Local statistics register an item to its nearest neighbors in the category space, as in Experiments 3, 4, and 5. Frequency is simply the summed experience with an item, as was manipulated in Experiments 4 and 5. Recency, or activation of a representation, was manipulated in Experiment 6. These are all obviously related to familiarity, as they all influence the likelihood of recall. They all increase the mnemonic fluency, or availability, of an item. They also all increase induction strength, with the exception of perceptual recency, which has no effect at best and a slight inhibitory effect at worst. Finally, there are characteristics of the stimulus, such as font legibility in Experiment 7. While these can increase or decrease fluency, that effect on fluency is manifestly unrelated to prior experience. Therefore, people tend to overcorrect for the influence of spurious fluency cues.

Clearly, there is a variety of influences on induction. Some of these are the kinds of statistical properties that are well known and have been incorporated into previous models. Others, like frequency, are certainly statistical in nature, but have not been accounted for by existing theories. Still others cannot at all by described in statistical terms, suggesting that even a broad view of statistical reasoning is insufficient to capture induction behavior.

It should not be overly surprising that statistics do not tell the whole story about category-based induction. While the statistical approach to categorization has a rich history and many successes, it is also clearly incomplete. Role-governed categories are defined by relational structures, not a set of common features. Some abstract categories,

like *cooperation*, have a high degree of coherence despite lacking any obvious statistical structure of exemplars (Loewenstein, 2009). Statistical similarity has difficulty capturing theoretical causal knowledge (Murphy & Medin, 1985). Even simple statistical relationships interact with conscious explanations of category information (Williams & Lombrozo, 2009; Wisniewski & Medin, 1994). Beyond classification and generalization, categories are used for many other functions and can be influenced by non-statistical properties of the learning environment (Markman & Ross, 2003).

One open question is how the results reported here relate to experts' more sophisticated induction. Medin, Coley, and colleagues have found that many experts' inductive reasoning has a character distinct from novices'. For instance, tree experts such as landscapers and park maintenance workers consider an item to be a good example of the *tree* category if it has more extreme values on ideal dimensions like height and weediness (Lynch et al., 2000). However, when told that two trees have two different diseases, the relative idealness of the trees has no influence on which disease the experts expect to be present in all trees. Interestingly, the relative central tendency also has no influence. Instead, tree experts rely on causal ecological knowledge, like a tree's relative susceptibility to disease and which trees are likely to be nearby (Proffitt, Coley, & Medin, 2000). Similarly, indigenous Maya, who are experts in their native ecology, consider turkeys to be the best example of birds because of ideal features like large size, obvious markings, and being a plentiful food source (Atran, 1999). Nevertheless, their induction about the presence of disease in birds is based on ecological knowledge, not ideals or central tendency (Bailenson, Shum, Atran, Medin, & Coley, 2002).

There are three principal differences between this work on experts' ideal-represented categories and the work reported in this paper. First, the experts have a great deal of functional knowledge, but trees and birds are still natural kind categories. It is therefore not obvious where they fall in regard to the feature-based/role-governed distinction or how they compare to some of artificial category structures used here. Second, as Medin, Coley, and colleagues have noted, experts have a rich set of causal knowledge that novices do not have, even for familiar everyday categories. Third and perhaps most critically, when induction is performed with non-blank predicates like diseases, this vast knowledge can be used for complex causal reasoning that is not possible for blank predicates that are unrelated to this knowledge base. It is possible that experts rely on central tendency and other familiarity cues for blank predicates, just as novices have been shown to reason causally about natural categories and non-blank predicates (E. E. Smith et al., 1993). Still, it seems prudent to suggest that expertise places boundary conditions on the effect of familiarity on induction. More generally, the effect of familiarity is likely mitigated when the reasoner has knowledge about the domain, motivation to make an accurate judgment, and sufficient resources to use more complex reasoning strategies. Weak methods like familiarity heuristics are useful in a variety of circumstances and may be the default, but should not be considered the only method in the cognitive toolkit.

