BEHAVIORAL AND BRAIN SCIENCES (2001) **24,** 629–640
Printed in the United States of America

Generalization, similarity, and Bayesian inference

Joshua B. Tenenbaum and Thomas L. Griffiths
Department of Psychology, Stanford University, Stanford, CA 94305-2130
jbt@psych.stanford.edu gruffydd@psych.stanford.edu
http://www-psych.stanford.edu/~jbt
http://www-psych.stanford.edu/~gruffydd/

Abstract: Shepard has argued that a universal law should govern generalization across different domains of perception and cognition, as well as across organisms from different species or even different planets. Starting with some basic assumptions about natural kinds, he derived an exponential decay function as the form of the universal generalization gradient, which accords strikingly well with a wide range of empirical data. However, his original formulation applied only to the ideal case of generalization from a single encountered stimulus to a single novel stimulus, and for stimuli that can be represented as points in a continuous metric psychological space. Here we recast Shepard's theory in a more general Bayesian framework and show how this naturally extends his approach to the more realistic situation of generalizing from multiple consequential stimuli with arbitrary representational structure. Our framework also subsumes a version of Tversky's set-theoretic model of similarity, which is conventionally thought of as the primary alternative to Shepard's continuous metric space model of similarity and generalization. This unification allows us not only to draw deep parallels between the set-theoretic and spatial approaches, but also to significantly advance the explanatory power of set-theoretic models.

Keywords: additive clustering; Bayesian inference; categorization; concept learning; contrast model; features; generalization; psychological space; similarity

1. Introduction

Consider the hypothetical case of a doctor trying to determine how a particular hormone, naturally produced by the human body, affects the health of patients. It seems likely that patients with too little of the hormone in their blood suffer negative effects, but so do patients with too much of the hormone. Assume that the possible concentration levels of this hormone can be represented as real numbers between 0 and 100 on some arbitrary measuring scale, and that one healthy patient has been examined and found to have a hormone level of 60. What other hormone levels should the doctor consider healthy?

Now imagine a baby robin whose mother has just given it its first worm to eat. The worms in this robin's environment vary in level of skin pigmentation, and only worms with some intermediate density of pigmentation are good to eat; too dark or too light worms are unhealthy. Finally, suppose for simplicity that robins are capable of detecting shades of worm coloration between 0 and 100 on some arbitrary scale, and that the first worm our baby robin has been given scores a skin pigmentation level of 60. Assuming the mother has chosen a worm that is good to eat, what other pigmentation levels should our baby robin consider good to eat?

These two scenarios are both cases of Shepard's (1987b; 1994) ideal generalization problem: given an encounter with a single stimulus (a patient, a worm) that can be represented as a point in some psychological space (a hormone level or pigmentation level of 60), and that has been found to have some particular consequence (healthy, good to eat), what other stimuli in that space should be expected to have the same consequence? Shepard observes that across a wide variety of experimental situations, including both human and animal subjects, generalization gradients tend to fall off approximately exponentially with distance in an appropriately scaled psychological space (as obtained by multidimensional scaling, or MDS). He then gives a rational probabilistic argument for the origin of this universal law, starting with some basic assumptions about the geometry of natural kinds in psychological spaces, which could be expected to apply equally well to doctors or robins, or even aliens from another galaxy. The argument makes no distinction in principle between conscious, deliberate, "cognitive" inferences, such as the healthy hormone levels scenario, and unconscious, automatic, or "perceptual" inferences, such as the good-to-eat worms scenario, as long as they satisfy the conditions of the ideal generalization problem.

In the opening sentences of his first paper on the uni-

Joshua B. Tenenbaum is Assistant Professor of Psychology at Stanford University. In 1999, he received a Ph.D. in Brain and Cognitive Sciences from MIT. His research focuses on learning and inference in humans and machines, with specific interests in concept learning and generalization, similarity, reasoning, causal induction, and learning perceptual representations.
Thomas L. Griffiths is a doctoral student in the Department of Psychology at Stanford University. His research interests concern the application of mathematical and statistical models to human cognition.

© 2001 Cambridge University Press 0140-525X/01 $12.50

versal law of generalization, Shepard (1987b) invokes Newton's universal law of gravitation as the standard to which he aspires in theoretical scope and significance. The analogy holds more strongly than might have been anticipated. Newton's law of gravitation was expressed in terms of the attraction between two point masses: every object in the universe attracts every other object with a force directed along the line connecting their centers of mass, proportional to the product of their masses and inversely proportional to the square of their separation. However, most of the interesting gravitational problems encountered in the universe do not involve two point masses. In order to model real-world gravitational phenomena, physicists following Newton have developed a rich theory of classical mechanics that extends his law of gravitation to address the interactions of multiple, arbitrarily extended bodies.

Likewise, Shepard formulated his universal law with respect to generalization from a single encountered stimulus to a single novel stimulus, and he assumed that stimuli could be represented as points in a continuous metric psychological space. However, many of the interesting problems of generalization in psychological science do not fit this mold. They involve inferences from multiple examples, or stimuli that are not easily represented in strictly spatial terms. For example, what if our doctor observes the hormone levels of not one but three healthy patients: 60, 30, and 50. How should that change the generalization gradient? Or what if the same numbers had been observed in a different context, as examples of a certain mathematical concept presented by a teacher to a student? Certain features of the numbers that were not salient in the hormone context, such as being even or being multiples of ten, now become very important in a mathematical context. Consequently, a simple one-dimensional metric space representation may no longer be appropriate: 80 may be more likely than 47 to be an instance of the mathematical concept exemplified by 60, 30, and 50, while given the same examples in the hormone context, 47 may be more likely than 80 to be a healthy level. Just as physicists now see Newton's original two-point-mass formulation as a special case of the more general classical theory of gravitation, so would we like a more general theory of generalization, which reduces to Shepard's original two-points-in-psychological-space formulation in the appropriate special cases, but which extends his approach to handle generalization from multiple, arbitrarily structured examples.

In this article we outline the foundations of such a theory, working with the tools of Bayesian inference and in the spirit of rational analysis (Anderson 1990; Chater & Oaksford 1998; 1999; Marr 1982). Much of our proposal for extending Shepard's theory to the cases of multiple examples and arbitrary stimulus structures has already been introduced in other papers (Griffiths & Tenenbaum 2000; Tenenbaum 1997; 1999a; 1999b; Tenenbaum & Xu 2000). Our goal here is to make explicit the link to Shepard's work and to use our framework to make connections between his work and other models of learning (Feldman 1997; Gluck & Shanks 1994; Haussler et al. 1994; Kruschke 1992; Mitchell 1997), generalization (Heit 1998; Nosofsky 1986), and similarity (Chater & Hahn 1997; Medin et al. 1993; Tversky 1997). In particular, we will have a lot to say about how our generalization of Shepard's theory relates to Tversky's (1977) well-known set-theoretic models of similarity. Tversky's set-theoretic approach and Shepard's metric space approach are often considered the two classic – and classically opposed – theories of similarity and generalization. By demonstrating close parallels between Tversky's approach and our Bayesian generalization of Shepard's approach, we hope to go some way towards unifying these two theoretical approaches and advancing the explanatory power of each.

The plan of our article is as follows. In section 2, we recast Shepard's analysis of generalization in a more general Bayesian framework, preserving the basic principles of his approach in a form that allows us to apply the theory to situations with multiple examples and arbitrary (nonspatially represented) stimulus structures. Sections 3 and 4 describe those extensions, and section 5 concludes by discussing some implications of our theory for the internalization of perceptual-cognitive universals.

