Available online at www.sciencedirect.com
SCIENCE DIRECT®
COGNITION
ELSEVIER
Cognition 93 (2004) 199–224
www.elsevier.com/locate/COGNIT

How surprising is a simple pattern? Quantifying "Eureka!"

Jacob Feldman

Department of Psychology, Center for Cognitive Science, Busch Campus, Rutgers University-New Brunswick, 152 Frelinghuysen Rd, Piscataway, NJ 08854, USA

Received 8 August 2002; revised 30 July 2003; accepted 19 September 2003

Abstract

Simple patterns are compelling. When all the observed facts fit into a simple theory or "story," we are intuitively convinced that the pattern must be *real* rather than random. But how surprising is a simple pattern, really? That is, given a pattern of featural data, such as the properties of a set of objects, how unlikely would the pattern be if they were actually generated at random? In conventional statistics dealing with patterns of numbers, this type of question would be answered by reference to a *null distribution* such as the *t* distribution. This paper gives the analogous answer in the realm of concept learning, that is, the formation of generalizations from patterns of featural data. Using a formal but psychologically valid definition of complexity, I derive and exhibit the *distribution of subjective complexity under the hypothesis of no pattern*. This leads directly to a number of applications, including a statistical test indicating whether an observed pattern is sufficiently simple that it is not likely to have been an accident: literally, the "significance of simplicity."

Keywords: Complexity; Concepts; Bayes; Pattern; Simplicity

Grand juries don't like coincidences.

-Anonymous legal aphorism

Simple patterns are compelling. When a police detective finds that the fingerprints, eyewitness testimony, forensic evidence, motive, etc. are all well-explained by the simple hypothesis *the butler did it*, the case will be convincing to the jury. After all, such a large

E-mail address: jacob@ruccs.rutgers.edu (J. Feldman).

0022-2860/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2003.09.013

amount of evidence is not likely to fit by accident into such a simple story. Like the solution to a crossword puzzle, the simple explanation makes all the clues pop into place. Eureka!

The same set of clues might also be explained by the hypothesis *the chauffeur did it, but framed the butler by bribing witnesses and concocting physical evidence*…But this theory's complexity and arbitrariness makes it very unconvincing. After all, almost *any* pattern of evidence could probably be squeezed into a theory that complicated. No Eureka.

Of course, if the evidence is not so clean (perhaps one eyewitness saw the chauffeur rather than the butler leaving the crime scene), the detective may have to settle for a somewhat more complex theory of the crime (*the butler did it, but one witness is mistaken*). In this case, the moderate complexity of the theory makes it moderately compelling. Medium Eureka.

In the more mundane domain of categorization and concept learning, roughly the same unconscious reasoning seems to occur. Say a certain set of observed objects have certain properties *ab, ac, ad, ae*… (e.g. *red triangle, red square, red circle*…). Although the objects differ from each other in many ways, they all share the single common property *a* (*red*). Surely (the unconscious reasoning goes) this commonality cannot be a coincidence; it is too unlikely that so many randomly selected objects would all happen to have property *a*. Hence it must be a real pattern, or (in Horace Barlow's suggestive phrase) a *suspicious coincidence*: the objects all belong to a common category whose members typically (or always) have property *a* (*red things*). Eureka!

Or consider the example of a "Bongard problem" (Bongard, 1970; Fig. 1a). Here eight objects are defined over four binary features (shape = *square* or *circle*, size = *small* or *large*, color = *filled* or *not-filled*, and side = *left* or *right*). The observer can quickly ascertain that these eight objects all obey the simple theory *Squares on the left, circles on the right*; in Boolean notation, (left → square) ∧ (right → circle). Surely given such a large number of objects (8) each defined over so many features (4), such a simple pattern is not likely to be the result of a random process. Rather some non-random process—in this case, segregation by shape classes—must have occurred, which we as the observer will surely care to note. Eureka!

Conversely, if we "scramble" the problem by exchanging two of the same eight objects as in Fig. 1b, we get a more complicated situation: no simple description now applies. We can still describe the arrangement, but only using a more complex and long-winded phrase such as *Big shaded square and small unshaded square and big shaded triangle and small unshaded triangle on the left, small shaded square and*… and so on. Indeed, the description is nothing other than an item-by-item recitation of the contents of the scene. No Eureka.

This line of reasoning presupposes that the observer has a way of measuring or estimating exactly how unlikely a given pattern of features is as a function of the *complexity* of the pattern: if you will, the *degree* of "Eureka." In order to accomplish this, the observer needs (a) a computable measure C of pattern complexity that accurately reflects the subjective impression of a theory's "badness"; and (b) the expected distribution $p(C)$ of this complexity measure over *random* patterns—this latter being essential if the observer is to judge the probability that the pattern is too simple to be a random outcome. To my knowledge, the statistical properties of conceptual complexity, including the form of the null distribution, have never previously been discussed in the literature. This article builds on recent progress in

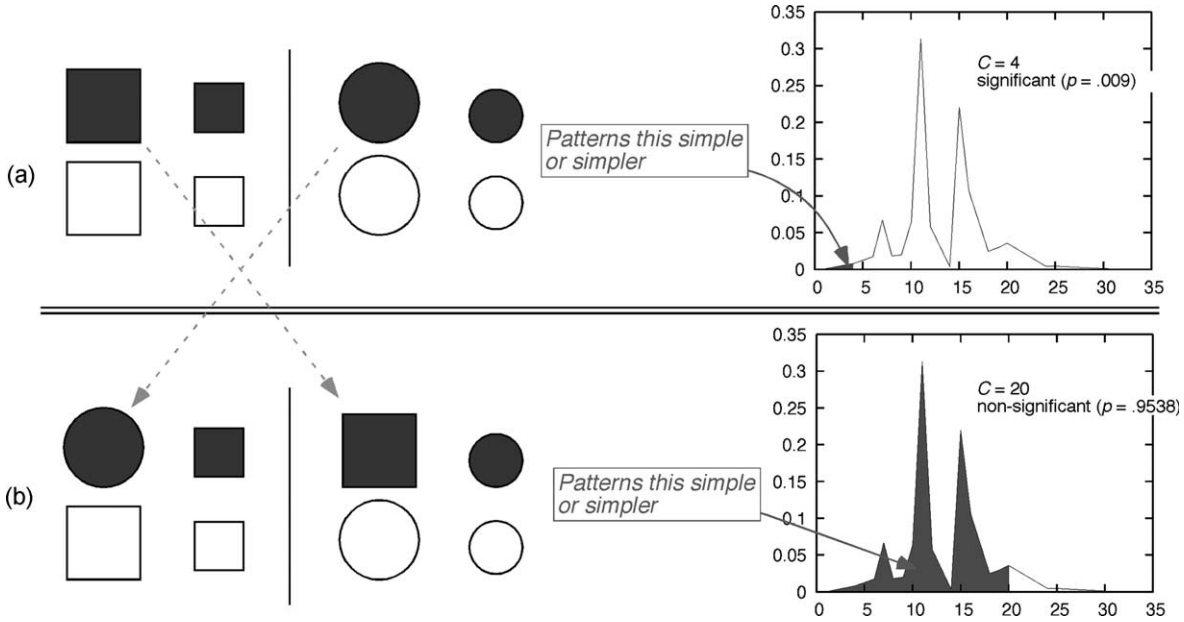

Fig. 1. (a) A "Bongard problem" (Bongard, 1970). Here, the categorical separation between the objects on the left and those on the right makes a "simple story" (complexity $C = 4$ literals), significantly simpler than a random pattern (see plot at right). In (b), two of the objects have been swapped, making a new problem with no simple solution. Complexity jumps to $C = 20$ literals, no longer simpler than a randomly constructed problem.

developing (a) (the complexity measure) by deriving and exhibiting (b) (the null distribution of complexity).

As illustrated below, the null distribution and the associated "significance test for simplicity" lead to a number of useful and concrete applications to problems in cognitive science. Broadly speaking, they are potentially relevant to any psychological process that is modulated by complexity, because they make it possible to connect degrees of complexity to degrees of probability. The premise is that this connection underlies the ineffable sense of "Eureka"—the subjective feeling that simple patterns bespeak non-random processes.

The mathematics governing the null distribution are somewhat involved, and many of the details are given in Appendices A and B. However, much of the math can be boiled down to a "thumbnail" or approximate version of the distribution, which gives nearly the right values with a minimum of fuss and calculation. The existence of this quick-and-dirty approximation is useful in practice, and moreover supports the idea that a rough probability assessment could be implemented in biological hardware, and could thus truly underlie our untutored intuitions about the genuineness of patterns.

1. Avoiding coincidences

Assume we have observed some set of objects, and coded them as a pattern with a given level of complexity C. Possessing the null complexity distribution would allow us to then say exactly how unlikely this degree of complexity is under the "null hypothesis" of *no pattern*. The inference that a given pattern is "real" (as opposed to random) corresponds to our subjective rejection of this null hypothesis.