In addition to induction *per* se, the current results also inform our understanding of the relationship between categorization and category-based induction. Considering the overlap in category knowledge used in both behaviors, it is tempting to think of them as

essentially the same process. In some tasks, categorization and induction behavior are remarkably similar and both can be predicted by a single similarity-based model (Heit & Hayes, 2008; Sloutsky & Fisher, 2004). In other tasks, however, these two behaviors are dissociated. Murphy and Ross (2005) showed that certainty of classification in the presence of contrast categories affects categorization, but not induction. In their work with American college students and indigenous Maya, Coley, Medin, and Atran (1997) showed that the basic level of categorization differed according to cultural expertise, but the genus level was inductively privileged for both groups. The experiments reported here show yet another way in which categorization behavior—e.g. example goodness judgments—and induction are dissociable.

One reason for this divergence is that example goodness is arguably a more complex judgment than induction. In addition to influences of availability and statistical relationships, exemplars also vary in their informative value to others. Attaching a category label to an item is fundamentally a communicative act (Brown, 1958). While an exemplar's idealness for a functional role, distance from a category boundary, or proximity to a global central tendency between clusters may not strongly influence familiarity, they nonetheless convey important information about the category.

Conclusion

These seven experiments provide two rejoinders to commonly-held beliefs about the nature of category-based induction. First, the well-documented typicality effect is actually a central tendency effect. Second, an item's familiarity strongly influences induction, and can be even more important than its location in the category space. This

influence of familiarity can occur entirely independently of any actual experience, though the reliability of a fluency cue as diagnostic of familiarity is critical and complex. These findings suggest that computational models of induction that simulate statistical reasoning are incomplete and should incorporate familiarity as a central construct.

Appendix A

Ideal and Central Tendency Characteristic Sets used in Experiments 1 and 7

Role-Governed Categories

Author

Ideal: Good writer, Funny, Books aren't too long, Clear, Relatable

Central Tendency: Middle-aged, Imaginative, Intellectual, Reclusive, Large Vocabulary

Customer

Ideal: Has a lot of money, Polite, Friendly, Nice, Knows what they want

Central Tendency: Female, Impatient, Has money, Demanding, Buying something

Drug

Ideal: No side effects, Inexpensive, Not addicting, Legal, Not Impairing

Central Tendency: Cures, Alters mind, Pain killer, Addicting, Helpful,

Friend

Ideal: Understanding, Intelligent, Sympathetic, Able to keep secrets, Compatible

Central Tendency: Nice, Helpful, Kind, Always there for you, Fun

Game

Ideal: Interesting, Challenging, Active, Changes frequently, Easy to learn

Central Tendency: Competitive, Rules, Athletic, Winners, Losers

Gift

Ideal: Practical, Something recipient wants, Liked by recipient, Hand-made, Reusable

Central Tendency: Wrapped, Card, Small, On special occasions, On birthdays

Guest:

Ideal: Clean, Courteous, Has manners, Fun, Unobtrusive

Central Tendency: Family, Friend, Kind, Dressed up, Invited

Home

Ideal: Looks good, Good location, Big, Comfortable, Pool

Central Tendency: Has a kitchen, Where family is, Roof, Windows, Made of brick

Hobby

Ideal: Inexpensive, Social, Intellectually stimulating, Exciting, Healthy

Central Tendency: Time-consuming, Relaxing, Done Alone, Requires a skill, Done regularly