2. A Bayesian framework for generalization

Shepard (1987b) formulates the problem of generalization as follows. We are given one example, x, of some consequence C, such as a "healthy person" or a "good-to-eat worm." We assume that x can be represented as a point in a continuous metric psychological space, such as the one-dimensional space of hormone levels between 0 and 100, and that C corresponds to some region – the *consequential region* – of that space. Our task is then to infer the probability that some newly encountered object y will also be an instance of C, that is, that y will fall in the consequential region for C. Formalizing this induction problem in probabilistic terms, we are asking for $p(y \in C|x)$, the conditional probability that y falls under C given the observation of the example x.

The theory of generalization that Shepard develops and that we will extend here can best be understood by considering how it addresses three crucial questions of learning (after Chomsky 1986):

1. What constitutes the learner's knowledge about the consequential region?
2. How does the learner use that knowledge to decide how to generalize?
3. How can the learner acquire that knowledge from the example encountered?

Our commitment to work within the paradigm of Bayesian probabilistic inference leads directly to rational answers for each of these questions. The rest of this section presents these answers and illustrates them concretely using the hormone or pigmentation levels tasks introduced above. Our main advance over Shepard's original analysis comes in introducing the *size principle* (Tenenbaum 1997; 1999a; 1999b) for scoring hypotheses about the true consequential region based on their size, or specificity. Although it makes little difference for the simplest case of generalization studied by Shepard, the size principle will provide the major explanatory force when we turn to the more realistic cases of generalizing from multiple examples (sect. 3) with arbitrary structure (sect. 4).

2.1. What constitutes the learner's knowledge about the consequential region?

The learner's knowledge about the consequential region is represented as a probability distribution $p(h|x)$ over an a priori-specified *hypothesis space* $\mathcal{H}$ of possible consequential regions $h \in \mathcal{H}$. $\mathcal{H}$ forms a set of exhaustive and mutually exclusive possibilities; that is, one and only one element of $\mathcal{H}$ is assumed to be the true consequential region for C (although the different candidate regions represented in $\mathcal{H}$ may overlap arbitrarily in the stimuli that they include). The learner's background knowledge, which may include both domain-specific and domain-general components, will often translate into constraints on which subsets of objects belong to $\mathcal{H}$. Shepard (1994) suggests the general constraint that consequential regions for basic natural kinds should correspond to connected subsets of psychological space. Applying the connectedness constraint to the domains of hormone levels or worm pigmentation levels, where the relevant stimulus spaces are one-dimensional continua, the hypothesis spaces would consist of intervals, or ranges of stimuli between some minimum and maximum consequential levels. Figure 1 shows a number of such intervals which are consistent with the single example of 60. For simplicity, we have assumed in Figure 1 that only integer stimulus values are possible, but in many cases both the stimulus and hypothesis spaces will form true continua.

At all times, the learner's knowledge about the consequential region consists of a probability distribution over $\mathcal{H}$. Prior to observing x, this distribution is the prior probability $p(h)$; after observing x, it is the posterior probability $p(h|x)$. As probabilities, $p(h)$ and $p(h|x)$ are numbers between 0 and 1 reflecting the learner's degree of belief that h is in fact the true consequential region corresponding to C. In Figure 1, $p(h|x)$ for each h is indicated by the thickness (height) of the corresponding bar. The probability of any h that does not contain x will be zero, because it cannot be the true consequential region if it does not contain the one observed example. Hence, Figure 1 shows only hypotheses consistent with $x = 60$.

2.2. How does the learner use that knowledge to decide how to generalize?

The generalization function $p(y \in C|x)$ is computed by summing the probabilities $p(h|x)$ of all hypothesized consequential regions that contain y:[1]

$$p(y \in C|x) = \sum_{h:y \in h} p(h|x). \tag{1}$$

We refer to this computation as *hypothesis averaging*, because it can be thought of as averaging the predictions that each hypothesis makes about y's membership in C, weighted by the posterior probability of that hypothesis. Because $p(h|x)$ is a probability distribution, normalized to sum to 1 over all $h \in \mathcal{H}$, the structure of Equation 1 ensures that $p(y \in C|x)$ will always lie between 0 and 1. In general, the hypothesis space need not be finite or even countable. In the case of a continuum of hypotheses, such as the space of all intervals of real numbers, all probability distributions over $\mathcal{H}$ become probability densities and the sums over $\mathcal{H}$ (in Equations 1 and following) become integrals.

The top panel of Figure 1 shows the generalization gradient that results from averaging the predictions of the integer-valued hypotheses shown below, weighted by their

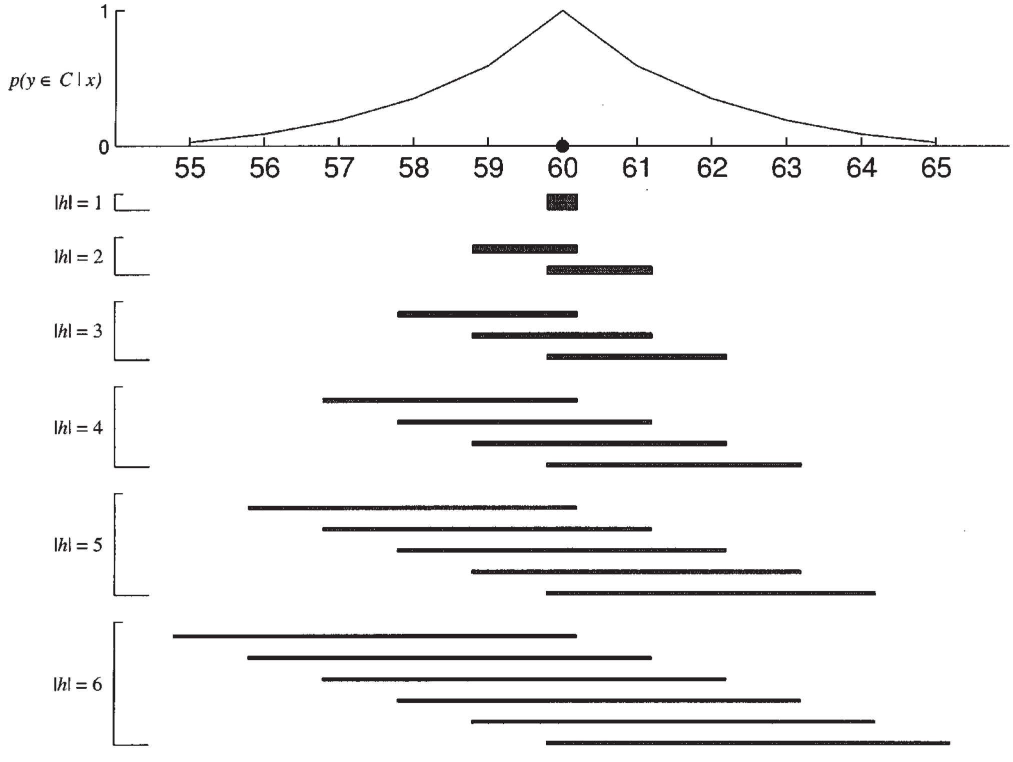

Figure 1. An illustration of the Bayesian approach to generalization from $x = 60$ in a one-dimensional psychological space (inspired by Shepard 1989, August). For the sake of simplicity, only intervals with integer-valued endpoints are shown. All hypotheses of a given size are grouped together in one bracket. The thickness (height) of the bar illustrating each hypothesis h represents $p(h|x)$, the learner's degree of belief that h is the true consequential region given the observation of x. The curve at the top of the figure illustrates the gradient of generalization obtained by integrating over just these consequential regions. The profile of generalization is always concave regardless of what values $p(h|x)$ takes on, as long as all hypotheses of the same size (in one bracket) take on the same probability.

probabilities. Note that the probability of generalization equals 1 only for $y = x$, when every hypothesis containing x also contains y. As y moves further away from x, the number of hypotheses containing x that also contain y decreases, and the probability of generalization correspondingly decreases. Moreover, Figure 1 shows the characteristic profile of Shepard's "universal" generalization function: concave, or negatively accelerated as y moves away from x. If we were to replace the integer-valued interval hypotheses with the full continuum of all real-valued intervals, the sum in Equation 1 would become an integral, and the piecewise linear gradient shown in Figure 1 would become a smooth function with a similar concave profile, much like those depicted in the top panels of Figures 2 and 3.

