The "Rational" Number *e*: A Functional Analysis of Categorization

Ángel Cabrera
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332-0170
phone: (404) 853 0192
angel@psy.gatech.edu

Abstract

Category formation is constrained by three factors: the perceptual structure of the domain being categorized, the limitations and biases of the learner, and the goals that trigger the learning process in the first place. Many studies of categorization have paid attention to the effects of the structure of the world and some to the biases due to the learner's prior knowledge. This paper explores the third factor: how the goals of the agent at the time of the learning episode affect what categories are formed. In particular it presents an information theoretical account that views categories as a means to increase the agent's chances of achieving its goals. One of the predictions of the theory is that information gain, the average reduction of uncertainty induced by a category, is maximized when the domain is partitioned into about 3 categories, the closest integer to the irrational number *e*. This prediction is confirmed by evidence derived from anthropological studies of folk classifications of animal and plants by different societies from around the world, and also by an informal observation of the behavior of cognitive scientists. Interestingly, *e* also emerges from optimization analyses of memory search as well as from experimental work on memory retrieval.

Introduction

"My problem," George Miller admitted in his magical 1956 paper, "is that I have been persecuted by an integer." My problem is even worse. I have been persecuted by an irrational number. The number *e*, base of the natural logarithms, made its first apparition as I was developing a functional analysis of categorization. Soon after, I found that, far from being an isolated event, the number *e* had also appeared, in no disguise, in other theoretical and empirical areas of psychology. Not wanting to draw premature conclusions from this precise coincidence, I looked at data obtained in anthropological studies of biological folk classification and, imagine that, there was *e* again. At this point it has become too hard for me not to think of some underlying pattern behind these occurrences. Either I am suffering from Miller's syndrome or this coincidence is actually telling us something. I have decided that submitting my symptoms for public scrutiny may be the only way to solve this dilemma. So be it.

Three factors constraining category acquisition

Category formation is affected by the structure of the stimuli being categorized, the agent's limitations and biases, and the goals leading the agent to form the categories. The effect of the structure of the world on human categories has received extensive theoretical and empirical attention. Feature matching (Rosch & Mervis, 1975; Tversky, 1977), similarity to exemplars (Medin & Schaffer, 1978; Nosofsky, 1984), correlations among features (Billman, 1989), and inter-feature predictability (Anderson, 1990; Corter & Gluck, 1992) are some of the characteristics of the input that have been shown to constrain the process of category formation and, consequently, determine which categories are most likely to be acquired. The biases of the learner (Keil, 1990) can be the result of innate, wired-in preferences (e.g. Spelke, 1990) or they can be imposed by different kinds of prior knowledge such as implicit domain theories (Pazzani, 1991) or even the language spoken by the learner (Cabrera & Billman, in press). Finally, categories are also affected by the goals of the agent. People can build categories in the service of specific goals even when the category members have little in common perceptually (Barsalou, 1983).

These three factors are jointly responsible for what categories end up being formed and how easily they are formed. They can be seen as three forces that push the process of category formation in different directions until some equilibrium is reached. Although there may well be cases where one particular factor becomes predominant, in general, constraints of the three kinds will have an effect on the resulting categories.

The functional view

The functional view of categories is an attempt to isolate the constraints imposed by the goals of the learner on category formation (Cabrera, 1994). It is based on the idea that categories are formed by an agent in order to increase its chances of achieving some goal. Imagine a person in a certain context who is trying to decide which action α_i to perform out of a range of n possible actions. A possible strategy that the person could use to optimize his choices would consist of using his prior experience in this situation to estimate the probabilities of success of each action,

$P(\alpha_i)$. The person could then distribute his choices proportionally to the probabilities of success of each action or simply perform the action with the highest probability[1]. There is no question that the knowledge of these probabilities will be beneficial to the agent. However, there is no guarantee that the selected action will in fact be the right choice. In information theory, this situation is described in terms of uncertainty, a measure of the amount of additional information the person would need in order to be certain of choosing the right action.

