Subjective Complexity of Categories Defined over Three-Valued Features

Cordelia D. Aitkin (cdaitkin@ruccs.rutgers.edu)
Department of Psychology, Rutgers University – New Brunswick, 152 Frelinghuysen Rd.
Piscataway, NJ 08854 USA

Jacob Feldman (jacob@ruccs.rutgers.edu)
Department of Psychology, Center for Cognitive Science, Rutgers University – New Brunswick, 152 Frelinghuysen, Rd.
Piscataway, NJ 08854 USA

Abstract

Many studies in the last four decades have investigated the relative difficulty in learning of concepts defined by various logical forms. Historically, most such studies have used Boolean concepts, that is, categories formed by logical combinations of binary-valued variables. In previous work, we have found that in such categories subjective difficulty is well predicted by Boolean complexity, that is, the length of the shortest propositional formula equivalent to the concept. However, this formalization does not extend easily to concepts defined using features with more than two values. Such categories are of particular interest because they are not easily handled by any contemporary theories based on continuous metric spaces. In more recent work, we have developed a representational formalism suitable for representing such categories. This theory provides a measure of conceptual complexity for such categories, called *algebraic complexity*. Here we report an experiment testing the learnability of a set of a set of categories defined over two three-valued features. The results show that algebraic complexity gives a good account of the subjective learnability of these concepts.

Keywords: Categorization, complexity, learning

Introduction

Categories help us to organize a complex world by grouping objects and entities in a coherent fashion. But when it comes to coherence, not all categories are created equal. Some collections (*pet cats*) seem well-organized and coherent; others (*snips, snails, and puppy-dog tails*) seem disjoint, heterogeneous, and incoherent. Studies since the 1950s have investigated the factors influencing subjective coherence, as manifested in the ease with which categories with various types of logical form can be learned by subjects. Simple conjunctive categories (*big red things*) can be learned accurately from few examples, while disjunctive categories (*big red or small blue things*) require more examples before subjects can acquire them (see Bourne, 1970, for a summary of this extensive literature). However the spectrum of categorical coherence cannot be reduced to the simple dichotomy of conjunction vs. disjunction. In a more complex setting involving three Boolean features, Shepard, Hovland and Jenkins (1961) found a reliable ordering of learning difficulty among six logical forms, whose structural differences cannot be described solely in terms of the number of conjunctions or dimensions involved in the concept.

The variety of possible conceptual forms in Boolean spaces ranges far beyond the six classes used by Shepard et al. (1961). An extensive study of learning difficulty in our laboratory (Feldman, 2000) considered 35 more such classes, testing the accuracy with which subjects could learn each classification after a fixed duration of learning. Each of these concept forms, like Shepard et al.'s famous types, involved a combination of Boolean features that is not isomorphic in logical form (congruent) to any of the other classes. And also like the famous set of six, the classes studied represented an exhaustive survey of one part of the space of Boolean forms: in the Shepard et al. set, all concepts in 3-dimensional Boolean space with four positive examples; in the Feldman (2000) set, all concepts in 3- or 4-dimensional Boolean space with two, three, or four positive examples (see Feldman, 2003).

Considering this wide array of types, a simple empirical trend emerged: the subjective learning difficulty of each concept type was well-predicted by its *Boolean complexity*. The Boolean complexity of a Boolean concept is simply the length of the shortest propositional formula equivalent to the original concept regarded as a propositional formula, measured in *literals* (mentions of a variable name). Like its more famous cousin Kolmogorov complexity (see Li & Vitanyi, 1997), Boolean complexity reflects the inherent *incompressibility* of the concept in question. The ease of learning is well predicted by the degree to which the concept can be faithfully compressed into a more compact form. This finding thus represents a kind of simplicity principle at work in human learning (cf. Pothos & Chater, 2002). The implication is that learners seek to compress the examples they have observed into a more compact representation, and learn more effectively to the extent that the training examples are, in fact, faithfully compressible.

