COGNITIVE SCIENCE 9, 309–339 (1985)

Languages and Designs for Probability Judgment*

GLENN SHAFER
School of Business
University of Kansas

AMOS TVERSKY
Department of Psychology
Stanford University

Theories of subjective probability are viewed as formal languages for analyzing evidence and expressing degrees of belief. This article focuses on two probability langauges, the Bayesian language and the language of belief functions (Shafer, 1976). We describe and compare the semantics (i.e., the meaning of the scale) and the syntax (i.e., the formal calculus) of these languages. We also investigate some of the designs for probability judgment afforded by the two languages.

INTRODUCTION

The weighing of evidence may be viewed as a mental experiment in which the human mind is used to assess probability much as a pan balance is used to measure weight. As in the measurement of physical quantities, the design of the experiment affects the quality of the result.

Often one design for a mental experiment is superior to another because the questions it asks can be answered with greater confidence and precision. Suppose we want to estimate, on the basis of evidence readily at hand, the number of eggs produced daily in the U.S. One design might ask us to guess the number of chickens in the U.S. and the average number of eggs laid by each chicken each day. Another design might ask us to guess the

* This research has been supported in part by NSF grants MCS-800213 and 8301282 to the first author and by ONR Grant NR197-058 to the second author. The article has benefited from the comments of Jonathan Baron, Morris DeGroot, Persi Diaconis and David Krantz.

Correspondence and requests for reprints should be sent to Glenn Shafer at the School of Business, University of Kansas, Lawrence, KS 66045.

number of people in the U.S., the average number of eggs eaten by each person, and some inflation factor to cover waste and export. For most of us, the second design is manifestly superior, for we can make a reasonable effort to answer the questions it asks.

As this example illustrates, the confidence and precision with which we can answer a question posed in a mental experiment depends on how our knowledge is organized and stored, first in our mind and secondarily in other sources of information available to us.

The quality of the design of a mental experiment also depends on how effectively the answers to the individual questions it asks can be combined to yield an accurate overall picture or accurate answers to questions of central interest. An analogy with surveying may be helpful. There are usually many different ways of making a land survey—many different angles and lengths we may measure. When we design the survey we consider not only the accuracy and precision with which these individual measurements can be made but also how they can be combined to give an accurate plot of the area surveyed (Lindley, Tversky, & Brown, 1979). Singer (1971) shows how a mental experiment may be designed to give a convincing estimate of the total value of property stolen by heroin addicts in New York City. Other examples of effective designs for mental experiments are given by Raiffa (1974).

One way to evaluate competing designs for physical measurement is to apply them to instances where the truth is known. But such empirical evaluation of final results is not always possible in the case of a mental experiment, especially when the experiment is designed to produce only probability judgments. It is true that probability judgments can be interpreted as frequences. But as we argue below, this interpretation amounts only to a comparison with a repeatable physical experiment where frequencies are known. How the comparison is made—what kind of repetitions are envisaged—is itself one of the choices we make in designing a mental experiment. There may not be a single set of repetitions to which the design must be referred for empirical validation.

Since empirical validation of a design for probability judgment is problematic, the result of carrying out the mental experiment must be scrutinized in other ways. The result of the whole experiment must be regarded as an argument, which, like all other arguments, is open to criticism and counterarguments.

Understanding and evaluating a design for probability judgment is also complicated by problems of meaning. When we are simply guessing the answer to a question of fact, such as the number of eggs produced daily in the U.S., the meaning of the question seems to be independent of our design. But when we undertake to make probability judgments, we find that we need a theory of subjective probability to give meaning to these judgments.

In the first place, we need a numerical scale or at least a qualitative scale (practically certain, very probable, fairly probable, etc.) from which to choose degrees of probability. We also need canonical examples for each

degree of probability on this scale—examples where it is agreed what degree of probability is appropriate. Finally, we need a calculus—a set of rules for combining simple judgments to obtain complex ones.

Using a theory of subjective probability means comparing the evidence in a problem with the theory's scale of canonical examples and picking out the canonical example that matches it best. Our design helps us make this comparison. It specifies how to break the problem into smaller problems that can be more easily compared with the scale of canonical examples and how to combine the judgments resulting from these separate comparisons.

Thought of in this way, a theory of subjective probability is very much like a formal language. It has a vocabulary—a scale of degrees of probability. Attached to this vocabulary is a semantics—a scale of canonical examples that show how the vocabulary is to be interpreted and psychological devices for making the interpretation effective. Elements of the vocabulary are combined according to a syntax—the theory's calculus.

Proponents of different theories of subjective probability have often debated which theory best describes human inductive competence. We believe that none of these theories provide an adequate account of people's intuitive judgments of probability. On the other hand, most of these theories can be learned and used effectively. Consequently, we regard these theories as formal languages for expressing probability judgments rather than as psychological models, however idealized.

The usefulness of one of these formal languages for a specific problem may depend both on the problem and on the skill of the user. There may not be a single probability language that is normative for all people and all problems. A person may find one language better for one problem and another language better for another. Furthermore, individual probability judgments made in one language may not be directly translatable into another.

This article studies the semantics and syntax of two probability languages, the traditional Bayesian language and the language of belief functions, and it uses these languages to analyze several concrete examples. This exercise can be regarded as a first step toward the general study of design for probability judgment. It illustrates the variety of designs that may be feasible for a given problem, and it yields a classification of Bayesian designs that clarifies the role of Bayesian conditioning. Our treatment is incomplete, however, because it does not provide formal criteria or lay out general empirical procedures for evaluating designs. The choice of design is left to the ingenuity of the user.

1. EXAMPLES

With the help of some simple examples we illustrate several designs for probability judgments. We will return to these examples in sections 3 and 4.

The Free-Style Race

We are watching one of the last men's swim meets of the season at Holsum University. We have followed the Holsum team for several seasons, so we watch with intense interest as Curt Langley, one of Holsum's leading free-stylers, gets off to a fast start in the 1650-yard race. As Curt completes his first 1000 yards, he is swimming at a much faster pace than we have seen him swim before. His time for the first 1000 yards is 9 min and 25 s. His best previous times for 1650 yards have been around 16 min and 25 s, a time that translates into about 9 min and 57 s at 1000 yards. The only swimmer within striking distance of him is a member of the visiting team named Cowan, whom we know only by name. Cowan is about half a lap (about 12 yards or 7 s) behind Curt.

Will Curt Win the Race? The first question we ask ourselves is whether he can keep up his pace. Curt is known to us as a very steady swimmer—one who knows what he is capable of and seldom, if ever, begins at a pace much faster than he can keep up through a race. It is true that his pace is much faster than we have seen before—much faster than he was swimming only a few weeks ago. It is possible that there has been no real improvement in his capacity to swim—that he has simply started fast and will slow down before the race is over. But our knowledge of Curt's character and situation encourages us to think that he must have trained hard and greatly increased his endurance. This is his senior year, and the championships are near. And he must have been provoked to go all out by Jones, the freshman on the team, who has lately overshadowed him in the long-distance races. We are inclined to think that Curt will keep up his pace.

If Curt does keep up his pace, then it seems very unlikely that Cowan could have enough energy in reserve to catch him. But what if Curt cannot keep up his pace? Here our vision becomes more murky. Has Curt deliberately put his best energy into the first part of the race? Or has he actually misjudged what pace he can keep up? In the first case, it seems likely he will soon slow down, but not to a disastrously slow pace; it seems to be a toss-up whether Cowan will catch him. On the other hand, if he has misjudged what pace he can keep up, then surely he has not misjudged it by far, and so we would expect him to keep it up almost to the end and, as usually happens in such cases, "collapse" with exhaustion to a very slow pace. There is no telling what would happen then—whether Cowan would be close enough or see the collapse soon enough to take advantage of the situation.

There are many different designs that we might use to assess numerically the probability of Curt's winning. There is even more than one possible Bayesian design. The Bayesian design suggested by our qualitative discussion assesses the probabilities that Curt will keep up the pace, slow down, or collapse and the conditional probabilities that he will win under each of

these hypotheses and then combines these probabilities and conditional probabilities to obtain his overall probability of winning. We call this a *total-evidence design* because each probability and conditional probability is based on the total evidence. In section 3 we will formalize and carry out this total-evidence design. We will also carry out a somewhat different Bayesian total-evidence design for the problem. In section 4 we will carry out a belief-function design for the problem.

