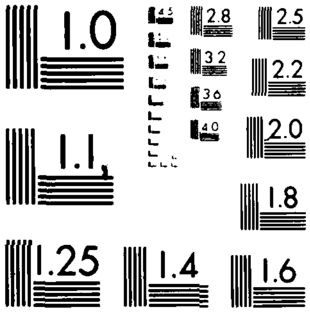
AD-A131 475 WEIGHING EVIDENCE: THE DESIGN AND COMPARISON OF PROBABILITY(U) STANFORD UNIV CA DEPT OF PSYCHOLOGY G SHAFER ET AL. JUN 83 N00014-79-C-0077 1/1

UNCLASSIFIED F/G 5/10 NL

END
DATE
FILMED
9-83
DTIC


1.0
4.5
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

12

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A131475	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Weighing Evidence: The Design and Comparison of Probability Thought Experiments		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Glenn Shafer and Amos Tversky		8. CONTRACT OR GRANT NUMBER(s) N00014-79-C-0077 NR 197-058
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research 800 North Quincy Street Arlington, VA 22217		12. REPORT DATE June 1983
		13. NUMBER OF PAGES 72
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

probability languages, thought experiments, weight of evidence

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The assessment of probability on the basis of evidence is viewed as a thought experiment that yields an expression of degree of belief. Theories of subjective probability are viewed as tools or languages for analyzing evidence and expressing degree of belief. This article focuses on two probability languages: the classical 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 analyze the designs of thought experiments afforded by the two languages and discuss their implications.

Weighing Evidence: The Design and Comparison of Probability Thought Experiments

Glenn Shafer Amos Tversky

A

Abstract

The assessment of probability on the basis of evidence is viewed as a thought experiment that yields an expression of degree of belief. Theories of subjective probability are viewed as tools or languages for analyzing evidence and expressing *degree of belief.* This article focuses on two probability languages: the classical 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 analyze the designs of thought experiments afforded by the two languages and discuss their implications.

1. Introduction

The assessment of probability on the basis of evidence may be viewed as a thought experiment. It involves asking questions of our mind, much as physical experiments ask questions of nature. And the design of the experiment, the choice of questions asked, is of crucial importance.

Often one design is superior to another simply because the questions it asks can be answered with greater confidence and precision. Suppose we want to estimate, on the basis of evidence readily on hand, the number of eggs produced daily in the United States. One design might ask us to guess the number of chickens in the country and the average number of eggs laid by each chicken each day. Another design might ask us to guess the number of people in the country, 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 stab at answering the questions it asks.

When we are simply guessing about matters of fact, we may not find it necessary to worry about subtle points of language. But when we turn from guessing facts to making probability judgments--when we ask, for example, how probable it is that more than 100 million eggs are produced each day in the United States--we inevitably face such subtleties.

In order to make judgments of probability we need a theory of probability. 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. And we need canonical examples for each degree of probability in this scale--examples where it is agreed that that degree of probability is appropriate. For complex judgments we also need a calculus--a set of rules for combining simple judgments to obtain complex ones.

Using a theory of probability means, essentially, comparing the problem with which we are concerned with the theory's scale of canonical examples and picking out the canonical example that matches it best. This comparison may be a *complicated process*. Ingenuity may be required in order to make the canonical examples and their similarity to our problem vivid to our imagination. And it may be necessary to break the comparison down into separate comparisons of restricted aspects of our problem or our evidence with the scale of canonical examples. It is in recombining the judgments resulting from these separate comparisons that the theory's calculus comes into play.

Thought of in this way, a theory of probability is very much like a language. At its base is 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 as a means of matching real problems to the scale. And elements of the vocabulary are combined according to a syntax--the theory's calculus.

In this essay we shall call different theories of probability "probability languages." This way of speaking encourages, we think, sensible attitudes towards the comparison of different theories of probability. It encourages us to think of theories of probability as flexible tools, whose success in providing insight in particular problems will often depend on the skill of the user. And it encourages us to recognize that the individual judgments made in one theory of probability are not always directly translatable into another.

Once we have set out a theory of probability or a probability language, we must still address the problem of design. How do we use the language to construct probability judgments in a particular problem? The basic theme of this essay is that we should study different probability languages in terms of the designs they make possible.

1.1 Evaluating and Comparing Designs

How can we evaluate designs for making probability judgments?

As we have already seen, in thinking about the estimation of egg production, a fundamental consideration is our ability to answer the questions the design asks us.

An analogy with surveying may be helpful. (Lindley, Tversky & Brown, 1979). There are usually many different ways of designing a land survey: there are many choices as to what angles and lengths we measure. We make these choices on the basis of the accuracy and precision with which we can make the different measurements.

In the case of thought experiments, whether and how well we can answer the questions a design asks us is a matter of psychology and practicality--it is a matter of how our knowledge and experience is organized, first in our mind and secondarily in other sources of information available to us. Examples of insightful ways of assessing uncertain quantities are discussed by Raiffa (1974), and Singer (1971) describes an intriguing example of a thought experiment regarding the total worth of property stolen by heroin addicts in New York City in the course of one year.

In many cases, the quality of a design for a thought experiment cannot be assessed until the experiment is at least partially carried out. Only then can we see to what extent we have been able to answer the design's questions, to what extent these questions capture intuitive insights we already had, and to what extent the experiment creates new insights.

The result of carrying out a probability thought experiment with a given design is an argument, or an analysis of one's evidence. The preceding paragraphs suggest that this argument or analysis, rather than the probability language itself or even the design, is the ultimate unit of comparison in evaluating and comparing probability languages. The ultimate question, when we apply different probability languages to the same problem, is which argument or analysis is most cogent, insightful, thorough, and reliable in its treatment of the evidence.

