Psychological Review
1999, Vol. 106, No. 1, 62–88

0033-295X/99/$3.00

Naive Probability: A Mental Model Theory of Extensional Reasoning

P. N. Johnson-Laird
Princeton University

Paolo Legrenzi
Universita di Milano

Vittorio Girotto
Université de Provence

Maria Sonino Legrenzi
Universita di Padova

Jean-Paul Caverni
Université de Provence

This article outlines a theory of naive probability. According to the theory, individuals who are unfamiliar with the probability calculus can infer the probabilities of events in an *extensional* way: They construct mental models of what is true in the various possibilities. Each model represents an equiprobable alternative unless individuals have beliefs to the contrary, in which case some models will have higher probabilities than others. The probability of an event depends on the proportion of models in which it occurs. The theory predicts several phenomena of reasoning about absolute probabilities, including typical biases. It correctly predicts certain cognitive illusions in inferences about relative probabilities. It accommodates reasoning based on numerical premises, and it explains how naive reasoners can infer posterior probabilities without relying on Bayes's theorem. Finally, it dispels some common misconceptions of probabilistic reasoning.

> The defence were permitted to lead evidence of the Bayes Theorem in connection with the statistical evaluation of the DNA profile. Although their Lordships expressed no concluded view on the matter, they had very grave doubts as to whether that evidence was properly admissible . . . their Lordships had never heard it suggested that a jury should consider the relationship between such scientific evidence and other evidence by reference to probability formulae
>
> —*The Times* (London), May 9, 1996, reporting the Court of Appeal's judgment on *Regina v. Adams.*

You think about probabilities because what you do depends on what you infer is likely to happen. If you speed on the freeway, are you likely to be caught by the police? If you choose surgery for a cancer, is it likely to be successful? If you put your money in the local bank, is it likely to be safe? Psychologists have studied thinking about probabilities, but they disagree about the process (e.g., Gigerenzer, 1996; Kahneman & Tversky, 1996). Similarly, like philosophers and statisticians, they disagree about the proper interpretation of the probability calculus. It embodies several self-evident principles, notably, the *extensional* notion that the probability of an event equals the sum of the probabilities of the different ways in which the event can occur. What are these probabilities? In theory, they can be interpreted as relative frequencies (e.g., von Mises, 1957), partial logical entailments (e.g., Keynes, 1943), or degrees of belief (e.g., Savage, 1954). Each of these, and other interpretations, has its defenders, and there is no consensus.

Our concern is not the disagreements of experts but the competence of naive individuals. We use the term *naive* to refer to people who have not acquired an explicit mastery of the probability calculus, but we do not impugn their intelligence. Many great intellects—Aristotle, for example, and perhaps the judges in the British Court of Appeal—reason about probabilities without benefit of calculus. By definition, they are naive. The demarcation between naive probabilistic reasoning and expert reasoning is not clear cut, and a naive ability is a precursor to acquiring a knowledge of the calculus. However, few people are real experts; nearly

P. N. Johnson-Laird, Department of Psychology, Princeton University; Paolo Legrenzi, Istituto di Psicologia, Facolta di Lettere e Filosofia, Universita di Milano, Milan, Italy; Vittorio Girotto and Jean-Paul Caverni, Centre de Recherche en Psychologie Cognitive, Université de Provence (CREPCO), Aix-en-Provence, France; Maria Sonino Legrenzi, Dipartimento di Psicologia Generale, Universita di Padova, Padua, Italy.

This research was supported in part by ARPA (CAETI) Contracts N66001-94-C-6045 and N66001-95-C-8605.

We thank Ruth Byrne for her help in developing the model theory; Fabien Savary for his help in studies of illusory inferences; Gil Harman, Danny Kahneman, and the late, and much missed, Amos Tversky for some stimulating advice; and the participants of the International Symposium on Mental Models in Cognitive Science, London, October 1996, for their constructive remarks. We also thank Gerd Gigerenzer, Michel Gonzalez, Rich Gonzalez, Ray Nickerson, Eldar Shafir, Shinsuke Shimojo, and David Over for their helpful comments on an earlier version of the article. P. N. Johnson-Laird is also grateful to the members of his laboratory for many useful ideas: Patricia Barres, Victoria Bell, Zachary Estes, Yevgeniya Goldvarg, Bonnie Meyer, Mary Newsome, Kyung-Soo Do, Lisa Torreano, and Isabelle Vadeboncoeur. The experiments were carried out at CREPCO at the University of Aix-en-Provence, and those of us who were visitors are grateful to the faculty, students, and staff for their hospitality.