The reasoning is extremely familiar from the domain of conventional diagnostic statistics. Imagine we have sampled a numeric measure (say, the lengths of seed pods) from two populations (e.g. species of plant), and would like to know whether the observed mean difference is "real" or "random." Are the two samples different enough to allow us to infer that the two distributions are genuinely different? (In proper statistical language, we say "the population means are not the same" instead of saying "the two distributions are genuinely different.") In this situation, we might use the t distribution: we would calculate the t value corresponding to the observed mean difference, and then look up this value on a t table, which would give the probability of observing such a large t under the null hypothesis that the populations were truly the same. If the probability is sufficiently low, we reject the null hypothesis, and conclude that the populations are in fact different. An analogous step would be possible in the domain of pattern detection, if only we had a model of the null distribution of complexity analogous to the null distribution of the sample mean given by the t distribution.

2. A measure of subjective complexity

The idea that human observers have a preference for simple or regular patterns has deep roots in psychology (Barlow, 1974; Hochberg & McAlister, 1953; Kanizsa, 1979)

and philosophy (Quine, 1965; Sober, 1975). But this idea has been slow to take hold in what is perhaps its most natural domain of application, namely the formation of generalizations and the induction of categories, or, as it is often called in the psychological literature, concept learning. Four decades ago Neisser and Weene (1962) suggested that human learners tended to prefer the logically simplest category consistent with their observations, but their suggestion was not met with enthusiasm in the learning literature. More recent models of human concept learning often have a very different flavor, emphasizing the used of stored examples ("exemplars") as a basis for generalization, with no explicit bias towards simplicity in the induced rule, and in fact no "induced rule" at all.[1]

Much of the trouble has stemmed from the difficulty in formulating the correct definition of "simplicity" and "complexity." Mathematicians beginning in the 1960s have converged on a solution to this problem: simplicity is the degree to which a given object can be *faithfully compressed* (an idea usually referred to as Kolmogorov complexity; see Li & Vitányi, 1997). ("Faithfully" here means "without loss of information.") Intrinsically complex objects are those that cannot be communicated without virtually quoting them verbatim, which takes as much information as the original object actually contains. Simple objects, by contrast, are those that can be expressed in a form much more compact then they are themselves.

In earlier work I have used a similar measure of complexity to study human learners' intuitions about concepts to be learned. In the simplest situation, learners are presented with a collection of sample objects, each of which can be represented by a set of Boolean (binary-valued) features. In this situation, the set of sample objects can be thought of as a long propositional formula, essentially the disjunction of the objects themselves. A simple measure of the complexity of these objects, then is the length of the *shortest equivalent formula*, that is, the length of the most compact way of faithfully describing the same set of object; this number is called the *Boolean complexity* (Givone, 1970; Wegener, 1987). Boolean complexity is most conveniently expressed in terms of the number of *literals*, or instances of variable names that appear in the shortest formula. For example the object set *red square or red circle* is logically equivalent to *red* (assuming shapes are always squares or circles), and so has Boolean complexity 1; exactly one variable name appears in the compressed expression. Conversely *red square or blue circle* cannot be compressed at all, so it has Boolean complexity 4.

Studies in my laboratory (Feldman, 2000b) have shown that the Boolean complexity correlates well with the difficulty subjects have learning the given concept—that is, that this number gives a good account of subjective or psychological complexity (see also Feldman, in press). Human concept learners—like police detectives—prefer patterns of data that can be summarized succinctly, i.e. "simple theories."

2.1. Algebraic complexity

In what follows below, I will actually use a related but different measure of subjective complexity, called the *algebraic complexity* (Feldman, 2001). Like Boolean complexity,

[1] Exemplar-based categorization models may in fact tend to produce simple abstractions, but only as it were "epiphenomenally," that is, not as an overt aspect of their design.

the algebraic complexity of a set of examples can be understood as the length of the shortest faithful representation of the observed objects, and is also measured in literals. However, rather than expressing formulae in the logicians' conventional basis of $\{\wedge, \vee, \neg\}$, the algebraic complexity uses a more subtle but more psychologically motivated basis. A synopsis of algebraic complexity and the concept algebra behind it is given in Appendix A. It should be noted, however, that the general argument below could be applied to any concrete, computable complexity measure, in whatever domain one happened to be interested in, and would probably yield generally similar results—although this is *not* so with uncomputable measures such as Kolmogorov complexity, for reasons discussed below.

Briefly, the idea behind algebraic complexity is to reduce each concept (or set of examples) to a structured representation, called the *implicational power series*, which expresses the concept in terms of all the regularities it obeys, expressed in a maximally compact form. The representation commits to a particular choice of what counts as a "regularity" (see Appendix A), assigning brief descriptions to those concepts that obey such regularities, and long descriptions to those that do not. The key idea is that the power series is stratified—overtly broken down by levels of component complexity. Thus the series gives a kind of spectral breakdown of the concept, the same way the Fourier decomposition breaks a function down into components at different frequencies (here, complexity levels); Fig. 2 shows several examples of concepts and their spectra. The spectral representation makes explicit how much of the concept's structure can be explained by simple rules, and how much can only be explained by more complex rules. The algebraic complexity itself is then simply the mean spectral power, indicating where in the complexity spectrum most of the concept's structure lies.

Because the representation overtly breaks the concept down into simple and complex components (and every level in between), it in effect expresses where the concept falls in the spectrum between "theory" and "exceptions" (cf. Nosofsky, Palmeri, & McKinley, 1994). The more regular it is, the more spectral power at the low end, and thus the lower the final complexity score C. The more internally "exceptional" it is, the more spectral power at the high end, and thus the higher the final complexity score (see Fig. 2). As C ranges from low to high, the associated concepts run the gamut from primarily rule-like to primarily "exceptional" in nature.

As with any concrete, computable complexity measure, the commitment to a particular regularity language means that some seemingly "regular" concepts are deemed complex because the regularities they obey are not of the type recognized by the language. An extreme example is the parity function (with $D = 2$ called exclusive-or; with $D = 3$ equivalent to Shepard, Hovland, and Jenkins (1961)'s type VI). This concept exhibits a neat alternating pattern and thus seems regular from a certain point of view, but nevertheless scores as complex and incompressible in any measure that doesn't know about that particular kind of alternation, such as the concept algebra (as well as many other languages; see Feldman, 2003; Schöning & Pruim, 1998 for discussion). Unlike Boolean complexity, algebraic complexity can be readily computed for concepts defined over non-Boolean discrete features, that is, features that have more than two possible values (e.g. *shape* = {*square*, *circle*, *triangle*,…}), though I will consider only Boolean features in