As has often been remarked in the literature, though, Boolean features represent a particularly artificial space in which to study categorization. One would like to extend this simplicity principle to a wider and perhaps more natural type of feature space. The notion of Boolean complexity, however, does not extend easily to non-Boolean features. Features with three or more discrete values (e.g., *shape* = {*square, circle, triangle*} or *species* = {*dog, elephant, llama*}), while intuitively seeming like a simple extension of the Boolean features studied in the 1960s, do not lend themselves to a simple propositional representation.

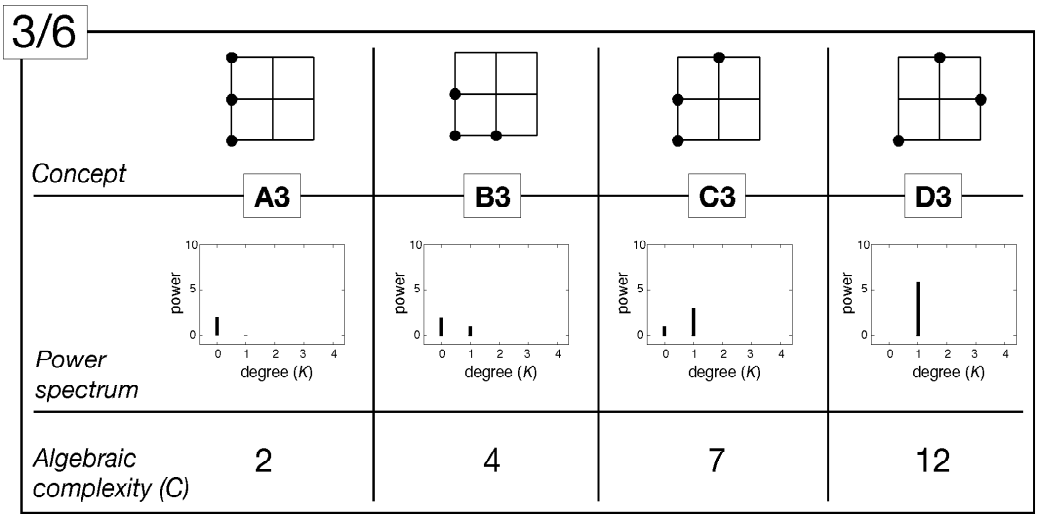

Figure 1: Abstract concept forms, algebraic power spectra, and algebraic complexity values for concepts from the 3/6 family.

In more recent work, however, we have developed a conceptual representation formalism called the concept algebra (Feldman, 2004; Feldman, in press) that is suitable for the expression of concepts defined over features with arbitrary numbers of levels per feature. This newer approach provides a natural complexity measure for concepts defined in such spaces, called *algebraic complexity*, which we will describe below. The main aim of this paper is to investigate the relationship between algebraic complexity and empirical subjective learning difficulty for concepts defined over two features with three levels each (a 3×3 grid).

Non-binary discrete-featured concepts are of particular theoretical interest, apart from their relation to algebraic representations. As Lee and Navarro (2002) have recently pointed out, such concepts are not well handled by most contemporary theories, which are predominantly based on similarity measures defined over continuous metric spaces. Strictly speaking, such a metric is not defined in a Boolean space, but one can easily be induced simply by imposing a continuous space between the terminal 0 and 1 corresponding to false and true values of each Boolean variable. But this is not possible with three or more categorical values per feature, due to the absence of a natural order (much less a metric) among the distinct values. In their study, Lee and Navarro studied such concepts, attempting to fit their data using the ALCOVE model of Kruschke (1992), which assumes a metric space. They found that ALCOVE only gave a good account of human data when a similarity space was imposed on the object set in such a way as to respect the qualitative pattern among the discrete features. Once suitably equipped with such a metric, and only then, ALCOVE was indeed successful in modeling human performance. But Lee and Navarro concluded that the nature of the qualitative pattern present in the concepts is psychologically critical, exerting a strong influence on the success of learning. Such qualitative patterns are the primary focus of interest in the concept algebra, and algebraic complexity itself is expressly designed to reflect the expressive complexity of exactly such patterns. Hence such patterns, difficult to handle in conventional accounts, but "home turf" for the concept algebra, constitute a particularly critical test case to establish the empirical validity of the concept algebra, and, more generally, to better understand the nature of subjective conceptual complexity.