Figure 1 demonstrates that Shepard's approximately exponential generalization gradient emerges under one particular assignment of $p(h|x)$, but it is reasonable to ask how sensitive this result is to the choice of $p(h|x)$. Shepard (1987b) showed that the shape of the gradient is remarkably insensitive to the probabilities assumed. As long as the probability distribution $p(h|x)$ is *isotropic,* that is, independent of the location of h, the generalization function will always have a concave profile. The condition of isotropy is equivalent to saying that $p(h|x)$ depends only on $|h|$, the size of the region h; notice how this constraint is satisfied in Figure 1.

2.3. How can the learner acquire that knowledge from the example encountered?

After observing x as an example of the consequence C, the learner updates her beliefs about the consequential region from the prior $p(h)$ to the posterior $p(h|x)$. Here we consider how a rational learner arrives at $p(h|x)$ from $p(h)$, through the use of Bayes' rule. We will not have much to say about the origins of $p(h)$ until section 5; Shepard (1987b) and Tenenbaum (1999a; 1999b) discuss several reasonable alternatives for the present scenarios, all of which are isotropic and assume little or no knowledge about the true consequential region.

Bayes' rule couples the posterior to the prior via the *likelihood,* $p(x|h)$, the probability of observing the example x given that h is the true consequential region, as follows:

$$p(h|x) = \frac{p(x|h)p(h)}{p(x)} \tag{2}$$

$$= \frac{p(x|h)p(h)}{\sum_{h' \in \mathcal{H}} p(x|h')p(h')}. \tag{3}$$

What likelihood function we use is determined by how we think the process that generated the example x relates to the true consequential region for C. Shepard (1987b) argues for a default assumption that the example x and consequential region C are sampled independently, and x just happens to land inside C. This assumption is standard in the machine learning literature (Haussler et al. 1994; Mitchell 1997), and also maps onto Heit's (1998) recent Bayesian analysis of inductive reasoning. Tenenbaum (1997; 1999a) argues that under many conditions, it is more natural to treat x as a random positive example of C, which involves the stronger assumption that x was explicitly sampled from C. We refer to these two models as weak sampling and strong sampling, respectively.

Under weak sampling, the likelihood just measures in a binary fashion whether or not the hypothesis is consistent with the observed example:

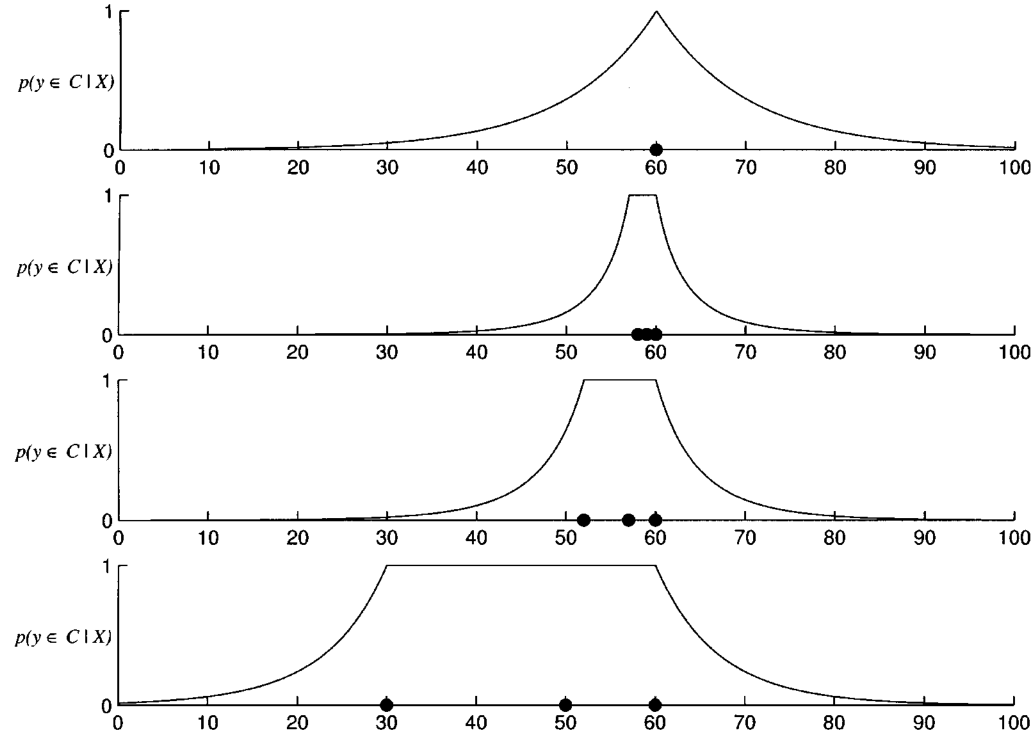

Figure 2. The effect of example variability on Bayesian generalization (under the assumptions of strong sampling and an Erlang prior, $\mu = 10$). Filled circles indicate examples. The first curve is the gradient of generalization with a single example, for the purpose of comparison. The remaining graphs show that the range of generalization increases as a function of the range of examples.

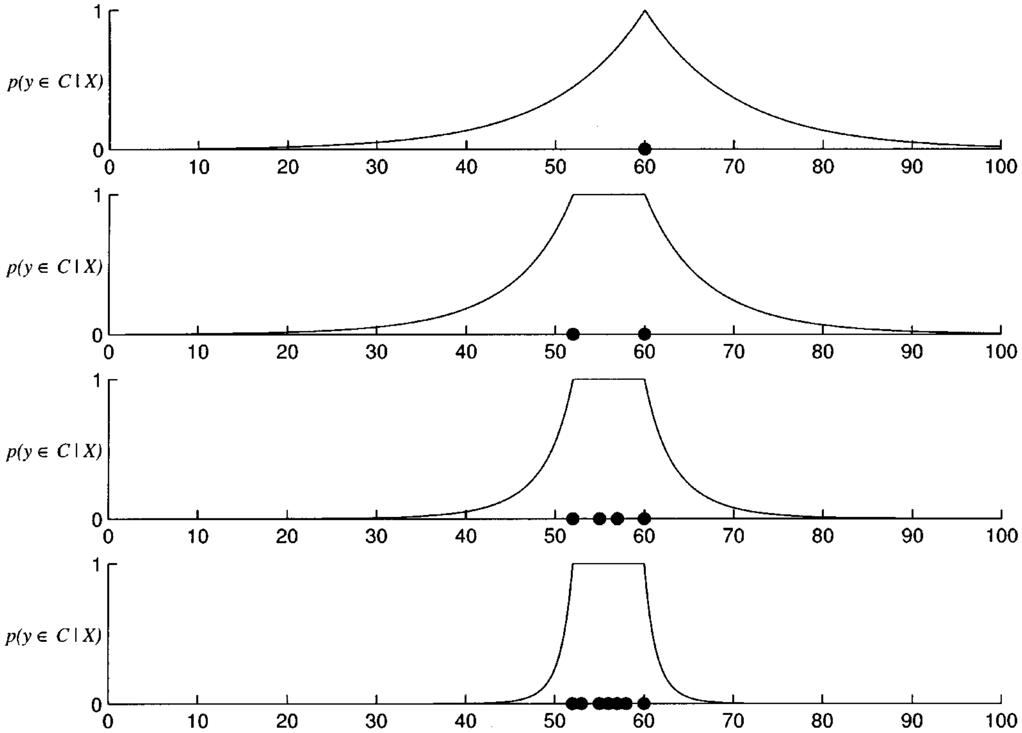

Figure 3. The effect of the number of examples on Bayesian generalization (under the assumptions of strong sampling and an Erlang prior, $\mu = 10$). Filled circles indicate examples. The first curve is the gradient of generalization with a single example, for the purpose of comparison. The remaining graphs show that the range of generalization decreases as a function of the number of examples.