The evaluation of a probability thought experiment is internal in a sense in which the evaluation of a physical experiment need not be. One way to evaluate competing designs for a physical experiment, like a land survey, is to apply them to instances where the truth is known; the results will show which design is best. But such empirical evaluation of final results is not always possible in the case of probability judgments. The evaluation of a probability thought experiment should, in a sense, be empirical--it should be an evaluation of how well the thought experiment worked. But it is necessarily an evaluation of the cogency of the whole process rather than an evaluation of the accuracy of the final result. In the case of probability we cannot always separate theoretical argument from empirical validation.

We shall study two different theories of probability in this essay: the Bayesian theory and the theory of belief functions. We think of these theories as alternative languages, languages in which probability judgments can be created. We think of them, that is to say, as tools. This point of view leads us to address different questions than those that are commonly raised in the philosophical and psychological literature.

We do not focus here on the question of whether these theories are accurate descriptions of how people think. We ask instead whether people are capable of using the theories and whether they can use them to given ends--whether, for example, they can use them to formalize and refine certain intuitively reasonable patterns of thought. Nor do we address here the question of whether the theories are "normative"--it it seems premature to prescribe the use of a given tool before we have an adequate understanding of how well and to what ends it can be used. Furthermore, the prescriptions can only be relative to a given set of alternative tools and to a usually ill-defined set of problems.

It fits our talk about tools to call our view of probability "constructive." We do not think that people come to the task of probability judgment with well-structured beliefs hidden in their psyche and waiting to be "elicited". No doubt they come with some beliefs already formulated. But the process of judgment, when successful, gives new and greater structure to one's beliefs and also tends to render less structured initial beliefs obsolete. So it seems more illuminating to talk about constructing probabilities than to talk about eliciting them. (Shafer, 1981).

Different theories of probability are often treated as either-or alternatives. But it is clear that the same person can use more than one language or tool, and so our view leads us to ask whether a person might find a design based on one theory better for one problem and a design based on another theory better for another. This possibility will be at the back of our minds throughout this essay. Which, we will ask, are the most successful Bayesian designs? Where do these designs work best? And where are they less successful than designs based on other theories?

1.2 Examples

With the help of some simple examples we can indicate in a general way what we mean when we write about different designs for probability judgment. In the first of the following examples, we ask whether a swimmer is likely to win a half-completed race. In the second, two scientists assess the probability of competing paleontological hypotheses. We will return to these examples again in Sections 3 and 4 below.

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 season, 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 9:25. His best previous times for 1650 yards have been around 16:25, a time which translates into about 9:57 at about 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 seconds) 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, in particular, than he was swimming only a few weeks ago. And 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 improved his endurance. This is, after all, 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 we are wrong? What if he cannot keep up his pace?

Here our vision becomes more murky, and we search for an understanding of what might be going on. 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?

If he is acting deliberately, then it seems likely he will soon slow down, but not to a disastrously slow pace; and in this case it seems to be a toss-up whether Cowan will catch him. If he has misjudged what pace he can keep up. on the other hand, 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, to try so hard to keep it up to the very end that he "collapses" with exhaustion to a very slow pace. And 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 is a design that 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 below 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 Richara E. T. Leakey discuss the hominid fossils that have recently been discovered in the region east of Lake Turkana in Kenya. These fossils, which are 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 that have been discovered in the upper member of the Koobi Fora Formation in East Turkana are of three forms: (I) A "robust" form that had large cheek teeth and massive jaws. These fossils show wide-fanning cheekbones, very large molar and premolar teeth and smaller incisor and canines. The brain case has an average capacity of about 500 cubic centimeters, 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 there has been general agreement that they should all be classified as members of the species *Australopithecus robustus*. (II) A smaller and slenderer (or more "gracile") form that lacks the wide-flaring cheekbones of A but has similar cranial capacity and only slightly less massive molar and premolar teeth. (III) A large-brained (c. 850 cubic centimeters) and small-jawed form that can confidently identified with the *Homo erectus* specimens found in the 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, enormously 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 the gracile form 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 the 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 co-exist 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 I, while the opposite is true in III.

(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 some 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 hypotheses 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 the type I have not been found in Java and China, where specimens of the 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 China and then condition this distribution on their absence there. It is natural to all this a *conditioning design*. It is not a total-evidence design, because the initial (or "prior") probabilities for whether the specimens occur in China will be based on only part of the evidence.

In Section 3 below 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 find examples where numerical degrees of belief can be agreed upon 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 !wo 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 will pay the greatest attention to the Bayesian language. Since it is the probability language that is most familiar to most readers, it is the most convenient vehicle for introducing the idea of design for probability thought experiments. We will study the language of belief function as well in order to provide a contrast to Bayesian ideas and in order to emphasize that our constructive view of probability, while not implying that all probability languages have equal normative claims, does leave 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. This idea can be discerned in the axiomatic foundations of modern Bayesian "personalism". (See, for example, Savage, 1954, pp. 33-40.) And it is sometimes invoked by practical Bayesian statisticians. In a recent article by G. E. P. Box (1980), for example, we find the comment that the adoption of given Bayesian probability distribution means that "current belief ... would be calibrated with adequate approximation by a *physical simulation* involving random sampling" from the distribution.

We believe, however, that the constructive aspect of the comparison with chance setups is not sufficiently emphasized in current Bayesian thinking. The personalist axioms assume, in effect, that every problem can be compared successfully to a scale where the truth is generated by known chances. Similarly, Box's formulation may give the impression that a well structured system of beliefs exists before the comparison, which is needed only for calibration. On the other hand, our position is compatible with the approach of Diaconis and Zabell (1982) who treat probability assessment as a constructive process and discuss useful Bayesian designs.

