Journal of Experimental Psychology: General
1998, Vol. 127, No.1, 3-21

0096-3445/98/$3.00

Individuation, Counting, and Statistical Inference: The Role of Frequency and Whole-Object Representations in Judgment Under Uncertainty

Gary L. Brase, Leda Cosmides, and John Tooby
University of California, Santa Barbara

Evolutionary approaches to judgment under uncertainty have led to new data showing that untutored subject reliably produce judgments that conform to may principles of probability theory when (a) they are asked to compute a frequency instead of the probability of a single event, and (b) the relevant information is expressed as frequencies. But are the frequency-computation systems implicated in these experiments better at operating over some kinds of input than others? Principles of object perception and principles of adaptive design led us to propose the *individuation hypothesis*: that these systems are designed to produce well-calibrated statistical inferences when they operate over representations of "whole" objects, events, and locations. In a series of experiments on Bayesian reasoning, we show that human performance can be systematically improved or degraded by varying whether a correct solution requires one to compute hit and false-alarm rates over "natural" units, such as whole objects, as opposed to inseparable aspects, views, and other parsings that violate evolved principles of object construal.

The ability to make well-calibrated probability judgments depends, at a very basic level, on the ability to count. The ability to count depends on the ability to individuate the world: to see it as composed of discrete entities. Research on how people individuate the world is, therefore, relevant to understanding the statistical inference mechanisms that govern how people make judgments under uncertainty.

Computational machinery whose architecture is designed to parse the world and make inferences about it is under intensive study in many branches of psychology: perception, psychophysics, cognitive development, cognitive neuroscience, evolutionary psychology, and others (for review, see Hirschfeld & Gelman, 1994). In consequence, well-articulated internal principles for individuating rigid objects and modeling their interactions are beginning to emerge (e.g., Leslie, 1988, 1994; Shepard, 1984; Spelke, 1988, 1990; Talmy, 1988). We will argue that these principles can illuminate the study of judgment under uncertainty, and report the results of a series of experiments that were designed to test some of the predictions derived.

Gary L. Brase, Department of Psychology, University of California, Santa Barbara; Leda Cosmides, Department of Psychology and Center for Evolutionary Psychology, University of California, Santa Barbara; John Tooby, Department of Anthropology and Center for Evolutionary Psychology, University of California, Santa Barbara.

The preparation of this article was supported in part by generous grants to John Tooby from the James S. McDonnell Foundation and the National Science Foundation Grant BNS9157-449.

We would like to warmly thank Roger Shepard, whose pioneering work on how the mind represents the world inspired many of the ideas in this article, and Gerd Gigerenzer, whose intellectual influence can be seen throughout. We would also like to thank Clark Barrett, Sandy Brase, Lorraine Daston, Randy Gallistel, Gernot Kleiter, Alan Leslie, Catrin Rode, Steve Pinker, Ilavenil Subbiah, and Amos Tversky, who provided enlightening insights, and Tenesa Garrison and Betsy Jackson for their assistance in conducting the studies. We are also deeply indebted to the many members of the Center for Evolutionary Psychology's regular seminar, with whom many of these ideas were discussed

Correspondence concerning this article should be addressed to Leda Cosmides, Department of Psychology, University of California, Santa Barbara, California 93106. Electronic mail may be sent via Internet to tooby@alishaw.ucsb.edu.

Biased humans versus Bayesian bumblebees: A paradox in the study of judgment under uncertainty

Disputes about whether human beings are "rational" have been erupting in philosophy for over two millennia. Recently, psychologists have joined the debate, with the view that this ancient argument could be settled by empirical evidence. Given clear-cut criteria for what counts as rational thinking, they argued, the question could be decided by assessing human performance on inductive and deductive reasoning tasks. A vigorous research program ensued, especially within cognitive psychology. Its premise: human thought processes are rational to the extent that they produce answers that conform to the strictures of normative theories drawn from mathematics, probability theory, or logic.

Psychologists studying judgment under uncertainty in humans have used this standard for more than two decades. Until recently, however, the results of this research program have provided little comfort for defenders of human rationality. Sound decisions often depend on the ability to estimate the probability of uncertain events with some accuracy. So one might expect a "rational" mind to be equipped with computational mechanisms that embody normative principles drawn from probability theory, and to routinely apply these to problems that require statistical inference. In study

after study, however, such tasks elicited performance that seemed to violate a variety of these principles (e.g., Kahneman, Slovic, & Tversky, 1982).

Such findings led many psychologists to conclude that human reasoning faculties are riddled with crippling defects: heuristics, biases, and fallacious principles that violate canons of rationality derived from logic, mathematics, and philosophy. Leading researchers began to argue that the human cognitive architecture has "mental limitations" that prevent people from applying rational methods. People apply heuristics instead because these "reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations", and "make them tractable for the kind of mind that people happen to have" (Kahneman, Slovic, & Tversky, 1982, pp. xi-xii; Tversky & Kahneman, 1974, p. 1124).

During the same two decades, the study of animal behavior was heading in the opposite direction. Evolutionary biologists were exploring judgment under uncertainty in nonhuman animals in order to test various mathematical models from optimal foraging theory (e.g., Stephens & Krebs, 1986). Their experiments were showing that animals with truly minuscule nervous systems, such as bumblebees, make judgments under uncertainty during foraging that manifest exactly the kind of well-calibrated statistical induction that the human brain was widely thought of as "too limited" to perform (e.g., Gallistel, 1990; Real, 1991; Real & Caraco, 1986; Staddon, 1988). It was beginning to look like Homo sapiens did not deserve to be called "the rational animal", but that bumblebees, birds, and other organisms did. What was happening?

Bumblebees appeared rational when humans did not because they were tested under ecologically valid conditions (Tooby & Cosmides, 1992, in press). When one does the same for human subjects, they too perform like good intuitive statisticians.

It turns out that human performance in probabilistic reasoning tasks is remarkably sensitive to the format in which information is presented and answers asked for. Most experiments that elicited "non-normative" performance asked subjects to judge the probability of a single event (e.g., "What is the chance that a person who tests positive for the disease actually has it?"). However, many purported biases and fallacies disappear when people are asked to judge a frequency instead (e.g., "How many people who test positive for the disease will actually have it?"). For example, on a Bayesian problem about medical diagnosis, this trivial change in the wording of the final question boosted performance on otherwise identical problems from 36% correct to 64% correct (Cosmides & Tooby, 1996). By presenting still other elements in the problem as frequencies, Cosmides and Tooby were able to push performance even higher: to 76% correct in a purely verbal version of the problem and to 92% correct in a version where the subject had to actively construct a visual representation of the relevant frequencies. The same problems elicited base rate neglect when subjects were given the likelihood information as a proportion (e.g., "5% of healthy people test positive"), and asked to judge the probability of a single event. With a frequency format, subjects not only used the base rate information, but they used it fully, producing answers that conform to the strictures of Bayes's rule. Gigerenzer and Hoffrage (1995) have obtained similar results on other Bayesian problems. The conjunction fallacy can also be eliminated by presenting problem information as frequencies and asking for answers as frequencies (Fiedler, 1988; Tversky & Kahneman, 1983). So can the overconfidence bias (Gigerenzer, Hoffrage, & Kleinbolting, 199 1). (For review, see Cosmides & Tooby, 1996; Gigerenzer, 1991.)

These results suggest that humans, like other animals, have inductive reasoning mechanisms that embody certain rational principles, but that the design of these mechanisms requires representations of event frequencies to operate properly (see also Christensen-Szalanski & Beach, 1982). Prior experiments did not reveal their existence, because these mechanisms cannot "read" input in other formats. By analogy, a spreadsheet program will appear to lack algorithms for converting dollars to pounds, unless one enters the numbers in the proper format. Its inability to correctly compute pounds from a dollar amount that was formatted as a "time of day" is not a design defect, nor is it evidence that the spreadsheet lacks the necessary algorithms.

Shepard (1984, 1987) has argued that many long-enduring invariances in the world will, through natural selection, become instantiated in the mind, and that appreciating this will help psychologists predict and explain many otherwise puzzling features of how human minds represent the world. In this spirit, we shall first consider why a well-engineered statistical inference machine would be designed to operate on representations with a frequency format. Second, we shall see whether principles of object perception impose additional constraints on the class of representations that such mechanisms can take.

The connection between these two questions is as follows. A frequency computation system requires input from mechanisms that count which, in turn, require input from mechanisms that parse the world into countable entities. By considering invariant properties of rigid objects, Shepard and others have discovered that the ways in which people conceive of objects and model their interactions is governed by a rich set of interlocking principles, which Leslie has dubbed a "theory of bodies", or ToBy[1] (e.g., Baillergeon, 1986; Leslie, 1988, 1994; Shepard, 1984; Spelke, 1988, 1990; Talmy, 1988). One of ToBy's primary functions is to parse surfaces into representations of discrete objects. Surfaces that it does not individuate should not be available as input to a counting routine. ToBy's operations should

[1]ToBy can be thought of as a set of functionally integrated computational machines (Leslie, 1994), whose procedures embody "principles", at least from the point of view of the scientist studying them. These principles need not take the form of declarative or propositional representations. For example, the principle that objects move along paths that are spatially and temporally continuous is embodied in the mechanisms that give rise to apparent motion (Shepard, 1984), but we doubt that these mechanisms contain a propositional representation of this principle.

thereby constrain the set of problems for which a frequency-computation system can render accurate judgments. We report a series of experiments that explore this claim. Based on the results, we propose that the internal principles by which ToBy generates representations of objects and their parts can explain some otherwise puzzling features of judgment under uncertainty -- in particular, why some probability problems are easily solved whereas other, mathematiccally equivalent ones, are intractable.

Engineered in ancestral environments

> [In discussing sonar in bats] ... I shall begin by posing a problem that the living machine faces; then I shall consider possible solutions to the problem that a sensible engineer might consider; I shall finally come to the solution that nature has actually adopted (Richard Dawkins, 1986, pp. 21-22).

To understand why we consider frequency representations to be more "ecologically valid" than proportions and single-event Probabilities, one needs to consider the ecological situations -the selection pressures -- that caused statistical inference mechanisms to evolve in humans and other animals.

On evolutionary grounds, one expects a mesh between the structure of a biological machine and its function. This is because natural selection is a hill-climbing process, in which a design feature that solves an adaptive problem well can be outcompeted by a new design feature that solves it better. This process has produced exquisitely engineered biological machines -- the vertebrate eye, photosynthetic pigments, efficient foraging algorithms, color constancy systems – whose performance is unrivaled by any machine yet designed by humans.

Invariant (or statistically recurrent) features of ancestral environments tightly constrain evolutionary hypotheses. For any given species, an *adaptive problem* is defined as a problem (e.g., finding food, avoiding predators) that recurred over many generations *in the environments* in which that species evolved, and whose solution tended to promote the reproduction of an organism or its kin *in those environments*. For example, our color-constancy mechanisms are calibrated to natural changes in terrestrial illumination. As a result, grass looks green at both high noon and sunset, even though the spectral properties of the light it reflects have changed dramatically. The same mechanisms fail, however, in parking lots lit by sodium vapor lamps: evolutionarily novel devices that cast an unearthly spectrum (Shepard, 1992). This does not mean that the algorithms responsible for color constancy are "irrational", defective, or poorly designed. Biological machines work well under conditions that resemble the ancestral ones that shaped their design. They are calibrated to those environments, and they embody information about the stably recurring properties of these ancestral worlds.

When natural selection is the engineer, a computational machine will be designed to recognize information in the form in which it regularly presented itself in the environments in which our ancestors evolved. To understand the design of statistical inference mechanisms, then, one needs to examine what form inductive reasoning problems – and the information relevant to solving them – regularly took in ancestral environments.

Why frequencies?