The concept algebra and algebraic complexity

The concept algebra is based on the idea of *regularities* among discrete features. If all objects in a set share an attribute (a specific value of a feature), for example, we acknowledge a particularly simple regularity. Six llamas, among an otherwise diverse set of animals, share a salient commonality, of which an observer would surely take note. This "regularity" in this example is simple because it only involves one feature. More complex regularities can be constructed from larger numbers of features. The building blocks in the concept algebra are regularities with an implicational form,

$$\sigma_1(v_1) \rightarrow \neg\sigma_2(v_2), \qquad (1)$$

to be read as "if feature σ_1 has value v_1, then feature σ_2 doesn't have value v_2"—a quasi-causal "law" observed among the observed objects. We speak of a set of objects as "obeying" such a regularity if all of its members are consistent with it (The strictness of this formulation can be relaxed in various ways, which for simplicity of presentation we don't pursue here.) In the above regularity, there is one antecedent (i.e. one attribute to the left of the implication $\rightarrow$). More generally there may be a conjunction of K of them,

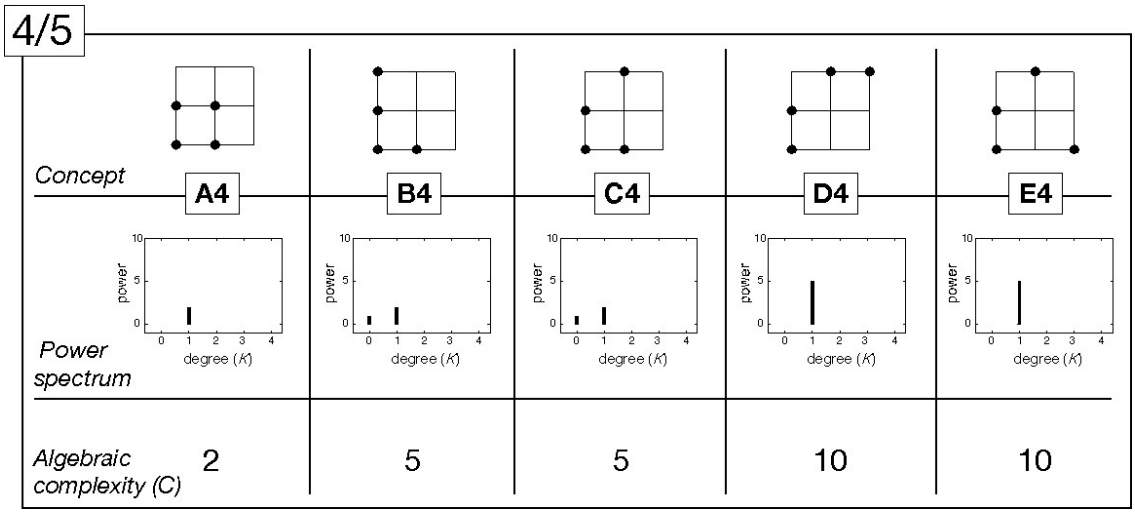

Fig. 2. Abstract concept forms, algebraic power spectra, and algebraic complexity values for concepts from the 4/5 family.

$$\sigma_1(v_1) \ldots \sigma_K(v_K) \rightarrow \neg \sigma_D(v_D), \quad (2)$$

where D is the maximum number of features involved, the dimensionality of the space in question. The number K, called the *degree* of the regularity, reflects the dimension of the regularity in question, because any such regularity can be thought of as "prohibiting"—ruling out objects within—a region of dimension K in the D-feature space. The generic regularity in Eq. 2 says, in words, that if an object has value v_1 of feature σ_1, and also value v_2 of feature σ_2 (etc. up to K), then it will *not* have value v_D of feature σ_D. Geometrically, a $K = 2$ regularity prohibits a "plane" (regardless of D) within the matrix of possible objects, meaning that it defines a K-dimensional region in which objects do not occur. Similarly, a $K = 3$ regularity prohibits a volume (3-dimesional region), and so forth. Hence the number K reflects in a very straightforward way the intrinsic complexity of the regularity in question.