Correspondence concerning this article should be addressed to P. N. Johnson-Laird, Department of Psychology, Princeton University, Princeton, New Jersey 08544. Electronic mail may be sent to phil@clarity.princeton.edu.

everyone goes wrong with certain problems (see, e.g., Tversky & Kahneman, 1983). And, a computationally tractable system of probabilistic reasoning is impossible. Problems soon outstrip the computational power of any feasible system (Osherson, 1996). Yet, naive individuals *are* able to reason about probabilities, and they often reach correct conclusions.

Our goals are to formulate a new theory of naive probabilistic reasoning and to examine its consequences. The theory purports to explain how naive individuals reason about probabilities in an extensional way, where we take *extensional* to mean inferring the probability of an event from the different possible ways in which it could occur. The probability calculus is accordingly a normative theory of extensional reasoning about probabilities. This way of reasoning aims to be deductive; that is, its conclusions must be true given that the premises are true, though naive reasoners make mistakes. It can be contrasted with *nonextensional* reasoning about the probability of an event, which relies on some relevant heuristic, index, or evidence, as, for example, when reasoners infer that because an exemplar is typical of a category, it has a high probability of being a member of the category (D. Kahneman, personal communication, January 26, 1994). Kahneman and his colleague, the late Amos Tversky, obtained abundant evidence that many such inferences depend on heuristics (see, e.g., Tversky & Kahneman, 1973, 1983). This reasoning is inductive; that is, its conclusions could be false even if their premises are true. Many inferences in daily life depend on a mixture of extensional and nonextensional reasoning. Nevertheless, we focus on extensional reasoning in cases where probabilities are implicit in the situation or expressed numerically.

Most psychological theories of probabilistic reasoning, such as Tversky and Koehler's (1994) "support" theory, concern nonextensional reasoning. However, there are theories of extensional reasoning relating to Bayes's theorem and other advanced topics. We describe these theories in due course, but only after we have considered more elementary matters. The new theory has critical implications for a number of assumptions that psychologists sometimes make:

1. Probabilistic reasoning is an inductive process.
2. It is seldom, if ever, extensional.
3. Extensional reasoning, insofar as it occurs, depends on a tacit knowledge of the probability calculus.
4. It also depends on premises about the frequencies of events.
5. Cognitive illusions disappear when individuals reason extensionally.

If the new theory is correct, then these assumptions are all mistaken.

In fact, many inferences about probabilities *are* deductive (contrary to the first assumption). Suppose you carry out a binomial test on the frequency of some event, say, you tossed a coin 10 times and it came down heads on each occasion, and the test shows that the chance probability of this observation, given an unbiased coin, is less than 1 in 1,000. Your inference is deductive. Your thinking becomes inductive only if you follow Fisher's (1932) procedure and infer the falsity of the "null" hypothesis that the 10 heads occurred by chance with an unbiased coin. This conclusion could be false.

Naive individuals, contrary to the second assumption, do reason about probabilities extensionally. Here is an example:

> A certain sign of a particular viral infection—a peculiar rash—never occurs in patients who do not have the infection, but some people with the infection do not have the rash. Which is more likely to occur: the rash or the viral infection?

The answer is that the viral infection is more likely to occur than the rash, because the viral infection could occur without the rash, whereas the rash cannot occur without the viral infection. This inference is deductively valid; that is, its conclusion must be true granted that its premise is true.

How do people reason extensionally? According to the third assumption, they tacitly follow the probability calculus. Most people, however, have never encountered the calculus, and so they are unlikely to have acquired its principles. Our proposed theory of naive probability provides an alternative explanation, accounting for both the mental representations and the processes of extensional reasoning. The theory, contrary to the fourth assumption, allows that extensional reasoning can occur even if the premises do not refer to frequencies. Likewise, contrary to the fifth assumption, it predicts the existence of cognitive illusions in extensional reasoning.