The unobservability of single events

Asking for the probability of a single event seems unexceptionable in the modem world, where we are bombarded with numerically expressed statistical information, such as weather forecasts telling us there is a 60% chance of rain today. Against this background, it is easy to forget that our hominid ancestors did not have access to the modem system of socially organized data collection, error-checking, and information accumulation which has produced, for the first time in human history, reliable, numerically expressed statistical information about the world beyond individual experience. In ancestral environments, the only external database available from which to reason inductively was one's own observations and, possibly, those communicated by the handful of other individuals with whom one lived.

The "probability" of a single event cannot be observed by an individual, however. Single events either happen or they don't -- either it will rain today or it will not. Natural selection cannot build cognitive mechanisms designed to reason about, or receive as input, information in a format that did not regularly exist.[2]

An individual can, however, observe the frequency with which events occur, for example, that it rained on 6 out of the last 10 days with cold winds and dark clouds, or that we were successful 5 out of the last 20 times we hunted in the north canyon. Our hominid ancestors were immersed in a rich flow of observable frequencies that could be used to improve decision-making, given procedures that could take advantage of them. So if humans have adaptations for inductive reasoning, one might expect them to be good at picking up frequency information incidentally (Hasher & Zacks, 1979; Hintzman & Stern, 1978) and using it to make probability judgments.

Natural sampling as an environmental invariant.

Certain methods of acquiring information - e.g., certain sampling methods -- are common across species and time. Counting events as one encounters them is a widespread sampling scheme across species, for example. Kleiter (1994) and Aitchison & Dunsmore (1975) call this form of information acquisition *natural sampling* (in contrast with *designed experiments*). Kleiter has argued that natural sampling is an invariant that has come to be reflected in the design of evolved inference mechanisms. Furthermore, he has shown that when information is acquired through natural sampling, base rate information is redundant, and base-rate neglect is an optimal strategy (Kleiter, 1994).

---

[2] Sometimes evolved mechanisms *can* process data in a format that did not exist during a species evolutionary history, but when this happens, it is a byproduct of a design that was shaped by selection to do something else.

Consider the sampling scheme and data depicted in Figure 1 (inspired by Kleiter, 1994). You are walking through a forest, looking for fruit trees. The foliage is dense, so the color of fruit is sometimes visible before its shape can be discerned. You are wondering how often trees with red fruit (*R*) are apple trees (*A*)—i.e., *p*(A|R). As you pass each tree, you note whether or not it is an apple tree, and whether or not it has red fruit. So far you've seen 36 trees (i.e., taken one sample of *N* trees). Twelve were apple trees, 24 were trees of other species. There was red fruit on 9 of the apple trees, and on 4 of the other trees –that is, 13 of the trees you saw had red fruit. This means that, using red fruit as a cue, you had 9 hits (red fruit on apple trees) and 4 false alarms (red fruit on other trees). So *p*(A|R) = 9/13 – (number of hits)/(number of hits + number of false alarms). No base rate information was necessary for calculating the conditional probability: under conditions of natural sampling, all you need to know is the absolute number of hits and false alarms. Base rates are needed only in designed experiments: ones in which two samples are taken, and the size of each is *set in advance*. In designed experiments, likelihoods contain no information about base rates.

Kleiter (1994) has noted that the knowledge bases for most animals, and for ancestral hominids:

> \. . . only contained data acquired through their own experiences. Such a highly individual knowledge base is constrained by a specific structural feature: the total number of observations made decomposes top-down in a strictly additive way. [Perhaps humans] tend to ignore base rates because we are well-adjusted to natural sampling conditions. If we process the information of our episodic memory, which by definition is a highly personal knowledge base of our own experiences, we are rational without base rates. (p. 385)

This analysis applies to other animals as well as humans, and it suggests why many animals maintain representations of the absolute frequencies of biologically relevant events (Gallistel, 1990; Kleiter, 1994). It can also explain findings by Christensen-Szalanski and Beach (1982) on Bayesian diagnosis problems in which subjects sequentially experienced event frequencies. They were able to solve the problem correctly when they experienced frequencies of hits and false alarms in a natural sampling scheme but not when they experienced only the base rate as a frequency (having been given diagnostic information verbally, as a percent).

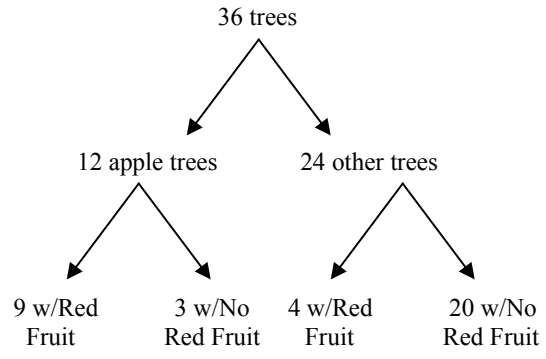

Figure 1. Natural sampling (inspired by Kleiter, 1994).

Storing probabilistic information in a frequency format

Once acquired, there are advantages to maintaining frequency representations. Important information is lost when an encountered frequency (e.g., "9 out of the last 13 trees with red fruit were apple trees") is converted to and stored as a proportion or single event probability ("there is a 69% chance that a tree with red fruit is an apple tree"). When this happens, the absolute frequencies of the two component events cannot be recovered. As a result, (a) it is difficult to update the database as one encounters new instances; (b) the sample size is lost, and with it a basis for indexing how reliable one's estimate is (300 observations provides a more reliable database than 3; indeed on Gallistel's (1990) model of classical conditioning, animals use sample size to compute the statistical uncertainty associated with an estimate of the rate of the unconditioned stimulus (US) in the presence of a conditioned stimulus CS), and their monotonically increasing learning curve reflects the decrease in this uncertainty as the sample size increases); (c) more data needs to be stored (including likelihoods and base rates); and (d) the original data cannot be recategorized to construct novel reference classes after the fact, as they become useful.[3] (These issues are discussed more fully in Cosmides & Tooby, 1996; Tooby & Cosmides, 1992, in press; and Gigerenzer & Hoffrage, 1995.)

Naturally, organisms need to make decisions about single events (e.g., should I take my umbrella *today*?). Forecasts about single instances can, however, be based on frequencies (e.g., in the past, how often did it rain on cloudless days like today?). Choices are generated by decision rules, and these can take frequency representations as input (as in signal-detection theory). Moreover, when fed into an appropriate decision rule, a frequency representation can easily produce a subjective degree of confidence, making me, for example, quite sure that I'll find an apple tree where l saw red fruit in the distance." In some frequency-based psychological models, *confidence levels* reflect an internal assessment of the reliability of the cues upon which one's judgment was based (Gigerenzer, Hoffrage, & Kleinbolting, 1991). (For these reasons, the fact that people routinely report experiencing subjective degrees of confidence that an event will occur does not weaken the claim that the machinery that underlies them operates along frequentist principles.)

Evidence that the human cognitive architecture includes a statistical inference system that requires data in frequency

[3] It is computationally trivial to reconstruct a reference class according to new criteria given frequency representations of events. The database will be versatile if individual instances, richly encoded by a broad band filter, are stored – as in episodic memory systems. If the human mind has a system designed to use episodic memory as a database for extracting frequencies associated with new reference classes, this system would need a procedure for using new cues to access the episodic memory base. Perhaps Tversky & Kahneman's (1974) "availability heuristic" reflects the operation of such a procedure. If so, then it would be more appropriately seen as a component of a well-designed system rather than a satisficing rule-of thumb.

formats in order to operate properly raises an obvious question: Frequencies of what?

Frequencies of What?

Individuation, Counting. and Natural Sampling.

Computing probabilities via natural sampling requires the ability to count, and the ability to count requires the ability to individuate events. In an extensive review of the animal literature, Gallistel (1990) has shown that a wide variety of animals are able to count; that is, to produce a one-one mapping from the numerosity of a set to behavior-controlling entities that, because of the formal properties of the computations into which they can enter, can be thought of as representations of number. Every organism that can learn via classical or operant conditioning can estimate rates – number per unit time – and use them to compute conditional probabilities in what amounts to a natural sampling scheme (Gallistel, 1990; Kleiter, 1994; Rescorla, 1967, 1968; Staddon, 1988). An association in space and time between, e.g., a tone and meat, is not sufficient to produce conditioning. The animal needs to be able to detect a contingency, such that the CS predicts a change in the rate of occurrence of the US: a dog will not salivate in response to a tone (CS) unless $p(\text{meat}|\text{tone}) > p(\text{meat}|\text{no tone})$. (See Gallistel, 1990; Rescorla, 1967, 1968; Staddon, 1988.)

These computations can be made because animals have mechanisms that parse the sensory stream, breaking it up into countable events: food pellets, lever presses, brief sounds, flashes of light. Conditioning in the laboratory is successful when the experimenter chooses stimuli and rewards that are adoptively important to the animal and can be individuated by it.

But what counts as a countable event differs from species to species. Some species parse the world in ways that others cannot. For example, on the basis of ultraviolet patterns, bees distinguish flowers that look identical to human eyes. Moreover, the cognitive architecture of a species need not contain machinery designed to count every class of events that it can parse. A bird who can assess the rate of return of berries on a bush may not be able to assess the frequency of cars or dog ears, even though it is capable of seeing them. For example, even though rats can compute the probability of electric shocks given the onset of a colored light and of nausea given the taste of a food, they cannot compute the probability of shocks given food or of nausea given lights (or if they do, the nofrmaitn is not used to regulate avoidance behavior; Garcia & Koelling, 1966). The female digger wasp, *Ammophila campestris*, distributes her larva among several different, burrows, which she provisions every day with paralyzed insects. Every morning she assesses the quantity of food in each burrow, and adjusts her provisioning accordingly -- this can be shown by experimentally altering the contents of her burrow. But these alterations affect her behavior only if they are done in the morning. Changes made at other times of day have no effect on her behavior (Baerends, 1941). So, although she can count paralyzed insects, the machinery that allows this is turned on only in the morning, " . . . almost as though it was a costly, power-consuming instrument" (Dawkins, 1986, p. 50). This counting mechanism is content-specific and situation specific: It can be applied only to insects (and possibly a few other items), and only in the morning.

Some of the mechanisms that assess quantities in humans are content-specific as well. Gallistel (1990) points out, for example, that the retina of a student who has never taken calculus computes the second derivative of the local distribution of light intensity (p. 332). In other words, the machine that carries out this computation can operate on representations of light, but not on representations with the content and/or format typically found in math books.

What can frequency computation systems count?

It is therefore reasonable to ask, To what extent are human inductive reasoning mechanisms content-specific? Can they operate on any kind of frequency information, or are there privileged contents? Does the design of our cognitive architecture favor certain ways of individuating objects and events over others?

Research from cognitive development suggests that it does. For example, a newborn's brain has response systems that "expect" faces to be present in the environment: babies less than 10 minutes old turn their eyes and head in response to face-like patterns, but not to scrambled versions of the same pattern with identical spatial frequencies (Johnson & Morton, 1991). Infants make strong ontological assumptions about how the world works and what kinds of things it contains -- even at 2 1/2 months (i.e., as soon as their visual systems have matured). They assume, for example, that it contains rigid objects that are continuous in space and time, and they have preferred ways of parsing the world into separate objects. Ignoring shape, color, and texture, they treat any surface that is cohesive, bounded, and moves as a unit as a single object (Spelke, 1988, 1990). When one solid object appears to pass through another, these infants are surprised (Baillargeon, 1986). Yet a system with no privileged hypotheses – a truly "open-minded" system – would be undisturbed by such displays. In watching objects interact, babies less than a year old distinguish causal events from non-causal ones that have similar spatio-temporal properties (Leslie, 1988); they distinguish objects that move only when acted upon from ones that are capable of self-generated motion (the inanimate-animate distinction; Leslie, 1994); they assume that the self-propelled movement of animate objects is caused by invisible internal states – goals and intentions -- whose presence must be inferred, since internal states cannot be seen (Baron-Cohen, 1995). When an adult utters a word-like sound while pointing to a novel object, toddlers assume the word refers to the whole object, rather than to one of its parts or the material it is made of or its superordinate category (Markman, 1989). Toddlers can count individual animals, but not kinds of animals; dots of different colors, but not the number of colors. They spontaneously count teddy bears, but not teddy bear ears: toddlers do not count the parts of intact objects. But if those same parts have been broken off of the parent object, and are

therefore capable of independent motion, toddlers will count them (Shipley & Shepperson, 1990).