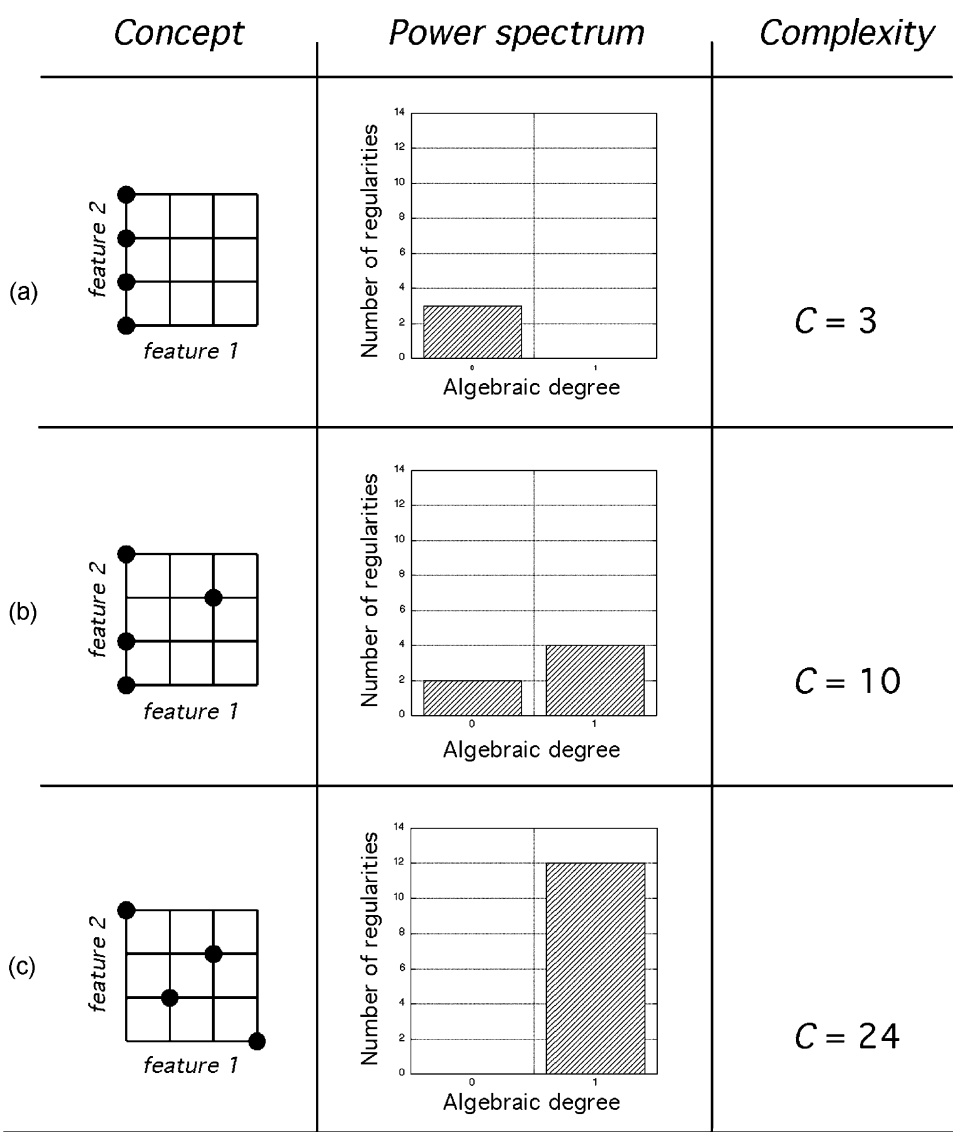

Fig. 2. Three concepts and their algebraic power spectra. As complexity increases, the spectra shift to the right and become more "exceptional." Each concept here is defined over two features, each having four possible values; heavy dots at the appropriate vertices indicating the members of the concept. With a simple concept (a), all the spectral power is at minimal degree, indicating unanimous satisfaction of a simple regularity. In (b), one of the four objects now fails to obey the regularity (an exception), and consequently some of the spectral power shifts to higher degree. In (c), none of the four objects obeys any common regularities, and much of the spectral power shifts to higher degrees, yielding a high complexity value. The final complexity value is the sum of spectral power weighted by degree plus one. For more explanation of how the algebraic power spectrum is computed, see Appendix A and Feldman (2001).

this paper (with the exception of the concepts illustrated in Fig. 2, defined over 4-valued features).

There are several reasons for using algebraic complexity rather than Boolean complexity as the measure in which to develop the null distribution. First, the actual computation of Boolean complexity is heuristic, using a variety of minimization tricks (because exact computation of it is computationally intractable); as a result it is difficult to model analytically. Second, computation of Boolean complexity, even in its heuristic approximation, is relatively intractable, which limits our ability to compute it to small numbers of features. For example, the Boolean complexity of random set of objects defined by five features would already be impractical to compute, whereas algebraic complexity of concepts with five or six or more features are readily computable (as will be seen below). Third, although the mathematics underlying algebraic complexity may seem more difficult, it is better suited to modeling human inferences, in that it more directly captures the main idea of building a more compact representation by extracting true regularities from the observed data. Finally, algebraic complexity fits the human data somewhat better than Boolean complexity (see Fig. 3).

The key points necessary to understand the following are as follows. Given a set of objects $\mathbf{x} = x_1, x_2, \ldots x_n$, we can compute its complexity $C(\mathbf{x})$, which is the length measured in literals of the shortest description of $\mathbf{x}$. Low values of C mean very simple theories, e.g. $C = 1$ means the whole set of objects can be described by a single property. High values of C mean complex sets of observations that cannot be compressed. For example with D features and n objects, one can list all the objects verbatim with a description of length Dn,

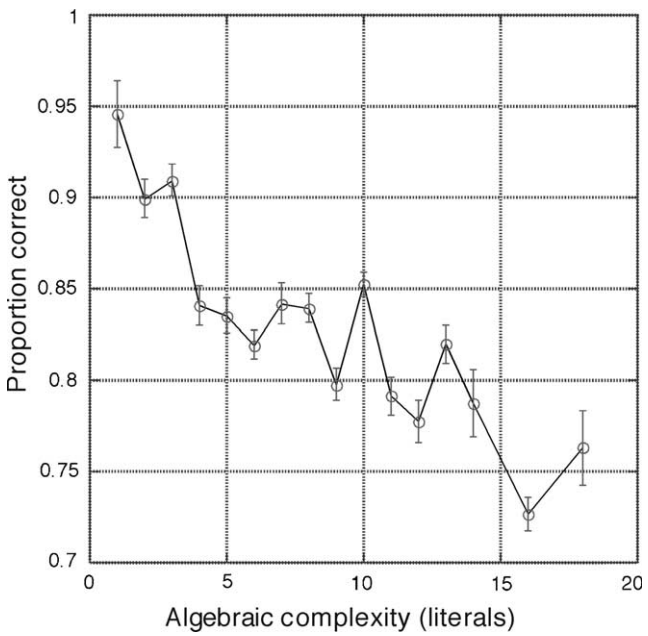

Fig. 3. As algebraic complexity increases, human performance in a learning task steadily decreases; human learners prefer simple generalizations. See Feldman (2000b) for details on the source of these data.

so a complexity value this high would mean that the observations could not be compressed at all—a very unsatisfying theory.

Now, with the complexity measure in place, we turn to the main question: if we generate object sets **x** at random, what will be the expected distribution of complexity $p(C)$?

3. The null distribution of complexity

As mentioned above, the mathematics required to explicitly model the distribution $p(C)$ is rather involved, but simply exhibiting and inspecting it is easy: all we need to do is to generate many object sets at random, compute their complexities, and tabulate the results. Here "at random" means that the binary features are all independent, with each taking each of its possible values with probability 1/2—that is, with feature values decided by flips of a fair coin. Of course, different assumptions about the nature of random concepts, e.g. different feature probabilities would lead to different distributions.

Fig. 4 shows the results of such a simulation, tabulating the complexities of large samples of randomly generated concepts for several values of D (= the number of features). Actually for low values of D we do not need to estimate the distribution via this Monte Carlo technique; we can measure it precisely. Boolean concepts only come in a finite variety of "flavors" or basic types (see Feldman, 2003). For $D = 2$, for example, there are only three types: affirmation (a, $C = 1$); conjunction/disjunction ($a \vee b$ or $a \wedge b$, $C = 2$); and exclusive-or ($(a \wedge \neg b) \vee (\neg a \wedge b)$, $C = 4$). All other two-feature concepts have essentially the same logical structure, and thus the same complexity, as one of these three types. This means, perhaps surprisingly, that for $D = 2$ complexity C can only take on the values 1, 2 or 4. The exact proportion of all concepts that fall into each of these three categories can be computed exactly, thus allowing the distribution $p(C)$ itself to be computed exactly. This means that the somewhat "jaggy" distributions shown for $D = 3$ and 4 in the figure are not actually poor approximations, but are themselves the real distributions: the "truth" itself is jaggy. Unfortunately, this strategy cannot be pursued for higher values of D, where the taxonomy, albeit similar, becomes much more complicated (see Feldman, 2003); hence the use of Monte Carlo simulations to estimate the distribution in these cases.

Randomly generated object sets can have different numbers n of objects (ranging from 0 to 2^D); the curves in Fig. 4 aggregate complexities calculated from sets of all sizes. Instead, given a sample of n objects, one might like to evaluate it with respect to the expected complexity for sets of equal size. These distributions are again easy to tabulate, but not as easy to inspect as the aggregate curves, because for each D there are 2^D distinct curves, which tend to overlap. Fig. 5 shows the distributions for $D = 4$ and 5 broken down into various levels of n (not all levels of n are shown, but rather a sample drawn from n close to 2^{D-1}). Each of these individual curves gives the expected distribution of complexity for a set of n objects defined over D features. As can be seen in the figure, these curves are much "jaggier" and overlap each other substantially, making it hard to distinguish them visually. (As before, the jagginess is "real" and not the result of undersampling.) I show them anyway because these curves constitute arguably the most

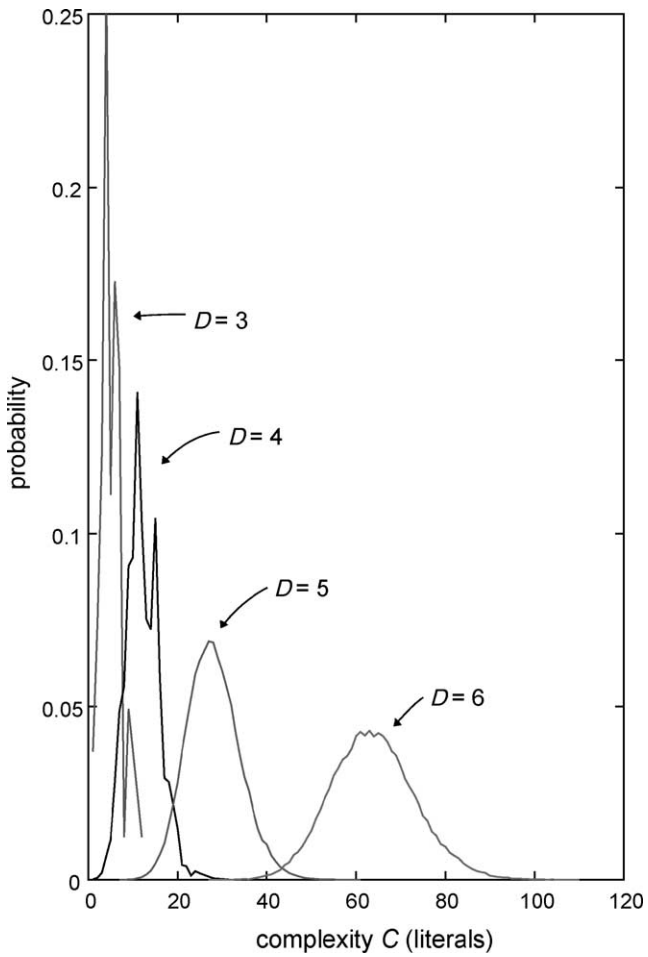

Fig. 4. The null distribution of complexity for $D = 3$, 4, 5 and 6. Each distribution tabulates the proportion of randomly selected objects that have each given level of algebraic complexity. The curves for $D = 3$ and 4 are complete tabulations, not samples: the variety of basic distinct types in those cases is small enough that they can be counted exactly. The curves at $D = 5$ and $D = 6$ are each estimates based on tabulation of 150,000 random concepts.

appropriate standard against which to judge the surprisingness of the complexity of a given observed sample of n objects. Conversely when one wishes to consider the expected complexity more abstractly *without* reference to a specific object set, the aggregate curves make a convenient summary.