$$p(x|h) = \begin{cases} 1 & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \quad \text{[weak sampling].} \qquad (4)$$

Under strong sampling, the likelihood is more informative. Assuming x is sampled from a uniform distribution over the objects in h, we have:

$$p(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \quad \text{[strong sampling],} \qquad (5)$$

where $|h|$ indicates the size of the region h. For discrete stimulus spaces, $|h|$ is simply the cardinality of the subset corresponding to h. For continuous spaces such as the hormone or pigmentation levels, the likelihood becomes a probability density and $|h|$ is the *measure* of the hypothesis – in one dimension, just the length of the interval.[2] Equation 5 implies that smaller, more specific hypotheses will tend to receive higher probabilities than larger, more general hypotheses, even when both are equally consistent with the observed consequential stimulus. We will call this tendency the *size principle.* It is closely related to principles of *genericity* that have been proposed in models of visual perception and categorization (Feldman 1997; Knill & Richards 1996). Figure 1 depicts the application of the size principle graphically.

Note that both Equations 4 and 5 are isotropic, and thus the choice between strong sampling and weak sampling has no effect on Shepard's main result that generalization gradients are universally concave. However, as we now turn to look at the phenomena of generalization from multiple stimuli with arbitrary, nonspatially represented structures, we will see that the size principle implied by strong sampling carries a great deal of explanatory power not present in Shepard's original analysis.

3. Multiple examples

In this section, we extend the above Bayesian analysis to situations with multiple consequential examples. Such situations arise quite naturally in the generalization scenarios we have already discussed. For instance, how should our doctor generalize after observing hormone levels of 60, 30, and 50 in three healthy patients? We first discuss some basic phenomena that arise with multiple examples and then turn to the extension of the theory. Finally, we compare our approach to some alternative ways in which Shepard's theory has been adapted to apply to multiple examples.

3.1. Phenomena of generalization from multiple examples

We focus on two classes of phenomena: the effects of example variability and the number of examples.

3.1.1. Example variability. All other things being equal, the lower the variability in the set of observed examples, the lower the probability of generalization outside their range. The probability that 70 is a healthy hormone level seems greater given the three examples {60, 50, 30} than given the three examples {60, 57, 52}, and greater given {60, 57, 52} than given {60, 58, 59}. Effects of exemplar variability on generalization have been documented in several other categorization and inductive inference tasks (Fried & Holyoak 1984; Osherson et al. 1990; Rips 1989).

3.1.2. Number of examples. All other things being equal, the more examples observed within a given range, the lower the probability of generalization outside that range. The probability that 70 is a healthy hormone level seems greater given the two examples {60, 52} than given the four examples {60, 52, 57, 55}, and greater given {60, 52, 57, 55} than given {60, 52, 57, 55, 58, 55, 53, 56}. This effect is most dramatic when there is very little variability in the observed examples. Consider the three sets of examples {60}, {60, 62, 61}, and {60, 62, 61, 62, 60, 62, 60, 61}. With just two more examples, the probability of generalizing to 70 from {60, 62, 61} already seems much lower than given {60} alone, and the probability given {60, 62, 61, 62, 60, 62, 60, 61} seems close to zero.

3.2. Extending the theory

Let $X = \{x_1, \ldots x_n\}$ denote a sequence of n examples of some consequence C, and let y denote a novel object for which we want to compute the probability of generalizing, $p(y \in C|X)$. All we have to do to make the theory of section 2 applicable here is to replace "x," wherever it appears, with "X," and to adopt the assumption of strong sampling rather than Shepard's original proposal of weak sampling. The rest of the formalism is unchanged. The only complication this introduces comes in computing the likelihood $p(X|h)$. If we make the simplifying assumption that the examples are sampled independently of each other (a standard assumption in Bayesian analysis), then Equation 5 becomes:

$$p(X|h) = \prod_i p(x_i|h) \tag{6}$$

$$= \begin{cases} \frac{1}{|h|^n} & \text{if } x_1, \ldots, x_n \in h \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Hence the size principle of Equation 5 has been generalized to include the influence of n: smaller hypotheses receive higher likelihoods than larger hypotheses, by a factor that increases exponentially with the number of examples observed. Figures 2 and 3 depict the Bayesian gradients of generalization that result for several different numbers and ranges of examples, assuming $p(X|h)$ based on strong sampling and an Erlang distribution (Shepard 1987b) for $p(h)$. In addition to showing the universal concave profile, these gradients display the appropriate sensitivity to the number and variability of examples.

To understand how the size principle generates these effects, consider how Equation 7 weights two representative hypotheses: h_0, the smallest interval containing all the examples in X, and h_1, a broader interval centered on h_0 but extending by $d/2$ units on either side, so that $|h_1| = |h_0| + d$. After observing n examples, the relative probabilities are proportional to the likelihood ratio:

$$\mathcal{L} = \frac{p(X|h_1)}{p(X|h_0)} = \left[\frac{1}{1 + d/|h_0|}\right]^n. \tag{8}$$

$\mathcal{L}$ is always less than 1, because d and $|h_0|$ are both positive. As $|h_0|$ increases, but the other quantities remain fixed, $\mathcal{L}$ increases. Thus, as we see in Figure 2, the relative probability that C extends a given distance d beyond the examples increases as the range spanned by the examples increases. As n increases while the other quantities remain fixed, $\mathcal{L}$ quickly approaches 0. Thus, as we see in Figure 3, the probability that C extends a distance d beyond the examples rapidly decreases as the number of examples increases within a fixed range. The tighter the examples, the smaller $|h_0|$ is, and the faster $\mathcal{L}$ decreases with increasing n, thus accounting for the interaction between these two factors pointed to earlier.

We can also now see why Shepard's original assumption of weak sampling would not generate these phenomena. Under weak sampling, the likelihoods of any two consistent hypotheses are always both 1. Thus $\mathcal{L} = 1$ always, and neither the range nor the number of examples have any effect on how hypotheses are weighted. In general, we expect that both strong sampling and weak sampling models will have their uses. Real-world learning situations may often require a combination of the two, if some examples are generated by mere observation of consequential stimuli (strong sampling) and others by trial-and-error exploration (weak sampling).

Figure 4 illustrates an extension to generalizing in two separable dimensions, such as inferring the healthy levels of two independent hormones (for more details, see Tenenbaum 1999b). Following Shepard (1987b), we assume that the consequential regions correspond to axis-aligned rectangles in this two-dimensional space, with independent priors in each dimension. Then, as shown in Figure 4, the size principle acts to favor generalization along those dimensions for which the examples have high variability and to restrict generalization along dimensions for which they have low variability. Tenenbaum (1999b) reports data from human subjects that are consistent with these predictions for a task of estimating the healthy levels of two biochemical compounds. More studies need to be done to test these predictions in multidimensional perceptual spaces of the sort with which Shepard has been most concerned.