Privileged parsings in statistical inference

The existence of sciences that investigate the properties of subatomic particles, gases, and waves shows that the human mind is capable of individuating an astonishing array of events. But some ways of parsing the world may be more natural than others -- faster, more automatic, and cross-culturally universal, emerging without conscious deliberation and in the absence of explicit instruction. Moreover, different adaptive problems might have required different ways of parsing the world. When you are sharing food with another person, it might make sense to think of your apple as having two halves; but when you are assessing the productiveity of apple trees while foraging, it is the frequency of apples that is relevant, not the frequency of half apples. To construct a 3-D representation, the visual system needs to know that two percepts are different views of the same apple tree; but we doubt there is any adaptive context in which the visual system would need to individuate and count up the number of different angles from which you have viewed that tree or the number of times you have viewed it from each of these angles.

Statistical inference mechanisms that embody content-independent rational principles are most useful when applied to adaptive problems whose solution requires recent samples of local information about rapidly-decaying reference class interrelationships -- changes in the spatial and temporal distributions of game, plant foods, predators, people, weather conditions, and so on. To do this, they should be designed to pick up incidental information about the frequencies of whole objects, actions, and events as a function of time, location, and the presence of other objects, actions, and events. By *whole object*, we mean cohesive, bounded entities that move as a unit, independent of other surfaces: the definition that human infants automatically apply (Spelke, 1988, 1990). The more closely a "part" of an object conforms to this definition, the easier it should be to count. For example, apples – which, though attached to the tree, can move somewhat independently of it – should be easier to count than 4 square-inch patches of bark on the tree's trunk. Furthermore, what gets classified as an individual object (e.g., tree versus apple) should shift meaningfully as a function of the adaptive problem at hand. (See Jackendoff, 1983, for a more general discussion of how the human mind individuates not only objects, but actions, events, locations, and paths. For evidence that infants can individuate and count actions and sounds, as well as objects, see Wynn, 1995; Starkey, Spelke, & Gelman, 1990.)

There are some aspects of the world that one would not expect such statistical inference mechanisms to spontaneously count. One example is the frequency of events that have remained stable over many generations (e.g., how often the sun rises in the east; how often the day-night cycle lasts 24 hours; how often solid objects fail to pass through one another). The probability of these events can be phylogenetically given (Shepard, 1987; Staddon, 1988). Nor should they count aspects of the world that are adaptively irrelevant – for example, the number of different colors in a scene or the number of sides of leaves on a tree. Indeed, inseparable aspects of objects, views of objects, or their orientations should be particularly difficult to count, because they are seldom individuated (e.g., top versus bottom side of a leaf, southwest face of a stone, right versus left ear of a cat), and when they are, their frequency often can be derived from information about the frequency of their parent object (oak leaves have two sides, cats have two ears). For similar reasons, one would not expect arbitrary chunks of intact objects -- chunks with no causal import -to be spontaneously counted: When looking at an intact fork, we see one fork, not six (or 10 or 100) fork parts. In fact, given the way our minds privilege "moves as a unit" as a dimension for construing objects, the less capable of independent motion an aspect of an object is, the more difficult it should be to individuate and count" (e.g., rock flanks should be more difficult than rabbit ears). In other words, the principles whereby ToBy parses the world should influence performance on any task that requires counting. This includes statistical inference tasks, if the frequency-natural sampling view discussed previously is correct.

Arbitrary parsings in statistical inference.

Given this perspective, consider the following problem (from Bar-Hillel & Falk, 1982):

> Three cards are in a hat. One is red on both sides (the red-red card). One is white on both sides (the white-white card). One is red on one side and white on the other (the red-white card). A single card is drawn randomly and tossed into the air.
>
> What is the probability that the red-red card was drawn, assuming that the drawn card lands with a red side up?

Most people answer "½" (66% in Bar-Hillel & Falk, 1982; 79% in Bar-Hillel, 1989). In doing so, they apparently reason as follows: "The card landed with a red side up, so it's not the white-white card. There are only two cards left, the red-white and the red-red. These are equally probable, so the probability that the drawn card is the red-red one must be ½" (see p.119 of Bar-Hillel & Falk, 1982). In this line of reasoning, the presence of a red side is used as a cue indicating which of several (whole) objects are potentially involved, and the subject computes the probability over whole objects (i.e., cards). The trouble is, these two cards – red-white and red-red – are not equally likely to land with a red side up.

To answer the 3 -card problem correctly, one has to count up sides of cards, rather than whole cards. We know the card in question landed with a red side up; the red-red card has two red sides, whereas the red-white card has only one red side. That means there are three red sides total, two of which

[4] All else equal. Other cues -- such as the sharp light-dark transition associated with many boundaries -- are highly correlated with surfaces that move as a unit. Ontogenetically, as these cues come to be used to distinguish objects, inseparable aspects of objects that happen to share them may become easier to individuate and count as well (e.g., polka dots on a piece of cloth).

are from the red-red card. Therefore, the answer is 2/3. Very few subjects give this answer: only 6% and 9%, respectively, in Bar-Hillel's studies. The 3-card problem appears to be intractable -- it reliably elicits the same wrong answer from most people. Why?

If our statistical inference mechanisms are designed to operate over frequencies of whole objects, then the 3-card problem is twice-cursed: (1) it is not posed in a frequency format, and (2) to solve it correctly, one needs to count views of objects -- sides of cards -- rather than whole objects. If whole objects are the natural unit of analysis for our statistical inference mechanisms, then people should have difficulty with any problem whose solution requires one to count arbitrary parsings of intact whole objects -- views, inseparable aspects, random chunks, nonfunctional fragments, and so on.

The Individuation Hypothesis

The experiments reported herein were designed to test the hypothesis that the computational structure of human mechanisms for assessing relative frequencies is better designed for operating over whole objects than arbitrary parsings of them. For convenience, we will refer to this as the *individuation hypothesis*. If it is correct, then the 3-card problem and others like it should remain difficult, even when they are posed in a frequency format. Furthermore, one should be able to systematically improve or depress performance on otherwise similar problems by varying whether a correct answer requires one to count over whole objects or arbitrary parsings of objects.

In the experiments that follow, the hit and false-alarm rates are not explicitly given: they must be inferred from what one knows about the structure of objects and their relationships, as in the 3-card problem. In that problem, the event of interest is whether the red-red card was drawn, and the presence of a red side is the *conditioning event:* the event upon which the probability of another event is conditioned. Bar-Hillel and Falk made the following observation:

> Outside the never-never land of textbooks, however, conditioning events are not handed out on silver platters. They have to be inferred, determined, extracted. In other words, real-life problems (or textbook problems purporting to describe real life) need to be modeled before they can be solved formally. (p. 121)

In this light, another way of expressing the individuation hypothesis is this: The human cognitive architecture privileges some parsings of the world over others. Inferring and extracting the appropriate conditioning event will be more difficult when this event represents an arbitrary parsing of the world than when it represents a more natural parsing.

Experiment 1

Experiment 1 was designed with three purposes in mind. First, we wanted to see if we could replicate Bar-Hillel's findings for the 3-card problem (Bar-Hillel, 1989; Bar-Hillel & Falk, 1982). Second, we wanted to see whether performance would improve if subjects were given a version of the problem in frequency format. Third, we wanted to explore our conjecture that arbitrary parsings impede performance. In the frequency version, subjects still have to assess the frequency of sides of cards. If the individuation hypothesis is correct, then posing this problem in frequency terms should produce little or no improvement. If it is incorrect, then the frequency version should elicit high levels of performance.

Method

Subjects. The subjects in all of our experiments were undergraduates at the University of California, Santa Barbara. Some participated to fulfill a class requirement, others were paid volunteers. Subjects were randomly assigned to conditions, and their average age was 19.1 years. Fifty-two participated in Experiment 1: 27 in Condition 1 and 25 in Condition 2.

Procedure. The procedure was identical in every experiment. Each subject was given a booklet that consisted of an instruction page followed by a word problem. They had no time limit, and most subjects completed the problem in less than 10 min. The instructions directed subjects to give the "typical" answer if they thought the answer to the word problem would vary over trials.

Materials for Condition 1. We used the same problem tested by Bar-Hillel and Falk (1982) and listed above. The only difference is that in our version, the colors were black and white rather than red and white.

Materials for Condition 2. This tested a frequency version of the 3-card problem. In the single event version, there is only one event in which a card is drawn. In the frequency version, this event occurred 30 times (sampling with replacement). The subject was then asked for the answer as a frequency, rather than as the probability of a single event. The text read as follows:

> Three cards are in a hat. One is black on both sides (the black-black card). One is white on both sides (the white-white card). One is black on one side and white on the other (the black-white card). Without looking, you draw a single card from the hat and toss it onto a table. Someone else records whether the card is black-black, black-white, or white-white. That person also records the color of the side which landed face up. Afler all this information is recorded for a card, the card is put back in the hat.
>
> The next day, you draw a card from the hat again, and the same information is recorded for that card. Each day you repeat this procedure, until you have drawn a card from the hat a total of 30 times.
>
> Of the cards which land on the table with a black side up, how many will be black-black cards? ___ out of ___ cards

Results

Any version of "⅔" was scored as a correct answer: "2 out of 3," "10 out of 15," "2/3," "66%," "2 to 1," and so on. The correct answer was elicited from 7% of subjects in the single event version, and 28% of subjects in the frequency version. The difference between these two conditions is significant (7% vs. 28%: $Z = 1.96$, $p = .025$), and the effect size, Φ, is .27.

Discussion

First, we were able to replicate Bar-Hillel's results quite precisely: the original, single event version of the 3-card problem elicited the correct answer from 7% subjects in

this study, versus 6% and 9% of subjects in the studies by Bar-Hillel (Bar-Hillel, 1989; Bar-Hillel & Falk, 1982).

Second, more subjects gave the correct answer in response to a frequency version of the 3card problem. This replicates, using different problem content, the results of other experiments showing that frequency versions of Bayesian problems elicit more correct answers than single event versions (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995).

Third, the degree of improvement – as measured either by the percent correct (28%) or the effect size (.27) – was modest. A minority of subjects answered the 3-card problem correctly, even when it had a frequency format: this is what one would expect if extracting the proper conditioning event is difficult for arbitrary parsings. In contrast, 76% of subjects responded with the correct answer on comparable versions of a medical diagnosis problem (effect sizes: .36 to .45; Cosmides & Tooby, 1996) and ~50% did on the more difficult problems tested by Gigerenzer & Hoffrage (1995).

This lower level of performance on the 3-card problem is not simply a function of differences in the subject populations tested. We found that the performance of subjects drawn from the same pool as those in Experiment 1 can be driven much higher, as will be shown in the experiments that follow.

To correctly answer the 3-card problem, one needs to compute the frequency of *sides* of cards. Little or no improvement on frequency versions of this problem is what one would expect if the frequency computation system were capable of counting *nothing but* whole objects. In contrast, a modest improvement (like the one found) is what one would expect if it is capable of counting other kinds of parsings, such as inseparable aspects or views of objects, but rarely receives this kind of input from a conceptual system that privileges whole objects. On this view, the conceptual system that generates inferences about the physical world – our "theory of bodies" – does not automatically individuate views and other inseparable aspects of objects. If they are not individuated, they won't be counted. The key act of insight in solving these problems is apprehending which aspects of the world need to be individuated and counted. Realizing that the conditioning event is the frequency of sides of cards should be difficult, however, for an architecture whose first and most important cut on the world is whole objects.