The plan of this article is as follows: We begin with an outline of psychological theories of reasoning based on formal rules and show how, in principle, those theories could accommodate the probability calculus. We then describe the mental model theory of reasoning and show how it forms the basis of a new theory of probabilistic reasoning. We report some studies that corroborate the theory's prediction that extensional reasoning can occur in the absence of data about frequencies. The theory also predicts some biases that should occur in extensional reasoning, and we describe a further study corroborating these predictions. We then turn to extensional inferences about relative probabilities. The theory correctly predicts the occurrence of cognitive illusions. In certain cases, for example, reasoners infer that one event is more probable than another, even though the event in question is impossible. We show how the model theory accounts for conditional probabilities and for Bayesian inferences to posterior probabilities. Finally, we consider some pedagogical implications of the theory.

Formal Rule Theories and the Probability Calculus

Psychological theories of reasoning are often based on formal rules of inference (see, e.g., Braine & O'Brien, 1991; Rips, 1994). *Rule theories,* as we will henceforth refer to them, postulate that reasoners match the logical forms of premises to relevant rules and that the derivations of conclusions are akin to the steps of a proof. These theories also postulate rules for each of the major sentential connectives. For example, the rule of *modus ponens,* which applies to *if,* is as follows:

If p then q

p

$\therefore$ q.

In general, the greater the number of steps in the formal derivation of an inference, the harder the inference should be, though other factors cause difficulty, for example, the availability of rules and their complexity. Most of the evidence in favor of rule theories derives from studies of deductive inferences that are based on

sentential connectives (see Braine, Reiser, & Rumain, 1984; Rips, 1994).

Current rule theories apply to deductions yielding necessary conclusions, but some systems of rules have been formulated for limited sets of inferences about what is possible (Osherson, 1976) and what is plausible (Collins & Michalski, 1989). In contrast, the following problem concerns the relative probabilities of two events:

> If there is a red or a green marble in the box, then there is a blue marble in the box.
>
> Which is more likely to be in the box: the red marble or the blue marble?

The problem has a valid answer: The blue marble is more likely to be in the box than the red marble. Psychologists have not studied such inferences, and rule theories have not been formulated to apply to them. Yet, these theories could be extended to cope with relative probabilities. Given the premise above, reasoners could make a supposition, that is, an assumption for the sake of argument:

> There is a red marble in the box.

Next, they could use the following variant of *modus ponens* postulated by certain theorists (e.g., Braine & O'Brien, 1991):

If p or q then r

p

$\therefore$ r,

in order to infer the conclusion

> There is a blue marble in the box.

Likewise, from the supposition

> There is a green marble in the box

they can derive the same conclusion as before from another variant of the rule. A system of "bookkeeping" could keep track of the respective possibilities; that is, whenever there is a red marble, there is a blue marble, but not vice versa. It follows that a blue marble is more likely to be in the box than a red marble.

Mathematical axiomatizations of the probability calculus depend on an apparatus of sentential connectives or the equivalent Boolean algebra of set theory. Three axioms formalize absolute probabilities by providing a measure for probabilities: The first axiom specifies that probabilities are real numbers greater than or equal to zero. The second axiom specifies that the probability of a tautology is one. The third axiom is the extensional principle that the probability of a disjunction of two mutually exclusive alternatives equals the sum of their respective probabilities. *Conditional probabilities*, which are probabilities that depend on the truth of some other proposition, can be defined in terms of the system, or specified in a fourth axiom. This axiom asserts that the conditional probability, $p(A|B)$, that is, the probability of A given that B is true, corresponds to the subset of cases of B in which A also holds:

$$p(A|B)=\frac{p(A \text{ and } B)}{p(B)}.$$

Rule theories in psychology eschew axioms in favor of rules of inference (see, e.g., Rips, 1994), and so the development of these theories to deal with probability would express the axioms as rules. For instance, the following extensional rule of inference,

If X and Y are independent, then p(X and Y)=p(X)p(Y),

could be used to infer the probability of conjunctions from the probabilities of their conjuncts. For example, if the probabilities of A and B are independent and both equal to 1/2, then the probability of *A and B* equals 1/4. No rule theory has been proposed by psychologists to deal with naive probability, so we do not pursue the details further. For our purposes, it suffices to know that a rule theory for probabilistic reasoning is at least feasible. One topic that we must consider, however, is the vexed business of Bayesian reasoning, because psychologists have proposed extensional theories for this domain.