There is a tendency among some personalist Bayesians to leave aside altogether the comparison to a scale of chance examples and to define "personal probabilities" in terms of a person's preference among bets. But the definition of probability in terms of bets does not address the problem of constructing probabilities. Once we admit that this job has not already been done by some genie in the back of the mind--once we admit that coherent preferences among a myriad of bets are not hidden in the mind waiting to be elicited--we must also admit that in order to construct these personal probabilities we must do more than query ourselves about the attractiveness of various bets. We must study the evidence and compare it to evidence in situations where we have a good understanding of what bets are reasonable--i.e., games of chance.

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 this 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 *fre-*

quency 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 the perspective from which we are approaching the question is not, so traditional. 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 may depend on the particular problem and the particular evidence. And even in a particular case we do not pretend 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 much less valuable.

The applicability and usefulness of both the frequency and the propensity semantics are clearly contingent on the nature of our evidence. It may or may not be the case that our evidence can be interpreted in terms of estimated or conjectured frequencies. And it may or may not be the case that our evidence lends itself to interpretation in terms of a causal model. But it is, we believe, precisely when our evidence can be cast in the form of relevant frequencies or causal models that the Bayesian probability language can give us insight into the force of that evidence. 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 doubt as to their validity, and there is all too much room in causal thinking for over-optimism. Thus 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 interpretation of probability one must distinguish logical from psychological considerations. Ramsey (1931) and his 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. Thus, the gambler chooses to bet on Team A rather than on Team B because he believes that A is more likely to win. He does not commonly infer this 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. But 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 has the same syntax as the theory of chance. Its syntax consists, that is to say, of 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 probability one. And in general probabilities add: if A and B are incompatible propositions, then P(A 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

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

As we shall see in Section 3 below, 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 compares evidence to canonical examples where the meaning of a message depends on known chances, or so-called objective probabilities.

By this we mean examples of the following sort. We know a chance experiment has been carried out. We know that the possible outcomes of the experiment are $o_{1,...,}o_n$ and that the chance of o_i is p_i. We are not told the actual outcome and we receive a message 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 that the truth is in A_i.

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

$$\mathrm{Bel}\,(A) = \sum\{m(B) \,|\, B \text{ implies } A\}.$$

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."

When we use the language of belief functions to report on what the evidence has to say about a particular proposition A, we often report both Bel(A) and

$$\mathrm{Pl}(A) = 1\text{-}\mathrm{Bel}(\text{not } A).$$

We call Pl(A) the *plausibility* of A. It measures how plausible A remains in light of the evidence.

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., that 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 non-additivity: the two complementary propositions E and not E have degrees of belief $Bel(E) = s<1$ and $Bel\ (\text{not } E) = 0$.

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 are unable to predict 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 skip 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_1r_2 that none of the message is reliable. If we set.

$$p_1 = (1-r_1)(1-r_2), \qquad A_1 = E_1 \;\&\; E_2,$$

$$p_2 = (1-r_1)r_2, \qquad A_2 = E_1,$$

$$p_3 = r_1 r_2 \qquad\qquad A_3 = \Theta,$$

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 result. We know the list of codes and the chance of each code being chosen--say the list is $o_{1,...,}o_n$, 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 arbitrary belief functions.

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 simple, more elementary judgments. These rules 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, 1982).

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

A Bayesian total-evidence design, we have said, is any design that determines a probability distribution by making probability judgments all based on the total evidence. 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.

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 $P(A \mid B_i)$, 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

$$P(A) = P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + P(B_3)P(A \mid B_3). \tag{4}$$

She did this qualitatively at the time, but she offers, in retrospect, the quantitative judgments indicated in Table 1. These numbers yield P(A) = .87 by (4).

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

Insert Table 1 about here

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 does allow us to decompose our evidence. The question of what the next

Table 1

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

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 $P(A \mid B_2)$ -- 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 (i) maintain his pace, (ii) slow down, but less than 3%, (iii) slow down more than 3%, or (iv) collapse. (Whether Curt slows down 3% is significant because this is how much he he would have to slow down for Cowan to catch him without speeding up.) We assess probabilities for whether Cowan (i) can speed up significantly, (ii) can only maintain his pace, (iii) 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 4x3 = 12 hypotheses about what Curt will do and what Cowan can do.

Table 2 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

$$(.85\text{x}.10\text{x}.5)+(.85\text{x}.70\text{x}1.0)+\cdots\approx.88$$

that Curt will win.

Notice that this design asks for judgments about what Cowan can do rather than judgments about what he will do. This is because our evidence about Cowan consists merely of our general knowledge about swimmers in the league. The numbers .10, .70, and .20 are based on our general notion 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.

Insert Table 2 about here

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. And 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 2

		Cowan		
		can speed up significantly	can only maintain pace	cannot maintain pace
Curt		.10	.70	.20
maintains pace	.85	.5	1.0	1.0
slows less than 3%	.03	.2	1.0	1.0
slows 3% or more	.07	0	0	.5
collapses	.05	.2	.7	.8

When we compare the design and analysis of Table 2 with the design we carried out earlier, we see that we have profited from the new design's focus on our evidence about Cowan. Having made the analysis, 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 2; 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 tree would not have a causal interpretation, but it 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 seven years," say--then he asks first how frequently such projects are completed and then how frequently projects that are completed take less than seven 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.

The probability of completing a project within seven years can, of course, be assessed using propensity instead of frequency semantics. Instead of comparing our project to other similar projects, we could concentrate on our knowledge of the capabilities of the group undertaking the project and assess the propensity of the group to complete the project and its propensity to do so within seven years if it does so at all. Alternatively, we could use a more complicated causal model that takes account of the steps involved in the project. The anecdote related by Kahneman and Tversky suggests, however, that frequency semantics is superior to propensity semantics for this problem because it is less likely to produce unrealistically optimistic results.

Total-Evidence Designs for Distributions of Random Quantities

Spetzler and Staël von Holstein (1975) have discussed in detail the problem of design for the construction of probability distribution for unknown quantities.