3.1. Poisson and Gaussian approximations

The relatively smooth aggregate distributions (Fig. 4) resemble a *Poisson* distribution. A Poisson can be thought of as the expected number of "successes" in a large set of

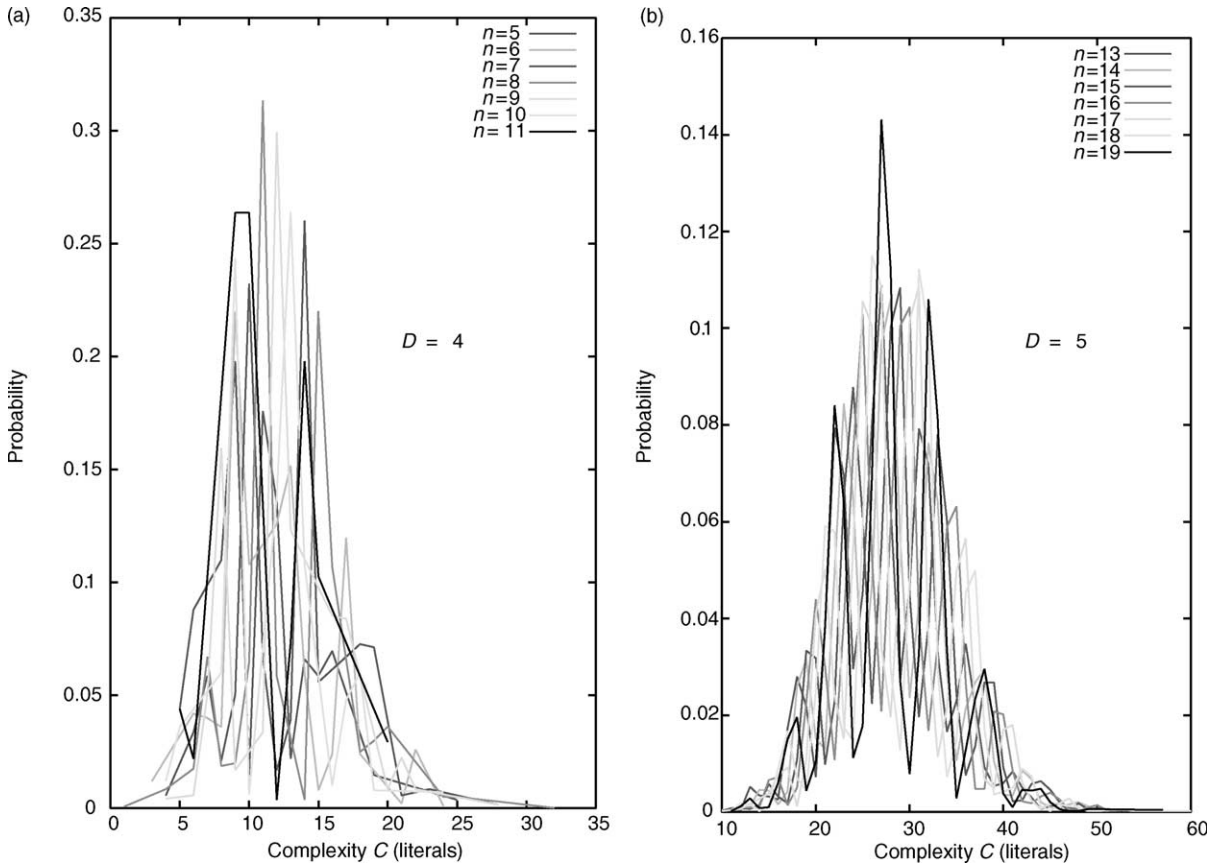

Fig. 5. The null distributions of complexity for (a) $D = 4$ and (b) $D = 5$, broken down into various levels of n (the number of objects). For each D the figures show n's near 2^{D-1}, i.e. near half the total objects.

independent trials of a low-probability event (e.g. the total number of "heads" on a series of tosses of a very tails-heavy coin; see Wickens, 1982 for an introduction). Here the rare event in question is roughly that a random set of observations will submit to a representation that includes a particular single literal in its most compact description. This observation is pursued more thoroughly in Appendix B, where the distribution is discussed in greater detail.

A Poisson distribution asymptotically resembles a Gaussian (normal) with identical mean and variance, in which case the single parameter is usually called λ (again see Wickens, 1982). Hence Fig. 4 resembles an ensemble of normal distributions;[2] as D increases, both the mean μ_D and variance σ_D^2 of expected complexity increase. These curves are very close to normality for $D \geq 5$: at $D = 5$ the correlation with a Gaussian of the same mean and variance is $r = 0.9956$, and at $D = 6$ it is $r = 0.9976$. Hence in each case the error in the normal approximation $(1 - r^2)$ is less than 1%. Moreover the means and variances of the distributions are almost perfectly correlated ($r = 0.9994$ in our sample of $D = 2$ through 6), supporting the claim that the distribution is approximately Poisson.

The approximately normal form of the aggregate distribution means that a typical random concept defined over, say, four features will tend to have complexity of about 12 literals, with about 68% falling within one standard deviation, i.e. between 8.5 and 15.5. Occasionally, such an object set will turn out more simple just by accident, but this will happen more rarely—the precise probability is given by the curve. Probabilities of *ranges* of complexity values—e.g. "simpler than $C = 4$"—can be computed by integrating the area under the curve, in the manner familiar from statistics books. In practice, because each curve (above $D = 4$) is approximately normal, this is easily accomplished by using a normal distribution look-up table found in any statistics book. This fact will be used below to help create a quick-and-dirty recipe for computing the significance of a given observed pattern.

The increase of $\mu_D (= \sigma_D^2 = \lambda)$ with D is approximately exponential (see Appendix B for some discussion of why this is so). Thus for $D = 3$ the distribution is centered at about $C = 5$ (s.d. about 2.2), for $D = 4$ at about $C = 12$ (s.d. about 3.6), for $D = 5$ about $C = 28$ (s.d. about 5); and for $D = 6$ about $C = 64$ (s.d. about 7.9); and so forth (Fig. 6). These values are well fit by a simple exponential growth model

$$\mu_D \approx \alpha e^{\beta D}. \tag{1}$$

The agreement between this model and the known means is very good ($R^2 = 0.99978$), with estimated parameters

$$\alpha = 0.36287, \tag{2}$$

$$\beta = 0.86686. \tag{3}$$

[2] The approximation by a normal distribution here cannot be completely correct, because a Gaussian has infinite tails in both directions, whereas complexity can never be negative. However, this is a numerically small error, as the probability assigned by the Gaussian model to negative complexities is vanishingly small, for example less than 0.000001 at $D = 5$ and decreasing with D. Nevertheless it should be kept in mind that the Gaussian model is only an approximation.

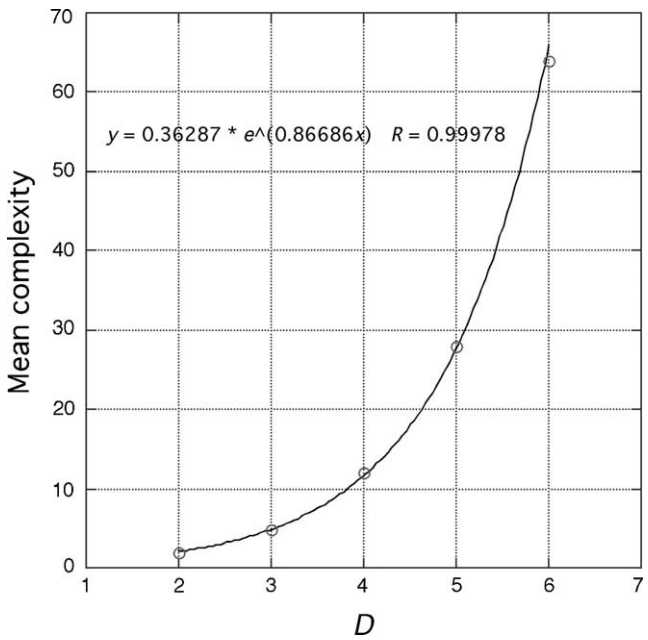

Fig. 6. Plot showing the exponential rise in mean expected complexity C as a function of D.