Bayes's Theorem and Studies of Bayesian Reasoning

Bayes's theorem is a valid equation in the probability calculus that allows one conditional probability to be inferred from the values of other probabilities. The simplest version of the equation can be expressed in terms of a hypothesis, H_1, and data, d:

$$p(H_1|d)=\frac{p(d|H_1)p(H_1)}{p(d)}. \quad (1)$$

Thus, the posterior probability of a hypothesis, H_1, given data, d, depends on the prior probability of the data, the prior probability of the hypothesis, and the conditional probability of the data given the truth of the hypothesis. If a set of hypotheses are exhaustive and mutually exclusive, then they are known as a "partition"; if each member of a partition has the same probability, then the probability distribution is "uniform." If there is a partition of *n* hypotheses, then p(d) can be expressed in terms of the equation

$$p(d)=p(d|H_1)p(H_1)+p(d|H_2)p(H_2)\ldots+p(d|H_n)p(H_n). \quad (2)$$

It follows that Bayes's theorem can be reexpressed as

$$p(H_1|d)=\frac{p(d|H_1)p(H_1)}{\sum_{i=1}^{n} p(d|H_i)p(H_i)}. \quad (3)$$

As a simple example of the use of Bayes's theorem, imagine that one of two bags is selected at random: One bag contains 70 red chips and 30 blue chips, and the other bag contains 30 red chips and 70 blue chips. From the selected bag, a chip is drawn at random, and it is, say, red. Hence, the probability that it is drawn from the preponderantly red bag, according to Bayes's theorem in Equation 3, is

$$p(H_{red}|d)=\frac{(.7)(.5)}{(.7)(.5)+(.3)(.5)}=.7. \quad (4)$$

Early studies of Bayesian inference used bags of chips or other materials that isolated the problem from the participants' everyday knowledge (e.g., Phillips & Edwards, 1966). The overwhelming result was that naive reasoners were far too conservative; that is, their estimates of the posterior probabilities were less extreme than

those calculated according to Bayes's theorem. As Edwards and Tversky (1967) commented, "Relative to Bayes's theorem, people are consistently conservative information processors, unable to extract from data anything like as much certainty as the data justify" (p. 123). There were various schools of thought about the source of the error. One view was that individuals were able to grasp the impact of a single observation but had trouble aggregating the joint impact of a sample of several observations (see von Winterfeldt & Edwards, 1986). In retrospect, it is clear that naive individuals are unlikely to be carrying out the calculations required by Bayes's theorem.

The Neglect of Base Rates

The debate about conservatism in Bayesian inference was soon overtaken by a more contentious issue: the question of whether individuals neglect the base rate of an event, that is, its prior probability. The pioneering experiments by Kahneman and Tversky (1973) showed that naive participants can estimate base rates but that when they judge the probability of category membership, they are governed more by nonextensional considerations, such as their knowledge of the representativeness of an instance as within a category. Indeed, in such circumstances, there is a strong correlation between explicit judgments of representativeness and judgments of probability. In one of Kahneman and Tversky's (1973) studies, for example, the participants were given the following cover story, which we here abbreviate:

> A panel of psychologists . . . administered personality tests to 30 engineers and 70 lawyers . . . . On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