Most featural patterns cannot be described by such a simple, uniform rule. But it turns out that any pattern can be fully described as a finite *combination* of such rules. The concept algebra describes how more complex patterns can be represented as combinations (literally, conjunctions) of implicational regularities, each of which has the form in Eq. (2). (The combinatoric nature of the representation scheme, whereby patterns are constructed out of combinations of simple patterns, is what makes the theory "algebraic.") The fact that this is possible for any featural pattern, captured in a "representation theorem" detailed in Feldman (in press), is at the heart of the algebraic approach.

Critically, the simplest representation of a given pattern in terms of component regularities—called its *power series expansion*—may, and usually does, contain regularities of various degrees. The power series thus gives a kind of "spectral breakdown" of the pattern, analogous to the Fourier expansion of a complex periodic function into sine and cosine components. The power series expansion decomposes the pattern into its constituent regular components, enumerating which regularities—simple (low-degree), complex (high-degree), and in-between—jointly constitute the pattern.

The number of regularities at each degree ($K = 1, 2 \ldots D$) contained in a pattern's power series is referred to as its power spectrum. The power spectrum of a featural pattern contains an enormous amount of information about what is going on in the pattern. The spectrum can be easily plotted, giving a useful graphical display of the given pattern's regularity structure (Figs. 1, 2). Patterns with most of their power at low degrees are "simple," in that most of their structure can be described using small numbers of features at a time. At the other extreme, patterns with most of their power at high degree are "complex," meaning that most of what is going in them can only be described by referring to many (or all) of the features at once. Most patterns are somewhere in the middle (Feldman, 2004).

Finally, *algebraic complexity* C is simply the total spectral power weighted by degree. This is a single scalar number that summarizes where in the spectrum of complexity the bulk of the pattern's power lies. Thus this number expresses the difficulty of expressing the pattern as an algebraic combination of regularities, that is, its complexity in algebraic terms. Several choices of weighting scheme are possible (see Feldman, in press, for discussion). A particularly simple scheme, which we use in this paper, is to weigh each regularity of degree K by $K + 1$,

$$C = \sum (K+1)\lambda_K \quad (3)$$

where λ_K is the spectral power at K, that is, the number regularities of degree K that the pattern obeys. Because each regularity of degree K requires $K+1$ literals to express (see Eq. 2), weighting power this way means that C is the total number of literals in the pattern's complete power series.

This means that the resultant complexities are commensurate with Boolean complexities, in that they are both measured in literals. Note, though, that while Boolean complexity is limited to binary variables, algebraic complexity can tolerate arbitrary numbers of levels per feature.

A Matlab package for computing algebraic power series, spectra, and algebraic complexity values, is freely available at ruccs.rutgers.edu/~jacob/demos/algebra.html.

Concepts chosen for study

As mentioned, the studies of Shepard et al. (1961) and Feldman (2000) had used "exhaustive" concept sets, containing every non-congruent class given a fixed range of dimension and number of examples. We followed a similar strategy here, using two three-valued discrete features as our space. The total number of objects with D features having V values each is V^D, here 3^2 or 9. We used concepts designating either three or four of these objects as positive, respectively referred to as the 3/6 concepts (Fig. 1) and the 4/5 concepts (Fig. 2), as well as their complements. As with Boolean concepts, many of the possible concept structures in this space are congruent, meaning that they are actually equivalent after the labels have been swapped or permuted. With the 3/6 concepts there are exactly four distinguishable types, and with the 4/5 concepts there are five. These nine concepts, along with their complements, were the objects of study in our experiments. The figures show the concept in canonical form (arbitrarily chosen from among the equivalent forms), along with their power spectra and algebraic complexities. Each category type is given a standard label (A3, B3, …, A4, ...), which we use for reference in the remainder of the paper.