This formula allows us to estimate the expected complexity for values of D greater than those directly simulated here.

As mentioned, a more acute comparison for the complexity of a given sample of n objects is actually given by the particular curve for that level of n (and D). Unfortunately, these curves cannot be approximated via anything as simple as a Gaussian or Poisson model. First, as mentioned, these distributions are inherently "jaggy." In part this is because these curves contain a substantial "periodic" or fluctuating component superimposed over their primarily unimodal shape. The period and phase of the fluctuation depend on n, so while these components smooth out when ns are combined, they are non-negligible for each individual n. An even more serious deviation from unimodality stems from the fact that some complexities can only occur for particular values of n; for example only concepts with $n = 2^{D-1}$ can exhibit $C = 1$ (see Feldman, 2003). For these reasons when we wish to employ a quick-and-dirty approximation to the null distribution of complexity, we can refer only to the aggregate curves, although for a given set of n objects they admittedly make a less perfectly apposite standard.

4. A significance test for simplicity

Given a set of observations on D features with computed complexity C, one would like to be able to estimate how unlikely this level of complexity would have been if the observations were actually random. If the answer to this question is "very

unlikely", then we can "reject the null hypothesis" and conclude that our theory[3] is pretty good.[4] If the answer is "moderately likely", we might conclude that the observations might have been random after all. Thus we have a quantitative tool to help answer the police detective's quandary: is my best theory of the crime good enough to take to the grand jury?

Ideally, we would like to conduct this significance testing using the true null distri butions. In practice, the exact distribution may not be available, and one would like to substitute an easily computed approximation. As discussed above, for curves specific to each level of D and n, this is not generally possible. However, for the aggregate curves at each level of D, the Gaussian approximation affords a convenient and accurate approximation. As mentioned, for a given level of D the distribution of complexity is approximately Gaussian (normal). Hence an easy way to perform a significance test is via a z-test, using the mean and variance of the null distribution as estimated by Eq. (1) (recall that because the distribution is approximately Poisson, the mean and variance are approximately equal).

Step-by-step recipe. Putting this all together, here's how to conduct a quick-and-dirty significance test for simplicity, using the methods given in any introductory statistics book for conducting a z-test.

1. Compute the complexity C of your sample.[5]
2. Compute the mean μ_D of the distribution of complexity at the given number of features D, using Eqs. (1)–(3).
3. Compute the z-score of your sample's complexity C via

$$z = \frac{C - \mu_D}{\sqrt{\mu_D}}. \tag{4}$$

4. Finally, evaluate the significance (one-tailed) of this z by looking at a table of z-scores in the back of any statistics book.

If the test is significant, we can reject the null hypothesis that the pattern is random.

4.1. Examples

As an example, consider the simple Bongard problem from Fig. 1a. This was defined with four features and has eight objects, so we focus on the curve for $D = 4$, $n = 8$. First we compute the significance using the curve specific to this level of n, and then repeat the test with the "quick-and-dirty" version that uses the Gaussian approximation to the aggregate distribution.

[3] Here "our theory" is just that the observations weren't generated randomly, i.e. the negation of the null hypothesis. Negating this doesn't support any specific affirmative theory of what *did* generate the data; see below for more discussion of this point.

[4] It is intriguing that some of the earliest foundational work on the mathematical theory of complexity, that of Martin-Löf (1996), focused on the tendency of random strings to pass arbitrary "significance tests" for regularity.

[5] Unfortunately, there is no easy way to perform this step: see Feldman (2001) for all the gory details. An on-line version of the algorithm is under development.

The complexity score for the pattern was $C = 4$. Fig. 1 shows the area (shaded) under the $D = 4$, $n = 8$ complexity curve corresponding to patterns this simple or simpler: only 0.00901 of the curve. So our pattern is significant at the $p = 0.009$ level. Eureka!

In contrast, now consider the scrambled problem from Fig. 1b, whose complexity was $C = 20$. This is much larger than the mean of 12.8 for $D = 4$, $n = 8$ patterns, so we can immediately see that it is not simpler than a random concept. In fact its "significance" is 0.9538, meaning that it is more complex than over 95% of all comparable concepts. Definitely no Eureka.

The original, unscrambled Bongard problem (Fig. 1a) was pretty convincing. What would happen if one of the examples were simply deleted? Would one still believe the pattern with a bit less evidence? If so, how much less strongly? We can answer this question with another significance test. In the abridged problem, complexity C is now 7, rather than 4 as in the original problem—the pattern is no longer quite as "clean" as it was. This level of complexity is now only "marginally" significant at $p = 0.0936$. You might believe the pattern, but you couldn't publish it.

Now the quick-and-dirty z-test, which uses the expected distribution of complexity for $D = 4$ patterns generally. The predicted expected complexity at $D = 4$ is $\mu_4 = 0.36287 \times \exp[0.86686 \times 4]$ or about 11.63 literals, and the standard deviation σ_4 is the square root of this, or about 3.41. (A look at Fig. 4 confirms that this is about right as the peak of the $D = 4$ curve.) So our measured complexity of $C = 4$ is about $2.24(= (11.63 - 4)/\sqrt{11.63})$ standard deviations less than the mean. Consulting a table of z-scores, we see that this z's deviation from zero is significant at $p = 0.0125$. This matches reasonably well with the true value of $p = 0.009$ we computed above by integrating under the non-aggregate curve, corroborating the accuracy of the quick-and-dirty test. By either test, we can feel confident that the Bongard problem has been solved; the pattern is too simple to be an accident. We are ready to take it to the grand jury!

5. Applications

Now let's put a famous Boolean pattern to the test. The logic I'll use is similar to that used to solve a notorious problem in archaeology: the mystery of the supposedly collinear patterns of paleolithic standing stones in Britain. These patterns, called ley lines, had struck many as *too* collinear to be accidental, suggesting intentional planning on the part of ancient Britons. But were they, perhaps, mere accidental alignments? To answer this, you'd have to know just how many collinearities would be expected from a random distribution of points in the plane—that is, the null distribution. Unfortunately the statistics of spatial distributions, like the statistics of complexity, did not lend themselves directly to any preexisting mathematical models. The problem was solved by the statisticians Broadbent (1980) and Kendall and Kendall (1980), who provided the necessary mathematics by inventing an essentially new field of statistics, now called spatial statistics. The answer, by the way, is that the angles comprised in known ley lines appeared to be well within the null distribution, and thus probably just coincidences.

In a similar spirit, we ask here just how surprisingly simple is a famous Boolean pattern. We consider the pattern *Only eat oysters during months containing the letter R*.[6] We can code the 12 months using two Boolean features to represent *contains an R* and *safe to eat oysters*,[7] and three additional Boolean features as dummy variables to distinguish the months (note that there are $8 = 2^3$ months that contain R). The complexity of the resulting pattern is $C = 8$. At $D = 5$, the expected complexity is about $\mu_5 = 27.7$, so our observed complexity of 8 is about $3.74(= (27.7 - 8)/\sqrt{27.7})$ standard deviations below the mean, which is highly significant ($p = 0.000091$). Unusually simple—and thus unusually memorable, as simplicity leads directly to ease of memorization (Feldman, 2000b). At $p < 0.00001$ this is a memorable pattern indeed—which is perhaps why it has been remembered by countless generations of seafood lovers.

A more serious application concerns the evaluation of concepts widely used as target concepts in experiments in the concept learning literature. For historical reasons, certain specific concepts—particular combinations of Boolean features—have been used over and over by different researchers in their experiments (mostly to allow easy comparisons of data across methodologies, etc.). Much of the data has supported one broad class of models, exemplar models, and the result is now that many contemporary researchers in human concept learning feel that they provide the best overall account of human concept learning (Nosofsky & Johansen, 2000; Palmeri & Nosofsky, 2001).

Recently, though, Smith and Minda (2000) have alleged that the impression of a good fit between exemplar models and the data is actually an accident, substantially due to the somewhat arbitrary range of concepts that have been frequently tested. Specifically, they argue that many of the most-tested concepts are unusually complex and irregular. Such concepts are exactly the type on which one would expect exemplar models to do well. Exemplar models are based around the storage of particular examples rather than on the extraction of common tendencies or regularities, as in prototype-based theories. When faced with a complex or irregular concept, there is no point in trying to extract common tendencies or regularities from it: such concepts—by definition—*don't have* any common tendencies or regularities. (From the point of view of complexity theory, that's exactly what makes them complex—the lack of any orderly structure that could form the basis of a compression scheme.) With such concepts, exemplar models, which simply store examples *without* trying to extract regularities, represent a more useful strategy. Thus testing primarily random concepts would tend to artificially inflate exemplar models' fit to the data.

But is it really true that the concepts that have been widely tested are relatively complex or irregular ones? This is impossible to say unless you can quantify complexity—which the complexity measure C allows you to do. But furthermore, you also need to be able to quantify where the observed level of complexity stands in relation to the range of possible levels—that is, place the given complexity C in the distribution $p(C)$. This of course is exactly what we are trying to do in the current paper.

[6] Or is it *Never eat oysters during months containing the letter R*? I can never remember. But either way, the complexity is the same, since the two versions differ only by a negation.