Uncertainty can be computed mathematically according to the expression

$$-\sum_{i=1}^{n} P(\alpha_i)\log_2 P(\alpha_i) \quad (1)$$

whose result is given in bits, the standard unit of information. Uncertainty is maximum when the probabilities $P(\alpha_i)$ are uniformly distributed, and zero when one of the probabilities is 1 and the rest are 0. This is consistent with the intuition that if the person knew that there was only one successful choice, he would need no additional information to behave optimally, whereas if all the alternatives appeared equally good the person would need some additional information before being able to make a decent choice.

Suppose now that there is a relationship between some variable aspect of the environment and the probability of each of the actions being helpful for the person's goals. Let us refer to the different forms that that aspect of the environment could take as stimuli, and to the set of all stimuli, as the domain. If the person knew the exact relationship between every possible stimulus and the probability of success of each action, his chances of success could increase considerably. In practice however, the large (generally infinite) size of the domain is likely to preclude the person from being able to experience every stimulus at least once during his lifetime. Even if this were possible, the person's memory might not be large enough to store all that information.

A less ideal but more feasible strategy would consist of partitioning the domain into a set of m categories of stimuli, C_j, and estimating the probabilities of success of each action given stimuli from each of the categories, $P(\alpha_i|C_j)$. This strategy may not completely eliminate the person's uncertainty, but it has the double advantage of reducing the storage requirements to a few sets of probabilities (one set per category), as well as to allow the person to produce informed guesses for stimuli never experienced before if category membership can be determined on the basis of some perceptual characteristic. The remaining uncertainty not captured by these conditional probabilities can be quantified as

$$-\sum_{i=1}^{n} P(\alpha_i|C_j)\log_2 P(\alpha_i|C_j) \quad (2)$$

If for each category, only one of the conditional probabilities were non zero --in other words, if category membership reduced the choices to one-- uncertainty would be eliminated. In general, even when things are not that ideal, uncertainty will nevertheless be reduced whenever the conditional probabilities $P(\alpha_i|C_j)$ are less uniformly distributed than the prior probabilities $P(\alpha_i)$. In other words, the categories will be helpful to the agent if the probabilities of success of the different actions within each category are more unequal than they are across the entire domain.

By combining the expressions for uncertainty prior to the formation of the categories (Equation 1) and conditional upon the categories (Equation 2), we can estimate the information gain (IG) associated with a category C_j as the *average* reduction in uncertainty induced by that category. Given $P(C_j)$, the relative frequency of occurrence of the category, information gain is simply[2]:

$$IG(C_j) = P(C_j)\sum_{i=1}^{n}\left[P(\alpha_i|C_j)\log_2 P(\alpha_i|C_j) - P(\alpha_i)\log_2 P(\alpha_i)\right] \quad (3)$$

Maximizing information gain

The central hypothesis underlying the functional view of categories (Cabrera, 1994) is that learners will always try to maximize the information gain of the categories they form, within the constraints, of course, of the world's structure and the learner's limitations. In order to isolate the consequences of this functional constraint, we first need to make some simplifying assumptions that keep the other constraints constant. We will assume, for instance, that the structure of the world is such that any possible category is equally salient perceptually, that stimuli are uniformly distributed, and that the probabilities of success of the different actions are uniformly distributed. These assumptions are not intended to be representative of any real situation in particular: they simply try to isolate the effects of the information maximizing bias.

Let us assume a worst case initial scenario of maximum uncertainty. In other words, let us assume that, in the absence of any knowledge about the environment, all actions are equally likely to be useful to the agent. If n is the number of possible actions α_i, this assumption amounts to saying that $P(\alpha_i) = \frac{1}{n}$ for all α_i. According to Equation 1

[1] For the purposes of the analyses presented here it does not matter what exact strategy (probability maximizing, probability matching, etc) the agent uses.

[2] This expression is identical in form to a measure of category utility proposed by Corter and Gluck (1992). However, whereas Corter and Gluck's measure was meant to capture inter-feature predictability, information gain is defined with respect to the agent's goals and actions, and is independent of the perceptual structure of the categories.

the total uncertainty in a situation like this will be $\log_2 n$. As we could expect, uncertainty increases monotonically with the number of alternatives.