We included complementary versions of each concept in our study because, while on the one hand they have very similar logical structure to the originals, studies of Boolean concepts have found a reliably inferior learning of them, termed the "parity effect" in Feldman (2000), with the "normal" form of each concept (smaller half presented as positive) termed "up" parity, and the complementary form (larger half presented as positive) termed "down" parity.

Algebraic complexity is generally different for complementary categories, in contrast to Boolean complexity, which is always identical for categories and their complements (which differ in propositional expression only by a negation, not affecting complexity). However we suspected that subjects might focus on the smaller half (in the spirit of the standard explanation of the parity effect itself), and thus in what follows use the algebraic complexity value derived from the smaller part regardless of parity.

Experiment

Subjects

Fourteen undergraduates at Rutgers University received class credit for participation. All confirmed having normal color vision.

Design, Materials and Procedure

For our two features, we chose color and shape. The three values for color were {*red, blue, yellow*}, while the three values for shape were {*triangle, square, circle*}. In each case, the three values lack any salient natural order. Thus, the basic stimuli were the nine objects *(red triangle, red square, red circle, blue triangle, blue square, blue circle, yellow triangle, yellow square, yellow circle*), corresponding to the nine points on the 3×3 grids shown in the figures. Stimuli were drawn on a computer screen. Each object occupied a space on the screen that was approximately 1/9 of the screen-width square, at varying locations on the screen as explained below.

Shepard, Hovland and Jenkins (1961) showed that learning tasks based on memorization produce the same rank ordering between logical structures (Experiments II and III). Thus, the learning phase consisted of a presentation of all nine objects on the screen. The screen was divided in half horizontally; members from the randomly generated category were located above the line with the label "In the category," while the remaining objects were located below the line with the label "Not in the category." Following results from Feldman (2000), subjects were given a fixed time to study the category: 15s for the 3/6 categories, and 20s for the 4/5-object categories.

The testing phase followed the learning period. The learning screen was cleared, and each object was presented one at a time in random order. Subjects had to press the "1" key if they thought the object was in the category and the "2" key if they thought the object was not. If the subjects pressed the wrong key, a beep indicated to them that they had made a mistake. The computer recorded each response and time it took to respond. After all the objects had been classified once, a screen appeared telling the subjects to press any key when they were ready to start the next trial, which would be a new category.

Each subject learned each category in three different randomized instantiations. Here an "instantiation" means an assignment of dimensions and values to specific meanings. Category **A3**, Up parity, for example, might be instantiated as *all red objects*, or *all blue objects*, or *all triangles*, or *all squares*, etc. Thus, each subject learned a total of 54 [(5 types for the 4-object category + 4 types for the 3-object category) × 2 parities × 3 different presentations] categories during his or her session. The computer randomly generated the order in which the 54 categories were presented. Each session took approximately 45-60 minutes.

Results

Fig. 3 shows measured performance (log proportion correct) as a function of algebraic complexity C. Data are collapsed over parity. A linear regression confirms a decreasing trend in performance as complexity increases: for 3/6 categories (Fig. 3), $R^2 = 0.06$, $F(1,110) = 7.310$, $p=.008$; for 4/5 categories (Fig. 4), $R^2 = 0.11$, $F(1,138) = 17.922$, $p < .001$. As with Boolean categories, subjects' performance was

better in the up parity cases than down—that is, when the smaller subset of objects was labeled as positive. But the parity effect was not significant in this experiment (t = .842, *ns.*). This presumably reflects the relatively small difference in the numerosity of positive and negative examples (here 3/6 and 4/5, but as extreme as 4/12 in the Feldman (2000) experiments), reducing the tendency to focus on the smaller part of the training set.