[7] Here I assume that the dictum is in fact correct, and that eating oysters during R months is in fact safe! The question is whether this pattern is too simple to appear to be accident.

Let's take two examples that have been at the forefront of the pro-exemplar argument since its inception. Exemplar models were invented by Medin and Schaffer (1978), who based their original argument around several experiments involving particular fixed concepts, each defined by a particular combination of four Boolean features. The influence of this paper was such that more recent papers have often used the same concepts (either retesting them or modeling earlier data). Medin and Schaffer's Exps. 2 and 3 concept, for example, has been used by Medin, Altom, Edelson, and Freko (1982), Nosofsky et al. (1994), and Pavel, Gluck, and Henkle (1988). This concept, which has $D = 4$, $n = 5$, has algebraic complexity $C = 14$. But how unusual is this level exactly? Consulting the (non-aggregate) distribution for $D = 4$, $n = 5$, we see that this is about the 85th complexity percentile (i.e. the area under the curve at or to the left of $C = 14$ is 0.85 of the total area). Thus this concept is more complex and irregular than more than three-quarters of all comparable concepts. Medin and Schaffer (1978)'s Exp. 4 concept ($D = 4$, $n = 6$), also front-and-center in the case for exemplar models, has complexity $C = 12$, which puts it in the 58th percentile. Another often-used concept was used by Medin, Altom, and Murphy (1984), and then re-used by McKinley and Nosofsky (1993) and Pavel et al. (1988). This one (at $D = 4$, $n = 4$) has complexity 8, which is at the 45th percentile, i.e. very close to average for a random concept.

Hence Smith and Minda (2000)'s worry was well-founded. These concepts, around which much of the empirical case in favor of exemplar models is based, are drawn from very random territory: average or even well above-average in complexity for randomly generated concepts. Exemplar models may account well for how human learners handle them, but this says little about how they might handle less complex concepts, where more regular structure is available to be extracted. This is certainly not evidence *against* exemplar concepts, but it does argue, as Smith and Minda (2000) suggested, that the evidence is not as strong as has been claimed.

6. Trouble at the right tail?

One aspect of the shape of the null distribution is a bit paradoxical, or at least, puzzling: it's symmetrical. This means that just as "simple stories" are unlikely to occur by accident, so are *complex* stories—indeed, high-complexity patterns are just as rare as low-complexity patterns. But obviously simple stories are treated very differently from complex ones psychologically. What's up?

Just as a simple pattern can only happen by accident when many elements line up just so, an *extremely* incompressible pattern can only happen when all the elements line up just so as to *break* any possible incipient regular trend. Just as a very simple theory of a crime is possible only when all the cues line up, a very complex theory is only necessary when all the cues *conflict*—each, say, pointing to a different suspect—and this only happens by accident a small proportion of the time.

So why do we as observers tend to focus on simple theories and not on complex ones? Simple theories' low probability from random sources cannot be the *only* reason we prefer them, because that property is shared by complex theories. One answer is that simple theories ought to be assigned *higher prior probability* than complex ones. This is

a common proposal in the Bayesian literature (often called an *Occam factor;* see Duda, Hart, and Stork (2001) for an introduction). Indeed even human subjects in shape classification experiments assign highest probability to the most symmetrical and regular shapes (Feldman, 2000a).

Why assign simple theories higher priors? When criminals perpetrate crimes, the clues generally tend to cluster into a simple theory, if only we had all the clues and could discern the pattern. But there is no similar source for complex theories: generally, the evidence is not created by an elaborate collusion designed to render it maximally complex. This is an asymmetry between simple and complex theories. When a very simple pattern is seen, it has low likelihood under the null hypothesis, and high likelihood under some high-prior-probability simple model. But when a very complex pattern is seen, it too has low likelihood under the null, but there *isn't* any high-pior-probability hypothesis that would explain it. Hence it is interpreted as a random pattern, albeit an usually complex one.

The symmetry of the null complexity distribution is in stark contrast to the classical theory of Kolmogorov complexity (Li & Vitányi, 1997), which entails that as D increases, a larger and larger proportion of all patterns will have near-maximal complexity. In the standard picture, as the size of the patterns under consideration increases, a higher and higher fraction of the total number of patterns have complexity above any fixed threshold. Inevitably, this means that the distribution of complexity does not tail off at all on the righthand side, but rather climbs inexorably until it reaches an abrupt cliff at the maximum possible value (approximately equal to the size of the patterns considered). Why is our distribution $p(C)$ so different?

The key difference is that classical Kolmogorov complexity is a universal measure, meaning that the complexity value assigned to a pattern is the lowest value available in *any* description system or computer language. (This also entails that its actual value is uncomputable; see Schöning & Pruim, 1998.) By contrast here we are dealing with a *fixed* description language (the concept algebra); briefer descriptions in other languages don't count. With a universal measure, if a supposedly complex pattern were very rare, it could ipso facto be referred to nearly uniquely, and hence described very compactly. For example, at the extreme, the single most complex pattern could be uniquely identified by the very brief phrase "the single most complex pattern"—and thus paradoxically have low Kolmogorov complexity (cf. the "Berry paradox," due to Bertrand Russell; Chaitin, 1995). But a fixed language such as the concept algebra doesn't necessarily allow such a description, so its brevity doesn't automatically confer low complexity. An extreme case of this, mentioned above, is the parity function, which is unique, and thus has low probability; it has a long description in the concept algebra notwithstanding its uniqueness and consequent brief descriptions in other languages.

Another way of seeing the same point is that in the conventional picture, longer descriptions automatically apply to more objects, because they allow more objects to be uniquely identified; this yields the ever-rising right tail of the Kolmogorov null distribution. But again this doesn't apply to complexity as measured in any fixed language, where uniqueness of reference doesn't necessarily relate to brevity of description. The result of all this is that very rare classes of patterns automatically have low Kolmogorov complexity, but may well have high complexity in any fixed code, thus allowing the vanishing right tail (without any paradox).

Real computing organisms, of course, have fixed description languages (or, equivalently, must choose from a fixed and finite set of description languages), and don't have arbitrary access to alternative forms of representation; that's what makes complexity computable. Hence it's worth considering that the Kolmogorov conception of complexity, based on a universal code, misses an important part of the picture in subjective (computable) complexity measures, embodied by the vanishing right tail.

7. Occam meets Bayes

All this raises the question: are human concept learners significance testers? The question of whether even *experimental psychologists* ought to be significance testers has quietly become controversial among statisticians in recent years (Dixon, 1993; Loftus, 1991). Conventional diagnostic statistics, many statisticians now argue, is in fact missing a large part of the picture of inferential statistics—specifically, it explicitly considers the probabilistic consequences of only one hypothesis, the null hypothesis, while disregarding others, e.g. target hypotheses of interest to the scientists. A more complete picture, arguably, is provided by Bayesian theory (Jaynes, 1983). Anderson (1991) has suggested that human categorization might follow Bayesian principles (viz. optimal use of available information), a proposal recently expanded and refined by Tenenbaum (1999).

But in order for the story advanced in this paper to bear Bayesian fruit, the null distribution—complexity due to random processes—must be partnered with other distributions—namely, of complexity due to *regular* processes. Technically, these would be class-conditional likelihood functions giving probability as a function of complexity for some number of regular classes or data sources. There are various ways of constructing such distributions, depending on exactly how one defines "regular," and as a result these distributions do not have the uniquely well-defined status possessed by the null. Understanding these distributions is of fundamental importance, because unlike the null distribution developed above, they reflect the observer's *affirmative* model of patterns extant in the environment. I will not further develop this point in the current paper, but simply put in a call for progress in this direction.

Hence the distribution of complexity under the null hypothesis is not the whole story: we also need to look at the distribution of complexity under particular regular hypotheses. The major point is that the subjective probability of categorical hypotheses is strongly modulated by their complexity, at least in human minds, as attested by the behavioral data (Feldman, 2000b). The null hypothesis is only one among many hypotheses that needs a probability assigned to it, but it's a particularly important one—in Polya's (1954)'s phrase, "the ever-present rival conjecture." Hence the larger significance of the null distribution developed in this paper is probably its role in a more complete theory of inference based on both simplicity and Bayesian principles.

8. Intuitions of randomness

In considering the distribution of complexity in randomly generated patterns, it must be remarked somewhere that *subjective* expectations about random processes are notoriously

inaccurate (see Falk & Konold, 1997). In the most famous illustration of this, Gilovich, Vallone, and Tversky (1985) showed that subjects expect random runs of Bernoulli trials (e.g. basketball shots or coin flips) to exhibit fewer long runs and more alternations than they really do, the so-called hot-hand illusion.[8] Recently Griffiths and Tenenbaum (2001) have shown that human judgments of randomness can, in fact, be well-modeled if you make some assumptions about their subjective probabilistic expectations about patterns produced by *regular* processes, as mentioned above. This debate is not strictly relevant here, because we are talking about categories of objects in Boolean space, rather than sequential patterns per se: the math is not directly applicable. The point remains however that just because complexity C is distributed as some function $p(C)$ doesn't necessarily entail that human intuitions will correctly recognize this. The distribution developed in the current paper rather plays the role played by the binomial distribution in the hot-hand literature: the *objective* distribution, albeit in this case of a subjective measure, complexity. It is impossible to begin an empirical investigation into human intuitions on this point, though, until the objective distribution is established, which is the goal of the current paper.