3.3. Alternative approaches

A number of other computational models may be seen as alternative methods of extending Shepard's approach to the case of multiple examples, but only the framework we describe here preserves what we take to be the two central features of Shepard's original analysis: a hypothesis space of possible consequential regions and a Bayesian inference procedure for updating beliefs about the true consequential region. The standard exemplar models of classification (e.g., Nosofsky 1986; 1998a) take Shepard's exponential law of generalization as a primitive, used to justify the assumption that exemplar activation functions decay exponentially with distance in psychological space. A different approach is based on connectionist networks (Gluck 1991; Shanks & Gluck 1994; Shepard & Kannapan 1990; Shepard & Tenenbaum 1991), in which input or hidden units represent consequential regions, and error-driven learning – rather than Bayesian inference – is used to adjust the weights from consequential region inputs to response outputs. A third class of models (Kruschke 1992; Love & Medin 1998) combines aspects of the first two, by embedding Shepard's exponential law within the activation functions of hidden units in a connectionist network for classification learning.

Space does not permit a full comparison of the various alternative models with our proposals. One important point of difference is that for most of these models, the generalization gradients produced by multiple examples of a given consequence are essentially just superpositions of the exponential decay gradients produced by each individual example. Consequently, those models cannot easily explain

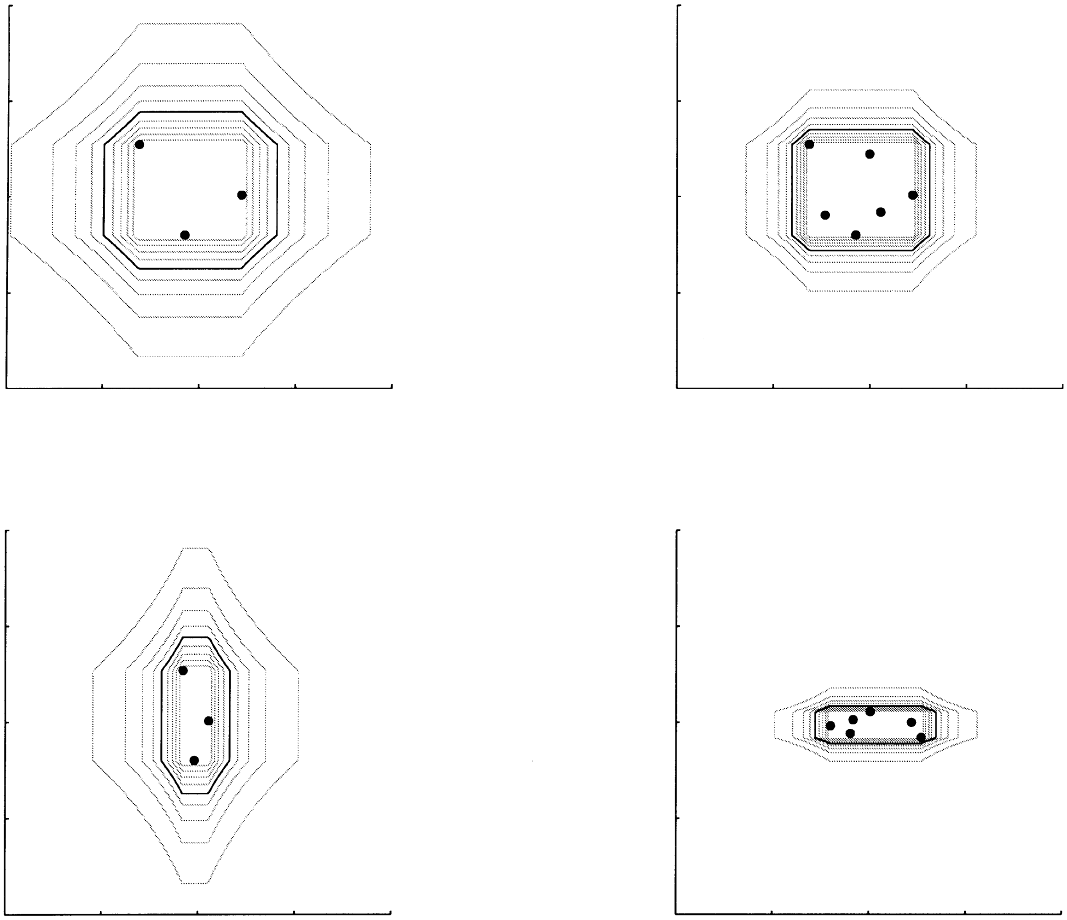

Figure 4. Bayesian generalization from multiple examples in two separable dimensions. Examples are indicated by filled circles. Contours show posterior probability, in increments of 0.1. Black contours illustrate the points at which $p(y \in C|X) = 0.5$. The range of generalization is affected by both the number of examples and the variability along a given dimension.

the phenomena discussed above, in which encountering additional consequential stimuli causes the probability of generalizing to some new stimulus to *decrease,* even when the additional examples are more similar to the new stimulus than the original example was. Exemplar and exemplar/connectionist hybrid models are frequently equipped with variable "attentional weights" that scale distances along a given input dimension by a greater or lesser amount, in order to produce variations in the contours of generalization like those in Figure 4. Such models could account for our phenomena by postulating that a dimension's length scale is initially large and decreases as the number of examples increases or the variability of the examples decreases, but nothing in the formal structure of these models necessarily implies such a mechanism. Our Bayesian analysis, in contrast, necessarily predicts these effects as rational consequences of the size principle.

4. Arbitrary stimulus structure

Shepard (1987b) assumed that objects can be represented as points in a continuous metric psychological space, and that the consequential subsets correspond to regions in that space with some convenient properties, such as connectedness or central symmetry. In general, though, we do not need to assume that the hypothesized consequential subsets correspond to regions in any continuous metric space; the notion of a consequential subset is sufficient for defining a Bayesian account of generalization. In this section we examine how arbitrary, nonspatially represented stimulus structures are modeled within the Bayesian framework.

Several authors, including Shepard himself, have described extensions of the original theory of generalization to conjunctive feature structures, in which objects are represented in terms of the presence or absence of primitive binary features and the possible consequential subsets consist of all objects sharing different conjunctions of features. For these cases, generalization gradients can still be shown to follow an exponential-like decay function of some appropriately defined distance measure (Gluck 1991; Russell 1988; Shepard 1989; 1994). However, the Bayesian analysis of generalization is more widely applicable than this. As we will show here, the analysis applies even when there is no independent notion of distance between stimuli and nothing like an exponential gradient emerges from the sum over consequential regions.

To motivate our analysis, consider a new generalization scenario. A computer has been programmed with a variety of simple mathematical concepts defined over the integers 1–100 – subsets of numbers that share a common, mathematically consequential property such as "even number," "power of two," or "square number." The computer will select one of these subsets at random, choose one or more numbers at random from that subset to show you as examples, and then quiz you by asking if certain other numbers belong to this same concept. Suppose that the number 60 is offered as one example of a concept the computer has chosen. What is the probability that the computer will accept 50? How about 51, 47, or 80? Syntactically, this task is almost identical to the hormone levels scenario above. But now, instead of generalization following a monotonic function of proximity in numerical magnitude, it seems more likely to follow some measure of mathematical similarity. For instance, the number 60 shares more mathematical properties with 50 than with 51, making 50 perhaps a better bet than 51 to be accepted given the one example of 60, even though 51 is closer in magnitude to 60 and therefore a better bet for the doctor trying to determine healthy hormone levels.