Discussion

As predicted, classification learning becomes progressively more difficult as algebraic complexity increases. While other unknown factors clearly contribute to learning, Figs. 3 and 4 demonstrate that learning performance closely tracks the intrinsic complexity of the concept forms, as quantified by the algebraic complexity. As in Feldman (2000), subjects have more difficulty learning concepts with more complex internal forms, reflecting a simplicity principle at work in human learning. The results here confirm that this effect does not in any substantial way depend on the details of Boolean representations or Boolean complexity, which are not even computable here, nor on the use of binary-valued features. The simplicity bias, we would argue, is more basic than any of these particulars, provided only that one has the means to quantify it. The concept algebra provides such means, in a way that is apparently approximately empirically correct.

Note further that the computation of algebraic complexity has no free parameters or "fudge factors" to improve its fit (other than the slope and intercept of the regression line itself, required to map the algebraic complexity scale to the empirical measurements). The number C in this sense reflects directly the quantity of internal regularity inherent in the featural pattern obeyed by the concept's members.

As mentioned, alternative accounts of these data are not easily available. Conventional accounts based on metric similarity spaces cannot be applied directly. As discussed, Lee & Navarro (2002) were able to induce ALCOVE to predict human performance on these concepts, but only by somewhat artificially expanding the three values of each feature into a set of binary features (e.g., red absent/present; blue absent/present, etc.), and then inducing a metric between objects by counting shared features. This somewhat roundabout procedure would be progressively more computationally expensive as the number of features and values increases. Additionally, such a method does not allow for subjects to utilize the structural similarity between the categories "All blue objects" and "All squares." This and other methodological differences may account for the different rank ordering found by Lee & Navarro (2002) in the 3/6 categories. Moreover, as we argued above, the very need for it only goes to illustrate the importance of incorporating the qualitative feature pattern in each concept more directly into the model's representation, as is the goal in the algebraic approach.

More broadly, the influence of pattern regularity (i.e., simplicity) on human concept learning is notable because such an effect is not, in principle, predicted by many

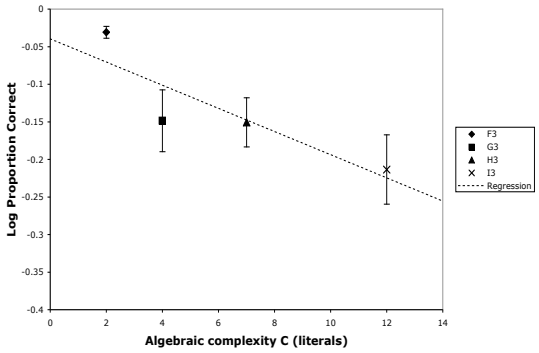

Figure 3: Results from the 3/6 categories, showing log proportion correct as a function of algebraic complexity.

conventional theories. Exemplar theories (Medin & Schaffer, 1978; Nosofsky, 1988), including ALCOVE, are based on the idea of overt storage of training examples, and lack any notion of the induction of a central tendency in the observations. Such theories are not sensitive to the presence of regularity, in a general sense, because they do not attempt in any way to extract regularities from the observations. In practice, such methods *are* in fact influenced by conceptual complexity, but the magnitude of this influence is difficult to quantify in advance, but rather is an epiphenomenon of process of exemplar comparison and depends heavily on the details of various parameters. By contrast, in the concept algebra, the influence of regularity on performance can be quantified analytically, as it is a direct reflection of the how compactly each concept can be expressed in algebraic language.

Prototype theories, by contrast, do predict an influence of conceptual complexity on learning, as concepts ought in principle to be difficult in accordance with how well or how poorly they conform to the prototype schema—for example, tightly clustering about a mean vs. highly dispersed throughout a space. But again prototype theories cannot be applied directly to our categorical feature space, as virtually all accounts form prototypes by some kind of vector averaging.