Experiment 2

Two factors are relevant to the individuation hypothesis: (a) whether a problem has a frequency or single-event format, and (b) whether the conditioning event is the frequency/presence of a whole object versus an inseparable aspect, view, or other arbitrary parsing of one. If the individuation hypothesis is correct, then we should be able to systematically change performance levels by creating new problems that vary in these dimensions.

Experiment 2 was designed to test these predictions. If the variables we are interested in are consequential, then we should be able to create a set of problems with new content that (a) elicits results parallel to those found for the 3-card problem, and (b) elicits enhanced performance on whole-object versions. So for Experiment 2, we created a new problem that uses candy canes (the straight kind) instead of cards. This allowed us to create a set of problems that are structurally isomorphic from the point of view of Bayes's rule, but which vary on the relevant dimensions. There were four problems, which fell into the following categories: frequency/whole object; frequency/arbitrary parsing; single event/whole object; and single event/arbitrary parsing. To create these categories, the problems had to differ a little in surface content; within this constraint, however, we tried to create problems whose surface content was as similar as possible.

These four problems allow us to conduct what is, in effect, two experiments: one testing the hypothesis that frequency formats enhance performance, and another testing the hypothesis that whole object parsings enhance performance. The same data can also reveal whether these two factors interact.

In these problems, the frequency or probability of the conditioning event is not given to the subject "on a silver platter", as Bar-Hillel & Falk (1982) put it. Hit rates and false-alarm rates must be derived by the subject from information in the problem. To do so, one needs a model of the problem space in which the correct events are individuated and their relationships to one another preserved. Inferring what these are and extracting their relationships is the crucial act of insight needed to arrive at a correct solution.

In the arbitrary parsing versions, whole candy canes of different colors/flavors played the role of whole cards; the ends of these candy canes played the role of sides of cards. Extracting the frequency of *ends* of intact candy canes, independent of the whole candy canes on which these ends exist, should be particularly difficult on the individuation hypothesis for a number of reasons:

1. A straight candy cane is a radially symmetrical object that is internally undifferentiated. This means it is difficult to think of it as having parts. The question, "How many parts does a candy cane have?" has no well-defined answer. It can be broken into pieces along a number of different axes, these pieces needn't be of equal length, and a single cane can be broken or ground up to yield an arbitrarily large number of pieces. So how many parts does it have? Two? Three? An infinite number, each corresponding to a distinct but infinitesimally small location? Ends are locations on a whole object that is difficult to think of as having parts.
2. Ends are inseparable aspects of candy canes. Unlike the handle on a mug (for example), they cannot exist independently of a whole object. This is because they are locations on the boundary of an object. When a small bit is broken off the "end" of a candy cane, the result is a candy cane that still has two ends, plus a left-over fragment.
3. Individuating boundaries (or areas on boundaries) of objects should be difficult for a cognitive system that privileges whole objects. It should be especially difficult for a system like ours, which uses boundaries to decide what counts as an individual object.

4. Even if ends are defined as the circular boundaries on a cylinder, there is nothing to differentiate the two ends of a candy cane from one another. They are symmetrical locations on a symmetrical object.
5. The boundary definition has other odd properties as well. Break a candy cane in half, and there will be four ends instead of two; break these in half again, and the same cane will have yielded eight ends, and so on. Moreover, the surface area of every daughter end will be equal to that of the parent ends. In contrast, breaking a mug does not yield an indefinitely large number of handles. At best, it will yield parts of handles, and the more one breaks it, the less "handle-like" the resulting pieces are.

In short, the "ends" of an intact candy cane are not the kind of surface that we would expect ToBy to (token) individuate.

Method

Subjects. There were 116 subjects in this experiment, with 29 in each of the four conditions.

Materials for Condition 1. This problem falls into the single-event/arbitrary parsing category. The event to be predicted is the presence of an all-pink peppermint stick; the conditioning event is the presence of a peppermint end. The text read as follows:

> At the grocery store, there are three jars of straight candy canes.
>
> The first jar contains only pink candy canes. These are peppermint.
>
> The second jar contains only yellow candy canes. These are lemon.
>
> The third jar contains candy canes that change flavor and color in the middle. One half of each stick is pink and tastes like peppermint; the other half of each stick is yellow and tastes like lemon.
>
> You just bought a large number of candy canes. You put three equal-sized handfuls of them into a bag. You took one handful from each of the three jars.
>
> When you got home, your friend closed her eyes, reached into the bag, and pulled out one candy cane. While still keeping her eyes closed, she tasted one end of it. The end she tasted was peppermint.
>
> What is the probability that the stick she tasted was one of the all-pink peppermint sticks? ____

Note that to solve this problem, one needs to realize that the probability of discovering a peppermint end on an all-pink peppermint stick (PP) is greater than the probability of discovering one on a two-toned stick (PL): p(P end|PP) = 1; p(P end|PL) = ½. The value to be computed is a posterior probability, p(PP|P end).

Materials for Condition 2. In this frequency/arbitrary parsing problem, the event to be predicted is the frequency of all-pink peppermint sticks, and the conditioning event is the frequency with which one encounters peppermint ends. The first five paragraphs are identical to those in condition 1; only the last three differ – they transform the prior problem into one with a frequency format.

> At the grocery store, there are thee jars of straight candy canes.
>
> The first jar contains only pink candy canes. These are peppermint.
>
> The second jar contains only yellow candy canes. These are lemon.
>
> The third jar contains candy canes that change flavor and color in the middle. One half of each stick is pink and tastes like peppermint; the other half of each stick is yellow and tastes like lemon.
>
> You just bought a large number of candy canes. You put three equal-sized handfuls of them into a bag. You took one handful from each of the three jars.
>
> When you got home, your friend reached into the bag and pulled out a big bunch of candy canes. She tasted one – and only one – end of each candy cane in that bunch.
>
> 30 of the ends she tasted were peppermint.
>
> How many of these do you expect were from all-pink peppermint sticks? ___ out of ___

Materials for Condition 3. This is a single event/whole object problem. The event to be predicted is which jar the drawn candy cane originated in; the conditioning event is the presence of a peppermint candy cane. Note that a peppermint candy cane is a whole object, and ajar is a whole object. We preserved as many elements of the problem as possible; the only changes in the text were ones necessary to transform the conditioning event from an arbitrarily parsed entity into a whole object (naturally, this required a change in the event to be predicted as well–in all four problems, however, this event was defined over whole objects.)

> At the grocery store, there are three jars of straight candy canes.
>
> The first jar contains only peppermint candy canes (the "Peppermint Only" jar).
>
> The second jar contains only lemon candy canes (the "Lemon Only" jar).
>
> The third jar contains peppermint candy canes and lemon candy canes, in equal proportions (the "Mixed" jar).
>
> You just bought a large number of candy canes. You put three equal-sized handfuls of them into a bag. You took one handful from each of the three jars.
>
> When you got home, your friend picked a peppermint candy cane out of your bag.
>
> What is the probability that the peppermint candy cane came from the "Peppermint Only" jar? ____

To solve this problem correctly, one needs to realize that a handful of sticks from the *peppermint only* jar has twice as many peppermint sticks as an equal sized handful from the *mixed* jar. Hence the probability that a peppermint stick (P stick) originated in the peppermint only jar (PO) is greater than the probability that it originated in the mixed jar. As always, the value to be computed is a posterior probability, in this case p(PO|P stick).

Materials for Condition 4. This frequency/whole object problem is identical to condition 3, except for the last two paragraphs, which transform the problem into a frequency format.

> At the grocery store, there are three jars of straight candy canes.
>
> The first jar contains only peppermint candy canes (the "Peppermint Only" jar).
>
> The second jar contains only lemon candy canes (the "Lemon Only" jar).
>
> The third jar contains peppermint candy canes and lemon candy canes, in equal proportions (the "Mixed" jar).
>
> You just bought a large number of candy canes. You put an equal-sized handfuls of them into a bag. You took one handful from each of the three jars.
>
> When you got home, your friend picked 30 peppermint candy canes out of your bag.
>
> How many of these peppermint candy canes do you expect came from the "Peppermint Only" jar? ___ out of ____

Results and Discussion

The percent correct for each condition is depicted in Figure 2.

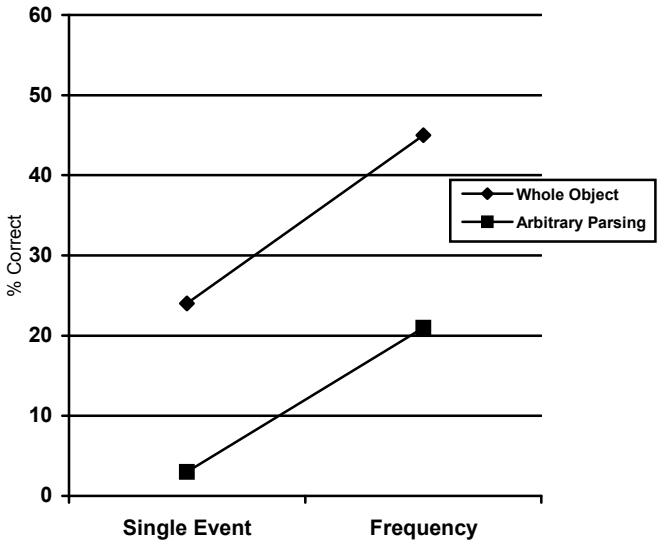

Figure 2. Whole-object versus arbitrary parsing problems: candy cane series.

Do frequency formats elicit higher levels of performance, all else equal? The arbitrary parsing problems tested in Conditions I and 2 were identical, except that one had a single event format and the other a frequency format. As predicted, the frequency version elicited higher levels of performance, and the difference was significant (3% vs. 21%: Z = 2.02, Φ = .26, p = .022).

The whole object problems tested in Conditions 3 and 4 were also identical, varying only in format. As predicted, the frequency version elicited higher levels of performance, and the difference was significant (24% vs. 45%: Z=1.66, Φ =.22, p =.049). In other words, as long as parsing was held constant, frequency versions elicited higher perfor-mance than single event ones. Moreover, the size of the frequency effect is similar for both parsings (whole-object parsing: .22; arbitrary parsing, .26).

We note that the frequency effect size for the arbitrary parsing condition (.26) is extremely similar to that found for the corresponding 3-card problems tested in Experiment 1 (.27). The absolute levels of performance were quite similar as well: 7% correct and 3% correct, respectively, for the single event versions; 28% correct and 24% correct for the frequency versions. The similarity between these two patterns of performance is what one would expect if our invocation of the individuation hypothesis to explain performance on the 3-card problem were correct. Like the original version of the 3-card problem, the candy cane problem with a single event format *and* an arbitrary parsing is twice cursed.

Is there support for the individuation hypothesis? The individuation hypothesis predicts that, when format is held constant, a problem will elicit higher levels of performance if the conditioning event can be parsed as a whole object (whole peppermint stick) rather than an arbitrary aspect of an intact object (end of a peppermint stick). When the problem was posed in a single event format, the whole object parsing elicited significantly higher performance than the arbitrary one: 24% correct versus 3% correct (Z=2.28, Φ =.30, p=.011). Performance for the whole object parsing was also higher than for the arbitrary one when the problem was posed in a frequency format: 45% correct versus 2l% correct (Z = 1.96, Φ = .26, p =.025). The size of the whole-object effect was similar, regardless of format: .30 for single event versions and .26 for frequency ones.