One way to construct a probability distribution is to specify percentiles, beginning with the median, then the quartiles, etc. Spetzler and Staël von Holstein call this design "the interval technique." One begins by thinking about a number and adjusts that number up or down until one feels that the unknown quantity is as likely to be greater than the number as it is to be less. This defines the median. The interval below the median is similarly divided to yield the first quartile, etc. The simplicity of this design is attractive, but a number of experiments, beginning with Alpert and Raiffa in 1969, have reported that the initial focus on the median tends to lead to a distribution that is too tightly concentrated around that median.

The design favored by Spetzler and Staël von Holstein runs roughly as follows: (i) specify upper and lower bounds for the unknown quantity; (ii) test these bounds by considering the possibility of even more extreme values, and adjust the bounds if necessary; (iii) consider various values between these bounds, in a haphazard order, and for each value, assess the probability that the unknown quantity is less than that value, checking that each assessment is consistent with the ones already made; (iv) continue this process until a cumulative distribution function for the unknown quantity has been constructed in sufficient detail; (v) check the cumulative distribution function by setting it aside and using the interval technique to assess the median and quartiles.

What semantics is used by Spetzler and Staël von Holstein's design? How do they match their belief that an unknown quantity is less than a certain value to the Bayesian scale of canonical examples? They report that they do so in a way direct and graphic way. They use a "probability wheel"--a disk resembling the spinner used in children's games of chance. The disk has adjustable blue and orange sectors, and the person making the probability judgment is asked to adjust the sectors so as to match the event that the unknown quantity is less than the given value with the event that the pointer will end up in the orange sector after the disk is open. They sometimes phrase the question in betting terms--they ask the person making the judgment to match events in the sense that they would be equally willing to bet on either. This use of betting language seems too superficial, however, to be called an example of betting semantics. The essential semantics is simply a direct comparison to a game of chance.

Spetzler and Staël von Holstein's design must be considered a total-evidence design; all the judgments are based on the total evidence. The weakness of the design is that it does not particularly focus attention on that evidence.

Total-Evidence Designs Based on Expectations?

From a purely mathematical point of view, it is often possible to construct a probability distribution from knowledge of expectations. And expectations, though they appear only as derived quantities when probabilities are interpreted as frequencies or causal propensities, have a direct betting interpretation. So if we took the betting semantics for the Bayesian theory seriously, we would want to consider Bayesian total-evidence designs based on direct judgments of expectation.

Consider, for example, a design based on moments. The expected value of the $i^{\underline{th}}$ power of a random quantity X is denoted by $E(X^i)$ and called the $i^{\underline{th}}$ moment of X. Mathematical theory tells us that if a random quantity X has a finite range, then its distribution can be approximated to any desired degree of accuracy from knowledge of a finite number of the moments. (See, for example, Kendall and Stuart, 1977, pp. 89-90.) Hence we can imagine a design that constructs a probability distribution for an unknown quantity X from judgments

$E(X)$, $E(X^2)$, . . . , $E(X^n)$. We would assess each $E(X^i)$ by asking ourselves what price our total evidence seems to justify for a contract that would return the unknown amount X^i.

In practice, designs based on moments or other expectations do not seem to be very useful, and this fact can be attributed to the weakness of the betting semantics. The personalist view of probability, with its emphasis on introspection and the "elicitation" of probabilities and expectations, encourages the idea that an expectation is as easy to "elicit" as a probability--both probabilities and expectations are prices for gambles and both can be determined by introspection about one's gambling preferences. But as soon as we begin to look at evidence, expectations seem much less accessible. Usually evidence can be related to frequencies or propensities much more readily than to prices for gambles.

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 obtained after probabilities are constructed. In partitioning designs we begin our process of probability judgment with all our evidence in hand, but we deliberately partition this evidence into "old evidence" and "new evidence", assess proabilities 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 of course the first is the more difficult one. 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' theorem. Bayes's theorem, we recall, says that if $B_1, \ldots, B_n$ are incompatible propositions, one of which must be true, then

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

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 $P(A \mid B_i)$. 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.

Moreover, we are carrying out 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 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 the *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) as our new evidence A. Thus we will assess probabilities $P(B_1), \ldots, P(B_5)$ and conditional probabilities $P(A|B_1), \cdots, P(A|B_5)$ and then calculate $P(B_i|A)$, which equals $P(B_i)P(A|B_i)$ divided by the sum $P(B_1)P(A|B_1)+P(B_2)P(A|B_2)$

$+ P(B_3)P(A\,|\,B_3) + P(B_4)P(A\,|\,B_4) + P(B_5)P(A\,|\,B_5)$. The apparent complexity of this expression 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 $P(A\,|\,B_i)/P(A\,|\,B_j)$. It is sufficient to assess their ratios $L(A\,|\,B_i{:}B_j)$.

One further elaboration of this design seems useful. Our new evidence A can be thought of as the event that types I, II and III should be so disparate (items of evidence (ii) and (iii)) *and* 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). We can write

$$A = A_1 \text{ and } A_2,$$

where A_1 is the event that the types should be so disparate and A_2 is the event that types I and II should be absent from the Far East. 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:

$$P(A \mid B_i) = P(A_1 \mid B_i)P(A_2 \mid B_i).$$

Substituting this into (6), we obtain

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

or

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

where $L(A_1 \mid B_i{:}B_j) = P(A_1 \mid B_i)/P(A_1 \mid B_j)$ and $L(A_2 \mid B_i{:}B_j) = P(A_2 \mid B_i)/P(A_2 \mid B_j)$.

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, and this consists of their article itself, which provides internal evidence as to their integrity and the cogency of their thought, our knowledge of the standards of *Scientific American*, our knowledge of the nature and history of this area of science, etc.) It will be instructive, nonetheless, to put ourselves in the shoes of Walker and Leaky and to try 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 take the form of 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 3 represent a compromise.