More recently, a number of authors have proposed "hybrid" models that combine aspects of prototype formation with exemplar storage (e.g. Love, Medin & Gureckis, 2004; Nosofsky, Palmeri & McKinley, 1994). Again, these theories cannot be directly applied here because of the lack of similarity metric upon which their components depend. More broadly, such theories recognize the need for some degree of regularity extraction, as they differentiate between the more "regular" elements of the training set, from which they induce a rule of some kind, and the less regular elements, which they store overtly. This dichotomy is more similar in spirit to the concept algebra, which overtly distinguishes between more and less regular aspects of the training set, as captured by the measurement of degree: low for regular components and broad trends, high for irregular components and narrow details. In the concept algebra, the mixture of such components embodied

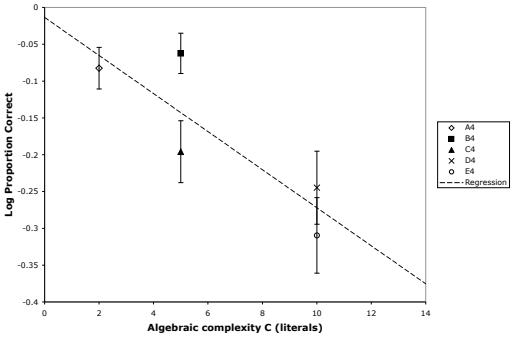

Figure 4: Results from the 4/5 categories, showing log proportion correct as a function of algebraic complexity**.**

by each concept is itself the object of study, and the algebraic complexity itself is the quantification of where in this spectrum each concept actually lies. In this light one could say that the algebraic complexity effect reflects the mixed presence of both types of learning strategy, with the net performance influenced by the degree to which each concept actually requires memorization.

Limitations and Future Work

The concepts tested here manifested a fairly limited variety of concept forms, and thus a fairly narrow range of complexity values. The concept algebra formalism makes it easy to generate analogous predictions in arbitrary cases of arbitrary dimension. Hence one very obvious avenue for future work is to expand the range of concepts tested, for example four- and five-dimensional structures, with more than three values per feature. Such concepts would, of course, provide more strenuous tests of the predictions of the concept algebra, but moreover would perhaps heighten the contrast with any alternative predictions such as those derived from induced metrics. Future studies along these lines are currently being planned in our laboratory.

Also of interest will be to see how ordering structures other than those considered here, such as ordering of feature values along each feature, influence performance. As explained above, in the current experiment we chose features with no obvious ranking. However, many feature scales of natural objects do have such natural orders, which may be psychologically influential apart from any metric scale derived from them. In contrast to the concepts tested here, concepts built around such features would not have natural representations in the concept algebra without some extension of the theory. Such extension will, it is hoped, allow the human bias towards conceptual simplicity to be understood in the broadest context possible.

Acknowledgments

This research is supported by an NSF graduate fellowship (CDA) and by NSF SBE-0339062 (JF).

References

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review*, 77.

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

Feldman, J. (2003) A catalog of Boolean concepts. *Journal of Mathematical Psychology*, 47(1).

Feldman, J. (2004). How surprising is a simple pattern? Quantifying "Eureka!" *Cognition*, 93.

Feldman, J. (in press). An algebra of human concept learning. *Journal of Mathematical Psychology*.

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

Medin, D. L. & Schaffer, M. M. (1978) Context model of classification learning. *Psychological Review*, 85.

Lee, M. D. & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review,* 9.

Li, M. & Vitanyi, P. (1997) *An introduction to Kolmogorov complexity and its applications*. New York: Springer.

Love, B., Medin, D. L. & Gureckis, T. (2004). SUSTAIN: a network model of category learning. *Psychological Review,* 111(2*)*.

Nosofsky, R. M. (1988) Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14(4)*.

Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. (1994) Rule-plus-exception model of classification learning. *Psychological Review*, 101(1).

Pothos, E. M. & Chater, N. (2002) A simplicity principle in unsupervised human categorization. *Cognitive Science*, 26.

Shepard, R. N., Hovland, C. L. & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, General and Applied*, 75.