Job

Ideal: High paying, Flexible hours, Good boss, Good co-workers, Fun

Central Tendency: Boring, Long hours, Low wages, Make money, Time consuming

Pet

Ideal: Playful, Easy to care for, Loyal, Obedient, Friendly

Central Tendency: Furry, Loveable, Loving, Soft, Fun

Predator

Ideal: Smart, Strong, Agile, Cunning, Camouflaged

Central Tendency: Mean, Sharp teeth, Big, Has Claws, Hungry

Feature-Based Categories

Beer

Ideal: Tastes good, Inexpensive, Non-fattening, Healthy, Flavorful

Central Tendency: Bottled, or Canned, Carbonated, Yellow, Alcoholic

Bicycle

Ideal: Comfortable, Fast, Looks Good, Cheap, Durable

Central Tendency: Handlebars, Seat, Made of Metal, Pedals, Has reflectors

Cell Phone

Ideal: Camera, Durable, Easy to use, Inexpensive, Has lots of memory

Central Tendency: Screen, Key pad, Ringtones, Buttons, Flip phone

Chair

Ideal: Adjustable, Reclining, Wheels, Leather, Rolls around/has wheels

Central Tendency: Wooden, Four legs, Backrest, Arms, Seat

Fridge

Ideal: Spacious, Energy efficient, Goes well with room, Looks good, Has a water dispenser

Central Tendency: Has a freezer, Stores food, White, Has two doors, Has an automatic light inside

Knife

Ideal: Easy to hold, Comfortable, Cuts well, Durable, Light weight

Central Tendency: Dangerous, Metal blade, Silver blade, Shiny, Black handle

Microwave

Ideal: Quiet, Fast, Energy efficient, Inexpensive, Easy to use

Central Tendency: Black, Heats food, Buttons, Turning plate, Light inside

Shoes

Ideal: Look good, Inexpensive, Match everything, Cute, Last a long time

Central Tendency: Laces, Rubber, White, Soles, Leather

Table

Ideal: Sturdy, Looks good, Strong, Adjustable, Big

Central Tendency: Four legs, Flat top, Round, or Rectangular, Glass top

Television

Ideal: Clear reception, High Definition, Big, Light weight, Inexpensive

Central Tendency: Black, Color picture, Square, Remote control, Screen

Truck

Ideal: Fuel efficient, Powerful, Carries a lot, Roomy, Strong

Central Tendency: Four wheels, Bed, Gas-guzzling, Two-door, Loud

Website

Ideal: Easy to use, Looks good, Useful, Entertaining, Fast-loading

Central Tendency: Links, Colorful, Pictures, Ads, Search

Appendix B

Experiment 1 subjects analysis ANOVA table

Effect	*F*	df	*p*
Judgment task	22.75	1,45	< .001
Category type	10.64	1,45	.002
Exemplar type	19.482	1,45	< .001
Judgement task x Category type	.25	1,45	.619
Judgment task x Exemplar type	5.42	1,45	.025
Category type x Exemplar type	21.53	1,45	< .001
Judgment task x Category type x Exemplar type	4.7	1,45	.035

Experiment 1 items analysis ANOVA table

Effect	*F*	df	*p*
Judgment task	138.01	1,22	< .001
Category type	2.262	1,22	.147
Exemplar type	29.12	1,22	< .001
Judgement task x Category type	0.309	1,22	0.584
Judgment task x Exemplar type	20.47	1,22	< .001
Category type x Exemplar type	12.057	1,22	.002
Judgment task x Category type x Exemplar type	12.591	1,22	.002

Experiment 2 ANOVA table

Effect	*F*	df	*p*
Judgment task	46.522	1,63	< .001
Idealness	5.132	2,126	.007
Central tendency	41.506	2,126	< .001
Label relevance	.038	1,63	.846
Judgment task × Idealness	6.061	2,126	.003
Judgment task × Central tendency	5.675	2,126	.004
Judgment task × Label relevance	.004	1,63	.949
Idealness × Central tendency	1.445	4,252	.22
Idealness × Label relevance	3.491	2,126	.033
Central tendency × Label relevance	1.802	2,126	.169
Judgment task × Idealness × Central tendency	3.192	4,252	.014
Judgment task × Idealness × Label relevance	0.655	2,126	.521
Judgment task × Central tendency × Label relevance	2.872	2,126	.06
Idealness × Central tendency × Label relevance	0.702	4,252	.591
Judgment task × Idealness × Central tendency × Label relevance	2.572	4,252	.038