Is a frequency-based inference system incapable of counting over arbitrary parsings? (i.e., Is there an interaction between format and parsing?) Kleiter (1994) proposed that statistical inference mechanisms in humans are designed to work well given natural samples, 16 and that they do so by computing the absolute number of hits and false alarms. According to the individuation hypothesis, this should be easiest when counting the frequency of whole objects, and most difficult when solving single event problems that involve arbitrary parsings. The results accord with this prediction: the highest level of performance was elicited by the problem for which the conditioning event was the frequency of a whole object (45%), and the lowest level was elicited by the single event/arbitrary parsing problem (3%). Moreover, the effect size for this comparison, 0.48, is a substantial one (Z = 3.68, Φ = .48, p = .00012). Using the single event/arbitrary parsing problem as a baseline, one can see that applying a frequency format to an arbitrary parsing improved performance by 18 percentage points, whereas applying it to a whole-object parsing improved performance by 42 percentage points. This kind of enhancement is what one would expect if whole objects were the natural unit of analysis for frequency-computation systems. However, there are (at least) two different ways that the system could be built, both of which would lead to this result.

It could be that the frequency-computation system is simply incapable of taking anything other than whole objects as input. If this were the case, then there would be a strong interaction between format and parsing: Frequency format would have a much *smaller* effect for conditioning events parsed arbitrarily than for ones parsed as whole objects. This does not appear to be the case. Format (frequency vs. single event) and parsing (whole object vs. arbitrary) had independent effects. To test for an interaction between these two variables, subjects' responses were coded as either 1 (*correct*) or 0 (*incorrect*), and an analysis of variance (ANOVA) was conducted on the coded data according to the procedures specified by Rosenthal and Rosnow (1984). There was a significant main effect for each independent variable, but no interaction between them (parsing: $F(1,112) = 10.25$, *eta* = .29, p = .002; format: $F(1,112) = 7.51$, *eta* = .25, p = .007; parsing x format: $F(1,112) = .00053$, *eta* = .0022, p = .98).

This result suggests an alternative view. The input to a frequency computation system must be representations of tokens (rather than types). This system is, in principle, capable of counting tokens of arbitrarily parsed entities. Such parsings are rarely fed into it as input, however. This is because representations of the physical world are built by other systems (e.g., ToBy), and these systems were designed to individuate and imagine transformations of whole objects, not arbitrary parsings of them. These systems can produce type and token representations of whole objects (e.g., "canes" and "this cane"). They might also be able to construct type representations of arbitrary parsings (e.g.,

"canes have ends"). This does not imply, however, that they automatically produce token representations of arbitrary parsings ("this cane end"). If the output of such systems is not tokens of arbitrary parsings – if, in fact, the system is not well designed for producing such tokens – then they will not usually be available as input to the frequency-computation system.

What kind of representations does ToBy produce? A large body of evidence (some of which was reviewed earlier) suggests that the architecture of our minds has systems designed to automatically construct and operate over both type and token representations of whole objects. Infants construe a surface that is bounded, cohesive and moves as a unit as an "object" (token representation; Spelke, 1990); toddlers privilege whole objects (type representations) when they infer the referent of a novel word (Markman, 1989); toddlers spontaneously count whole objects (tokens) but not undivided parts of them (Shipley & Shepperson, 1990). There is evidence of systems that are designed to represent not only the presence of whole objects, but their movement and interactions as well. As early as 7 months, for example, infants distinguish events in which one object "launches" another (causes it to move by hitting it) from other events with similar spatiotemporal properties (Leslie, 1988). Through experiments on apparent motion, "mental rotation", and related phenomena, Shepard and his colleagues have shown that representations of the movement of objects are constrained by procedures that reflect evolutionarily long-enduring properties of the world– even when these representations occur in the absence of an external stimulus. Consequently, this system represents translations and rotations that are, in many ways, functionally isomorphic to the translations and rotations of rigid objects through three-dimensional space (e.g., Shepard, 1984, 1987). In other words, the mental models it produces reflect the world with some accuracy. Results of this kind have prompted Leslie (1994) to propose that a basic component of the human cognitive architecture is a system called ToBy, which stands for "theory of bodies". ToBy embodies principles for defining objects (types and tokens), representing their movements and interactions, distinguishing animate objects from inanimate ones, and so on. The representations of motion it produces can be individuated either as types of events (e.g., "launchings") or as tokens (e.g., "the launching event just seen").

On this view, ToBy and related systems produce representations of the physical world, and these are fed, as input, into a frequency computation system. Because the actual number of events is critical to probability estimates, these representations would need to be of individual events -- of tokens. In this light, one can think about how the results of Experiment 2 might have been produced.

How do ToBy and the frequency computation system interact? In the typical Bayesian word problem, the frequency of the conditioning event is given on a silver platter—often as explicit hit and false alarm rates. But in these experiments, they must be inferred and extracted.[5] This requires good models of the relationships among the physical objects (or aspects of objects) involved.

For example, in the whole object/frequency problem, one needs to model the *path* taken by individual *sticks* of candy as they *moved* from *jars* into a *bag* into your friend's *hand* (see Jackendoff, 1983, on the individuation of "paths"). This should not be very difficult, given that it is a model of the movement of whole objects to different spatial locations: modeling paths, objects, and movements is what ToBy does (see also Freyd, 1987, on dynamic mental representations). In fact, the paths of various peppermint sticks differ only in their point of origin: some came from the peppermint only jar and others came from the mixed jar. By running the "mental movie" ToBy constructed backwards, it is clear that any peppermint stick in your friend's hand had to originate in one of these two jars. ToBy automatically individuated peppermint sticks because they are whole objects, and this output can be fed into the frequency computation system, to be counted. By running the movie forwards again, the counting system can infer that one out of every two sticks drawn from the mixed jar is peppermint, whereas every stick drawn from the peppermint only jar was peppermint. The problem stipulates that an equal number of sticks were drawn from each jar. This is translated by the frequency-computation system into an absolute number drawn from each jar (say, 10), because that system is designed to represent and operate over absolute frequencies (for the reasons discussed under *Natural Sampling as an Environmental Invariant* earlier). By counting instances of hits and false alarms in this imagined sample, the system computes that there are 10 peppermint sticks from the Peppermint Only jar (10 hits) and 5 from the Mixed jar (5 false alarms). This means there is a total of 15 peppermint sticks, 10 of which came from the Peppermint Only jar. So the answer to *p*(PO| P stick) is "10 out of every l5" (or "20 out of 30", if one transforms the rate to match the absolute value of hits + false alarms [30] given in the problem).

Now consider how ToBy and the frequency computation system would interact on a frequency/arbitrary parsing problem. ToBy is designed to model whole objects. It uses information such as color, flavor, or texture to identify objects, so it would be very natural for it to use the presence of peppermint as a cue that identifies the object at hand as

[5] We assume this is why the absolute levels of performance in these experiments were a bit lower than those found for Cosrnides and Tooby's medical diagnosis problem, which involved calculations of similar complexity. (The calculations required for those tested by Gigerenzer and Hoffrage, 1995, were somewhat more complex.) One reviewer asked why absolute levels of performance were not higher, if our minds contain a well-designed frequency computation system. In our view, it is remarkable that they work on paper and pencil problems at all. A natural sampling system is designed to operate on actual events, counting hits and false alarms. Numbers of hits and false alarms are not given in the problems we tested. They need to be inferred on the basis of transformations of internal representations, and in the absence of the physical stimuli upon which these internal representations are based. In Cosmides and Tooby's (1996) experiments, those stimuli that prompted subjects to form more "perceptual" representations of the problem elicited higher performance, to a ceiling of 92% correct.

either the all pink stick or the pink-yellow stick. But stopping here will lead to the wrong answer (½), as in the card problem. Alternatively, ToBy could ignore this cue and simply look for all pink sticks, which would lead to a different wrong answer (⅓). When one looks at errors for the arbitrary parsing conditions, these were, in fact, the most common ones. In the single event problem, 20 out of 28 errors fell into these two categories (10 in each); in the frequency problem, 21 out of 23 errors did (11 answered "½," 10 answered "⅓"). In both cases, whole objects are being counted, but not ends.

To solve the arbitrary parsing problem correctly, ToBy would need to find a way of individuating ends. The end of a candy cane could, for example, refer to one of its boundaries. ToBy does use boundary information to define whole objects, but it does not automatically individuate undifferentiated areas on the boundary of an object and, therefore, would not feed tokens representing separate boundary areas into a frequency computation system. Yet solving the problem would require that the subject count such tokens; moreover, these individuated boundary areas would have to be counted up in a way that is independent of the intact objects whose boundaries they define (because one needs to compute that an all pink stick and a pink-yellow stick yield a total of three pink ends). Other ways of individuating ends are equally problematic. In fact, tasting only one "end" of a candy cane could refer to licking anywhere to either the left or right of the midpoint, or even to biting off a chunk near one of the circular boundaries of the cylinder. A defining feature of the stick that changes color and flavor in the middle, from peppermint pink to lemon yellow, is that there is a plane bisecting the stick where the color and flavor change abruptly: It can therefore be thought of as having "two halves" (3 month old infants ignore color and texture in defining objects and parts, but older children and adults do not; Spelke, 1988). From a mental image of this cane, it may be straightforward to read off that only one end – on any of the above definitions – is peppermint. But the all pink peppermint stick is a different matter. It is a completely symmetrical, undifferentiated object. To compute the frequency of hits and false alarms, ToBy would have to create an arbitrary distinction between various areas of this continuous cohesive object. This distinction would have to classify a homogenous object as one with two or more parts, and to classify two (and only two) of these parts as "ends". The fact that these "ends" are visually indistinguishable would, presumably, make it even more difficult to have them counted as two separate tokens by a frequency-computation system. And the system would have to add these two tokens to one derived from a different whole object, the pink-yellow stick.

In a whole object problem, ToBy is doing what it was designed to do: model the movement and interaction of rigid objects. It is also executing one of its design functions when it uses color or flavor as a cue to the identity of an object – but this output leads to an incorrect probability estimate in the arbitrary parsing problems. To solve these correctly, it needs to do something it was not designed for; even worse, it needs to parse the world in a way that *contradicts its own internal principles for defining rigid objects* – principles that cause a pink stick to be categorized as a single rigid object with no obviously differentiated parts.

Experiment 3

This analysis of how ToBy and the frequency-computation system might interact in the candy cane problems suggests the following: If one can cause subjects to categorize a candy cane as having two distinct ends, this should enhance performance on the problems in which the conditioning event is the probability/frequency of ends. A straightforward way of doing this would be to take advantage of ToBy's own internal principles for dividing the world into objects and objects into parts. These include cues such as "cohesive, bounded, entity" and "entity that can move independently of other entities". Breaking a candy cane in half creates two "pieces" -- two cohesive bounded entities, each of which can move independently of the other. This is the kind of entity that ToBy naturally individuates. Because these daughter objects are what need to be counted to solve the problem correctly, having subjects imagine this manipulation should improve performance on what were previously "arbitrary parsing" problems.

In Experiment 3 we tested this hypothesis, using slightly altered versions of the arbitrary parsing problems tested in Experiment 2. We tested two versions: one with a single-event format and one with a frequency form. On the individuation hypothesis, one would predict that performance on these two problems will be similar to that found for the whole object versions tested in Experiment 2.

Experiment 3 has an additional advantage: It allows one to see whether some extraneous surface feature of the whole-object problems tested in Experiment 2— which involved candy canes in jars – can account for the higher levels of performance that they elicited. The surface content of the broken cane problems tested in Experiment 3 is almost identical to that of the arbitrary parsing problems tested in Experiment 2. It is, in fact, an even closer match than that provided by Experiment 2's whole object, jar/cane conditions. Finding the same pattern of performance in Experiment 3 as we found for the whole object conditions of Experiment 2 would militate against the hypothesis that the improved performance for whole objects in Experiment 2 was caused by some extraneous feature of those problems.

Method

Subjects. There were 48 subjects in this experiment; 21 in the single event condition and 27 in the frequency condition.

Materials for Condition 1. In this single-event problem, each candy cane is broken into two halves. The event to be predicted is the presence of an all-pink peppermint stick; the conditioning event is the presence of a peppermint end (half stick). It is identical to the single event/arbitrary parsing problem tested in Experiment 2 (condition 1), except for the last three paragraphs.

> At the grocery store, there are three jars of straight candy canes.
>
> The first jar contains only pink candy canes. These are peppermint.