The Hominids of East Turkana

In the August, 1978 issue of *Scientific American,* Alan Walker and Richard E. T. Leakey discuss the hominid fossils that have recently been discovered in the region east of Lake Turkana in Kenya. These fossils, between a million and two million years of age, show considerable variety, and Walker and Leakey are interested in deciding how many distinct species they represent.

In Walker and Leakey's judgment, the relatively complete cranium specimens discovered in the upper member of the Koobi Fora Formation in East Turkana are of three forms: (I) A "robust" form with large cheek teeth and massive jaws. These fossils show wide-fanning cheekbones, very large molar and premolar teeth, and smaller incisors and canines. The brain case has an average capacity of about 500 cubic centimenters, and there is often a bony crest running fore and aft across its top, which presumably provided greater area for the attachment of the cheek muscles. Fossils of this form have also been found in South Africa and East Asia, and it is generally agreed that they should all be classified as members of the species *Australopithecus robustus.* (II) A smaller and slenderer (more "gracile") form that lacks the wide-flaring cheekbones of I, but has similar cranial capacity and only slightly less massive molar and premolar teeth. (III) A large-brained (c. 850 cubic cm) and small-jawed form that can be confidently identified with the *Homo erectus* specimens found in Java and northern China.

The placement of the three forms in the geological strata in East Turkana shows that they were contemporaneous with each other. How many distinct species do they represent? Walker and Leakey admit five hypotheses:

1. I, II, and III are all forms of a single, extremely variable species.
2. There are two distinct species: one, *Australopithecus robustus,* has I as its male form and II as its female form; the other, *Homo erectus,* is represented by III.
3. There are two distinct species: one, *Australopithecus robustus,* is represented by I; the other has III, the so-called *Homo erectus* form, as its male form, and II as its female form.

4. There are two distinct species: one is represented by the gracile-form II; the other, which is highly variable, consists of I and III.
5. The three forms represent three distinct species.

Here are the items of evidence, or arguments, that Walker and Leakey use in their qualitative assessment of the probabilities of these five hypotheses:

(i) Hypothesis 1 is supported by general theoretical arguments to the effect that distinct hominid species cannot coexist after one of them has acquired culture.
(ii) Hypotheses 1 and 4 are doubtful because they postulate extremely different adaptations within the same species: The brain seems to overwhelm the chewing apparatus in III, while the opposite is true in I.
(iii) There are difficulties in accepting the degree of sexual dimorphism postulated by hypotheses 2 and 3. Sexual dimorphism exists among living anthropoids, and there is evidence from elsewhere that hints that dental dimorphism of the magnitude postulated by hypothesis 2 might have existed in extinct hominids. The dimorphism postulated by hypothesis 3, which involves females having roughly half the cranial capacity of males, is less plausible.
(iv) Hypotheses 1 and 4 are also impugned by the fact that specimens of type I have not been found in Java and China, where specimens of type III are abundant.
(v) Hypotheses 1 and 3 are similarly impugned by the absence of specimens of type II in Java and China.

Before specimens of type III were found in the Koobi Fora Formation, Walker and Leakey thought it likely that the I and II specimens constituted a single species. Now on the basis of the total evidence, they consider hypothesis 5 the most probable.

What Bayesian design might we use to analyze this evidence? A total-evidence design may be possible, but it is natural to consider instead a design in which some of the evidence is treated as an "observation" and used to "condition" probabilities based on the rest of the evidence. We might, for example, first construct a probability distribution that includes probabilities for whether specimens of Type I and II should occur in Java and China and then condition this distribution on their absence there. It is natural to call this a *conditioning design*. It is not a total-evidence design, because the initial (or "prior") probabilities for whether the specimens occur in Java and China will be based on only part of the evidence.

Later in section 3, we will work this conditioning design out in detail. In section 4 we will apply a belief-function design to the same problem.

2. TWO PROBABILITY LANGUAGES

In order to make numerical probability judgments, we need a numerical scale. We need, in other words, a scale of canonical examples in which numerical degrees of belief are agreed upon. Where can we find such a scale?

The obvious place to look is in the picture of chance. In this picture, we imagine a game which can be played repeatedly and for which we know the chances. These chances, we imagine, are facts about the world: they are long-run frequencies, they can be thought of as propensities, and they also define fair betting rates—rates at which a bettor would break even in the long run.

There are several ways the picture of chance can be related to practical problems, and this means we can use the picture to construct different kinds of canonical examples and thus, different theories or probability languages. In this essay, we shall consider two such languages: the Bayesian language, and the language of belief functions. The Bayesian language uses a scale of canonical examples in which the truth is generated by chance and our evidence consists of complete knowledge of the chances. The language of belief functions uses a scale of canonical examples in which our evidence consists of a message whose meaning depends on known chances.

We emphasize the Bayesian language because it is familiar to most readers. We study the language of belief functions as well in order to emphasize that our constructive view of probability, while not implying that all probability languages have equal normative claims, leaves open the possibility that no single language has a preemptively normative status.

2.1 The Bayesian Language

As we see it, a user of the Bayesian probability language makes probability judgments in a particular problem by comparing the problem to a scale of examples in which the truth is generated according to known chances and deciding which of these examples is most like the problem. The probability judgment $P(A) = p$, in this language, is a judgment that the evidence provides support for A comparable to what would be provided by knowledge that the truth is generated by a chance setup that produces a result in A exactly p of the time. This is not to say that one judges the evidence to be just like such knowledge in all respects, nor that the truth is, in fact, generated by chance. It is just that one is measuring the strength of the evidence by comparing it to a scale of chance setups.

The idea that Bayesian probability judgment involves comparisons with examples where the truth is generated by chance is hardly novel. It can be found, for example, in Bertrand (1907) and in Box (1980). Box states that

the adoption of given Bayesian probability distribution means that "current belief...would be calibrated with adequate approximation by a *physical stimulation* involving random sampling" (p. 385) from the distribution. The Bayesian literature has not, however, adequately addressed the question of how this comparison can be carried out. One reason for this neglect may be the emphasis that twenthieth-century Bayesians have put on betting. When "personal probabilities" are defined in terms of a person's preferences among bets, we are tempted to think that the determination of probabilities is a matter of introspection rather than a matter of examining evidence, but see Diaconis and Zabell (1982).

Bayesian Semantics. The task of Bayesian semantics is to render the comparison of our evidence to the Bayesian scale of canonical examples effective—to find ways of making the scale of chances and the affinity of our evidence to it vivid enough to our imagination that we can meaningfully locate the evidence on the scale.

By concentrating on different aspects of the rich imagery of games of chance, we can isolate different ways of making the Bayesian scale of chances vivid, and each of these ways can be thought of as a distinct semantics for the Bayesian probability language. Three such semantics come immediately to mind: a frequency semantics, a propensity semantics, and a betting semantics. The *frequency semantics* compares our evidence to the scale of chances by asking how often, in situations like the one at hand, the truth would turn out in various ways. The *propensity semantics* makes the comparison by first interpreting the evidence in terms of a causal model and then asking about the model's propensity to produce various results. The *betting semantics* makes the comparison by assessing our willingness to bet in light of the evidence: at what odds is our attitude towards a given bet most like our attitude towards a fair bet in a game of chance?

It is traditional, of course, to argue about whether probability should be given a frequency, a propensity, or a betting interpretation. But from our perspective these "interpretations" are merely devices to help us make what may ultimately be an imperfect fit of our evidence to a scale of chances. Which of these devices is most helpful will depend on the particular problem. We do not insist that there exists, prior to our deliberation, some particular frequency or numerical propensity in nature or some betting rate in our mind that should be called the probability of the proposition we are considering.

Which of these three Bayesian semantics tends to be most helpful in fitting our evidence to the scale of chances? We believe that the frequency and propensity semantics are central to the successful use of the Bayesian probability language, and that the betting semantics is less useful. Good Bayesian designs ask us to make probability judgments that can be translated into well-founded judgments about frequencies or about causal structures.