9. Conclusion

Grand juries hate coincidences—and they are right to. Coincidences mean unsatisfying complexities, unexplained anomalies, and awkward exceptions that mar an otherwise elegant and coherent solution. When you've really solved the puzzle, all the clues ought to fall into place in a simple, coherent story. And when the story is simple, intuition says its unlikely to be an accident. The contribution of this paper is to quantify exactly *how* unlikely, using a psychologically realistic measurement of simplicity.

One of the major advances in the theory of induction in the last few decades has been a historical reconciliation of the ideas of *simplicity* and *truth*—the idea that the simplest hypothesis can be shown, under various assumptions, to be the most likely to be correct. Thus Rissanen (1978) has shown that under broad assumptions the least complex (minimum description length or MDL) hypothesis has the highest Bayesian posterior, and Chater (1996) that the simplest or most "Pragnänt" interpretation of a visual scene is likely the most veridical. The null distribution of complexity presented here takes a step towards a similar reconciliation in the realm of concept learning.

Acknowledgements

I am grateful to Tom Griffiths, Josh Tenenbaum, and two anonymous reviewers for helpful comments. Preparation of this manuscript was supported by NSF SBR-9875175.

[8] Note though that some researchers (e.g. Kubovy & Gilden, 1991) have disputed their explanation for this phenomenon, while others (e.g. Kareev, 1992) have argued that human judgments are as reasonable as can be expected given short-term memory limitations.

Appendix A. Algebraic complexity

This appendix gives a precis of an algebraic theory of concepts, explained more fully by Feldman (2001).

We begin with a set of observations or objects $\mathbf{x} = x_1, x_2, \ldots x_n$ defined over D features $\sigma_1 \ldots \sigma_D$; each object is defined by a certain conjunction of D literals (positive or negative variable symbols). Our goal is to extract from these observations the set of *regularities* ϕ that occur in it, where by "regularity" we mean a lawful relation that all n objects satisfy.

We wish to consider regularities of the form "if object x satisfies [some proposition], then x will have some property σ;" such rules have the form of a "causal law." Also, we wish to distinguish between regularities that involve small numbers of features from those that involve larger numbers of features, because the former entail more simple patterns, while the latter entail more complex ones. Thus we define each regularity ϕ^K as a formula of the form

$$\sigma_1 \sigma_2 \ldots \sigma_K \rightarrow \sigma_0, \tag{A1}$$

where we have arbitrarily renumbered the regularities in order to show that there are K of them on the left side of the implication and one of them on the right. The number K, giving the number of literals in the antecedent of the "law," is called the *degree* of the regularity. Thus σ (meaning, "the object has property σ") has degree $K = 0$ (there is no antecedent). A $K = 1$-degree regularity has the form $\sigma_1 \rightarrow \sigma_2$, meaning "if the object has property σ_1 then it has property σ_2." Again if a regularity ϕ appears in the description of $\mathbf{x}$ it means that all n objects obey it.

Now consider the set $\hat{S}(\mathbf{x})$ of all regularities satisfied by a given observation set $\mathbf{x}$. In its raw form, this representation of the objects is highly redundant, because some regularities are automatically entailed by others in the set. However, given certain notational conventions, there is a unique minimal set of regularities that contains only those that are essential in describing $\mathbf{x}$, which is denoted $S(\mathbf{x})$ and called the *power series expansion of* $\mathbf{x}$. $S(\mathbf{x})$ is in a sense the smallest set of regularities whose transitive closure is $\hat{S}(\mathbf{x})$, that is, that fully describes the original dataset.

The power series of $\mathbf{x}$ is so named because, like a Fourier or Taylor series, it decomposes the structure of $\mathbf{x}$ into components of differing degrees of complexity. All object sets can be fully expressed by regularities of maximal degree $K = D - 1$ (entailed by the so-called *representation theorem*), in the same way that all periodic signals may be expressed by sums of sine and cosine components at various frequencies (the Fourier decomposition), or an analytic function can be expressed by a weighted sum of its derivatives at various degrees (the Taylor series). A power series may or may not contain regularities of lower than maximal degree; it does so only if contains some orderly structure.

The function $|\phi|_{\mathbf{x}}(K)$ giving the *number* of regularities at each degree K contained in $S(\mathbf{x})$ is called the *power spectrum* of $\mathbf{x}$ (see Fig. 2). This function, giving the number of regularities of degree K that are necessary in a *minimal* description of $\mathbf{x}$, provides a very useful summary of $\mathbf{x}$'s regularity content. Because regularities at degree $K = 0$ are the very simplest in form, a power series that contains many of them—i.e. has much power at $K = 0$—has a relatively large amount of its structure explained by extremely simple rules.

Similarly, series with power at $K = 1$ have structure that can be accounted for by implicational rules; and so forth for higher degrees. If $S(\mathbf{x})$ contains power primarily at higher degrees it means that $\mathbf{x}$ is intrinsically complex in form; its structure can only be explained by positing relatively complex rules.

A useful numeric measure of the overall complexity of $\mathbf{x}$ is its total spectral power weighted by degree plus one $(K + 1)$,

$$C(\mathbf{x}) = \sum_{K=0}^{D-1} (K+1)|\phi|_{\mathbf{x}}(K), \tag{A2}$$

which gives the average amount of spectral power contained in $\mathbf{x}$'s spectrum. This is the measure of complexity used in the current paper.

$C(\mathbf{x})$ has a very direct interpretation as the total length in literals of the minimal representation $S(\mathbf{x})$, because each regularity of degree K contributes exactly $K + 1$ literals to the total expression. As mentioned in the text, on a wide range of concepts tested by Feldman (2000b), this number $C(\mathbf{x})$ gives very good quantitative prediction of human subjects' performance in a learning task (Fig. 3) accounting for more than half the variance in performance (Feldman, 2001), well over twice as much as competing learning models (e.g. ALCOVE, Kruschke, 1992). This result inspires the main premise of the current paper, that $C(\mathbf{x})$ makes a good model of subjective complexity.

Finally, note that a very useful consequence of partitioning the structure of $\mathbf{x}$ into bins by degree is it allows one to *truncate* the series, that is, to represent $\mathbf{x}$ using only terms below some fixed degree, discarding higher-order terms. When higher order terms are truncated, in general the observations $\mathbf{x}$ can no longer be represented precisely; the representation has been "simplified," bringing benefits for generalization. (Roughly, higher order terms are more liable to represent "noise" or accidental structure in $\mathbf{x}$, so discarding them avoids overfitting.) For example truncating at $K \leq 1$ (the so-called *linear terms*) amounts to representing the observations only in terms of their constant and implicational structure, and ignoring any regularities that require three or more symbols to express. There is good evidence (again see Feldman, 2001) that human learners adopt something like this linear truncation, termed a "bias towards linearity."

Appendix B. Elements of the null distribution of complexity

This appendix discusses some internal elements of the distribution $p(C)$. For reasons discussed below, the model given is approximate rather than analytic. However, several aspects of the internal workings of the distribution are interesting, and moreover shed light on why the distribution is approximately Poisson; thus I delve more deeply into the distribution here for the benefit of interested readers.

We seek to model the distribution $p(C[\mathbf{x}])$ of the parameter C when the object set $\mathbf{x} = x_1, x_2, \ldots, x_n$, defined over D Boolean features, is chosen at random. By "random" we mean that each of the 2^D possible objects is included in $\mathbf{x}$ with probability 1/2. This means that the number $n = |\mathbf{x}|$ of objects will itself be distributed binomially with mean

$2^D/2 = 2^{D-1}$. As explained in Appendix A and Feldman (2001), complexity C is the total number of literals in the minimal power series $S(\mathbf{x})$.

The situation can be broken down by degree K. For each value of K, there are a certain total number $N_{D,K}$ of regularities. To visualize how these can be counted, note that each regularity is defined by deleting (prohibiting) one cell from a column, side, or "hyper-side" (depending on K) of the Boolean D-cube. (The regularity asserts that observed objects do not fall in this cell.) For example, with $D = 3$, a $K = 1$ regularity prohibits one of the four corners of one two-dimensional side of the three-dimensional Boolean cube. (e.g. the $K = 1$ regularity $\sigma_1 \rightarrow \sigma_2$ inhabits the $\langle \sigma_1, \sigma_2 \rangle$ plane of the 3-cube, and prohibits the single cell $\sigma_1 \bar{\sigma}_2$.) In general, a hyperside is defined by choosing $K + 1$ of the D features (K for the features on the left side of Eq. (A1), one for the single feature on the right side). Given a hyperside, a regularity is defined by selecting one of its 2^{K+1} vertices to be prohibited. Hence we see that the total number of regularities at a fixed level of D and K is given by

$$N_{D,K} = \binom{D}{K+1} 2^{K+1}, \tag{B1}$$

where the notation $\binom{D}{K+1}$ denotes the binomial coefficient $D!/[(D - K - 1)!(K + 1)!]$, which gives the number of ways D objects can be chosen $K + 1$ at a time.