In our Bayesian framework, the difference between the two scenarios stems from the very different consequential subsets (elements of $\mathcal{H}$) that are considered. For the doctor, knowing something about healthy levels of hormones in general, it is quite natural to assume that the true consequential subset corresponds to some unknown interval, which gives rise to a generalization function monotonically related to proximity in magnitude. To model the number game, we can identify each mathematical property that the learner knows about with a possible consequential subset in $\mathcal{H}$. Figure 5 shows a generalization function that results under a set of 33 simple hypotheses, as calculated from the size principle (Eq. 5) and hypothesis averaging (Eq. 1). The generalization function appears much more jagged than in Figures 1–3 because the mathematical hypothesis space does not respect proximity in the dimension of numerical magnitude (corresponding to the abscissa of the figures). More generally, numerical cognition may incorporate both the spatial, magnitude properties as well as the nonspatial, mathematical properties of numbers. To investigate the nature of mental representations of numbers, Shepard et al. (1975) collected human similarity judgments for all pairs of integers between 0 and 9, under a range of different contexts. By submitting these data to an additive clustering analysis (Shepard & Arabie 1979; Tenenbaum 1996), we can construct the hypothesis space of consequential subsets that best accounts for people's similarity judgments. Table 1 shows that two kinds of subsets occur in the best-fitting additive clustering solution (Tenenbaum 1996): numbers sharing a common mathematical property, such as {2, 4, 8} and {3, 6, 9}, and consecutive numbers of similar magnitude, such as {1, 2, 3, 4} and {2, 3, 4, 5, 6}. Tenenbaum (2000) studied how people generalized concepts in a version of the number game that made both mathematical and magnitude properties salient. He found that a Bayesian model using a hypothesis space inspired by these additive clustering results, but defined over the integers 1–100, yielded an excellent fit to people's generalization judgments. The same flexibility in hypothesis space structure that allows the Bayesian framework to model both the spatial hormone level scenario and the nonspatial number game scenario there allows it to model generalization in a more generic context, by hypothesizing a mixture of consequential subsets for both spatial, magnitude properties and nonspatial, mathematical properties. In fact, we can define a Bayesian generalization function not just for spatial, featural, or simple hybrids of these representations, but for almost any collection of hypothesis subsets $\mathcal{H}$ whatsoever. The only restriction is that we be able to define a prior probability measure (discrete or continuous) over $\mathcal{H}$, and a measure over the space of objects, required for strong sampling to make sense. Even without a measure over the space of objects, a Bayesian analysis using weak sampling will still be possible.

4.1. Relations between generalization and set-theoretic models of similarity

Classically, mathematical models of similarity and generalization fall between two poles: continuous metric space models such as in Shepard's theory, and set-theoretic matching models such as Tversky's (1977) contrast model. The latter strictly include the former as a special case, but are most commonly applied in domains where a set of discrete conceptual features, as opposed to a low-dimensional continuous space, seems to provide the most natural stimulus representation (Shepard 1980). Our number game is such a domain, and indeed, when we generalize Shepard's Bayesian analysis from consequential regions in continuous metric spaces to apply to arbitrary consequential subsets, the model comes to look very much like a version of Tversky's set-theoretic models. Making this connection explicit allows us not only to unify the two classically opposing approaches to similarity and generalization, but also to explain

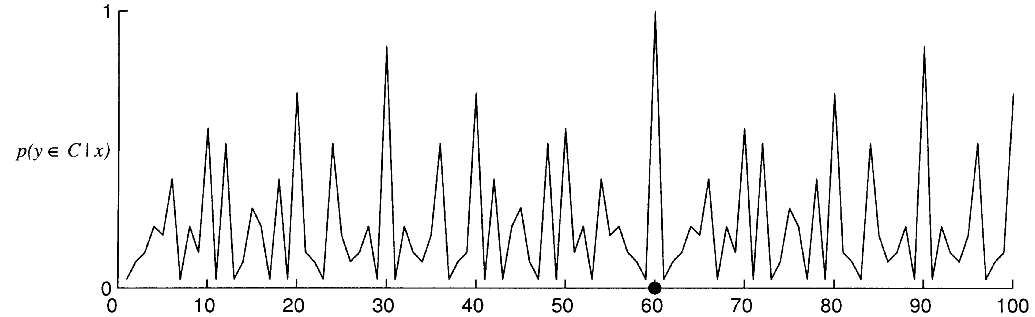

Figure 5. Bayesian generalization in the number game, given one example $x = 60$. The hypothesis space includes 33 mathematically consequential subsets (with equal prior probabilities): even numbers, odd numbers, primes, perfect squares, perfect cubes, multiples of a small number (3–10), powers of a small number (2–10), numbers ending in the same digit (1–9), numbers with both digits equal, and all numbers less than 100.

some significant aspects of similarity that Tversky's original treatment did not attempt to explain.

Tversky's (1977) contrast model expresses the similarity of y to x as

$$S(y,x) = \theta f(\mathcal{Y} \cap \mathcal{X}) - \alpha f(\mathcal{Y} - \mathcal{X}) - \beta f(\mathcal{X} - \mathcal{Y}), \tag{9}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the feature sets representing x and y, respectively, f denotes some measure over the feature sets, and θ, α, β are free parameters of the model. Similarity thus involves a contrast between the *common* features of y and x, $\mathcal{Y} \cap \mathcal{X}$, and their *distinctive* features, those possessed by y but not x, $\mathcal{Y} - \mathcal{X}$, and those possessed by x but not y, $\mathcal{X} - \mathcal{Y}$. Tversky also suggested an alternative form for the matching function, the *ratio* model, which can be written as

$$S(y,x) = 1/\left[1 + \frac{\alpha f(\mathcal{Y} - \mathcal{X}) + \beta f(\mathcal{X} - \mathcal{Y})}{f(\mathcal{Y} \cap \mathcal{X})}\right]. \tag{10}$$

The ratio model is remarkably similar to our Bayesian model of generalization, which becomes particularly apparent when the Bayesian model is expressed in the following form (mathematically equivalent to Eq. 1):

$$p(y \in C|x) = 1/\left[1 + \frac{\sum_{h:x \in h, y \notin h} p(h,x)}{\sum_{h:x,y \in h} p(h,x)}\right]. \tag{11}$$

Here, $p(h, x) = p(x|h)p(h)$ represents the weight assigned to hypothesis h in light of the example x, which depends on both the prior and the likelihood. The bottom sum ranges over all hypotheses that include both x and y, while the top sum ranges over only those hypotheses that include x but do not include y. If we identify each feature k in Tversky's framework with a hypothesized subset h, where an object belongs to h if and only if it possesses feature k, and if we make the standard assumption that the measure f is additive, then the Bayesian model as expressed in Equation 11 corresponds formally to the ratio model with $\alpha = 0$, $\beta = 1$. It is also monotonically related to the contrast model, under the same parameter settings.

Interpreting this formal correspondence between our Bayesian model of generalization and Tversky's set-theoretic models of similarity is complicated by the fact that in general the relation between similarity and generalization is not well understood. A number of authors have proposed that similarity is the more primitive cognitive process and forms (part of) the basis for our capacity to generalize inductively (Goldstone 1994; Osherson et al. 1990; Quine 1969; Rips 1975; Smith 1989). But from the standpoint of reverse-engineering the mind and explaining why human similarity or generalization computations take the form that they do, a satisfying theory of similarity is more likely to depend upon a theory of generalization than vice versa. The problem of generalization can be stated objectively and given a principled rational analysis, while the question of how similar two objects are is notoriously slippery and underdetermined (Goodman 1972). We expect that, depending on the context of judgment, the similarity of y to x may involve the probability of generalizing from x to y, or from y to x, or some combination of those two. It may also depend on other factors altogether. Qualifications aside, interesting consequences nonetheless follow just from the hypothesis that similarity somehow depends on generalization, without specifying the exact nature of the dependence.