Now consider A_1, the argument that the different types must represent three distinct species because of their diversity. The 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 3 labeled "$L(A_1 | B_i:B_5)$". These numbers reflect the great implausibility of the intra-species 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 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 probability $P(A_1 | B_5)$. We can think ab^out 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.

Insert Table 3 about here

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 3. Notice again that

Table 3

	$P(B_i)$	$L(A_1 \mid B_i{:}B_5)$	$L(A_2 \mid B_i{:}B_5)$	$P(B_i \mid A)$
B_1	.75	.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

we are spared the well-nigh meaningless task of assessing absolute probabilities: we do not have to say how likely it is that East African species of hominids should have failed to appear in the Far East.

As the last column of Table 3 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 seem 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?

The Role of Likelihood Ratios

In the preceding example we noted that it is sufficient, in a likelihood-based partitioning design, to assess likelihood ratios; absolute likelihoods need not be assessed. This point is further discussed by Edwards, et al., 1968.

In a likelihood-based observational design, on the other hand. we usually assess absolute likelihoods, not just likelihood ratios. 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 $L(A \mid B_i.B_j) = \dfrac{P(A \mid B_i)}{P(A \mid B_j)}$ and $L(\text{not } A \mid B_i{:}B_j) = \dfrac{P(\text{not } A \mid B_i)}{P(\text{not } A \mid B_j)}$. But since $P(A \mid B_i) + P(\text{not } A \mid B_i) = P(A \mid B_j) + P(\text{not } A \mid B_j) = 1$, the likelihood ratios $L(A \mid B_i{:}B_j)$ and $L(\text{not } A \mid B_i{:}B_j)$ fully determine the absolute likelihoods $P(A \mid B_i)$ and $P(A \mid B_j)$.

The Choice of New Evidence

How do we decide which evidence to take as new evidence in a partitioning design?

In the preceding example we identified certain evidence as new evidence because we found better grounds for probability judgment when we thought about the likelihoods of its happening than when we thought about it as a condition affecting the likelihoods of the possible answers to questions of direct interest.

Sometimes we treat evidence as new evidence because of its psychological salience. The salience of evidence 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. Spetzler and Staël von Holstein (1975, p. 346) give the following example:

> a company had to decide whether or not to introduce a new product that was considered to have a high demand potential. The product was test marketed and there was a slightly unfavorable outcome; the revised assessment of the market said there was a low demand. This revision was made in spite of past experiences with similar market tests that had been less than accurate in predicting the final market s'ze and in contrast to the strong prior judgment indicating a high demand.

The market test is salient because it is more specifically related to the new product than the other evidence is. Spetzler and Staël von Holstein suggest that the tendency to overvalue this salient evidence could be checked by using it as new evidence in a partitioning design.

Bayesian Statistical Theory

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 above, is one example. But Bayesians 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 can outgrow its preoccupation with observational designs and learn to deal explicitly with the problems involved in partitioning designs. More attention needs to be paid to the problem of framing in partitioning designs: the principles that should govern the selection of evidence that is to be treated as new evidence.

A Partitioning Design that is not Likelihood-Based

Our study of partitioning designs should include consideration of designs that are not likelihood-based.

Here is a problem that suggests a partitioning design that is not likelihood-based. Gracchus is accused of murdering Maevius. Maevius's 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's 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 felt had wronged him. The means to commit the murder is not at issue: Gracchus or anyone else could have easily have committed it. But we think it very unlikely that anyone else had reason to kill Maevius.

Our partitioning design uses the fact of Maevius's 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|not H) = .01;

P(SI) = .001, and SI is independent of GI;

P(GM|GI and SI) = .4, P(SM|GI and SI) = .4, P(NM|GI and SI) =.2;

P(GM|GI and not SI) = .8, P(NM|GI and not SI) = .2;

P(SM|SI and not GI) = .8, P(NM|SI and not GI) = .2;

$$P(NM|\text{not } GI \text{ and not } SI) = 1.$$

Combining these judgments, we obtain

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

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

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

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 before much of the other evidence. Only after Maevius's 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 a limited number 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 is the case where the total evidence is used only to construct probabilities $p_1, \ldots, p_n$ for a set of mutually incompatible and collectively exhaustive propositions $A_1, \ldots, A_n$, so that the final distribution P is determined by setting $P(A_i) = p_i$ and $P(B|A_i) = P_o(B|A_i)$ for every other proposition B that is considered. In this case we call the design a ***Jeffrey design.***

Here is an example of a Jeffrey design. Gracchus is accused of murdering Maevius, and the evidence against him is just 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 total evidence to assess directly whether we think Maevius has been murdered or not. Say we assess the probability of Maevius's having been murdered at .95. We then obtain a conditional probability from P_o: P_o(Gracchus did it |Maevius was murdered) = .98. The final result is a probability of .95x.98 = .931 for the event that Gracchus murdered Maevius.

For further examples of Jeffrey designs, see Shafer (1981).

3.4 Conclusion

Our study of Bayesian design has not produced any startling new discoveries. None of the particular designs we have considered will seem novel to applied statisticians or decision analysts. We hope, however, that the vocabulary we have introduced will help bring what is commonplace in Bayesian practice into the mainstream of Bayesian theory.

Our examples have illustrated the point that probability analyses are, in the end, only arguments. Our two analyses of the free-style race are more or less convincing arguments that Curt will win. But there is no absolute sense in which it can be said that either gives true or correct probabilities.

4. Belief-Function Design

Because of the importance of Dempster's rule of combination in the theory of belief functions, belief-function designs tend to emphasize the decomposition of evidence more than Bayesian designs. Here we illustrate this point by presenting belief-function designs for the free-style race and the hominid problem.

For further examples of belief-function design, see Shafer, (1981, 1982).

The Free-Style Race

The second of the two Bayesian total evidence designs that we gave for the free-style race (section 3.1 above) 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 70% 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 were 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.