Let us further assume that the agent partitions the domain into m contrasting categories that reduce uncertainty uniformly and maximally. This assumption has three implications. First, it implies that each category will reduce the choices from n initial actions to $k \leq n$. Second, once a category C_j has been determined, the remaining k actions will have equal conditional probabilities of success $P(\alpha_i|C_j) = \frac{1}{k}$ whereas the rest n-k actions will have a conditional probability $P(\alpha_i|C_j) = 0$. In information terms, this means that uncertainty will be reduced from $\log_2 n$ to $\log_2 k$. Third, all categories must be equally likely to occur: $P(C_j) = \frac{1}{m}$.

Let us consider the case where $m \leq n$ (fewer categories than actions). For the uniform uncertainty reduction assumption to hold, k must equal n/m. For example, if the agent has to decide among 8 possible actions but only forms 4 categories, uncertainty reduction will be maximized if each of the categories reduces the choices to 2 actions. From Equation 3, we can show that the average reduction of uncertainty in this case will be:

$$IG(C_j) = \frac{1}{m}(\log_2 n - \log_2 k) = \frac{\log_2 m}{m} \quad (4)$$

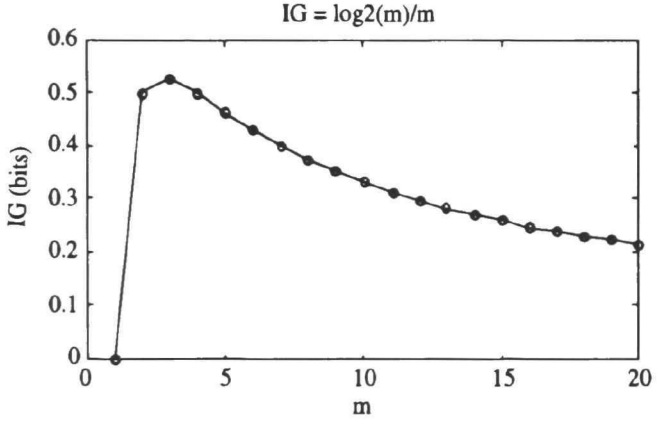

Figure 1. Information gain as a function of number of contrasting categories ($m \leq n$).

Surprisingly, this expression does not depend on the total number of actions: it only depends on the number of categories. The shape of this function is shown in Figure 1. It is 0 for m = 1, --obviously, having only one category that includes the whole domain does not reduce uncertainty at all--, it peaks at m = 3 and then decreases monotonically with m. The exact location of the maximum of this function can be obtained as the zero of its derivative, $\frac{1}{\ln 2}\frac{1-\ln m}{m^2}$, which happens to occur at $m = e$ (2.71828...). Since the number of categories must be a positive integer, the actual maximum is m = 3 (with an average information gain of .53 bits) followed very closely by m = 2 (.5 bits) and m = 4 (.5 bits).

In the case where $m \geq n$, the assumption of uniform reduction of uncertainty requires that each category reduce the alternative actions to one (k = 1). This translates into a reduction of uncertainty from $\log_2 n$ to zero, and therefore:

$$IG(C_j) = \frac{1}{m}(\log_2 n - 0) = \frac{\log_2 n}{m} \quad (5)$$

For any given number of alternative actions n, information gain decreases hyperbolically with the number of categories m. Since $m \geq n$, information gain will be maximum when $m = n$. Having more categories than alternative actions provides no additional reduction of uncertainty beyond that obtained with only one category per action. Instead, every additional category induces a cost that is reflected on an overall lower probability of occurrence of each individual category.