Each regularity has a certain probability $p_{D,K,n}$ of randomly occurring in (being satisfied by) a randomly selected object set of size $n = |\mathbf{x}|$; this probability is given by

$$p_{D,K,n} = \prod_{i=0}^{n-1} \left(\frac{2^D - 2^{D-K-1} - i}{2^D - i} \right), \tag{B2}$$

which after algebraic manipulation is equivalent to the closed-form expression

$$p_{D,K,n} = \frac{(2^D - 2^{D-K-1})!(2^D - n)!}{(2^D - 2^{D-K-1} - n)!(2^D)!}.$$

(To see where this formula comes from, consider that the product in Eq. (B2) multiplies together a series of n fractions, the numerators of which are the n integers running downwards from $2^D - 2^{D-K-1}$, and the denominators of which are the n integers running downwards from 2^D.)

The complex form of Eq. (B2) is due to the fact we are sampling objects without replacement; that is, once we have evaluated a single object with respect to the regularity, we move on to consider the remaining objects, which are reduced by one in number (this reduction is indexed by i in Eq. (B2)). At each step, we ask whether the observed object falls within the region that satisfies the regularity. Of the $2^D - i$ objects remaining, 2^{D-K-i} are prohibited by the regularity, and the rest are allowed, which yields the ratio given in the formula.

The number of regularities of degree K in the minimal power series of a random object set is just the number of successes among $N_{D,K}$ approximately independent Bernoulli trials with success probability $p_{D,K,n}$ and failure probability $1 - p_{D,K,n}$, and thus will follow a binomial distribution. Each regularity of degree K that occurs in the final representation (minimal power series) contributes $K + 1$ literals to the total complexity. Hence

the complexity distribution we seek is that of the random variable formed by the sum of a number of independent binomially distributed random variables. By the central limit theorem, such a distribution will tend to normality as the number of contributing distributions grows large. This number grows exponentially with D. (At each fixed value of D, there is one contributing distribution per value of K and n. There are D possible values of K, and 2^D possible values of n, for a total of $D2^D$ distributions all together.) Hence we can expect the distribution to grow more normal rapidly as D increases—exactly as can be seen from Fig. 4, in which the plotted curves for $D = 5$ and 6 deviate from normality by (respectively) less than one part in 200 and less than one part in 400.

In practice this approach to calculating the null distribution has several drawbacks. First, it is a very complicated summation, because as mentioned the number of binomial distributions we are summing up grows exponentially with D. Second, there are several imperfect assumptions in the above that are difficult to correct for. One is that different regularities (our Bernoulli trials) are not actually all independent. Second, and more importantly, many regularities that occur by chance will not appear in the minimal power series, because they are logically entailed by other regularities of smaller K. Thus the above substantially overestimates the size of the actual power series. Hence the above description is definitely *not* exactly correct.

Fortunately, there is a better way of getting a good approximation for $p(C)$. The probability $p_{D,K,n}$ here is usually small: any one regularity has little chance of being satisfied by a randomly chosen object set (for example with $D = 4$, $n = 8$, and $K = 0$ and 1, $p_{D,K,n}$ will be respectively 1/12, 870 and 0.0385). In this situation the distribution will tend to be Poisson (see Wickens, 1982): that is, approximately normal with equal mean and variance.[9] Again, this is visibly confirmed by the pattern of distributions in Fig. 4, in which the variance of the distributions scales up almost exactly linearly with the mean of the distribution (again, $r = 0.9994$ in our sample). This single parameter is usually called λ in the context of a Poisson distribution, which has general form

$$p(X) = \frac{e^{-\lambda}\lambda^X}{X!}. \tag{B3}$$

Hence our task is reduced to estimating this single parameter $\lambda = \mu_D = \sigma_D^2$, which can be thought as the expected (mean) number of literals in a typical object set containing a *typical* number n of objects at a *typical* degree K. Because we are assuming random objects generated with each binary feature having probability 1/2, the expected number of objects is 2^{D-1} (that is, half the maximum of 2^D); this is the "typical" value of n. Similarly, the "expected" value of K grows with D (recall that K ranges from 0 to $D - 1$). Thus the total number $N_{D,K}$ of regularities (Eq. (B1)), which is exponential in K, is also exponential in D. Hence we have good reason to expect the expected mean μ_D (which is proportional to $N_{D,K}$) to grow exponentially with D, as in Eq. (1), with exponential growth parameters α and β. It then remains only to estimate values of α and β from the known values of μ_D

[9] Even without considering the Poisson approximation, the approximate equality of the mean and variance follows from the fact that the probability of success $p_{D,K,n}$ is close to zero. The number of successes in N trials of probability p is approximately normal with mean Np and variance $Np(1 - p)$. When p is approximately zero, $1 - p$ is approximately 1, and Np is approximately the same as $Np(1 - p)$.

(2.00, 4.91, 12.11, 28.01, 63.87 for $D = 2, 3, 4, 5, 6$ respectively, plotted in Fig. 6), as is done in the text. This method has the advantage of avoiding all the modeling subtleties mentioned above, as these numbers reflect the true values of C, and hence μ_D, after all relevant subleties have been taken into account.

The predicted value of λ given by Eq. (1) can now be used to execute the quick-and-dirty z-test for pattern randomness (Eq. (4)), using $\mu = \lambda$ and $\sigma = \sqrt{\mu} = \sqrt{\lambda}$ as parameters, as illustrated in the text.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Barlow, H. B. (1974). Inductive inference, coding, perception, and language. *Preception*, *3*, 123–134.

Bongard, M. (1970). *Pattern recognition*. New York: Spartan Books.

Broadbent, S. (1980). Simulating the ley hunter. *Journal of the Royal Statistical Society A*, *143*(2), 109–140.

Chaitin, G. J. (1995). The Berry paradox. *Complexity*, *1*(1), 26–30.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566–581.

Dixon, P. (1993). Why scientists value *p* values. *Psychonomic Bulletin and Review*, *5*(3), 390–396.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301–318.

Feldman, J. (2000a). Bias toward regular form in mental shape spaces. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 1–14.

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

Feldman, J (2001). *An algebra of human concept learning*. Under review.

Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology*, *47*(1), 98–112.

Feldman, J (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, 12(6), 227–232.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314.

Givone, D. D. (1970). *Introduction to switching circuit theory*. New York: McGraw Hill.

Griffiths, T. L., & Tenenbaum, J. B. (2001). Randomness and coincidences: Reconciling intuition and probability theory. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 370–375.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural "goodness". *Journal of Experimental Psychology*, *46*, 361–364.

Jaynes, E. T. (1983). Confidence intervals vs Bayesian intervals. In R. D. Rosenkrant (Ed.), *E. T. Jaynes: Papers on probability, statistics and statistical physics* (pp. 757–804). Dordrecht: Reidel.

Kanizsa, G. (1979). *Organization in vision: Essays on Gestalt perception*. New York: Praeger.

Kareev, Y. (1992). Not that bad after all: Generation of random sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 1189–1194.

Kendall, D. G., & Kendall, W. S. (1980). Alignments in two-dimensional random sets of points. *Advances in Applied Probability*, *12*, 380–424.

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

Kubovy, M., & Gilden, D. (1991). Apparent randomness is not always the complement of apparent order. In G. Lockhead, & J. Pomerantz (Eds.), *The perception of structure: Essays in honor of Wendell Garner*. Washington, DC: MIT Press.

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

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, *36*(2), 102–105.

Martin-Löf, P. (1996). The definition of random sequences. *Information and Control*, *9*, 602–619.

McKinley, S. C., & Nosofsky, R. M (1993). *Attention learning in models of classification* (Cited in Nosofsky, Palmeri, and McKinley, 1994).

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 37–50.

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 333–352.

Medin, D. L., & Schaffer, M. M. (1978). Context model of classification learning. *Psychological Review*, *85*, 207–238.

Neisser, U., & Weene, P. (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology*, *64*(6), 640–645.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin and Review*, *7*(3), 375–402.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology*, *54A*(1), 197–235.

Pavel, M., Gluck, M. A., & Henkle, V. (1988). Generalization by humans and multi-layer networks. *Proceedings of the 10th Annual Conference of the Cognitive Science Society*.

Polya, G. (1954). *Mathematics and plausible reasoning*. Princeton, NJ: Princeton University Press.

Quine, W. (1965). On simple theories of a complex world. In M. H. Foster, & M. L. Martin (Eds.), *Probability, confirmation, and simplicity: Readings in the philosophy of inductive logic*. New York: Odyssey.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Schöning, U., & Pruim, R. (1998). *Gems of theoretical computer science*. Berlin: Springer.

Shepard, R., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(1), 3–27.

Sober, E. (1975). *Simplicity*. London: Oxford University Press.

Tenenbaum, J (1999). *A Bayesian framework for concept learning*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Wegener, I. (1987). *The complexity of Boolean functions*. Chichester: Wiley.

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology*. San Francisco: Freeman.