The Hominids of East Turkana

Recall that Walker and Leakey considered five hypotheses:

B_1 = One species.

B_2 = Two species, one composed of A (male) and B (female).

B_3 = Two species, one composed of C (male) and B (female).

B_4 = Two species, one composed of A and C.

B_5 = Three species.

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

(1) A theoretical argument for B_1.

(2) An argument that the three types are too diverse to be separate 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 A nor B specimens have been found among the C 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 $m_2(B_5) = .5$, $m_2(B_2 \text{ or } B_5) = .45$, $m_2(B_2 \text{ or } B_3 \text{ or } B_5) = .04$, and $m_2(\Theta)=.01$, and

(3) Bel_3, with $m_3(B_2 \text{ or } B_5) = .99$ and $m_3(\Theta) = .01$.

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

These belief-function results can be compared to the Bayesian results of Section 3.2, where we obtained $P(B_5) = .7993$ and $P(B_2 \text{ or } B_5) = .9992$. 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 as 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.

5. The Nature of Probability Judgment

Our understanding of proability begins, as this paper began, with the idea that probability judgment is a kind of mental experimentation. The comparison of our evidence in a particular problem to a scale of canonical examples is an experiment. Sometimes it is a sampling experiment--as when we search, in our 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 think through the ways a causal story might go. In any case it is an experiment whose result is not known and not determined in advance.

The act of probability judgment is constitutive. A probability judgment is a creation, not a discovery. it is useful, in this respect, to draw an analogy between "probability" and affective words 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.

Probability and Evidence

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 ɩrguments 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.

Mental and Extra-mental Experimentation

Though we have directed attention to the notion of mental experimentation, we want also to emphasize that when a person undertakes to make a probability judgment he is not necessarily limited to the resources of his memory and his imagination. He may also use paper, pencils, books, files, and computers. And he need not necessarily limit his sampling experiments to haphazard search of his memory and personal bookshelves. He may wish to extend his 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 techniques and tools of judgment that are strictly mental and ones that are extra-mental. 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. On the other hand, 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 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 involving both technological and judgmental aspects. The design and analysis of probability thought experiments therefore represents a challenge to both statisticians and psychologists.

University of Kansas

Stanford University

Acknowledgments

This research has been supported in part by NSF grant MCS-800213 to the first author and by ONR Grant N00014-79-C-0077 to the second author.

References

Alpert, M., and Raiffa, H.: 1982, 'A progress report on the training of probability assessors', pp. 294-305 of *Judgment Under Uncertainty: Heuristics and Biases*, Kahneman, Slovic, and Tversky (eds), Cambridge.

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., and 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., and Goodman, B. C.: 1968, 'Probabilistic information processing systems: Design and evaluation', *IEEE Transactions on Systems Science and Cybernetics SSC-4,* 248-265.

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

Kahneman, D., and Tversky, A.: 1982, 'Variants of uncertainty', *Cognition 11,* 143-157.

Kendall, M. G., and Stuart, A.: 1977, *The Advanced Theory of Statistics,* Vol. 1, Fourth Edition, Macmillan.

Lieblich, I., and Lieblich, A.: 1969, *Perceptual Motor Skills 29,* p. 467.

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

Raiffa, H.: 1974, 'Analysis for decision making', *An Audiographic, Self-Instructional Course,* Encyclopedia Britanica Educational Corporation.

Ramsey, F. P.: 1931, 'Truth and probability', of *The Foundations of Mathematics and Other Logical Essays,* R. G. Braithwaite (ed.), Routledge and Kegan Paul.

Richardson, H. R., and 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,* Wiley.

Shafer, G.: 1976, *A Mathematical Theory of Evidence,* Princeton.

Shafer, G.: 1981, 'Constructive probability', *Synthese 48,* 1-60.

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

Shafer, G.: 1982, '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., and Staël von Holstein, C. S.: 1975, 'Probability encoding in decision analysis', *Management Science 22,* 340-358.

Tversky, A., and Kahneman, D.: 1983, 'Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment', *Psychological Review* in press.

Walker, A., and Leakey, R. E. T.: 1978, 'The hominids of East Turkana', *Scientific American,* 54-66.

November 1982

OFFICE OF NAVAL RESEARCH

Engineering Psychology Group

TECHNICAL REPORTS DISTRIBUTION LIST

OSD

CAPT Paul R. Chatelier
Office of the Deputy Under Secretary
of Defense
OUSDRE (E&LS)
Pentagon, Room 3D129
Washington, D. C. 20301

Dr. Dennis Leedom
Office of the Deputy Under Secretary
of Defense (C^3I)
Pentagon
Washington, D. C. 20301

Department of the Navy

Engineering Psychology Group
Office of Naval Research
Code 442 EP
Arlington, VA 22217 (2 cys.)

Aviation & Aerospace Technology
Programs
Code 210
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Communication & Computer Technology
Programs
Code 240
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Physiology & Neuro Biology Programs
Code 441NB
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Depament of the Navy

Tactical Development & Evaluation
Support Programs
Code 230
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Manpower, Personnel & Training
Programs
Code 270
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Mathematics Group
Code 411-MA
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Statistics and Probability Group
Code 411-S&P
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Information Sciences Division
Code 433
Office of Naval Research
800 North Quincy Street
Arlington, VA 2217

CDR K. Hull
Code 230B
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Department of the Navy

Special Assistant for Marine Corps
Matters
Code 100M
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Dr. J. Lester
ONR Detachment
495 Summer Street
Boston, MA 02210

Mr. R. Lawson
ONR Detachment
1030 East Green Street
Pasadena, CA 91106

CDR James Offutt, Officer-in-Charge
ONR Detachment
1030 East Green Street
Pasadena, CA 91106

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, D. C. 20375

Dr. Michael Melich
Communications Sciences Division
Code 7500
Naval Research Laboratory
Washington, D. C. 20375