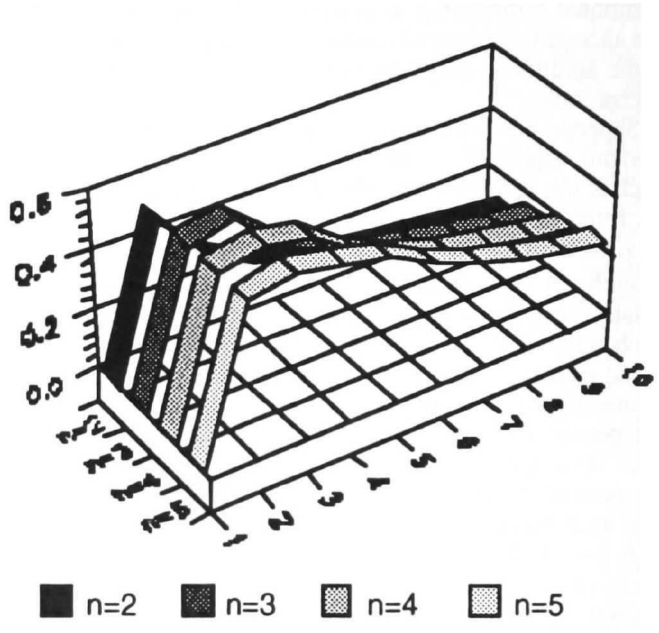

Figure 4. Information gain for two to four actions and one to ten categories.

Figure 4 shows the average information gain for n = 2 to 5 actions and m = 1 to 10 categories. For each n, information gain was computed according to Equation 4 for values of $m < n$ and according to Equation 5 for $m \geq n$. Notice how IG does not vary with n when $m \leq n$, but it does so (decreasingly) for larger m's. Also, with the exception of n = 2, information gain is maximum for m = 3 independently of n.

Empirical support

The previous analyses show that, if the prior probabilities of success are uniformly distributed across actions, and if categories are formed which reduce uncertainty uniformly, average functional utility is optimized when the domain is split into *e*, that is, about 3, *contrasting* categories. But we know that clear-cut categories like these tend to be the exception rather than the norm (Smith & Medin, 1981) and we also have no reason to expect that success probabilities will actually be distributed uniformly in real situations. Does this invalidate the conclusion? Not necessarily. The assumptions are made to isolate functional constraints by setting up ideal situations from the point of view of functional utility. In every particular situation, world structure as well as individual biases will moderate the effects of functional constraints and will therefore determine whether the categories that would be optimal from the point of view of the agent's goals are actually formed.

According to Anderson's *General Principle of Rationality* (1990, p.28), if having three categories is optimal for the individual's adaptation to the environment, we should expect the human cognitive system to have developed (philo- or ontogenetically) a tendency to organize knowledge in sets of about three contrasting categories. To test this prediction I turned to anthropological research. Ethnobiology is the branch of anthropology devoted to the study of how human societies view and use nature (Berlin, 1992). A great deal of ethnobiological research has compared how different societies classify the plants and animals in their natural environments and how those classifications relate to the taxonomies built by western scientists.

There is evidence suggesting that basic level categories of animals and plants may be determined by perceptual structure, whereas subordinate and superordinate categories tend to be formed to fulfill specific cultural needs (Malt, 1994; Berlin, 1978). If this is right, the functional analysis presented here would predict subordinate categories to partition basic categories into sets of around 3 contrasting categories and superordinate categories to group sets of around 3 contrasting basic categories. We should therefore expect folk taxonomies to be organized at different levels of abstraction in contrasting sets of about 3 categories on average.

It turns out that "a general principle of ethnobiological classification is that folk species most commonly occur in contrast sets of few (two or three) members" (Berlin, 1992, p. 122). Table 1 summarizes data reported by Berlin (1992, p. 126-128) on the frequency distribution of contrast sets with two or more categories in different biological taxonomies used by several linguistically unrelated traditional societies according to detailed ethnobiological inventories developed by himself and others. I have added in the last two columns the mean and median of the distributions of category set sizes corresponding to each classification system. Overall, the average number of contrasting categories was 2.982.

Table 1. Frequency distribution of biological category sets of different sizes.

Classification System	2	3	4	5	6	≥7	Mean	Median
Tzeltal plants (Mexico)	41	16	2	5	5	5	3.123	2
Aguaruna plants (Peru)	68	12	9	2	3	9	2.903	2
Wayampí plants (French Guyana)	47	13	4	2	3	7	2.974	2
Hanunóo plants (Philippines)	224	53	15	9	5	16	2.612	2
Tobelo plants (Indonesia)	142	13	22	7	3	12	2.754	2
Seri plants (Mexico)	32	7	8	3	3	1	2.907	2
Tobelo animals (Indonesia)	46	20	9	2	1	4	2.830	2
Tzeltal animals (Mexico)	25	18	5	3	2	1	2.926	3
Wayampí animals (F. Guyana)	55	22	14	5	1	3	2.840	3
Huambisa birds (Peru)	36	9	2	0	0	0	2.278	2
Huambisa fish (Peru)	8	4	1	2	2	1	3.389	3
Aguaruna mammals (Peru)	9	3	4	1	0	0	2.824	2
Cantonese fish (China)	6	7	5	3	1	9	4.410	4
*Cog. Sci. concepts (Int'l.)	21	22	8	10	4	4	3.770	3