The second jar contains only yellow candy canes. These are lemon.

The third jar contains candy caries that change flavor and color in the middle. One half of each stick is pink and tastes like peppermint; the other half of each stick is yellow and tastes like lemon.

You just bought a large number of candy canes. You put three equal-sized handfuls of them into a bag. You took one handful from each of the three jars.

When you got home, you broke every candy cane in half. You put all the pink halves into a pink jar and all the yellow halves into a yellow jar.

Your friend reached into the pink jar and pulled out one pink half stick.

What is the probability that it was originally from one of the all-pink peppermint sticks? ____

Materials for Condition 2. This is a frequency version the broken stick problem in condition 1. Except for the last three paragraphs, it is identical to the frequency/arbitrary parsing problem tested in Experiment 2 (condition 2).

At the grocery store, there are three jars of straight candy canes.

The first jar contains only pink candy canes. These are peppermint.

The second jar contains only yellow candy canes. These are lemon.

The third jar contains candy canes that change flavor and color in the middle. One half of each stick is pink and tastes like peppermint; the other half of each stick is yellow and tastes like lemon.

You just bought a large number of candy canes. You put three equal-sized handfuls of them into a bag. You took one handful from each of the three jars.

When you got home, you broke every candy cane in half. You put all the pink halves into a pink jar, and all the yellow halves into a yellow jar.

Your friend reached into the pink jar and pulled out 30 pink half sticks.

How many of these halves do you expect were originally from all-pink peppermint sticks? ___ out of ___

Results and Discussion

The percent correct for each condition is depicted in Figure 3. Because we are interested in whether the results of the broken stick problems more closely resemble those for the whole object than the arbitrary parsing problems in Experiment 2, the results of Experiment 3 are superimposed on a graph that depicts the results of both experiments.

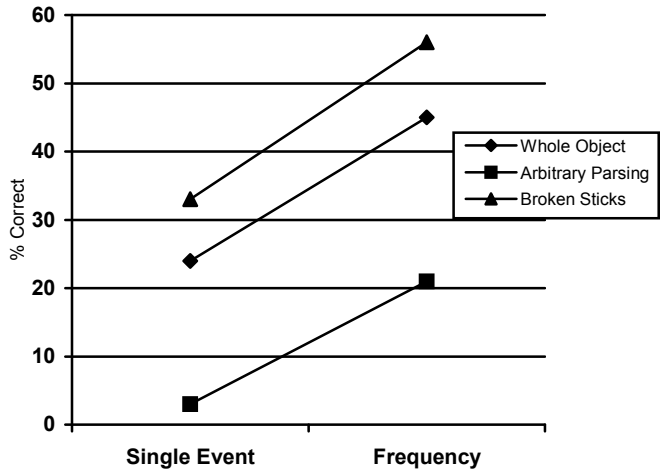

Figure 3. Transforming an inseparable aspect into a whole object: candy cane series.

Performance on the broken stick problems was clearly more similar to that for Experiment 2's whole object problems (jar/cane) than its arbitrary parsing problems (cane/end). The correct answer was elicited from 33% of subjects in the single event version, and 56% in the frequency version. These results are not significantly different from the 24% correct (single event) and 45% correct (frequency) found for the corresponding whole object conditions. They are very different, however, from the 3% correct (single-event) and 21% correct (frequency) found for the corresponding arbitrary parsing conditions (33% vs. 3%: $Z = 2.84$, Φ=.40, $p = .0022$; 56% vs. 21%: $Z = 2.69$, $\Phi = .36$, $p = .0035$). This is in spite of the fact that the surface content of the broken stick problems was, in many ways, more similar to that of the arbitrary parsing problems. The only dimension along which they differ is the one that is relevant to the individuation hypothesis. Moreover, this is the only dimension along which the broken stick problems more closely match the whole object ones. When a stick is broken, the ends are processed differently than when they are attached. They are treated more like whole objects than like arbitrarily parsed aspects.

We note that in both of the broken stick conditions, performance was ~10 percentage points higher than in the corresponding whole object versions. Although this difference is not statistically significant, it is interesting given our hypothesis about the ways in which ToBy and the frequency-computation system interact. Given the internal principles that govern ToBy's operation, there is a close causal relationship between a pink candy cane and the two daughter halves that are produced when it is broken: One object was broken in two. By running the "movie" backwards, it is easy to see that all half sticks originated on whole ones. By running it forwards, it is easy to see the following: Every pink half-stick was "born", so to speak, when a whole stick was broken; every whole stick produced two halves; every all-pink stick produced two pink half sticks; and every two-toned stick produced only one pink half stick. Given the dimensions along which ToBy construes causality for inanimate objects, there is tight and necessary causal relationship between parent object and daughter objects when one object is broken in two. In contrast, there is no necessary or causal connection between the three spatial locations concerned Oar, bag, and hand) when a whole stick moved from place to place in the whole object versions of Experiment 2. The mental model constructed would still have conformed to ToBy's internal principles had other locations been used.

The effect of frequency, holding parsing constant, appears in this data also. In fact, the size of the effect was almost identical to that for the matching whole object condition of Experiment 2. For the broken stick versions, there was a 23 percentage point difference between the single event and frequency conditions, and an effect size of .22; for the matching whole object conditions, it was a 21-point difference, with an effect size of .22 (56% vs. 33%: Z= 1.53,

Φ=.22, p =.063—the p value is slightly larger in this condition because the sample size is smaller—48 vs. 58).

Experiment 4

The logic of Experiment 4 is identical to that for Experiment 2. They differ only in surface content: The prior experiment involved candy canes, whereas this experiment involves fish. Psychologists have been surprised to find that changes in surface content frequently alter reasoning performance on problems that were thought to be structurally isomorphic (for discussion, see Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Murray, 1987; Wason & Johnson-Laird, 1972). In many cases, differences that were assumed to be "surface" turned out to be "structural". For example, two conditional rules with an identical logical structure (If P then Q) will elicit very different patterns of performance when the content of one of them happens to express the benefit/requirement relationships typical of social contracts (Cosmides, 1985; Cosmides & Tooby, 1989). It turns out that for social contracts, logical categories are better thought of as structural features, whereas benefit/requirement categories are better thought of as structural features (see, e.g., Cosmides & Tooby, 1992, on switched social contracts; Gigerenzer & Hug, 1992, on perspective change). Similarly, performance on rules with exactly the same surface content (e.g., "If a man eats cassava root, then he must have a tattoo on his face") is vastly different when the context causes the terms to be mapped onto the benefit/requirement contingencies of a social contract than when it maps them onto a contingent relationship that merely describes the co-occurrence of events.

For reasons like these, we take surface content very seriously. If the individuation hypothesis is correct, then it will allow us to predict which kinds of surface content will produce changes in performance, and which will not. The fish problems that follow were designed such that they would be structurally isomorphic to the candy cane ones, given the object primitives that are important to the individuation hypothesis. In the arbitrary parsing conditions, sides of a fish played the role of ends of a candy cane, and whole fish played the role of whole candy canes. In the whole-object conditions, fish stood in for candy canes, and lakes stood in for jars.

Method

Subjects. There were 118 subjects in this experiment, with 28 in three of the conditions, and 34 in one of them (single event/whole object).

Materials for Condition 1. This is a single event/arbitrary parsing problem. The event to be predicted is the presence of a fish of the blue-blue species, and the conditioning event is the sight of a fish in profile, which reveals that that side of the fish is blue.

> There are a large number of fish in a tank. One tank contains three different species of fish.
>
> One third of the fish are blue on both sides (the blue-blue species).
>
> One third of the fish are red on both sides (the red-red species).
>
> And one third of the fish are blue on one side and red on the other side (the blue-red species).
>
> While you are looking at the tank, a fish swims past you. You can only see one side of this fish, and note that the side you can see is blue.
>
> What is the probability that this fish is a member of the blue-blue species? ____

To solve this problem, one needs to realize that the probability of seeing a blue side on a fish of the blue-blue species is twice that of seeing a blue side on a fish of the blue-red species. The value to be computed is a posterior probability, p(BB species | B side).

Materials for Condition 2. This frequency/arbitrary parsing problem is virtually identical to condition 1, except for the changes needed to transform the problem into a frequency format.

> There are a large number of fish in a tank. The tank contains three different species of fish.
>
> One third of the fish are blue on both sides (the blue-blue species).
>
> One third of the fish are red t-n both sides (the red-red species).
>
> And one third of the fish are blue on one side and red on the other side (the blue-red species).
>
> You took a photo of the tank; some of the fish can be seen in the photo. These fish are in profile, so you can see only one side of each fish.
>
> In the photo, 30 of the fish profiles you see are blue.
>
> How many of these fish do you expect are members of the blue-blue species? ___out of ___

Materials for Condition 3. In this single event/whole object problem, the event to be predicted is which lake a fish came from, and the conditioning event is the presence of a blue fish. Note that a fish is a whole object, and a lake can be thought of as a container for fish in the same way that a jar can be thought of as a container for candy canes–when described in this way, it is an honorary object, bounded, internally cohesive and continuous, with its "insides" (the water) capable of moving independently of the surrounding banks. (Also, on many theories of conceptual structure, lakes are locations that can be individuated either as types or tokens; indeed, they would have to be if ToBy is to be able to model the movement of objects through paths that connect a point of origin to a destination; see Jackendoff, 1983). We preserved as many elements of the problem as possible; the only changes in the text were ones necessary to transform the conditioning event from an arbitrarily parsed entity into a whole object (naturally, this required a change in the event to be predicted as well -- in all four problems, however, this event was defined over an easily individuated entity).

> You sell fish. The fish you sell come from three different fishermen: Tom, Dick, and Harry. Each man fishes in a different lake.
>
> Tom fishes in a lake in which there are only blue fish. It is called Blue Lake.
>
> Dick fishes in a lake in which there are only red fish. It is called Red Lake.
>
> Harry fishes in a lake in which half the fish are blue and half the fish are red. It is called Blue-Red Lake.
>
> Today, you bought a large number of fish. You bought equal numbers from each of the fl= fishermen. You dumped all of these fish into one big ice chest.
>
> Later that day, a customer picked one blue fish out of your ice chest.
>
> What is the probability that this fish came from Blue Lake? ___

To solve this problem, one needs to realize that rate of return for blue fish when fishing in Blue-Red Lake is half the rate of return of

blue fish when fishing in Blue Lake. The value to be computed is *p*(Blue Lake|Blue fish).

Materials for Condition 4. This frequency/whole object problem is identical to condition 3 except for the last three paragraphs, which transform the problem into a frequency format.

> You sell fish. The fish you sell come from three different fishermen: Tom, Dick, and Harry. Each man fishes in a different lake.
>
> Tom fishes in a lake in which there are only blue fish. It is called Blue Lake.
>
> Dick fishes in a lake in which there are only red fish. It is called Red Lake.
>
> Harry fishes in a lake in which half the fish are blue and half the fish are red. It is called Blue-Red Lake.
>
> Today, you bought I 00 fish from each man. You dumped all of the blue fish into one ice chest and all of the red fish into a different ice chest.
>
> Later that day, a customer bought 30 blue fish.
>
> How many of these fish do you expect came from Blue Lake? ___ out of ___

Results and Discussion

The percent correct for each condition is depicted in Figure 4. Given the primitives relevant to the individuation and frequency hypotheses, the fish problems are structurally isomorphic to the candy cane problems, and should therefore elicit similar results, which they do. For the single-event problems, one sees the difference between arbitrary parsing and whole object versions predicted by the individuation hypothesis (7% vs. 26%: $Z = 1.98$, $\Phi = .25$, $p = .024$). This l9-percentage-point increase is similar to the 2l-point increase found for the comparable candy cane problems, and the effect sizes are similar as well (candy cane: .30; fish: .25). On frequency versions, the correct answer was elicited from 36% of subjects for the arbitrary parsing problem versus 50% of subjects for the whole object problem ($Z = 1.08$, Φ .14, $p = .14$). This 14-percentage-point increase in performance is in the right direction, but it is smaller than the 24 percentage point increase found in the matching conditions of the candy cane series, as is the effect size (fish: .14; candy cane: .26). However, frequency/whole-object versions of the fish and candy cane problems elicited almost the same percent correct (fish: 50%; candy cane: 45%), and performance on the frequency/arbitrary parsing versions was not statistically different (fish: 36%; candy cane: 21%. $Z=1.26$; $\Phi = .17$; $p = .10$). Nevertheless, there was a l5-percentage-point difference between the latter. This means that in the fish series, the smaller effect of parsing on frequency versions is due entirely to performance on the frequency/arbitrary parsing problem, which was slightly higher than that found on the comparable candy cane and card problems (36% vs. 21% and 28%).