The cover story also explained that the participants would be paid a bonus for assigning probabilities accurately, that is, in accordance with a panel of experts. Another set of participants was given the same story, except that the numbers were reversed so that now there were 70 engineers and 30 lawyers. Within both sets of participants, one group judged the probability that each description was of an engineer, and the other group judged the probability that each description was of a lawyer. The participants' judgments did not show much effect of the prior odds. The mean estimate of the probability of an occupation was 55% for the condition in which the majority was in that occupation, and the mean estimate was a probability of 50% in the condition in which the minority was in that occupation. This difference was reliable but far from the size that should have occurred if the participants were using a Bayesian method. The moral is clear: Naive individuals rely on their knowledge of typical engineers and typical lawyers in order to estimate the probabilities of category membership. Only in the following sixth test did the participants take the base rate fully into account:

> Suppose now that you are given no information whatsoever about an individual chosen at random from the sample.
>
> The probability that this man is one of the 70 engineers in the sample of 100 is _%.

Many subsequent studies showed similar effects.

In a recent review of the literature, Koehler (1996) wrote: "We have been oversold on the base rate fallacy in probabilistic judgment from an empirical, normative, and methodological standpoint" (p. 1). He argued that in many cases it is unclear whether Bayes's theorem is the appropriate normative model. These cases include problems that violate its assumptions, or at least are ambiguous with respect to them, and problems where the reasoners' goals may differ from those of the normative models. In other cases, he claimed, reasoners almost always take the base rate into account, though perhaps not to a sufficient degree. He also tried to isolate the conditions in which they are most likely to do so and pinpointed several, including the presentation of problems in terms of frequencies rather than probabilities.

Gigerenzer and his colleagues also argued that when problems are expressed in frequencies based on naturally occurring samples, cognitive illusions disappear and reasoners can make Bayesian inferences (see, e.g., Gigerenzer & Hoffrage, 1995). Similarly, Cosmides and Tooby (1996) proposed that evolution has led to the development of a specific inductive module that can make Bayesian inferences from frequency data. On this hypothesis of "frequentism," extensional reasoning depends on an innate embodiment of elements of the probability calculus. Perhaps people who have been taught the rudiments of the probability calculus reason in this way, but, as we suggested earlier, naive reasoners do not appear to rely on the probability calculus. We therefore return to frequentism after we have outlined two alternative approaches to extensional reasoning.

Intuitive Beliefs About Probabilities

Shimojo and Ichikawa (1989) carried out a pioneering study of errors in difficult Bayesian problems. They interviewed four graduate students, who had some statistical training but no detailed knowledge of Bayes's theorem, and elicited from them their intuitive beliefs (or "subjective theorems") about probabilities. Shimojo and Ichikawa carried out three experiments with mainly naive participants, whose task was to give estimates of probability and reasons justifying them for a series of problems. These problems were either the following "three prisoners problem" or variations on it:

> Three men, A, B, and C, were in jail. A knew that one of them was to be set free and the other two were to be executed. But he didn't know who was the one to be spared. To the jailer who did know, A said, "Since two of the three will be executed, it is certain that either B or C will be, at least. You will give me no information about my own chances if you give me the name of one man, B or C, who is going to be executed." Accepting this argument after some thinking, the jailer said, "B will be executed." Thereupon A felt happier because now either he or C would go free, so his chance had increased from 1/3 to 1/2. The prisoner's happiness may or may not be reasonable. What do you think?

The correct answer to this problem, granted certain plausible assumptions, is that the prisoner's chances of being executed remain equal to 1/3. In the section Mental Models and the Pedagogy of Bayesian Reasoning, we explain how to solve this problem and other Bayesian puzzles. There were three main intuitions elicited by Shimojo and Ichikawa:

1. "Number of cases": When the number of possible alternatives is N, the probability of each alternative is $1/N$. Reasoners who hold this belief infer that Prisoner A's chances rise to 1/2 because there are only two alternatives: either A or C will be executed.

2. "Constant ratio": When one alternative is eliminated, the ratio of probabilities for the remaining alternatives is the same as the ratio of prior probabilities for them. Reasoners who hold this belief also infer that Prisoner A's chances rise to 1/2. In a variant on the problem, A's chances of execution are 1/4, B's chances of execution are 1/2, and C's chances of execution are 1/4 (see the section Mental Models and the Pedagogy of Bayesian Reasoning). Given that C is not to be executed, those who follow the constant-ratio intuition infer that A's chances of execution are now in the ratio 1/4 to (1/4 + 1/2), that is, 1/3. Those who follow the number-of-cases intuition infer that A's chances rise to 1/2.