Appendix C

Table C1
Categories Used in Experiment 6 Judgment Tasks

Fruits	Goodness	CT	Vegetables	Goodness	CT
Apple[H]	8.7	4.5	Green beans[H]	7.6	4.4
Orange[H]	8.4	4.6	Spinach[H]	7.4	4.4
Strawberries	8.1	4.4	Corn	7.3	3.7
Banana	7.1	3.4	Zucchini[H]	7.2	4.7
Pear[H]	6.8	4.6	Carrot	7.2	3.9
Peach[H]	6.2	5.1	Peas	7.1	4.0
Plum	6.2	5.0	Broccoli[H]	6.7	4.5
Pineapple	6.2	3.6	Squash	6.6	4.7
Apricot[H]	6.1	4.9	Lettuce[H]	6.6	4.2
Nectarine	5.9	4.8	Asparagus	6.5	4.1
Tangerine	5.9	4.6	Cauliflower[L]	5.8	3.9
Grapes[L]	5.7	4.4	Cucumber	5.7	4.2
Cherry[L]	5.1	4.8	Celery	5.5	4.3
Watermelon[L]	5.1	2.9	Cabbage[L]	5.4	4.1
Berries[L]	4.7	4.4	Artichoke[L]	4.6	3.7
Lemon[L]	4.1	3.6	Potato[L]	4.5	3.4
Blueberries	4	4.4	Beans	3.7	3.9
Raisins[L]	3.3	3.7	Sprouts[L]	3.4	3.7
			Onions[L]	3.4	3.0

Birds	Goodness	CT	Insects	Goodness
Robin[H]	8.4	5.4	Fly[H]	9.8
Bluejay[H]	7.9	5.1	Mosquito[H]	9.8
Sparrow[H]	7.8	4.9	Ant[H]	9.4
Bluebird[H]	7.6	5.3	Wasp[H]	9.2
Blackbird[H]	7.6	5.2	Beetle[H]	9.0
Parakeet	7.2	4.7	Bee[H]	8.9
Parrot	7.2	4.3	Flea	8.8
Seagull	7.2	3.8	Moth	8.7
Canary[H]	7.1	4.7	Locust	8.6
Eagle	7.1	3.9	Firefly	8.6
Hummingbird	6.7	3.7	Grasshopper	8.5
Pigeon	6.3	4.7	Termite[L]	8.4
Hawk[L]	6.2	3.8	Butterfly	8.1

Cardinal	5.8	5.1	Caterpillar[L]	7.6
Dove	5.3	4.7	Centipede[L]	7.6
Oriole	5.1	5.1	Millipede[L]	7.5
Falcon	4.6	4.2	Spider	7.2
Condor[L]	4.4	3.4	Lice[L]	6.3
Finch	3.8	4.9	Tarantula	6.3
Chicken[L]	3.7	2.8	Silkworm[L]	6.2
Pelican[L]	3.7	2.6	Scorpion	5.3
Thrush	3.4	5.0	Leech	4.9
Ostrich[L]	2.4	2.2		
Penguin[L]	2.4	2.1		

Goodness = Exemplar goodness rating in original norming studies (Barsalou, 1985; McCloskey & Glucksberg, 1978); CT = Central tendency; H = High exemplar goodness items in judgment trials; L = Low exemplar goodness items; unmarked items were fillers on memory trials.

Table C2

Filler Categories for Experiment 6 Memory Task

Tools	Furniture	Clothing	Vehicles
Hammer	Couch	Pants	Car
Screwdriver	Chair	Shirt	Truck
Pliers	Sofa	Blouse	Bus
Wrench	Dresser	Jeans	Motorcycle
Saw	Table	Dress	Plane
Drill	Desk	Underwear	Jeep
Shovel	Coffee table	Sweater	Bike
Chisel	Dining table	T-shirt	Train
Socket wrench	Bed	Trousers	Moped
Nails	Rocking chair	Skirt	Boat
Crowbar	Stool	Shorts	Tractor
Knife	Cabinet	Bra	Skateboard
	Bed stand	Jacket	
		Shoes	
		Coat	
		Socks	
		Tie	
		Belt	
		Gloves	
		Hat	

References

Alter, A. L., & Oppenheimer, D. M. (2008). Effects of fluency on psychological distance and mental construal (or why New York is a large city, but New York is a civilized jungle). *Psychological Science*, *19*(2), 161–167.