NOTE: Following Berlin's own concerns about the origin of the data from Ndumba plants (mean = 4.800) and animals (3.733), I have excluded them from my analyses (1992, p. 283). *The "Cognitive Science concepts" data come from an informal inventory of numbered and alphabetized lists in the 1994 volume of the "Cognitive Science" journal (vol. 18).

For comparison's sake, I estimated the frequency distribution of category set sizes typically used by cognitive scientists in describing their theories and defending their claims. These data come from an informal inventory of all the articles published during 1994 by the "Cognitive Science" journal (Vol. 18). Specifically, it is based on the length of numbered and alphabetized lists that appear in each article. These data are particularly interesting because the theoretical concepts developed by cognitive scientists rarely rely on specific perceptual forms and may therefore be more subject to functional biases than biological categories. About 62.32% of the lists in the papers examined contained two or three items (the frequency of three item lists being slightly higher than two item lists), and 88.41% contained 5 or less (there were however two papers presenting classifications consisting of 15 and even 16 items!). The mean category set size was 3.770, slightly greater than the mean of the biological taxonomies reported by Berlin (1992) but within a similar range.

Converging Evidence and Conclusions

Category acquisition is constrained by (a) stimulus structure --things that look similar tend to be categorized together--, (b) the learner's innate or acquired preferences --things that we are programmed to categorize together or that we have categorized together in the past tend to be categorized together again--, and (c) the goals of the learner --things that require similar kinds of actions tend to be categorized together. When these constraints agree, acquisition will take place readily. Sometimes, one of the constraints may overshadow the rest, as stimulus structure may do in the case of biological basic categories (Berlin, 1978), prior beliefs in the case of illusory correlations (Chapman & Chapman, 1969), and the agent's goals in the case of ad-hoc categories (Barsalou, 1983). In general, however, categories will result from some sort of compromise among the three factors.

This paper has tried to isolate the consequences of the third kind of constraint: the connection between the categories and their function with respect to the agent's goals. Functional utility of a category was defined as the amount of information the category provides about the best actions to perform given some goal. Then, it was demonstrated that, under a few simplifying assumptions, having 3 contrasting categories (actually e) maximizes the average information gain obtained from each category. This prediction is supported by anthropological studies of classification of animal and plants in traditional societies around the world and also by the behavior of cognitive scientists while reporting their research (this paper is no exception).

Categorization is not the only aspect of cognition where e appears to be an optimizing factor. Dirlam (1972) demonstrated mathematically that a branching factor of 3 maximizes search efficiency in a hierarchical memory structure if efficiency is defined as the maximum number of items that need to be scanned in order to find a target piece of information. In fact, 3 appeared, as it did here, as the closest integer to the irrational e. Dirlam's prediction was later confirmed by a number of experimental studies of human memory (Broadbent, 1975). These studies, in combination, convinced Anderson (1993, p. 26) that the best chunk size for the declarative memory system of his ACT-R model might be three.

The optimizing power of e has also been noted in the area of psychological testing. Tversky (1964) showed that "given a fixed total number of alternatives for a multiple-choice type test, the use of three alternatives at each choice point will maximize discriminability, power and information of a test" (p. 386). Although Tversky's finding did not have a big influence in the testing community for a number of reasons[3], some of the points he raised in his paper tie very nicely into the discussion at hand. In particular, he suggested that his result might "shed some light on the study of information coding and processing" (p. 390). He cited data pointing to three as the optimal number of alternatives per variable in discrimination tasks and conjectured that, under some assumptions, "the use of three-level factors will minimize confusion and decrease memory load" (p. 391).