Since we readily think in terms of causal models, the propensity semantics often seems more attractive than the frequency semantics. But this attraction has its danger; the vividness of causal pictures can blind us to doubts about their validity. A simple design based on frequency semantics can sometimes be superior to a more complex design based on propensity semantics. We may, for example, obtain a better idea about how long it will take to complete a complex project by taking an "outside view" based on how long similar projects have taken in the past than by taking an "inside view" that attempts to assess the strength of the forces that could delay the completion of the project (Kahneman & Tversky, 1982).

The betting semantics has a generality that the frequency and propensity semantics lack. We can always ask ourselves about our attitude towards a bet, quite irrespective of the structure of our evidence. But this lack of connection with the evidence is also a weakness of the betting semantics.

In evaluating the betting semantics, one must distinguish logical from psychological and practical considerations. Ramsey (1931), Savage (1954), and their followers have made an important contribution to the logical analysis of subjective probability by showing that it can be derived from coherent preferences between bets. This logical argument, however, does not imply psychological precedence. Introspection suggests that people typically act on the basis of their beliefs, rather than form beliefs on the basis of their acts. The gambler bets on Team A rather than on Team B because he believes that A is more likely to win. He does not usually infer such a belief from his betting preferences.

It is sometimes argued that the prospect of monetary loss tends to concentrate the mind and thus permits a more honest and acute assessment of the strength of evidence than that obtained by thinking about that evidence directly. There is very little empirical evidence to support this claim. Although incentives can sometimes reduce careless responses, monetary payoffs are neither necessary nor sufficient for careful judgment. In fact, there is evidence showing that people are sometimes willing to incur monetary losses in order to report what they believe (Lieblich & Lieblich, 1969). Personally, we find that questions about betting do not help us think about the evidence; instead they divert our minds to extraneous questions: our attitudes towards the monetary and social consequences of winning or losing a bet, our assessment of the ability and knowledge of our opponent, etc.

Bayesian Syntax. It follows from our understanding of the canonical examples of the Bayesian language that this language's syntax is the traditional probability calculus. A proposition that a person knows to be false is assigned probability zero. A proposition that a person knows to be true is assigned probabilty one. And in general probabilities add: if A and B are incompatible propositions, then $P(A \text{ or } B) = P(A) + P(B)$.

The conditional probability of A given B is, by definition,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}. \tag{1}$$

If $B_1, \ldots, B_n$ are incompatible propositions, one of which must be true, then *the rule of total probability* says that

$$P(A) = \sum_{j=1}^{n} P(B_j)P(A|B_j), \tag{2}$$

and Bayes's theorem says that

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}. \tag{3}$$

As we shall see in section 3, both total-evidence and conditioning designs can use the concept of conditional probability. Total-evidence designs often use (2), while conditioning designs use (1). Some conditioning designs can be described in terms of (3).

2.2 The Language of Belief Functions

The language of belief functions uses the calculus of mathematical probability, but in a different way than the Bayesian language does. Whereas the Bayesian language asks, in effect, that we think in terms of a chance model for the facts in which we are interested, the belief-function language asks that we think in terms of a chance model for the reliability and meaning of our evidence.

This can be put more precisely by saying that the belief-function language compares evidence to canonical examples of the following sort. We know a chance experiment has been carried out. We know that the possible outcomes of the experiment are $o_1, \ldots, o_n$ and that the chance of o_i is p_i. We are not told the actual outcome but we receive a message concerning another topic that can be fully interpreted only with knowledge of the actual outcome. For each i there is a proposition A_i, say, such that if we knew the actual outcome was o_i then we would see that the meaning of the message is A_i. We have no other evidence about the truth or falsehood of the A_i and so no reason to change the probabilities p_i.

What degrees of belief are called for in an example of this sort? How strongly should be believe a particular proposition of A?

For each proposition A, set $m(A) = \sum \{p_i | A_i = A\}$. This number is the

total of the chances for outcomes that would show the message to mean A; we can think of it as the total chance that the message means A. Now let Bel(A) denote the total chance that the message implies A; in symbols, $\text{Bel}(A) = \sum\{m(B)|B \text{ implies } A\}$. It is natural to call Bel(A) our degree of belief in A.

We call a function Bel a *belief function* if it is given by the above equation for some choice of m(A). By varying the p_i and the A_i in our story of the uncertain message, we can obtain any such values for the m(A), and so the story provides canonical examples for every belief function.

We call the propositions A for which $m(A) > 0$ the *focal elements* of the belief function Bel. Often the most economical way of specifying a belief function is to specify its focal elements and their "m-values."

Semantics for Belief Functions. We have based our canonical examples for belief functions on a fairly vague story: We receive a message and we see, somehow, that if o_i were the true outcome of the random experiment, then the message would mean A_i. One task of semantics for belief functions is to flesh out the story in ways that help us compare real problems to it. Here we shall give three ways of fleshing out the story. The first leads to canonical examples for a small class of belief functions, called *simple support functions.* The second leads to canonical examples for a larger class, the *consonant support functions.* The third leads to canonical examples for arbitrary belief functions.

(i) A Sometimes Reliable Truth Machine. Imagine a machine that has two modes of operation. We know that in the first mode it broadcasts truths. But we are completely unable to predict what it will do when it is in the second mode. We also know that the choice of which mode the machine will operate in on a particular occasion is made by chance: There is a chance s that it will operate in the first mode and a chance 1-s that it will operate in the second mode.

It is natural to say of a message broadcast by such a machine on a particular occasion that it has a chance s of meaning what it says and a chance 1-s of meaning nothing at all. So if the machine broadcasts the message that E is true, then we are in the setting of our general story: The two modes of operation for the machine are the two outcomes o_1 and o_2 of a random experiment; their chances are $p_1 = s$ and $p_2 = 1\text{-}s$; if o_1 happened then the message means $A_1 = E$, while if o_2 happened the message means nothing beyond what we already know—i.e., it means $A_2 = \Theta$, where Θ denotes the proposition that asserts the facts we already know. So we obtain a belief function with focal elements E and Θ; $m(E) = s$ and $m(\Theta) = 1\text{-}s$.

We call such a belief function a *simple support function.* Notice its nonadditivity: the two complementary propositions E and not E have degrees of belief $\text{Bel}(E) = s < 1$ and $\text{Bel}(\text{not}E) = 0$.

It is natural to use simple support functions in cases where the message of the evidence is clear but where the reliability of this message is in question. The testimony of a witness, for example, may be unambiguous, and yet we may have some doubt about the witness's reliability. We can express this doubt by comparing the witness to a truth machine that is less than certain to operate correctly.

(ii) A Two-stage Truth Machine. Consider a sometimes reliable truth machine that broadcasts two messages in succession and can slip into its untrustworthy mode before either message. It remains in the untrustworthy mode once it has slipped into it. As before, we know nothing about whether or how often it will be truthful when it is in this mode. We know the chances that it will slip into its untrustworthy mode: r_1 is the chance it will be in untrustworthy mode with the initial message, and r_2 is the chance it will slip into untrustworthy mode after the first message, given that it was in trustworthy mode then.

Suppose the messages received are E_1 and E_2, and suppose these messages are consistent with each other. Then there is a chance $(1-r_1)(1-r_2)$ that the message "E_1 and E_2" is reliable, a chance $(1-r_1)r_2$ that the message "E_1" alone is reliable, and a chance r_1 that neither of the messages is reliable. If we set

$$\begin{aligned} p_1 &= (1-r_1)(1-r_2), & A_1 &= E_1 \,\&\, E_2, \\ p_2 &= (1-r_1)r_2, & A_2 &= E_1, \\ p_3 &= r_1, & A_3 &= \Theta, \end{aligned}$$

then we are in the setting of our general story: there is a chance p_i that the messages mean A_i.

Notice that A_1, A_2, and A_3 are "nested": A_1 implies A_2, and A_2 implies A_3. In general, we call a belief function with nested focal elements a *consonant support function.* It is natural to use consonant support functions in cases where our evidence consists of an argument with several steps; each step leads to a more specific conclusion but involves a new chance of error.