3. "Irrelevant, therefore invariant": If it is certain that at least one of several alternatives will be eliminated, the information specifying which alternative will be eliminated is irrelevant and does not change the probabilities of the other alternatives. Reasoners who hold this belief infer that A's chances of execution are unchanged by the jailer's message in both versions of the problem.

Shimojo and Ichikawa (1989) reported that most participants stuck to the same intuition throughout, but a few shifted from one intuition to another. They argued that the participants were not reasoning within the framework of Bayes's theorem, and so there may be a module for intuitive reasoning that is independent of formal mathematical reasoning. They pointed out that reasoners often overlook the context of events and that the problem is highly sensitive to the precise nature of the jailer's question (see also Nickerson's, 1996, analysis of the unstated assumptions in Bayesian problems). Thus, the appropriate partition is often unclear, and Shimojo and Ichikawa raise the key question, What determines the participants' partition of the problem?

Falk (1992) has also discussed the three prisoners problem. In a reanalysis of Shimojo and Ichikawa's results, she showed that the most prevalent intuition is "number of cases," and then the "irrelevant, therefore invariant" intuition. She renamed them as "uniformity" and "no-news, no-change," respectively. Only a tiny proportion of the participants' answers were not based on either of these two intuitions. Her own informal studies corroborate their ubiquity. And uniformity, she argued, is seldom questioned and generally prevails over no-news, no-change if the two should clash. "To assume uniformity," she wrote, "when we think we know nothing else but the list of possible outcomes, seems so natural that just describing the phenomenon seems sufficient to explain it" (p. 206). She suggested that it may reflect a preference for symmetry and fairness.

With hindsight, what is missing from these pioneering analyses is an account of how reasoners represent problems, particularly the partition for a problem, of how they make simple extensional inferences about probabilities, and of how they may err in such cases. The new theory of extensional reasoning, to which we now turn, aims to answer these questions. It accommodates the ideas of Shimojo and Ichikawa and of Falk, but it aims to go beyond them by giving a more general framework for extensional reasoning about probabilities.

A Model Theory of Naive Probability

The Model Theory of Sentential Reasoning

Our account of naive probability is based on the theory of mental models, which was originally developed to explain the comprehension of discourse and deductive reasoning (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). The theory postulates that when individuals understand discourse, perceive the world, or imagine a state of affairs, they construct mental models of the relevant situations. A mental model is defined as a representation of a possibility that has a structure and content that captures what is common to the different ways in which the possibility might occur. For example, when individuals understand a conjunction such as "There is triangle and there is circle," they represent its meaning (its intension), from which they can construct a representation of what it refers to (its extension). The representation of the extension takes the form of the mental model

○ △

in which ○ represents a circle, and △ represents a triangle. The model captures what is common to any situation in which there is a triangle and a circle; that is, it represents nothing about the size of the objects, their spatial relation, or other such matters. However, the two objects are represented by two mental tokens, which have properties that correspond to the properties of the two objects. The construction of models from a description is thus part of the process of verbal comprehension, and how this process occurs has been considered in detail elsewhere (see, e.g., Garnham, 1987; Johnson-Laird, 1983; Stevenson, 1993).

The theory postulates that reasoning is a semantic process rather than a formal one, because reasoners build mental models on the basis of their understanding of the premises and on any relevant knowledge. They can formulate a conclusion that is true in these models, and they can establish its validity by ensuring that there are no models of the premises in which the conclusion is false.

A fundamental principle of the theory is that mental models normally represent only what is true. In this way, there is a limit on the load on working memory. The principle is subtle, because it applies at two levels. First, individuals represent only true possibilities; second, they represent those literal propositions in the premises—affirmative or negative—that are true in the true possibilities. For example, an assertion of the form

There is a circle or there is a triangle, but not both

elicits two alternative models to represent the two true possibilities:

○

△

where each row denotes a model of a separate possibility. Each model represents only what is true in a particular possibility. Hence, the model shown on the first line represents that it is true that there is a circle, but it does not represent explicitly that it is also false in this case that there is a triangle. Similarly, the second model represents that there is a triangle, but it does not represent explicitly that it is also false that there is a circle. The theory postulates that reasoners try to remember the information about falsity, but that these "mental footnotes" on models are soon likely to be forgotten.