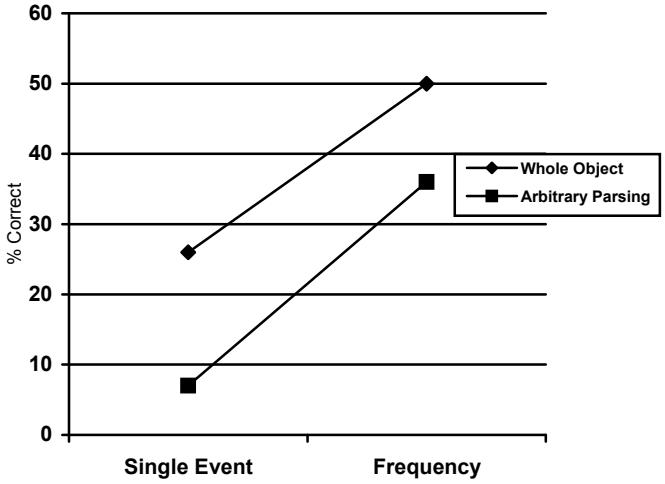

Figure 4. Whole-object versus arbitrary parsing problems: fish series.

On the problems with arbitrary parsings, the effect of frequency versus single event format is quite clear: 3 6% for the frequency format versus 7% for the single event format (Z = 2.6 1; phi .35; p=.0046). The effect of format was also clear on the whole object problems: 50% for the frequency format versus 26% for the single event format (Z = 1. 9 1; phi = .24; p = .028).

As expected, the frequency/whole object problem elicited the best performance in the fish series, and the most dramatic difference was between this problem and the single event/arbitrary parsing problem: 50% correct versus 7% correct, a difference of 43 percentage points ($Z=3.55$, $\Phi= .47$, $p=.0002$). This compares favorably to the difference of 42 percentage points in the comparable conditions of the candy cane series (45% vs. 3%); the effect sizes are comparable as well (fish: .47, candy cane: .48).

An ANOVA for the fish series shows a main effect for both format and parsing, but no interaction (parsing: $F(1, 114) = 4.31$, *eta* = . 19, $p = .04$; format: $F(1,114) = 10.36$, *eta* = .29, $p = .002$; Parsing x Format: $F(1,114) = .097$, *eta* = .03, $p = .86$). The values computed are similar to those for the candy cane series. As before, the lack of an interaction speaks against the hypothesis that frequency-computation systems are simply incapable of counting arbitrarily parsed aspects of the world. Instead, it favors the hypothesis that performance is lower on arbitrary parsing problems because ToBy is not designed to individuate the world in these ways, and therefore has trouble modeling the situation in a way that would provide the input necessary for these systems to arrive at a correct solution.

Thus the results of the first series nicely replicated those found for the candy cane series.

Experiment 5

The logic of Experiment 5 is identical to that for Experiment 3. Instead of breaking a single candy cane into two pieces, in this experiment we sliced a fish in half, creating two filets. These problems were designed as counterpoints for the arbitrary parsing problems of the fish series. In those problems, the fish was a whole object, and its sides were inseparable aspects of that object, indeed, boundaries of it. But slicing a fish in half creates two "pieces". If ToBy generates mental transformations that mirror physical ones, as Shepard has argued, then imagining this action will produce representations of two cohesive bounded entities, each of which can move independently of the other. As "newly formed" whole objects, these representations can be (token) individuated and counted, to arrive at a correct

solution. If the individuation hypothesis is correct, then, the results elicited by these filet problems will more closely resemble Experiment 4's whole object problems than its arbitrary parsing ones.

Method

Subjects. There were 52 subjects in this experiment, 26 in each condition.

Materials for Condition 1. In this single event problem, the event to be predicted is the presence of a fish of the blue-blue species, and the conditioning event is the presence of a blue filet. The latter is analogous to the sight of a blue side profile, which was the conditioning event in the arbitrary parsing version. The text is similar to that for the single event/arbitrary parsing problem tested in condition 1 of Experiment 4.

You sell fish filets. By splitting a fish down the middle of its spine, you can get two identical filets from each fish (i.e., each filet has one eye, one gill, its skin, etc.).

The fish you sell come from a fish farm, in which the fish are raised in an artificial lake. The lake contains three different species of fish.

One third of the fish are blue on both sides (the blue-blue species).

One third of the fish are red or. both sides (the red-red species).

And one third of the fish are blue, on one side and red on the other side (the blue-red species).

A random assortment of fish from the lake were brought to your shop today. You split each fish in half.

So each fish from the blue-blue species yielded two blue filets.

Each fish from the red-red species yielded two red filets.

And each fish from the blue-red species yielded one blue filet and one red filet.

You put all of the blue filets in one ice chest and all of the red filets in another ice chest.

A customer bought one blue filet today.

What is the probability that it came from the blue-blue species? ___

Materials for Condition 2. The text of this frequency version is identical to condition 1, except for the last two sentences.

You sell fish filets. By splitting a fish down the middle of its spine, you can get two identical filets from each fish (i.e., each filet has one eye, one gill. its skin, etc.).

The fish you sell come from a fish farm, in which the fish are raised in an artificial lake. One lake contains three different species of fish.

One third of the fish are blue on both sides (the blue-blue species).

One third of the fish are red on both sides (the red-red species).

And one third of the fish are blue on one side and red on the other side (the blue-red species).

A random assortment of fish from the lake were brought to your shop today. You split each fish in half.

So each fish from the blue-blue species yielded two blue filets.

Each fish from the red-red spears yielded two red filets.

And each fish from the blue-red species yielded one blue filet and one red filet.

You put all of the blue filets in one ice chest and all of the red filets in another ice chest.

A customer bought 30 blue filets today.

How many of these filets do Note expect came from the blue-blue species? ___ out of ___

Results and Discussion

The percent correct for each condition is depicted in Figure 5. These results are superimposed on a graph that depicts the results of Experiment 4, so that it is easy to see the relationship among them.

Is a filet problem more like a whole object problem or an arbitrary parsing problem? In both the filet problem and the arbitrary parsing problem in Experiment 4, subjects were asked to estimate p(blue-blue fish | blue side)— the problems differed mainly in whether the side was still connected to the fish. Yet frequency versions of the filet problem elicited the correct answer from considerably more subjects than the analogous arbitrary parsing one: 62% versus 36%, respectively (Z = 1.90; Φ = .26; p = .029). Indeed, the filet vs. arbitrary parsing difference was even larger than the whole fish vs. arbitrary parsing difference found in Experiment 4 (effect sizes: .26 vs. .14). In contrast, there was no significant difference in performance between the filet problem and the matching whole object version (62% vs. 50%: Z=0.85; Φ =.12; p =.20). In other words, performance was more similar for these two whole object conditions, even though the events to be predicted were quite different: p(lake|fish) vs. p(fish|filet).

Comparisons among single event versions yield analogous results. Performance on the filet and whole object problems was similar: 19% correct and 26% correct, respectively (19% vs. 26%: Z=0.66; Φ =.08; p =.26). This is what one would expect if filets were being processed as whole objects. In contrast, only '/% of subjects produced the correct answer for the matching arbitrary parsing problem (7% vs. 19%: Z = 1.32; Φ = . 18; p = .09). Although this 12-percentage-point difference for single-event versions is not significant at the .05 level, the results for the filet problem are clearly more similar to those for the whole object problem than the arbitrary parsing problem (filet vs. whole object: Φ = .08; filet vs. arbitrary parsing: Φ = . 18).

Performance on the filet problems was similar to that for the broken stick problems from the candy cane series. Frequency versions of the filet and broken-stick problems did not differ significantly (62% filet vs. 56% sticks:

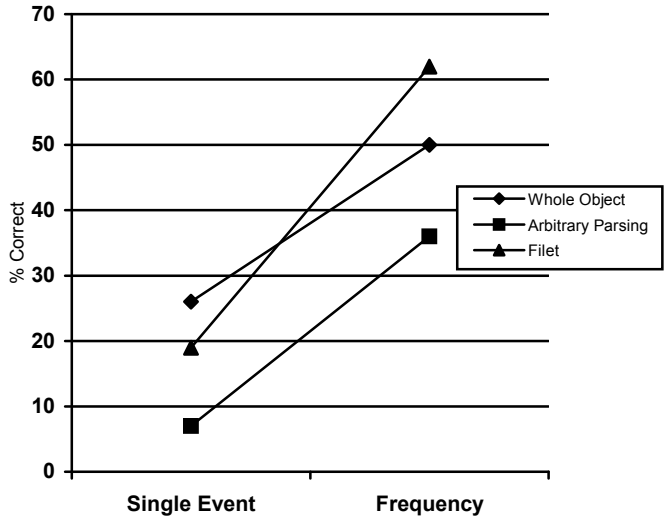

Figure 5. Transforming an inseparable aspect into a whole object: fish series.

Z = .44; Φ = .06; p = .33), nor did the single event versions (19% filet vs. 33% sticks: Z = 1. 10; Φ = . 16; p = . 14). Interestingly, in both cases, frequency versions of the problems involving sundered objects elicited slightly higher performance than their matching whole object problems did: an 11-point difference for the candy cane series, and a 12 point difference for tlie fish series. This is consistent with our prior suggestion that, when one object is broken in two, ToBy will construe the relationship between parent object and daughter objects as a close and necessary causal one—thereby clarifying the numerical relationship between the conditioning event and the event to be predicted.

Is there a frequency effect in the filet series? In the filet problems, which hold parsing constant, the effect of frequency format is large: 62% correct for the frequency version, 19% for the single event version (Z = 3.11, Φ = .43, p = .001).

Conclusions

In tasks requiring judgment under uncertainty, people appear to behave like good "intuitive statisticians" when (a) information is given and answers are asked for in frequencies, rather than proportions and single event probabilities, and (b) the conditioning event is the frequency of a whole object rather than an arbitrary parsing.

Experiments 1-5 tested the effect of frequency format on Bayesian problems with new content— cards, candy canes, fish— and found the same enhancement as in previous studies. In fact, the effect sizes produced for whole-object problems in these experiments were almost identical to those found by Gigerenzer and Hoffrage (1995) and similar to those found by Cosmides and Tooby (1996), in spite of the fact that the subject populations and problem contents were different. This obtained even though the appropriate hit and false alarm rates were given to the subjects on a "silver platter" in previous studies, but not in these problems. In these problems, the subjects had to construct the hit and false alarm rates, based on information about the relationships between rigid objects and their parts, or rigid objects and their locations.

In fact, we found an enhancing effect of frequency format even when the problems tested required subjects to compute conditional probabilities over arbitrary parsings of events— over sides of cards, ends of candy canes, sides of fish. Across seven independent comparisons (three arbitrary parsing pairs, four whole-object ones) the size of the frequency effect remained fairly constant.