4.1.1. The syntax of similarity. Most fundamentally, our Bayesian analysis provides a rational basis for the qualitative form of set-theoretic models of similarity. For instance, it explains why similarity should in principle depend on both the common *and* the distinctive features of objects. Tversky (1977) asserted as an axiom that similarity is a function of both common and distinctive features, and he presented some empirical evidence consistent with that assumption, but he did not attempt to explain why it should hold in general. Indeed, there exist both empirical models (Shepard 1980) and theoretical arguments (Chater & Hahn 1997) that have successfully employed only common or distinctive features. Our rational analysis (Eq. 11), in contrast, explains why both kinds of features should matter in general, under the assumption that similarity depends on generalization. The more hypothesized consequential subsets that contain both x and y (common features of x and y), relative to the number that contain only x (distinctive features of x), the higher the probability that a subset known to contain x will also contain y.

Along similar lines, the hypothesis that similarity depends in part on generalization explains why similarity may in principle be an asymmetric relationship, that is, why the similarity of x to y may differ from the similarity of y to x. Tversky (1977) presented compelling demonstrations of such asymmetries and showed that they could be modeled in his set-theoretic framework if the two subsets of distinctive features $\mathcal{X} - \mathcal{Y}$ and $\mathcal{Y} - \mathcal{X}$ have different measures under f and are given different weights in Equations 9 or 10. But Tversky's formal theory does not explain *why* those two subsets should be given different weights; it merely allows this as one possibility. In contrast, the probability of generalizing from x to y is intrinsically an asymmetric function, depending upon the distinctive features of x but not those of y. Likewise, the probability of generalizing from y to x depends only on the distinctive features of y, not those of x. To the extent that similarity depends on either or both of these generalization probabilities, it inherits their intrinsic asymmetry. Note that generalization can still be symmetric, when the distinctive features of x and y are equal in number and weight. This condition holds in the spatial scenarios considered above and in Shepard's work, which (not coincidentally) are also the domains in which similarity is found to be most nearly symmetric (Tversky 1977).

Finally, like Shepard's analysis of generalization, Tversky's contrast model was originally defined only for the comparison of two individual objects. However, our Bayesian framework justifies a natural extension to the problem of computing the similarity of an object y to a set of objects $X = \{x_1, \ldots x_n\}$ as a whole, just as it did for Shepard's theory in section 3. Heit (1997a) proposed on intuitive grounds that the contrast model should still apply in this situation, but with the feature set $\mathcal{X}$ for the examples as a whole identified with $\cap_{i=1}^{n} \mathcal{X}_i$, the intersection of the feature sets of all the individual examples. Our Bayesian analysis (replacing x with X in Eq. 11) explains why the intersection, as opposed to some other combination mechanism such as the union, is appropriate. Only those hypotheses consistent with all the examples in X – corresponding to those features belonging to the intersection of all the feature sets $\mathcal{X}_i$ – receive non-zero likelihood under Equation 7.

4.1.2. The semantics of similarity. Perhaps the most persistent criticisms of the contrast model and its relatives focus on semantic questions: What qualifies as a feature? What determines the feature weights? How do the weights change across judgment contexts? The contrast model has such broad explanatory scope because it allows any kind of features and any feature weights whatsoever, but this same lack of constraint also prevents the model from explaining the origins of the features or weights. Our Bayesian model likewise offers no constraints about what qualifies as a feature, but it does explain some aspects of the origins and the dynamics of feature weights. The Bayesian feature weight $p(h, x) = p(x|h)p(h)$ decomposes into prior and likelihood terms. The prior $p(h)$ is not constrained by our analysis; it can accommodate arbitrary flexibility across contexts but explains none of that flexibility. In contrast, the likelihood $p(x|h)$ is constrained by the assumption of strong sampling to follow the size principle.

One direct implication of this constraint is that, in a given context, features belonging to fewer objects – corresponding to hypotheses with smaller sizes – should be assigned higher weights. This prediction can be tested using additive clustering analyses, which recover a combination of feature extensions and feature weights that best fit a given similarity data set. For instance, the additive clustering analysis of the integers 0–9 presented in Table 1 is consistent with our prediction, with a negative correlation ($r = -0.83$) between the number of stimuli in each cluster and the corresponding feature weights. Similar relationships can be found in several other additive clustering analyses (Arabie & Carroll 1980; Chaturvedi & Carroll 1994; Lee, submitted; Tenenbaum 1996); see Tenenbaum et al. (in preparation) for a comprehensive study. Tversky (1977) proposed several general principles of feature weighting, such as the diagnosticity principle, but he did not explicitly propose a correlation between feature specificity and feature weight, nor was his formal model designed to predict these effects.

A second implication of the size principle is that certain kinds of features should tend to receive higher weights in similarity comparisons, if they systematically belong to fewer objects. Medin et al. (1993) have argued that primitive features are often not as important as are *relational* features, that is, higher-order features defined by relations between primitives. Yet in some cases a relation appears *less* important than a primitive feature. Consider which bottom stimulus, A or B, is more similar to the top stimulus in each panel of Figure 6 (inspired by Medin et al.'s comparisons). In the left panel, the top stimulus shares a primitive feature with B ("triangle on top") and a relational feature with A ("all different shapes"). In an informal survey, 8 out of 10 observers chose B – the primitive feature match – as more similar at first glance. In the right panel, however, a different relation ("all same shape") dominates over the same primitive feature (9 out of 10 different observers chose A as more similar). Goldstone et al. (1989) report several other cases where "same" relations are weighted more highly than "different" relations in similarity comparisons. If similarity depends in part upon Bayesian generalization, then the size principle can explain the relative salience of these features in Figure 6. Let m be the number of distinct shapes (square, triangle, etc.) that can appear in the three positions of each stimulus pattern. Then the consequential subset for "all same shape" contains exactly m distinct stimuli, the subset for "triangle on top" contains m^2 stimuli, and the subset for "all different shapes" contains $m(m-1)(m-2)$ stimuli. Thus feature saliency is inversely related to subset size, just as we would expect under the size principle. More careful empirical tests of this hypothesis are required, but we conjecture that much of the relative importance of relational features versus primitive features may be explained by their differing specificities.

A final implication arises from the interaction of the size principle with multiple examples. Recall that in generalizing from multiple examples, the likelihood preference for smaller hypotheses increases exponentially in the number of examples (Eq. 7). The same effect can be observed with the weights of features in similarity judgments. For instance, in assessing the similarity of a number to 60, the feature "multiple of ten" may or may not receive slightly greater weight than the feature "even number." But in assessing similarity to the set of numbers {60, 80, 10, 30} as a whole, even though both of those features are equally consistent with the full set of examples, the more specific feature "multiple of ten" appears to be much more salient.