(iii) A Randomly Coded Message. Suppose someone chooses a code at random from a list of codes, uses the chosen code to encode a message, and then sends us the results. We know the list of codes and the chance of each code being chosen—say the list is $o_1, \ldots, o_n$, and the chance of o_i being chosen is p_i. We decode the message using each of the codes and we find that this always produces an intelligible message. Let A_i denote the message we get when we decode using o_i. Then we have the ingredients for a belief function: a message that has the chance p_i of meaning A_i.

Since the randomly coded message is more abstract than the sometimes reliable truth machine, it lends itself less readily to comparison with real evidence. But it provides a readily understandable canonical example for an arbitrary belief function. (For other scales of canonical examples for belief functions, see Krantz and Miyamoto, 1983 and Wierzchon, 1984.)

Syntax for Belief Functions. Our task, when we assess evidence in the language of belief functions, is to compare that evidence to examples where the meaning of a message depends on chance and to single out from these examples the one that best matches it in weight and significance. How do we do this? In complicated problems we cannot simply look at our evidence holistically and write down the best values for the $m(A)$. The theory of belief functions provides, therefore, a set of rules for constructing complicated belief functions from more elementary judgments. These rules, which ultimately derive from the traditional probability calculus, constitute the syntax of the language of belief functions. They include rules for combination, conditioning, extension, conditional embedding, and discounting.

The most important of these rules is Dempster's rule of combination. This is a formal rule for combining a belief function constructed on the basis of one item of evidence with a belief function constructed on the basis of another, intuitively independent item of evidence so as to obtain a belief function representing the total evidence. It permits us to break down the task of judgment by decomposing the evidence.

Dempster's rule is obtained by thinking of the chances that affect the meaning or reliability of the messages provided by different sources of evidence as independent. Consider, for example, two independent witnesses who are compared to sometimes reliable truth machines with reliabilities s_1 and s_2, respectively. If the chances affecting their testimonies are independent, then there is a chance s_1s_2 that both will give trustworthy testimony, and a chance $s_1 + s_2 - s_1s_2$ that at least one will. If both testify to the truth of A, then we can take $s_1 + s_2 - s_1s_2$ as our degree of belief in A. If, on the other hand, the first witness testifies for A and the second testifies against A, then we know that not both witnesses are trustworthy, and so we consider the conditional chance that the first witness is trustworthy given that not both are: $s_1(1 - s_2)/(1 - s_1s_2)$, and we take this as our degree of belief in A. For further information on the rules for belief functions, see Shafer (1976, 1982b).

3. BAYESIAN DESIGN

We have already distinguished two kinds of Bayesian designs: *total-evidence* designs, in which all one's probability judgments are based on the total

evidence, and *conditioning* designs, in which some of the evidence is taken into account by conditioning. In this section we will study these broad categories and consider some other possibilities for Bayesian design.

3.1 Total-Evidence Designs

There are many kinds of probability judgments a total-evidence design might use, for there are many mathematical conditions that can help determine a probability distribution. We can specify quantities such as probabilities, conditional probabilities and expectations, and we can impose conditions such as independence, exchangeability, and partial exchangeability. Spetzler and Staël von Holstein (1975), Alpert and Raiffa (1982), and Goldstein (1981) discuss total-evidence designs for the construction of probability distributions for unknown quantities. Here we discuss total-evidence designs for a few simple problems.

Two Total-Evidence Designs for the Free-Style Race. The Bayesian design for the free-style race suggested by our discussion in section 1.2 above is an example of a total-evidence design based on a causal model. This design involves six possibilities:

A_1 = Curt maintains the pace and wins.
A_2 = Curt maintains the pace but loses.
A_3 = Curt soon slows down but still wins.
A_4 = Curt soon slows down and loses.
A_5 = Curt collapses at the end but still wins.
A_6 = Curt collapses at the end and loses.

The person who made the analysis (the story was reconstructed from actual experience) was primarily interested in the proposition

$$A = \{A_1 \text{ or } A_3 \text{ or } A_5\} = \text{Curt wins},$$

but her insight into the matter was based on her understanding of the causal structure of the swim race. In order to make the probability judgment P(A), she first made the judgments $P(B_i)$ and $P(A|B_i)$, where

$B_1 = \{A_1 \text{ or } A_2\}$ = Curt maintains his pace,
$B_2 = \{A_3 \text{ or } A_4\}$ = Curt soon slows down,
$B_3 = \{A_5 \text{ or } A_6\}$ = Curt collapses near the end,

and she then calculated P(A) using the rule of total probability—in this case, the formula

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3). \quad (4)$$

She did this qualitatively at the time, but she offers, in retrospect, the quantitative judgments indicated in Table I. These numbers yield $P(A) = .87$ by (4).

This example brings out the fact that the value of a design depends on the experience and understanding of the person carrying out the mental experiment. For someone who lacked our analyst's experience in swimming and her familiarity with Curt Langley's record, the design (4) would be worthless. Such a person might find some other Bayesian design useful, or he/she might find all Bayesian designs difficult to apply.

TABLE I
Component Judgments for the First Total-Evidence Design

$P(B_1) = .8$	$P(A|B_1) = .95$
$P(B_2) = .15$	$P(A|B_2) = .5$
$P(B_3) = .05$	$P(A|B_3) = .7$

Though it is correct to call the design we have just studied a total-evidence design, there is a sense in which its effectiveness does depend on the fact that it allows us to decompose our evidence. The question of what the next event in a causal sequence is likely to be is often relatively easy to answer precisely because only a small part of our evidence bears on it. When we try to decide whether Curt will still win if he slows down—i.e., when we assess $P(A|B_2)$—we are able to leave aside our evidence about Curt and focus on how likely Cowan is to maintain his own pace.

Here is another total-evidence design for the free-style race, one which combines the causal model with a more explicit judgment that Cowan's ability is independent of Curt's behavior and ability. We assess probabilities for whether Curt will (a) maintain his pace, (b) slow down, but less than 3%, (c) slow down more than 3%, or (d) collapse. (Whether Curt slows down 3% is significant because this is how much he would have to slow down for Cowan to catch him without speeding up.) We assess probabilities for whether Cowan (a) can speed up significantly, (b) can only maintain his pace, (c) cannot maintain his pace. We judge that these two questions are independent. And finally, we assess the probability that Curt will win under each of the $4 \times 3 = 12$ hypotheses about what Curt will do and what Cowan can do.

Table II shows the results of carrying out this design. The numbers in the vertical margin are our probability judgments about Curt, those in the horizontal margin are our probability judgments about Cowan, and those in the cells are our assessments of the conditional probability that Curt will win. These numbers lead to an overall probability of

$$(.85 \times .10 \times .5) + (.85 \times .70 \times 1.0) + \cdots \approx .88$$

that Curt will win.

Our judgments about Cowan are based on our general knowledge about swimmers in the league. The numbers .10, .70, and .20 reflect our impression that perhaps 20% of these swimmers are forced to slow down in the second half of a 1650-yard race and that only 10% would have the reserves of energy needed to speed up. We are, in effect, thinking of Cowan as having been chosen at random from this population. We are also judging that Curt's training and strategy are independent of this random choice. Curt's training has probably been influenced mainly by the prospect of the championships. We doubt that Cowan's ability and personality are well enough known to Curt to have caused him to choose a fast start as a strategy in this particular race.

TABLE II
Component Judgments for the Second Total-Evidence Design

		Cowan		
		Can Speed Up Significantly	Can Only Maintain Pace	Cannot Maintain Pace
Curt		.10	.70	.20
Maintains pace	.85	0.5	1.0	1.0
Slows less than 3%	.03	0.2	1.0	1.0
Slows 3% or more	.07	0.0	0.0	0.5
Collapses	.05	0.2	0.7	0.8

When we compare the design and analysis of Table II with the design we carried out earlier, we see that we have profited from the new design's focus on our evidence about Cowan. We feel that the force and significance of this evidence is now more clearly defined for us. On the other hand, we are less comfortable with the conditional probability judgments in the cells of Table II; some of these seem to be pure speculation rather than assessments of evidence.

Total-Evidence Designs Based on Frequency Semantics. In the two designs we have just considered the breakdown into probabilities and conditional probabilities was partly determined by a causal model. In designs that depend more heavily on frequency semantics, this breakdown depends more on the way our knowledge of past instances is organized.