Are all these findings mere coincidence? I do not think so. If a branching factor of three optimizes search efficiency in a hierarchically organized data set one would expect that same factor to maximize the information gained from every decision made at a decision point in the hierarchy. The two are, in a sense, different ways of saying the same thing.

Do these findings contradict Miller's number seven? I do not think so either. Miller's concern was about information processing capacity *limits* (the second type of constraint if you wish), not about *optimal* organization of knowledge. The fact that most people might be limited to discriminate a maximum of seven levels of loudness does not mean that people will tend to organize sounds into sevens. In turn, the idea that e optimizes information gain and search efficiency should not be interpreted as an absolute bound, but as an indication that there is an optimal number of contrasting categories, "that performance will deteriorate if one goes beyond it, and therefore, that the system will tend to organize itself in chunks [contrast sets] of that size" (Anderson, 1993; p. 27). If anything, these results are compatible with Miller's: people's discrimination capacity *allows* them to organize things in triads.

The number e has a long history of extraordinary appearances in mathematics and the natural sciences (Maor, 1994). In fact, e plays such an important role in differential calculus that the logarithm base e is known as the *natural* logarithm. Mathematicians, however, do not call e itself *natural* because it can not be expressed without decimal digits. They do not even consider it *rational* because it can not be expressed as the ratio between two integers. It is ironic that, as irrational as it is, e seems to be an optimizing factor in some *rational* analyses of cognition. If I am not suffering from delusions of persecution, this peculiar irrational number may deserve to be upgraded to the rank of "Rational" Number.

[3] It seems that the assumption of a fixed total number of alternatives is a rare constraint in the design of real tests (Nambury S. Raju, personal communication).

Acknowledgments

I am indebted to the Spanish Ministry of Education and Science and the U. S.. Information Agency (Institute of International Education) for their financial and logistic support through the Fulbright Scholarship program. I am also grateful to Dorrit Billman, Beth Cabrera, Alex Kirlik, Jack Marr, Ashwin Ram, Nambury Raju, and Tony Simon for their challenging comments and suggestions.

References

Anderson, J. R. (1990). The Adaptive Character of Thought. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (1993). Rules of the Mind. Hillsdale, NJ: Lawrence Erlbaum.

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

Billman, D. (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. Language and Cognitive Processes, 4, 127-155.

Berlin, B. (1978). Ethnobiological classification. In E. Rosch and B. Lloyd (Eds.), Cognition and Categorization. Hillsdale, NJ: Lawrence Erlbaum.

Berlin, B. (1992). Ethnobiological Classification: Principles of Categorization of Plants and Animals in Traditional Societies. Princeton, NJ: Princeton University Press.

Broadbent, D. E. (1975). The magic number seven after fifteen years. In A. Kennedy and A. Wilkes, (Eds.), Studies in Long Term Memory. London, UK.: John Wiley & Sons.

Cabrera, A. (1994a). Functional and conditional equivalence: Conceptual contributions from behavior analysis. Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Lawrence-Erlbaum.

Cabrera, A. & Billman, D. (in press). Language-driven concept learning: deciphering "Jabberwocky". Journal of Experimental Psychology: Learning, Memory & Cognition.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. Journal of Abnormal Psychology, 74, 271-280.

Corter, J. E. & Gluck, M. A. (1992). Explaining basic categories: feature predictability and information. Psychological Bulletin, 111, 291-303.

Dirlam, D. K. (1972). Most efficient chunk sizes. Cognitive Psychology, 3, 355-359.

Keil, F. C. (1990). Constraints on constraints: surveying the epigenetic landscape. Cognitive Science, 14, 135-168.

Malt, B. C. (1994). Category coherence in cross-cultural perspective. Cognitive Science Technical Report UIUC-BI-CS-94-03. The Beckman Institute, University of Illinois.

Maor, E. (1994). e : The Story of a Number. Princeton, N.J. : Princeton University Press.

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

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 63, 81-97.

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

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. Journal of Experimental Psychology: Learning, Memory and Cognition, 17, 416-432.

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

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

Spelke, E. S. (1990). Principles of object perception. Cognitive Science, 14, 29-56.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. Journal of Mathematical Psychology, 1, 386-391.

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