These experiments also allowed us to test the individuation hypothesis: that whole objects, rather than arbitrary parsings of objects, are the natural unit of analysis for the frequency-computation systems that are activated in this kind of experiment. Experiments 1-5 consistently supported this hypothesis: Holding format constant, whole-object problems reliably elicited higher levels of performance than arbitrary parsing problems. Moreover, the lowest levels of performance were elicited by single-event/arbitrary parsing problems, and the highest levels by frequency/whole-object problems. The difference between these two categories of problem was consistently large: In the candy series, performance increased by 42 and 53 percentage points, respectively, for the two whole object problems (whole cane and broken sticks). In the fish series, performance increased by 43 and 55 percentage points, respectively (whole fish and filet problems). This is what one would expect on the individuation hypothesis.

There are two alternative versions of the individuation hypothesis, which we also explored. The first holds that the frequency computation system can only operate on representations of whole objects. On this view, arbitrary parsing problems elicit lower performance because the frequency-computation system is simply not capable of counting tokens of arbitrarily parsed regions, even if ToBy were to generate them. If this were true, frequency format would not improve performance on problems in which subjects have to compute the frequencies of arbitrarily parsed forms, and one would see in interaction between the format and parsing variables. The data spoke very clearly against this hypothesis. We consistently found main effects for format and parsing variables, but we found no Format x Parsing interactions. Arbitrary parsing does not reduce the size of the frequency effect, even marginally: The mean frequency effect size for the three pairs of arbitrary parsing problems (.29) is almost identical to that for the four pairs of whole object ones (.28). Indeed, the only difference between these two classes of problems was that the absolute level of performance on the arbitrary parsing ones was, on average, about 24 percentage points lower than that found for their matching, whole-object analogs.

This data is more consistent with a different version of the individuation hypothesis: that the frequency-computation system is capable of counting tokens of arbitrarily parsed events, but that it is rarely fed that kind of information. On this view, there are conceptual systems, such as ToBy— the theory of bodies (Leslie, 1994)— that are designed to model physical objects and their interactions. The principles of object perception embodied by these systems privilege whole objects over arbitrary aspects (see page 8 for relevant senses of "arbitrary"). They can produce type and token representations of whole objects, and these whole-object tokens are routinely fed into the frequency-computation system. However, although these systems can produce type representations of some arbitrary aspects (e.g., sides of cards), there are many type representations that they are not designed to produce, because they would be adoptively irrelevant (e.g., "trapezoidal patches of bark near the bottom of tree trunks"). Moreover, although they sometimes produce type representations of arbitrary aspects, they rarely produce token representations of them. They are not designed to automatically produce tokens of arbitrarily parsed aspects because in many cases this information would be useless (e.g., tokens representing each side of each leaf on a tree), and in many others it would be positively harmful. For example, the memory capacity of the brain would be overwhelmed very quickly if the visual system stored *tokens* of every percept produced by every saccadic fixation on a familiar tree.

It should be possible to transform an arbitrary aspect of an object into an honorary whole object, if the internal principles by which ToBy and related systems individuate objects are known. Investigations by Spelke (1988, 1990), Leslie (1988), Shepard (1984) and others have elucidated some of the principles whereby rigid objects are perceived and their movements imagined. These led us to predict that arbitrary chunks of an intact object would be represented as whole objects if one imagines breaking them off of the original object. ToBy should classify these representations of daughter units as whole objects because it can now generate images of them moving independently of one another (in analog representations, the "motion" of imagined objects is constrained in the same ways as the motion of real objects; Shepard, 1984). We gave subjects problems in which they had to compute probabilities over arbitrary units of this kind, but varied whether these were detached from their parent object or still an undivided aspect of it. Subjects' performance on the detached, "honorary objects" improved markedly. This result also shows that low levels of performance on arbitrary parsing problems were not due to the arbitrary aspect's function, color, or other characteristics: The honorary objects had these same characteristics. The one characteristic that the honorary objects had that the others lacked, is freedom: They were no longer bound to the parent object, and therefore were capable of independent motion.

A computational system can produce well-calibrated statistical inferences only if its operations are based on an appropriate model of the situation at hand. We proposed that ToBy, a system that represents rigid objects and generates inferences about them, supplies such models to a frequency-computation system. Our experiments supported this conjecture. When the model needed to solve a probability problem was the kind that ToBy is designed to build, subjects did very well. But when it required elements that ToBy does not spontaneously individuate, subjects did poorly. These results implicate ToBy. In so doing, they show that it is not necessary to invoke judgmental heuristics— rules of thumb for estimating probabilities— to explain poor performance on tasks requiring judgment under uncertainty. Even a sophisticated computational system will produce errors when it is fed the wrong kind of information.

If our interpretation is correct, then computational machines that build models of the world are routinely activated in tasks requiring judgment under uncertainty. Probability tasks will appear intractable when their solution depends on individuating the world in a way that violates the internal principles of the representational system activated. If we know what these principles are, we should be able to predict in advance when people will produce systematic errors, and what those errors will be.

Jackendoff (1983) has proposed that, just as there are automatic, reliably developing, cross-culturally universal principles for individuating rigid objects, there are similar principles for individuating actions, events, and paths. Cosmides and Tooby (1989, 1992), Fiske (1991) and others have proposed that there are principles for individuating certain kinds of social situations. The individuation hypothesis (second version) should apply to the representations produced by these principles as well. For example, our minds parse a person's behavioral stream into *actions*, which *cause* "other" *actions*, where the mode of causation involves mental states, such as beliefs, desires, perceptions, and intentions (Baron-Cohen, 1995; Leslie, 1994). Probability problems that tap into the principles whereby we individuate actions and mental states should elicit patterns analogous to those we found for problems involving rigid objects. Experiments of this kind can serve two purposes:

1. They can provide further insight into the conditions that promote or impede judgment under uncertainty in different domains. To understand how frequency data is used in judging character or personal risk, for example, one needs to know how the frequency computation system interacts with a conceptual system that treats some aspects of a behavioral stream as arbitrary parsings while privileging others as "actions" that can be individuated, classified, and counted.
2. They can be used as a tool to evaluate competing hypotheses about conceptual structure -especially ones that posit nonconscious modes of construal, which cannot be studied via introspection. Because statistical inference requires individuation, these tasks can unobtrusively reveal the principles whereby our cognitive architecture carves a wide variety of ontological categories— actions, events, journeys, paths, and so on— into types and tokens.

After all, ToBy is just one component of the human cognitive architecture that produces models of the world. There are many others, and their properties are largely unknown.

References

Aitchison, J. & Dunsmore, 1. (1975). *Statistical prediction analysis.* Cambridge, UK: Cambridge University Press.

Baerends, G. (1941). Fortpflanzungsverhalten und Orientierung der Grabwespe *Ammophila campestris Jur.* [Reproductive behavior & orientation of the sand wasp] *Tijdschrifi voor Entomologie 84,* 68-275.

Baillargeon, R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month old infants. *Cognition, 23,* 21-41.

Bar-Hillel, M. (1989). Discussion: How to solve probability teasers. *Philosophy of Science, 56*, 348-358.

Bar-Hillel, M. & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition, 11*, 109-122.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Christensen-Szalanski, J., & Beach, L. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance, 29*, 270-278.

Cosmides, L. (1985). *Deduction or Darwinian Algorithms? An explanation of the "elusive" content effect on the Wason selection task*. Doctoral dissertation, Harvard University. (University Microfilms #86-02206)

Cosmides, L. (1989) The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187-276.

Cosmides, L. & Tooby, J. (1989). Evolutionary psychology and the generation of culture, Part II. Case study: A computational theory of social exchange. *Ethologv and Sociobiologv, 10*, 51-97

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind.- Evolutionary psychology and the generation of culture.* (pp. 163-228). New York: Oxford University Press.

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all?: Rethinking some conclusions of the literature on judgment under uncertainty. *Cognition, 58,* 1-73.

Dawkins, R. (1986). *The blind watchmaker*. New York: Norton.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50*, 123-129.

Fiske, A. (1991). *Structures of social life: The four elementary forms of human relations*. NY: Free Press.

Freyd, J. J. (1987). Dynamic mental representations. *Psychological Review,* 94, 427-438.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Garcia, J. & Koelling, R.A. (1966). Relations of cue to consequence in avoidance learning. *Psychonomic Science, 4*, 123-124.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond heuristics and biases. *European Review of Social Psychology, 2*, 83 -115.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Gigerenzer, G. Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528.

Gigerenzer, G. & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating and perspective change. *Cognition, 43,* 127-171.

Gigerenzer, G. & Murray, D. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Hasher, L. & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*, 356-388.

Hintzman, D. L. & Stern, L. D. (1978). Contextual variability and memory for frequency. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 539-549.

Hirschfeld, L. & Gelman, S. (Eds.). (1994). *Mapping the mind.- Domain specificity in cognition and culture*. NY: Cambridge University Press.

Jackendoff, R. (1983). *Semantics and Cognition*. Cambridge, MA: MIT Press.

Johnson, M. & Morton, J. (1991). *Biology and cognitive development: The case of face recognition.* Oxford, United Kinddom: Blackwell.

Kahneman, D., Slovic, P, & Tversky, A., (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, United Kingdom: Cambridge University Press.

Kleiter, G. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & D. Lang (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology.* (pp. 375-388). NY: Springer-Verlag.

Leslie, A. (1988). The necessity of illusion: Perception and thought in infancy. In L Weiskrantz (Ed.), *Thought without language*. (pp. 185-210). Oxford: Clarendon Press.

Leslie, A. (1994). ToMM, ToBY, and agency: Core architecture and domain specificity. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*. (pp. 119-148). New York: Cambridge University Press.

Markman, E. (1989). *Categorization and naming in children*. Cambridge, MA: MIT Press.

Real, L. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science, 253*, 980-986.

Real, L. & Caraco, T. (1986) Risk and foraging in stochastic environments: theory and evidence. *Annual Review of Ecology and Systematics, 17*, 371-390.

Rescorla, R. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review, 74*, 71-80.

Rescorla, R. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychologv, 66*, 1-5.

Rosenthal, R. & Rosnow, R. (1984). *Essentials of behavioral research*. New York: McGraw-Hill.

Shepard, R.N. (1984) Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review, 91*, 417-447.

Shepard, R.N. (1987). Evolution of a mesh between principles of the mind and regularities of the world. In J. Dupre (Ed.), *The latest on the best: Essays on evolution and optimality.* (pp. 251-275). Cambridge, MA: MIT Press.

Shepard, R. N. (1992). The three-dimensionality of color: An evolutionary accommodation to an enduring property of the world? In J. Barkow, L. Cosmides, & J. Tooby (Eds), *The adapted mind:Evolutionary psychology and the generation of culture*. (pp. 495-532). New York: Oxford University Press.

Shipley, E. & Shepperson, B. (1990). Countable entities: Developmental changes. *Cognition, 34*, 109-136.

Spelke, E. (1988). The origins of physical knowledge. In L. Weiskrantz (Ed.), *Thought without language*. (pp. 168-184). Oxford, United Kingdom: Clarendon Press.

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

Staddon, J. E. R. (1988) Learning as inference. In R. C. Bolles & M. D. Beecher (eds.), *Evolution and learning*. (pp. 59-77). Hillsdale, NJ: Erlbaum.

Starkey, P., Spelke, E., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition, 36,* 97-128.

Stephens, D. & Krebs, J. (I 986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*, 49-100.

Tooby, J. & Cosmides, L. (1992). *Ecological rationality and the multimodular mind: Grounding normative theories in adaptive problems* (Tech. Rep. No. 92-1). Santa Barbara: University of California, Center for Evolutionary Psychology.

Tooby, J. & Cosmides, L. (Eds.) (in press). *Evolutionary psychology: Foundational papers*. Cambridge, MA: MIT Press.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293-315.

Wason, P. & Johnson-Laird, P. (1972). *Psychology of reasoning.- Structure and content* Cambridge, MA: Harvard University Press.

Wynn, K. (1995). Origins of numerical knowledge. *Mathematical Cognition, 1*, 35 -60.

Received April 5, 1996
Revision received August 19, 1996
Accepted October 23, 1996 ■