Consider, for example, the problem of deciding what is wrong when an automobile fails to start. If a mechanic were asked to consider the possible causes for this failure, he might first list the major systems that could be at fault, (fuel system, ignition system, etc.) and then list more specific possible defects within each system. This would result in a "fault tree" that

could be used to construct probabilities. The steps in the tree would not have a causal interpretation, but the tree would correspond, presumably, to the way the mechanic's memory of the frequencies of similar problems is organized. Fischhoff, Slovic, and Lichtenstein (1978) have studied the problem of designing fault trees so as to make them as effective and unbiased as possible.

Here is another simple example based on an anecdote reported by Kahneman and Tversky (1982). An expert undertakes to estimate how long it will take to complete a certain project. He does this by comparing the project to similar past projects. And he organizes his effort to remember relevant information about these past projects into two steps: First he asks how often such projects were completed, and then he asks how long the ones that were completed tended to take. If he focuses on a particular probability judgment—"the probability that our project will be finished within 7 years" say—then he asks first how frequently such projects are completed and then how frequently projects that are completed take less than 7 years.

Why does the expert use this two-step design? Presumably because it facilitates his mental sampling of past instances. It is easier for the expert to thoroughly sample past projects he has been familiar with if he limits himself to asking as he goes only whether they were completed. He can then come back to the completed projects and attack the more difficult task of remembering how long they took.

The emphasis in this example is on personal memory. The lesson of the example applies, however, even when we are aided by written or electronic records. In any case, the excellence of a design depends in part on how the information accessible to us is organized.

3.2 Conditioning Designs

Bayesian conditioning designs can be divided into two classes: *observational* designs and *partitioning* designs. In observational designs, the evidence to be taken into account by conditioning is deliberately obtained after probabilities are constructed. In partitioning designs, we begin our process of probability judgment with all our evidence in hand, but we partition this evidence into "old evidence" and "new evidence," assess probabilities on the basis of the old evidence alone, and then condition on the new evidence.

It should be stressed that a conditioning design always involves two steps: constructing a probability distribution and conditioning it. The name "conditioning design" focuses our attention on the second step, but the first is more difficult. An essential part of any conditioning design is a subsidiary design specifying how the distribution to be conditioned is to be constructed. This subsidiary design may well be a total-evidence design.

Likelihood-Based Conditioning Designs. Bayesian authors often emphasize the use of Bayes's theorem. Bayes's theorem, we recall, says that if $B_1, \ldots, B_n$ are incompatible propositions, one of which must be true, then

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=i}^{n} P(B_j)P(A|B_j)}. \tag{5}$$

If A represents evidence we want to take into account, and if we are able to make the probability judgments on the right hand side of (5) while leaving this evidence out of account, then we can use (5) to calculate a probability for B_i.

When we use Bayes's theorem in this simple way, we are carrying out a conditioning design. Leaving aside the "new evidence" A, we use the "old evidence" to make probability judgments $P(B_i)$ and $P(A|B_i)$. Making these judgments amounts to constructing a probability distribution. We then condition this distribution on A. Formula (5) is simply a convenient way to calculate the resulting conditional probability of B_i.

This is a particular kind of conditioning design. The subsidiary design that we are using to construct the probability distribution to be conditioned is a total-evidence design that just happens to focus on the probabilities $P(B_i)$ and $P(A|B_i)$, where A is the new evidence and the B_i are the propositions whose final probabilities interest us. Since the conditional probabilities $P(A|B_i)$ are called "likelihoods," we may call this kind of conditioning design a *likelihood-based* conditioning design.

Both observational and partitioning designs may be likelihood-based. Bayesian theory has traditionally emphasized likelihood-based conditioning designs, and they will also be emphasized in this section. At the end of the section, however, we will give an example of a conditioning design that is not likelihood-based.

A Likelihood-Based Observational Design: The Search for Scorpion. The successful search for the remains of the submarine *Scorpion*, as reported by Richardson and Stone (1971), provides an excellent sample of a likelihood-based observational design. The search was conducted from June to October, 1968, in an area about 20 miles square located 400 miles southwest of the Azores. The submarine was found on October 28.

Naval experts began their probability calculations by using a causal model to construct a probability distribution for the location of the lost submarine. They developed nine scenarios for the events attending the disaster and assigned probabilities to those scenarios. They then combined these probabilities with conditional probabilities representing uncertainties in the submarine's course, speed, and initial position to produce a probability distribution for its final location on the ocean floor. They did not attempt to

construct this probability distribution for the final location in continuous form. Instead, they imposed a grid over the search area with cells about one square mile in size and used their probabilities and conditional probabilities in a Monte Carlo simulation to estimate the probability of *Scorpion* being in each of these approximately 400 cells. They then used these probabilities to plan the search: The cells with the greatest probability of containing *Scorpion* were to be searched first.

Searching a cell meant towing through the cell near the ocean bottom a platform upon which were mounted cameras, magnetometers, and sonars. The naval experts assessed the probability that this equipment would detect *Scorpion* if Scorpion were in the cell searched. So when they searched a cell and conditioned on the fact that *Scorpion* was not found there, they were, in effect, using a likelihood-based conditioning design to assess new probabilities for its location.

This example is typical of likelihood-based observational designs. The probabilities required by the design were subjective judgments, not known objective probabilities. (The assessed likelihood of detecting *Scorpion* when searching the cell where it was located turned out, for example, to be over optimistic.) But these judgments were made before the observation on which the experts conditioned was made. In fact, these judgments were the basis of deciding which of several possible observations to make—i.e., which cell to search.

A Likelihood-Based Partitioning Design: The Hominids of East Turkana. Let us now turn back to Walker and Leakey's discussion of the number of species of hominids in East Turkana one and a half million years ago. They begin, we recall, by taking for granted a classification of the hominids into three types: the "robust" type I, the "gracile" type II, and *Homo erectus,* type III. They were interested in five hypotheses as to how many distinct species these three types represent:

B_1 = One species.
B_2 = Two species, one composed of I (male) and II (female).
B_3 = Two species, one composed of III (male) and II (female).
B_4 = Two species, one composed of I and III.
B_5 = Three species.

We summarized the evidence they brought to bear on the problem under five headings:

(i) A theoretical argument for B_1.
(ii) Skepticism about such disparate types as I and III being variants of the same species.

(iii) Skepticism about the degree of sexual dimorphism postulated by B_2 and B_3.
(iv) Absence of type I specimens among the type III specimens in the Far East.
(v) Absence of type II specimens among the type III specimens in the Far East.

How might we assess this evidence in the Bayesian language?

Partitioning design seems to hold more promise in this problem than total-evidence design. Except for items (i) and possibly (ii), the evidence cannot be interpreted as an understanding of causes that generate the truth, and hence there is little prospect for a total-evidence design using propensity semantics. We also lack the experience with similar problems that would be required for a successful total-evidence design using frequency semantics. And since it is the diversity of the evidence that complicates probability judgments in the problem, a design that decomposes the evidence seems attractive.

Which of the items of evidence shall we classify as old evidence, and which as new? The obvious move is to classify (i) as old evidence and to treat (ii)-(v), taken together, as our new evidence A. This means we will need to assess probabilities, $P(B_1), \ldots, P(B_5)$ and conditional probabilities, $P(A|B_1), \cdots, P(A|B_5)$ and calculate $P(B_i|A)$, by (5). The apparent complexity of (5) is lessened if we divide it by the corresponding expression for B_j, obtaining

$$\frac{P(B_i|A)}{P(B_j|A)} = \frac{P(B_i)}{P(B_j)} \frac{P(A|B_i)}{P(A|B_j)}, \tag{6}$$

or

$$\frac{P(B_i|A)}{P(B_j|A)} = \frac{P(B_i)}{P(B_j)} L(A|B_i:B_j), \tag{7}$$

where $L(A|B_i:B_j) = P(A|B_i)/P(A|B_j)$ is called the *likelihood ratio* favoring B_i over B_j.