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219-235.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology-General*, *136*(4), 569–576.

Atran, S. (1999). Itzaj Maya folk-biological taxonomy. In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 119–204). Cambridge, MA: MIT Press.

Bailenson, J. N., Shum, M. S., Atran, S., Medin, D. L., & Coley, J. D. (2002). A bird's eye view: Biological categorization and reasoning within and across cultures. *Cognition*, *84*(1), 1-53.

Barr, R. A., & Caplan, L. J. (1987). Category representations and their implications for category structure. *Memory & cognition*, *15*(5), 397-418.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211-227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629-654.

Blok, S. V., Medin, D. L., & Osherson, D. N. (2007). Induction as conditional probability judgment. *Memory and Cognition*, *35*(6), 1353-1364.

Borkenau, P. (1990). Traits as ideal-based and goal-derived social categories. *Journal of Personality and Social Psychology*, *58*(3), 381-396.

Bornstein, R. F., & D'Agostino, P. R. (1994). The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attribution model of the mere exposure effect. *Social Cognition, 12*, 103-128.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65*(1), 14-21.

Bruce, V., Doyle, T., Dench, N., & Burton, M. (1991). Remembering facial configurations. *Cognition*, *38*(2), 109-144.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

Coley, J. D., Medin, D. L., & Atran, S. (1997). Does rank have its privilege? Inductive inferences within folkbiological taxonomies. *Cognition*, *64*(1), 73-112.

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*(1), 15-28.

Davis, T., & Love, B. C.How goals shape category acquistion: The role of contrasting categories. *Psychological Science*.

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 151–175). Washington, DC: American Psychological Association.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*(5), 608-628.

Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Memory & Cognition*, *31*(2), 169-180.

Goldwater, M. B., Stilwell, C. H., & Markman, A. B. (2008). *The ideal representation of role-governed categories*. Presented at the 30th Annual Conference of the Cognitive Science Society, Washington, DC.

Hadjichristidis, C., Sloman, S., Stevenson, R., & Over, D. (2004). Feature centrality and property induction. *Cognitive Science*, *28*(1), 45–74.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review*, *7*(4), 569-592.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248-274). Oxford: Oxford University Press.

Heit, E., & Feeney, A. (2005). Relations between premise similarity and inductive strength. *Psychonomic Bulletin & Review*, *12*(2), 340-344.

Heit, E., & Hayes, B. K. (2008). Predicting reasoning from visual memory. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 83-88). Presented at the Cognitive Science Society, Austin, TX: Cognitive Science Society.

Homa, D., Goldhardt, B., Burruel-Homa, L., & Smith, J. C. (1993). Influence of manipulated category knowledge on prototype classification and recognition.

Memory & Cognition, *21*(4), 529-538.

Inagaki, K. (1990). The effects of raising animals on children's biological knowledge. *British Journal of Developmental Psychology*, *8*(2), 119–129.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*(3), 306–340.

Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, *56*(3), 326–338.

Johnston, W. A., Dark, V. J., & Jacoby, L. L. (1985). Perceptual fluency and recognition judgments. *Journal of experimental psychology. Learning, memory, and cognition*, *11*(1), 3–11.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory and Cognition*, *31*(1), 155-165.

Klein, G. (1998). *Sources of power*. Cambridge, MA: MIT Press.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*(5140), 1747–1749.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 754–770.

Levering, K., & Kurtz, K. J. (2006). The influence of learning to distinguish categories on graded structure. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1681-1686). Vancouver, BC.