Dr. J. S. Lawson
Naval Electronic Systems Command
NELEX-06T
Washington, D. C. 20360

Dr. Robert E. Conley
Office of Chief of Naval Operations
Command and Control
OP-094H
Washington, D. C. 20350

CDR Thomas Berghage
Naval Health Research Center
San Diego, CA 92152

Department-of the Navy

Dr. Robert G. Smith
Office of the Chief of Naval
Operations, OP987H
Personnel Logistics Plans
Washington, D. C. 20350

Dr. Andrew Rechnitzer
Office of the Chief of Naval
Operations, OP 952F
Naval Oceanography Division
Washington, D. C. 20350

Combat Control Systems Department
Code 35
Naval Underwater Systems Center
Newport, RI 02840

Human Factors Department
Code N-71
Naval Training Equipment Center
Orlando, FL 32813

Dr. Alfred F. Smode
Training Analysis and Evaluation
Group
Orlando, FL 32813

CDR Norman E. Lane
Code N-7A
Naval Training Equipment Center
Orlando, FL 32813

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, CA 93940

Dean of Research Administration
Naval Postgraduate School
Monterey, CA 93940

Mr. H. Talkington
Ocean Engineering Department
Naval Ocean Systems Center
San Diego, CA 92152

Department of the Navy

Mr. Paul Heckman
Naval Ocean Systems Center
San Diego, CA 92152

Dr. Ross Pepper
Naval Ocean Systems Center
Hawaii Laboratory
P. O. Box 997
Kailua, HI 96734

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, D. C. 20380

Dr. L. Chmura
Naval Research Laboratory
Code 7592
Computer Sciences & Systems
Washington, D. C. 20375

HQS, U. S. Marine Corps
ATTN: CCA40 (Major Pennell)
Washington, D. C. 20380

Commanding Officer
MCTSSA
Marine Corps Base
Camp Pendleton, CA 92055

Chief, C^3 Division
Development Center
MCDEC
Quantico, VA 22134

Human Factors Technology Administrator
Office of Naval Technology
Code MAT 0722
800 N. Quincy Street
Arlington, VA 22217

Commander
Naval Air Systems Command
Human Factors Programs
NAVAIR 334A
Washington, D. C. 20361

Department of the Navy

Commander
Naval Air Systems Command
Crew Station Design
NAVAIR 5313
Washington, D. C. 20361

Mr. Philip Andrews
Naval Sea Systems Command
NAVSEA 03416
Washington, D. C. 20362

Commander
Naval Electronics Systems Command
Human Factors Engineering Branch
Code 81323
Washington, D. C. 20360

Larry Olmstead
Naval Surface Weapons Center
NSWC/DL
Code N-32
Dahlgren, VA 22448

Mr. Milon Essoglou
Naval Facilities Engineering Command
R&D Plans and Programs
Code 03T
Hoffman Building II
Alexandria, VA 22332

CDR Robert Biersner
Naval Medical R&D Command
Code 44
Naval Medical Center
Bethesda, MD 20014

Dr. Arthur Bachrach
Behavioral Sciences Department
Naval Medical Research Institute
Bethesda, MD 20014

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Groton, CT 06340

Department of the Navy

Head
Aerospace Psychology Department
Code L5
Naval Aerospace Medical Research Lab
Pensacola, FL 32508

Commanding Officer
Naval Health Research Center
San Diego, CA 92152

Commander, Naval Air Force,
U. S. Pacific Fleet
ATTN: Dr. James McGrath
Naval Air Station, North Island
San Diego, CA 92135

Navy Personnel Research and
Development Center
Planning & Appraisal Division
San Diego, CA 92152

Dr. Robert Blanchard
Navy Personnel Research and
Development Center
Command and Support Systems
San Diego, CA 92152

CDR J. Funaro
Human Factors Engineeing Division
Naval Air Development Center
Warminster, PA 18974

Mr. Stephen Merriman
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA 18974

Mr. Jeffrey Grossman
Human Factors Branch
Code 3152
Naval Weapons Center
China Lake, CA 93555

Human Factors Engineering Branch
Code 1226
Pacific Missile Test Center
Point Mugu, CA 93042

Department of the Navy

Dean of the Academic Departments
U. S. Naval Academy
Annapolis, MD 21402

Dr. S. Schiflett
Human Factors Section
Systems Engineering Test
Directorate
U. S. Naval Air Test Center
Patuxent River, MD 20670

Human Factor Engineering Branch
Naval Ship Research and Development
Center, Annapolis Division
Annapolis, MD 21402

Mr. Harry Crisp
Code N 51
Combat Systems Department
Naval Surface Weapons Center
Dahlgren, VA 22448

Mr. John Quirk
Naval Coastal Systems Laboratory
Code 712
Panama City, FL 32401

CDR C. Hutchins
Code 55
Naval Postgraduate School
Monterey, CA 93940

Office of the Chief of Naval
Operations (OP-115)
Washington, D. C. 20350

Professor Douglas E. Hunter
Defense Intelligence College
Washington, D. C. 20374

Department of the Army

Mr. J. Barber
HQS, Department of the Army
DAPE-MBR
Washington, D. C. 20310

Department of the Navy

Dr. Edgar M. Johnson
Technical Director
U. S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Director, Organizations and
Systems Research Laboratory
U. S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Technical Director
U. S. Army Human Engineering Labs
Aberdeen Proving Ground, MD 21005

Department of the Air Force

U. S. Air Force Office of Scientific
Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, D. C. 20332

AFHRL/LRS TDC
Attn: Susan Ewing
Wright-Patterson AFB, OH 45433

Chief, Systems Engineering Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, OH 45433

Dr. Earl Alluisi
Chief Scientist
AFHRL/CCN
Brooks Air Force Base, TX 78235

Foreign Addressees

Dr. Daniel Kahneman
University of British Columbia
Department of Psychology
Vancouver, BC V6T 1W5
Canada