Table 1. *Additive clustering of similarity judgments for the integers 0–9 (from Tenenbaum 1996)*

Rank	Weight	Stimuli in class	Interpretation
1	.444	2 4 8	powers of two
2	.345	0 1 2	small numbers
3	.331	3 6 9	multiples of three
4	.291	6 7 8 9	large numbers
5	.255	2 3 4 5 6	middle numbers
6	.216	1 3 5 7 9	odd numbers
7	.214	1 2 3 4	smallish numbers
8	.172	4 5 6 7 8	largish numbers

5. Conclusions: Learning, evolution, and the origins of hypothesis spaces

We have described a Bayesian framework for learning and generalization that significantly extends Shepard's theory in two principal ways. In addressing generalization from multiple examples, our analysis is a fairly direct extension of

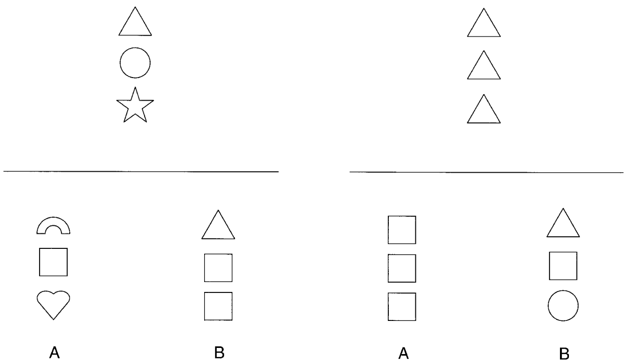

Figure 6. The relative weight of relations and primitive features depends on the size of the set of objects that they identify. Most observers choose B (the primitive feature match) as more similar to the top stimulus in the left panel, but choose A (the relational match) in the right panel, in part because the relation "all same shape" identifies a much smaller subset of objects than the relation "all different shapes."

Shepard's original ideas, making no substantive additional assumptions other than strong sampling. In contrast, our analysis of generalization with arbitrarily structured stimuli represents a more radical broadening of Shepard's approach, in giving up the notion that generalization is constrained by the metric properties of an evolutionarily internalized psychological space. On the positive side, this step allows us to draw together Tversky's set-theoretic models of similarity and Shepard's continuous metric space models of generalization under a single rational framework, and even to advance the explanatory power of Tversky's set-theoretic models using the same tools – chiefly, the size principle – that we used to advance Shepard's analysis of generalization. Yet it also opens the door to some large unanswered questions, which we close our article by pointing out.

In discussing similarity or generalization with arbitrarily structured stimuli, our Bayesian analysis explains only one piece of the puzzle of how features or hypotheses are weighted. Weights are always a product of both size-based likelihoods and priors, and while the size principle follows rationally from the assumption of strong sampling, the assignment of prior probabilities lies outside the scope of a basic Bayesian analysis. Thus, we can never say anything for certain about the relative weights of any two particular features or hypotheses merely based on their relative sizes; any size difference can always be overruled by a greater difference in prior probability.

The ability of prior probability differences to overrule an opposing size-based likelihood difference is hardly pathological; on the contrary, it is essential in every successful inductive generalization. Consider as a hypothesis in the number game that the computer accepts all multiples of ten, except 20 and 70. "Multiples of ten, except 20 and 70" is slightly more specific than "all multiples of ten," and thus should receive higher probability under the size principle given a set of examples that is consistent with both hypotheses, such as {60, 80, 10, 30}. But obviously, that does not happen in most people's minds. Our Bayesian framework can accommodate this phenomenon by stipulating that while the former hypothesis receives a somewhat higher likelihood, it receives a very much lower prior probability, and thus a significantly lower posterior probability when the prior and likelihood are combined.

It is by now almost a truism that without some reasonable a priori constraints on the hypotheses that learners should consider, there will always be innumerable bizarre hypotheses such as "all multiples of ten, except 20 and 70" that will stand in the way of reasonable inductive generalizations (Goodman 1955; 1972; Mitchell 1997). Trying to determine the nature and origin of these constraints is one of the major goals of much current research (e.g., Medin et al. 1993; Schyns et al. 1998). Shepard's original analysis of generalization was so compelling in part because it proposed answers to these questions: sufficient constraints on the form of generalization are provided merely by the representation of stimuli as points in a continuous metric psychological space (together with the assumption that hypotheses correspond to a suitable family of regions in that space), and our psychological spaces themselves are the products of an evolutionary process that has shaped them optimally to reflect the structure of our environment. In proposing a theory of generalization that allows for arbitrarily structured hypothesis spaces, we owe some account of where those hypothesis spaces and priors might come from. Evolution alone is not sufficient to explain why hypotheses such as "multiples of ten" are considered natural while hypotheses such as "all multiples of ten, except 20 and 70" are not.

The major alternative to evolution as the source of hypothesis space structure is some kind of prior learning. Most directly, prior experience that all and only those objects belonging to some particular subset h tend to possess a number of important consequences may lead learners to increase $p(h)$ for new consequences of the same sort. Unsupervised learning – observation of the properties of objects without any consequential input – may also be extremely useful in forming a hypothesis space for supervised (consequential) learning. Noting that a subset of objects tend to cluster together, to be more similar to each other than to other objects on some primitive features, may increase a learner's prior probability that this subset is likely to share some important but as-yet-unencountered consequence. The machine learning community is now intensely interested in improving the inductive generalizations that a supervised learning agent can draw from a few labeled examples, by building on unsupervised inferences that the agent can draw from a large body of unlabeled examples (e.g., Mitchell 1999; Poggio & Shelton 1999). We expect this to become a critical issue in the near future for cognitive science as well.

Our proposal that the building blocks of Shepard's "perceptual-cognitive universals" come into our heads via learning, and not just evolution, resonates with at least one other contribution to this issue (see Barlow's target article). However, we fundamentally agree with an earlier statement of Shepard's, that "learning is not an alternative to evolution but itself depends on evolution. There can be no learning in the absence of principles of learning; yet such principles, being themselves unlearned, must have been shaped by evolution" (Shepard 1995a, p. 59). Ultimately, we believe that it may be difficult or impossible to separate the contributions that learning and evolution each make to the internalization of world structure, given the crucial role that each process plays in making the other an ecologically viable means of adaptation. Rather, we think that it may be more worthwhile to look for productive synergies of the two processes, tools which evolution might have given us for efficiently learning those hypothesis spaces that will lead us to successful Bayesian generalizations. Such tools might include appropriately tuned stimulus metrics and topologies, as Shepard proposes, but also perhaps: unsupervised clustering algorithms that themselves exploit the size principle as defined over these metrics; a vocabulary of templates for the kinds of hypothesis spaces – continuous spaces, taxonomic trees, conjunctive feature structures – that seem to recur over and over as the basis for mental representations across many domains; and the ability to recursively compose hypothesis spaces in order to build up structures of ever-increasing complexity.

We believe that the search for universal principles of learning and generalization has only just begun with Shepard's work. The "universality, invariance, and elegance" of Shepard's exponential law (to quote from his article reprinted in this volume) are in themselves impressive, but perhaps ultimately of less significance than the spirit of rational analysis that he has pioneered as a general avenue for the discovery of perceptual-cognitive universals. Here we have shown how this line of analysis can be extended to yield what may yet prove to be another universal: the size principle, which governs generalization from one or more examples of arbitrary structure. We speculate that further universal principles will result from turning our attention in the future to the interface of learning and evolution.

ACKNOWLEDGMENTS

The writing of this article was supported in part by NSF grant DBS-9021648 and a gift from Mitsubishi Electric Research Labs. The second author was supported by a Hackett Studentship.

NOTES

[1] We derive Equation 1 as follows. Because $\mathcal{H}$ denotes an exhaustive and mutually exclusive set of possibilities, we can expand the generalization function as

$$p(y \in C|x) = \sum_{h \in \mathcal{H}} p(y \in C, h|x) \quad (12)$$

$$= \sum_{h \in \mathcal{H}} p(y \in C|h, x) p(h|x). \quad (13)$$

Note that $p(y \in C|h, x)$ is in fact independent of x. It is simply 1 if $y \in h$, and 0 otherwise. Thus we can rewrite Equation 13 in the form of Equation 1.

[2] Note that in a continuous space, when $|h| < 1$, $p(x|h)$ will be greater than 1 (for $x \in h$). This occurs because $p(x|h)$ is a probability density, not a probability distribution; probability density functions may take on values greater than 1, as long as they integrate to 1 over all x.