Expression (7) represents a real simplification of the design. Since the probabilities, $P(B_1|A), \ldots P(B_5|A)$ must add to one, they are completely determined by their ratios, $P(B_i|A)/P(B_j|A)$. Therefore, equation (7) tells us that it is not necessary to assess the likelihoods, $P(A|B_i)$ and $P(A|B_j)$. It is sufficient to assess their ratios, $L(A|B_i:B_j)$. (Cf. Edwards, Phillips, Hays, & Goodman, 1968).

One further elaboration of this design seems useful. Our new evidence A can be thought of as a conjuction: $A = A_1$ *and* A_2, where A is the event

that types I, II and III should be so disparate (items of evidence (ii) and (iii)) and A_2 is the event that specimens of types I and II should not be found along with the type III specimens in the Far East (items of evidence (iv) and (v). The two events A_1 and A_2 seem to involve independent uncertainties, and this can be expressed in Bayesian terms by saying that they are independent events conditional on any one of the five hypotheses:

$$P(A|B_i) = P(A_1|B_i)P(A_2|B_i).$$

Substituting this into (6), we obtain

$$\frac{P(B_i|A)}{P(B_j|A)} = \frac{P(B_i)}{P(B_j)} \frac{P(A_1|B_i)}{P(A_1|B_j)} \frac{P(A_2|B_i)}{P(A_2|B_j)},$$

or

$$\frac{P(B_i|A)}{P(B_j|A)} = \frac{P(B_i)}{P(B_j)} L(A_1|B_i:B_j)L(A_2|B_i:B_j).$$

We are not, of course, qualified to make the probability judgments called for by this design; it is a design for experts like Walker and Leakey, not a design for laymen. (If we ourselves had to make probability judgments about the validity of Walker and Leakey's opinions, we would need a design that analyzes our own evidence. This consists of their article itself, which provides internal evidence as to the integrity and the cogency of their thought, our knowledge of the standards of *Scientific American*, etc.) It will be instructive, nonetheless, to put ourselves in the shoes of Walker and Leakey and to carry out the design on the basis of the qualitative judgments they make in their article. As we shall see, there are several difficulties.

The first difficulty is in determining the prior probabilities $P(B_i)$ on the basis of the evidence (i) alone. This evidence is an argument for B_1, and so evaluation of it can justify a probability $P(B_1)$, say $p(B_1) = .75$. But how do we divide the remaining .25 among the other B_i? This is a typical problem in Bayesian design. In the absence of relevant evidence, we are forced to depend on symmetries, even though the available symmetries may seem artificial and conflicting. In this case, one symmetry suggests equal division among B_2, B_3, B_4, B_5, while another symmetry suggest equal division between the hypothesis of two species (B_2, B_3, B_4) and the hypothesis of three species (B_5). The $P(B_i)$ given in Table III represent a compromise.

Now consider A_1, the argument that the different types must represent three distinct species because of their diversity. Our design asks us, in effect, to assess how much less likely this diversity would be under the one-species hypothesis and under the various two-species hypotheses. Answers to these

questions are given in the column of Table III labeled "$L(A_1|B_i:B_5)$." These numbers reflect the great implausibility of the intraspecies diversity postulated by B_1 and B_4, the marginal acceptability of the degree of sexual dimorphism postulated by B_2, and the implausibility, especially in the putative ancestor of *Homo sapiens,* of the sexual dimorphism postulated by B_3. Notice how fortunate it is that we are required to assess only the likelihood ratios, $L(A_1|B_i:B_5) = P(A_1|B_i)/P(A_1|B_5)$ and not, say, the absolute likelihood $(P(A_1|B_5)$. We can think about how much less likely the observed disparity among the three groups would be if they represented fewer than three species, but we would be totally at sea if asked to assess the unconditional chance of this degree of disparity among three extinct hominid species.

TABLE III
Component Judgments for the Likelihood-Based Partitioning Design

	$P(B_i)$	$L(A_1\|B_i:B_5)$	$L(A_2\|B_i:B_5)$	$P(B_i\|A)$
B_1	.70	.01	.01	.00060
B_2	.05	.50	1.00	.19983
B_3	.05	.05	.01	.00020
B_4	.05	.01	.01	.00004
B_5	.10	1.00	1.00	.79933

Finally, consider A_2, the absence of specimens of type I or II among the abundant specimens of type III in the Far East. This absence would seem much less likely if I or II were forms of the same species as III than if they were not, say 100 times less likely. This is the figure used in Table III. Notice again that we are spared the well-nigh meaningless task of assessing absolute likelihoods.

As the last column of Table III shows, the total evidence gives a fairly high degree of support to B_5, the hypothesis that there are three distinct species. This is Walker and Leakey's conclusion.

How good an analysis is this? There seems to be two problems with it. First, we lack good grounds for some of the prior probability judgments. Second, the interpretation of the likelihoods seems strained. Are we really judging that the observed difference between I and III is 100 times more likely if they are separate species than if they are variants of the same species? Or are we getting this measure of the strength of this argument for separate species in some other way?

We should remark that it is a general feature of likelihood-based partitioning designs that only likelihood ratios need be assessed. In likelihood-based observational designs, on the other hand, we do usually need to assess absolute likelihoods. This is because in an observational design we must be prepared to condition on any of the possible observations. If, for example, the possible observations are A and *not* A, then we need to have in hand

both $L(A|B_i:B_j) = \frac{P(A|B_j)}{P(A|B_j)}$ and $L(not\ A|B_i:B_j) = \frac{P(not\ A|B_i)}{P(not\ A|B_j)}$. Since $P(A|B_i) + P(not\ A|B_i) = P(A|B_j) + P(not\ A|B_j) = 1$, these likelihood ratios fully determine the absolute likelihoods $P(A|B_i)$ and $P(A|B_j)$.

The Choice of New Evidence. Traditionally, Bayesian statistical theory has been concerned with what we have called likelihood-based observational designs. This is because the theory has been based on the idea of a statistical experiment. It is assumed that one knows in advance an "observation space"—the set of possible outcomes of the experiment—and a "parameter space"—the set of possible answers to certain questions of substantive interest. One assesses in advance both prior probabilities for the parameters and likelihoods for the observations.

Many statistical problems do conform to this picture. The search for Scorpion, discussed earlier, is one example. But Bayesians and other statisticians have gradually extended their concerns from the realm of planned experiments, where parameter and observation spaces are clearly defined before observations are made, to the broader field of "data analysis." In data analysis, the examination of data often precedes the framing of hypotheses and "observations." This means that the Bayesian data analyst will often use partitioning designs rather than genuine observational designs.

We believe that Bayesian statistical theory will better meet the needs of statistical practice if it will go beyond observational designs and deal explicitly with partitioning designs. In particular, we need more discussion of principles for the selection of evidence that is to be treated as new evidence. In the example of the hominids, we treated certain arguments as new evidence because we could find better grounds for probability judgment when thinking of the likelihood of their arising than when thinking about them as conditions affecting the likelihood of other events. In other cases, we may single out evidence because its psychological salience can give it excessive weight in total-evidence judgments. By putting such salient evidence in the role of new evidence in a partitioning design, we gain an opportunity to make probability judgments based on the other evidence alone. (Cf. Spetzler and Staël von Holstein, 1975, p. 346; and Nisbett and Ross, 1980, Chapter 3.) We need more discussion of such principles, and more examples.

A Partitioning Design that is not Likelihood-Based. Here is a problem that suggests a partitioning design that is not likelihood-based. Gracchus is accused of murdering Maevius. Maevius' death brought him a great and sorely needed financial gain, but it appears that Maevius and Gracchus were good friends, and our assessment of Gracchus' character suggests only a slight possibility that the prospect of gain would have been sufficient motive for him to murder Maevius. On the other hand, some evidence has come to

light to suggest that beneath the apparent friendship Gracchus actually felt a simmering hatred for Maevius, and Gracchus is known to be capable of violent behavior towards people he feels have wronged him. The means to commit the murder is not at issue: Gracchus or anyone else could have easily committed it. But we think it very unlikely that anyone else had reason to kill Maevius.

Our partitioning design uses the fact of Maevius' murder as the new evidence. We consider the propositions.

H = Gracchus hated Maevius,
GI = Gracchus intended to kill Maevius,
SI = Someone else intended to kill Maevius,
GM = Gracchus murdered Maevius,
SM = Someone else murdered Maevius,
NM = No one murdered Maevius.