Foreign Addressees

Dr. Kenneth Gardner
Applied Psychology Unit
Admiralty Marine Technology
Establishment
Teddington, Middlesex TW11 OLN
England

Director, Human Factors Wing
Defence & Civil Institute of
Environmental Medicine
Post Office Box 2000
Downsview, Ontario M3M 3B9
Canada

Dr. A. D. Baddeley
Director, Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF England

Other Government Agencies

Defense Technical Information Center
Cameron Station, Bldg. 5
Alexandria, VA 22314 (12 copies)

Dr. Craig Fields
Director, System Sciences Office
Defense Advanced Research Projects
Agency
1400 Wilson Blvd.
Arlington, VA 22209

Dr. M. Montemerlo
Human Factors & Simulation
Technology, RTE-6
NASA HQS
Washington, D. C. 20546

Dr. J. Miller
Florida Institute of Oceanography
University of South Florida
St. Petersburg, FL 33701

Other Organizations

Dr. Robert R. Mackie
Human Factors Research Division
Canyon Research Group
5775 Dawson Avenue
Goleta, CA 93017

Dr. Amos Tversky
Department of Psychology
Stanford University
Stanford, CA 94305

Dr. H. McI. Parsons
Human Resources Research Office
300 N. Washington Street
Alexandria, VA 22314

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard Street
Alexandria, VA 22311

Professor Howard Raiffa
Graduate School of Business
Administration
Harvard University
Boston, MA 02163

Dr. T. B. Sheridan
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139

Dr. Arthur I. Siegel
Applied Psychological Services, Inc.
404 East Lancaster Street
Wayne, PA 19087

Dr. Paul Slovic
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Harry Snyder
Department of Industrial Engineering
Virginia Polytechnic Institute and
State University
Blacksburg, VA 24061

Other Organizations

Dr. Ralph Dusek
Administrative Officer
Scientific Affairs Office
American Psychological Association
1200 17th Street, N. W.
Washington, D. C. 20036

Dr. Robert T. Hennessy
NAS - National Research Council (COHF)
2101 Constitution Avenue, N. W.
Washington, D. C. 20418

Dr. Amos Freedy
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

Dr. Robert C. Williges
Department of Industrial Engineering
and OR
Virginia Polytechnic Institute and
State University
130 Whittemore Hall
Blacksburg, VA 24061

Dr. Meredith P. Crawford
American Psychological Association
Office of Educational Affairs
1200 17th Street, N. W.
Washington, D. C. 20036

Dr. Deborah Boehm-Davis
General Electric Company
Information Systems Programs
1755 Jefferson Davis Highway
Arlington, VA 22202

Dr. Ward Edwards
Director, Social Science Research
Institute
University of Southern California
Los Angeles, CA 90007

Dr. Robert Fox
Department of Psychology
Vanderbilt University
Nashville, TN 37240

Other Organizations

Dr. Charles Gettys
Department of Psychology
University of Oklahoma
455 West Lindsey
Norman, OK 73069

Dr. Kenneth Hammond
Institute of Behavioral Science
University of Colorado
Boulder, CO 80309

Dr. James H. Howard, Jr.
Department of Psychology
Catholic University
Washington, D. C. 20064

Dr. William Howell
Department of Psychology
Rice University
Houston, TX 77001

Dr. Christopher Wickens
Department of Psychology
University of Illinois
Urbana, IL 61801

Mr. Edward M. Connelly
Performance Measurement
Associates, Inc.
410 Pine Street, S. E.
Suite 300
Vienna, VA 22180

Professor Michael Athans
Room 35-406
Massachusetts Institute of
Technology
Cambridge, MA 02139

Dr. Edward R. Jones
Chief, Human Factors Engineering
McDonnell-Douglas Astronautics Co.
St. Louis Division
Box 516
St. Louis, MO 63166

Other Organizations

Dr. Babur M. Pulat
Department of Industrial Engineering
North Carolina A&T State University
Greensboro, NC 27411

Dr. Lola Lopes
Information Sciences Division
Department of Psychology
University of Wisconsin
Madison, WI 53706

Dr. A. K. Bejczy
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91125

Dr. Stanley N. Roscoe
New Mexico State University
Box 5095
Las Cruces, NM 88003

Mr. Joseph G. Wohl
Alphatech, Inc.
3 New England Executive Park
Burlington, MA 01803

Dr. Marvin Cohen
Decision Science Consortium
Suite 721
7700 Leesburg Pike
Falls Church, VA 22043

Dr. Wayne Zachary
Analytics, Inc.
2500 Maryland Road
Willow Grove, PA 19090

Dr. William R. Uttal
Institute for Social Research
University of Michigan
Ann Arbor, MI 48109

Dr. William B. Rouse
School of Industrial and Systems
Engineering
Georgia Institute of Technology
Atlanta, GA 30332

Other Organizations

Dr. Richard Pew
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02238

Dr. Hillel Einhorn
Graduate School of Business
University of Chicago
1101 E. 58th Street
Chicago, IL 60637

Dr. Douglas Towne
University of Southern California
Behavioral Technology Laboratory
3716 S. Hope Street
Los Angeles, CA 90007

Dr. David J. Getty
Bolt Beranek & Newman, Inc.
50 Moulton street
Cambridge, MA 02238

Dr. John Payne
Graduate School of Business
Administration
Duke University
Durham, NC 27706

Dr. Baruch Fischhoff
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Andrew P. Sage
School of Engineering and
Applied Science
University of Virginia
Charlottesville, VA 22901

Denise Benel
Essex Corporation
333 N. Fairfax Street
Alexandria, VA 22314

Psychological Documents (3 copies)
ATTN: Dr. J. G. Darley
N 565 Elliott Hall
University of Minnesota
Minneapolis, MN 55455