López, A. (1995). The diversity principle in the testing of arguments. *Memory & Cognition*, *23*(3), 374-382.

López, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, *32*(3), 251-295.

López, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category-based induction. *Child Development*, *63*(5), 1070-1090.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Lynch, E. B., Coley, J. D., & Medin, D. L. (2000). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition*, *28*(1), 41-50.

Malt, B. C., Ross, B. H., & Murphy, G. L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 646–661.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592-613.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 329-358.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472.

McRae, K., Ferretti, T. R., & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, *12*(2), 137-176.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517-532.

Medin, D. L., & Shaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society*, *7*(3), 283-284.

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

Murphy, G. L., & Ross, B. H. (2005). The two faces of typicality in category-based induction. *Cognition*, *95*(2), 175-200.

Murphy, G. L., & Ross, B. H. (1999). Induction with cross-classified categories. *Memory & Cognition*, *27*(6), 1024-1041.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical

heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–299.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57.

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency in choice. *Journal of Marketing Research*, *44*(3), 347–356.

Oppenheimer, D. M. (2004). Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science, 15*, 100-105.

Oppenheimer, D. M., & Frank, M. C. (2008). A rose in any other font would not smell as sweet: Effects of perceptual fluency on categorization. *Cognition*, *106*(3), 1178–1194.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185-200.

Patalano, A. L., Chin-Parker, S., & Ross, B. H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of*

Memory and Language, *54*(3), 407–424.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353-363.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 811–828.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and cognition*, *8*(3), 338–342.

Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, *27*(5), 709-748.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264–314.

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition*, *91*(2), 113-153.

Rehder, B., & Kim, S. W. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 659-683.

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1261-1275.

Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition*, *108*(2), 543-556.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning & Verbal Behavior*, *14*(6), 665-681.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*, 1-20.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

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

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 491-502.

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 736–753.

Ross, N., Medin, D., Coley, J. D., & Atran, S. (2003). Cultural and experiential differences in the development of folkbiological induction. *Cognitive Development*, *18*(1), 25–47.

Rothbart, M., & Lewis, S. (1988). Inferring category attributes from exemplar attributes: Geometric shapes and social categories. *Journal of Personality and Social Psychology*, *55*(6), 861-872.

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, *14*(4), 332-348.

Shafir, E. B., Smith, E. E., & Osherson, D. N. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, *18*(3), 229-239.

Shafto, P., Coley, J. D., & Vitkin, A. (2007). Availability in category-based induction. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches* (p. 114). New York: Cambridge University Press.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, *109*(2), 175-192.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *119*, 231-280.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*(1), 1–33.

Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189-228.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166-188.

Smith, E. E., Rhee, J., Dennis, K., & Grossman, M. (2001). Inductive reasoning in Alzheimer's disease. *Brain and Cognition*, *47*(3), 494-503.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition*, *49*(1-2), 67-96.

Solso, R. L., & McCarthy, J. E. (1981). Prototype formation of faces: A case of pseudo-memory. *British Journal of Psychology*, *72*(4), 499–503.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207-232.

Whittlesea, B. W. A., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of experimental psychology. Learning, memory, and cognition*, *26*(3), 547–565.

Whittlesea, B. W., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language*, *29*(6), 716–732.

Williams, J. J., & Lombrozo, T. (2009). *Explaining promotes discovery: Evidence from category learning.* Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18*(2), 221-281.

VITA

Jonathan R. Rein attended High Technology High School in Lincroft, New Jersey, graduating in May 2001. He enrolled at Rutgers College within Rutgers University – New Brunswick in September 2001. He received his Bachelor of Arts degree in May 2005, majoring in Psychology and Philosophy with a minor in Cognitive Science. In September of that year, he enrolled in the Ph.D. program in Cognition & Perception in the Psychology Department at The University of Texas at Austin.

Permanent Address: jonathan.rein@gmail.com

This manuscript was typed by the author.