Using the old evidence alone, we make the following probability judgments:

$P(H) = .2$, $P(GI|H) = .2$, $P(GI|\textit{not}\ H) = .01$;
$P(SI) = .001$, and SI is independent of GI;
$P(GM|GI\ \textit{and}\ SI) = .4$, $P(SM|GI\ \textit{and}\ SI) = .4$, $P(NM|GI\ \textit{and}\ SI) = .2$;
$P(GM|GI\ \textit{and not}\ SI) = .8$, $P(NM|GI\ \textit{and not}\ SI) = .2$;
$P(SM|SI\ \textit{and not}\ GI) = .8$, $P(NM|SI\ \textit{and not}\ GI) = .2$;
$P(NM|\textit{not}\ GI\ \textit{and not}\ SI) = 1$.

Combining these judgments, we obtain

$$\begin{aligned} P(GI) &= P(GI|H)P(H) + P(GI|\ \textit{not}\ H)P(\textit{not}\ H) \\ &= (.2)(.2) + (.8)(.01) = .048, \\ P(GM) &= P(GM|\ \textit{not}\ GI)P(\textit{not}\ GI) + P(GM|GI\ \textit{and}\ SI)P(GI)P(SI) \\ &\quad + P(GM|GI\ \textit{and not}\ SI)P(GI)P(\textit{not}\ SI) \\ &= (0)(.952) + (.4)(.048)(.001) + (.8)(.048)(.999) \\ &= .03838 \end{aligned}$$

Similarly,

$$P(SM) = .00078 \text{ and } P(NM) = .96084$$

Finally we bring in the new evidence—the fact that Maevius was murdered. We find a probability

$$P(GM|\textit{not}\ NM) = \frac{.03838}{.03838 + .00078} = .98$$

that Gracchus did it.

One interesting aspect of this example is the fact that the "new evidence"—the fact that Maevius was murdered—is actually obtained be-

fore much of the other evidence. Only after Maevius' death would we have gathered the evidence against Gracchus.

3.3 Other Bayesian Designs

What other Bayesian designs are possible in addition to total-evidence and conditioning design?

A large class of possible designs is suggested by the following general idea. Suppose one part of our evidence lends itself to a certain design d, while the remainder of our evidence does not fit this design, but seems instead relevant to some of the judgments specified by a different design d′. Then we might first construct a distribution P_o using d and considering only the first part of the evidence, and then switch to d′, using the total evidence to make those judgments for which the second part of the evidence is relevant and obtaining the other judgments from P_o.

An interesting special case occurs when the total evidence is used only to construct probabilities $p_1, \ldots, p_n$ for a set of mutually incompatible and collectively exhaustive propositions $A_1, \ldots, A_n$, and the final distribution P is determined by setting $P(A_i) = p_i$ and $P(B|A_i) = P_o(B|A_i)$ for all B. Since such designs were considered by Jeffrey (1965), we may call them *Jeffrey designs.*

Here is an example of a Jeffrey design. Gracchus is accused of murdering Maevius and the evidence against him is the same as in the preceding example, except that it is not certain that Maevius has been murdered. Perhaps Maevius has disappeared after having been seen walking along a sea cliff. We partition our evidence into two bodies of evidence—the evidence that was used in the probability analysis above, and the other evidence that suggests Maevius may have been murdered. We use the first body of evidence to make the analysis of the preceding section, obtaining the probabilities obtained there: a probability of .03838 that Gracchus murdered Maevius, a probability of .00078 that someone else did, and a probability of .96084 that no one did. We label this probability distribution P_o. Then we use the total evidence to assess directly whether we think Maevius has been murdered or not. Say we assess the probability of Maevius' having been murdered at .95. We then obtain a conditional probability from P_o:P_o (Graccus did it| Maevius was murdered) $\approx .98$. The final result is a probability of $.95 \times .98 \approx .93$ for the event that Gracchus murdered Maevius. For further examples of Jeffrey designs, see Shafer (1981b) and Diaconis and Zabell (1982).

4. BELIEF-FUNCTION DESIGN

Belief-function design differs from Bayesian design in that it puts more explicit emphasis on the decomposition of evidence. As we have seen, total-evidence designs are basic to the Bayesian language. (Even conditioning and

Jeffrey designs must have subsidiary designs for the construction of initial distributions, and these subsidiary designs are usually total-evidence designs.) These total-evidence designs break down the task of judgment by asking us to answer several different questions. It is a contingent matter whether different items of evidence bear on these different questions, though this seems to be the case with the most effective total-evidence designs. The belief-function language, on the other hand, since it directly models the meaning and reliability of evidence, breaks down the task of judgment by considering different items of evidence. It is a contingent matter whether these different items of evidence bear on relatively separate and restricted aspects of the questions that interest us, but again, as we shall see, this seems to be the case with the most effective belief-function designs.

Here we shall explore the possibilities for belief-function design for Curt's swim race and Walker and Leakey's hominids. For further examples of belief-function design, see Shafer (1981a, 1981b, 1982a, 1982b).

4.1 The Free-Style Race

The second of the two Bayesian total-evidence designs that we gave for the free-style race (section 3.1) was based on independent judgments about Curt and Cowan. We gave Curt an 85% chance of maintaining his pace, a 3% chance of slowing less than 3%, a 7% chance of slowing more than 3%, and a 5% chance of collapsing. And we gave Cowan a 10% chance of being able to speed up, a 70% chance of only being able to maintain his pace, and a 20% chance of being unable to maintain his pace. Since we are using the Bayesian language, we compared our evidence to knowledge that the evolution of the race actually was governed by these chances. It is equally convincing, however, to interpret these numbers within the language of belief functions. We compare our knowledge about Curt to a message that has an 85% chance of meaning that he will maintain his pace, etc., and we compare our knowledge about Cowan to a message that has a 70% chance of meaning that he can only maintain his pace, etc.

Formally, we have a belief function Bel_1 that assigns degrees of belief .85, .03, .07, and .05 to the four hypotheses about Curt, and a second-belief function Bel_2 that assigns degrees of belief .10, .70, and .20 to the three hypotheses about Cowan. Judging that our evidence about Curt is independent of our evidence about Cowan, we combine these by Dempster's rule. If no further evidence is added to the analysis, then our resulting degree of belief that Curt will win will be our degree of belief that Curt will maintain his pace or slow less than 3% while Cowan is unable to speed up: $(.85 + .03)(.70 + .20) = .792$. And our degree of belief that Cowan will win will be our degree of belief that Curt will slow 3% or more and Cowan will be able to at least maintain his pace: $(.07)(.10 + .70) = .056$.

These conclusions are weaker than the conclusions of the Bayesian analysis. This is principally due to the fact that we are not claiming to have evidence about what will happen in the cases where our descriptions of Curt's and Cowan's behavior do not determine the outcome of the race. If we did feel we had such evidence, it could be introduced into the belief-function analysis.

We can also relax the additivity of the degrees of belief about Curt and Cowan that go into the belief-function analysis. Suppose, for example, that we feel our evidence about Curt justifies only an 85% degree of belief that he will maintain his pace, but we do not feel we have any positive reason to think he will slow down or collapse. In this case, we can replace the additive degree of belief .85, .03, .07, and .05 with a simple support function that assigns only degree of belief .85 to the proposition that Curt will maintain his pace. If we retain the additive degrees of belief .10, .70, and .20 for Cowan's behavior, this leads to a degree of belief

$$(.85)(.70 + .20) = .765$$

that Curt will win and a degree of belief zero that Cowan will win.

As this example illustrates, a belief-function design can be based on a causal structure like those used in Bayesian total-evidence designs. The belief-function design must, however, go beyond this causal structure to an explicit specification of the evidence that bears on its different parts.

4.2 The Hominids of East Turkana

Recall that Walker and Leakey considered five hypotheses:

B_1 = One species.
B_2 = Two species, one composed of I (male) and II (female).
B_3 = Two species, one composed of III (male) and II (female).
B_4 = Two species, one composed of I and III.
B_5 = Three species.

In our Bayesian analysis in Section 3.2, we partioned the evidence into three intuitively independent arguments:

1. A theoretical argument for B_1.
2. An argument that the three types are too diverse not to be distinct species. This argument bears most strongly against B_1 and B_4, but also carries considerable weight against B_3 and some weight against B_2.

3. The fact that neither I nor II specimens have been found among the III specimens in the Far East. This provides evidence against hypotheses B_1, B_3, and B_4.

Let us represent each of these arguments by a belief function. Making roughly the same judgments as in the Bayesian analysis, we have

1. Bel_1, with $m_1(B_1) = .75$ and $m_1(\Theta) = .25$,
2. Bel_2, with $m_2(B_5) = .5$, $m_2(B_2 \textit{ or } B_5) = .45$, $m_2(B_2 \textit{ or } B_3 \textit{ or } B_5) = .04$, and $m_2(\Theta) = .01$, and
3. Bel_3, with $m_3(B_2 \textit{ or } B_5) = .99$ and $m_3(\Theta) = .01$.

Combining these by Dempster's rule, we obtain a belief function Bel with $m(B_5) = .4998$, $m(B_2 \textit{ or } B_5) = .4994$, $m(B_2 \textit{ or } B_3 \textit{ or } B_5) = .0004$, $m(B_1) = .0003$, and $m(\Theta) = .0001$. This belief function gives fair support to B_5 and overwhelming support to B_2 *or* B_5: Bel $(B_5) = .4998$ and Bel $(B_2 \textit{ or } B_5) = .9992$.

These belief-function results can be compared to the Bayesian results of section 3.2, where we obtained $P(B_5) = .7993$ and $P(B_2 \textit{ or } B_5) = .9992$. The different results for B_5 can be attributed to the different treatments of the first item of evidence, the argument against coexistence of hominid species. In the belief-function analysis, we treated this argument simply by giving B_1 a 75% degree of support. In the Bayesian analysis, we had to go farther and divide the remaining 25% among the other four hypotheses. The belief-function analysis, while it reaches basically the same conclusion as the Bayesian argument, can be regarded as a stronger argument, since it is based on slightly more modest assumptions.

5. THE NATURE OF PROBABILITY JUDGMENT

We have suggested that probability judgment is a kind of mental experiment. Sometimes it is more like a physicist's thought experiment, as when mind or on a bookshelf, for examples on which to base a frequency judgment. Sometimes it is more like a physicist's thought experiment, as when we try to trace the consequences of an imagined situation.

Probability judgment is a process of construction rather than elicitation. People may begin a task of probability judgment with some beliefs already formulated. But the process of judgment, when successful, gives greater content and structure to these beliefs and tends to render initial beliefs obsolete. It is useful, in this respect, to draw an analogy between probability and affective notions such as love and loyalty. A declaration of love is not simply a report on a person's emotions. It is also part of a process whereby an intellectual and emotional commitment is created; so too with probability.

A probability judgment depends not just on the evidence on which it is based, but also on the process of exploring that evidence. The act of designing a probability analysis usually involves reflection about what evidence is available and a sharpening of our definition of that evidence. And the implementation of a design involves many contingencies. The probability judgments we make may depend on just what examples we sampled from our memory or other records, or just what details we happen to focus on as we examine the possibility of various scenarios (Tversky & Kahneman, 1983).

It may be helpful to point out that we do not use the word "evidence" as many philosophers do—to refer to a proposition in a formal language. Instead, we use it in a way that is much closer to ordinary English usage. We refer to "our evidence about Cowan's abilities," to "our memory as to how frequently similar projects are completed," or to "the argument that distinct hominid species cannot coexist." The references are, as it were, ostensive definitions of bodies of evidence. They point to the evidence in question without translating it into statements of fact in some language. This seems appropriate, for in all these cases the evidence involves arguments and claims that would fall short of being accepted as statements of fact.

Evidence, as we use the word, is the raw material from which judgments, both of probability and of fact, are made. Evidence can be distinguished in this respect from information. Information can be thought of as answers to questions already asked, and hence we can speak of the quantity of information, which is measured by the number of these questions that are answered. Evidence, in contrast, refers to a potential for answering questions. We can speak of the weight of evidence as it bears on a particular question, but it does not seem useful to speak of the quantity of evidence.

Though we have directed attention to the notion of mental experimentation, we want also to emphasize that when an individual undertakes to make a probability judgment that individual is not necessarily limited to the resources of memory and imagination. He or she may also use paper, pencils, books, files, and computers. And an individual need not necessarily limit his or her sampling experiments to haphazard search of memory and personal bookshelves. The individual may wish to extend sampling to a large-scale survey, conducted with the aid of randomization techniques.

There is sometimes a tendency to define human probability judgment narrowly—to focus on judgments people make without external aids. But it may not be sensible to try to draw a line between internal and external resources. Psychologists who wish to offer a comprehensible analysis of human judgment should, as Ward Edwards (1975) has argued, take into account the fact that humans are tool-using creatures. Moreover, statisticians and other practical users of probability need to recognize the continuity between apparently subjective judgments and supposedly objective statistical techniques. The concept of design that we have developed in this paper is meant

to apply both to probability analyses that use sophisticated technical aids and to those that are made wholly in our heads. We believe that the selection of a good design for a particular question is a researchable problem with both technical and judgmental aspects. The design and analysis of mental experiments for probability judgment therefore represents a challenge to both statisticians and psychologists.

REFERENCES

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Bertrand, J. (1907). *Calcul des Probabilités* (2nd ed.). Gauthier-Villars et Fils.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A, 143,* 383-430.

Diaconis, P., & Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association, 77,* 822-830.

Edwards, W. (1975). Comment on paper by Hogarth. *Journal of the American Statistical Association, 70,* 291-293.

Edwards, W., Phillips, L. D., Hays, W. L., & Goodman, B. C. (1968). Probabilistic information processing systems: Design and evaluation. *IEEE Transactions on Systems Science and Cybernetics, 4,* 248-265.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 330-344.

Goldstein, M. (1981). Revising previsions: A geometric interpretation. *Journal of the Royal Statistical Society, Series B, 43,* 105-130.

Jeffrey, R. (1965). *The logic of decision.* New York: McGraw-Hill.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11,* 143-157.

Krantz, D., & Miyamoto, J. (1983). Priors and likelihood ratios as evidence. *Journal of the American Statistical Association, 78,* 418-423.

Lieblich, I., & Lieblich, A. (1969). Effects of different pay-off matrices on arithmetic estimation tasks: An attempt to produce "rationality." *Perceptual and Motor Skills, 29,* 467-473.

Lindley, D. V., Tversky, A., & Brown, R. V. (1979). On the reconciliation of probability assessments. *Journal of the Royal Statistical Society, 142,* 146-180.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice Hall.

Raiffa, H. (1974). *Analysis for decision making. An audiographic, self-instructional course.* Chicago: Encyclopedia Britanica Educational Corporation.

Ramsey, F. P. (1931). Truth and probability. In R. G. Braithwaite (Ed.), *The foundations of mathematics and other logical essays.* Routledge and Kegan Paul.

Richardson, H. R., & Stone, L. D. (1971). Operations analysis during the underwater search for *Scorpion. Naval Research Logistics Quarterly, 18,* 141-157.

Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton: Princeton University Press.

Shafer, G. (1981a). Constructive probability. *Synthese, 48,* 1-60.

Shafer, G. (1981b). Jeffrey's rule of conditioning. *Philosophy of Science, 48,* 337-362.

Shafer, G. (1982a). Lindley's paradox. *Journal of the American Statistical Association, 77,* 325–351.

Shafer, G. (1982b). Belief functions and parametric models. *Journal of the Royal Statistical Society, Series B, 44,* 322–352.

Singer, M. (1971). The vitality of mythical numbers. *The Public Interest, 23,* 3–9.

Spetzler, C. S., & Staël von Holstein, C. S. (1975). Probability encoding in decision analysis. *Management Science, 22,* 340–358.

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

Walker, A., & Leakey, R. E. T. (1978). The hominids of East Turkana. *Scientific American, 238,* 54–66.

Wierzchón, S. T. (1984). *An inference rule based on Sugeno measure.* Unpublished paper, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.