In contrast to mental models, fully explicit models represent the

false components in each possibility. Thus, the fully explicit models of the exclusive disjunction above are as follows:

○ ¬△
¬○ △

where "¬" denotes negation. Thus, false affirmatives are represented by true negations in fully explicit models, and false negatives are represented by true affirmatives in fully explicit models. The two fully explicit models of the disjunction match the two rows that are true in its truth table, that is, the second and third rows in the truth table as depicted in Table 1.

The theory gives an analogous account of all the so-called "truth functional" connectives, that is, those for which the truth value of a sentence formed with the connective depends only on the truth values of the clauses it connects. Thus, the following inclusive disjunction is true if at least one of its disjuncts is true:

There is a circle or there is a triangle, or both.

It calls for three mental models:

○
△
○ △

which again represent only the true components of the true possibilities. A conditional, such as

If there is a circle, then there is a triangle,

elicits one explicit model of the possibility in which its antecedent and consequent are true. There are other possibilities, that is, those in which the antecedent of the conditional (there is a circle) is false, but individuals do not normally represent them explicitly. The theory accordingly proposes that these possibilities are represented in a single implicit model, a model that has no explicit content and that we symbolize with an ellipsis (three dots). Thus, the conditional has two mental models, one explicit and the other wholly implicit:

○ △
. . .

where the ellipsis denotes the wholly implicit model. Readers will notice the similarity to the model of the earlier conjunction. The difference is that the conditional has an implicit model allowing for possibilities in which the antecedent is false. A biconditional,

If, and only if, there is a circle, then there is a triangle,

has the same mental models as a conditional, but its implicit model corresponds to the possibility in which both the antecedent and the consequent of the conditional are false. If reasoners retain the mental footnotes about what is false in the possibilities represented by an implicit model, they can construct a set of fully explicit models for a conditional or biconditional. The evidence suggests that reasoners soon lose access to these mental footnotes and to the existence of an implicit model, especially with propositions that contain more than one connective (see, e.g., Johnson-Laird & Savary, 1996).

Table 1
A Truth Table for an Exclusive Disjunction

Circle	Triangle	Circle or triangle, but not both
True	True	False
True	False	True
False	True	True
False	False	False

Table 2 summarizes the mental models that are based on the five main sentential connectives, and it also shows the corresponding fully explicit models for these assertions in which the false cases are represented as true negations. Readers may worry that the theory is based on dubious concepts, such as mental footnotes and ellipses, and on hidden assumptions that can be added or dropped as needed to account for experimental results. In fact, the theory postulates that individuals normally reason using mental models but that, in simple cases, individuals can flesh out their models to make them fully explicit. All the specific predictions about reasoning with sentential connectives derive from Table 2. Mental footnotes, which are not shown in the table, indicate what is exhaustively represented in the mental models, and the notion underlies the computer programs modeling both propositional and syllogistic reasoning (see Johnson-Laird & Byrne, 1991). We have sometimes used square brackets to represent mental footnotes, but there is no need for them in the present article, and so we forgo them. The ellipses represent wholly implicit models, that is, models that serve as "placeholders" representing other possibilities that as yet have no explicit content. To explain certain tasks going beyond deduction, such as Wason's (1966) selection task, we have made additional assumptions. Likewise, in the present article, we make additional explicit assumptions to explain naive probability.

There is much evidence in support of the model theory. Deductions that call for a greater number of mental models are more difficult, taking longer and leading to more errors. Erroneous conclusions tend to correspond to individual mental models of premises (for a review, see Johnson-Laird & Byrne, 1991). Reasoners tend to focus on what is explicit in their models and thus to be susceptible to various "focusing effects," including an influence of the verbal framing of premises on deductive reasoning (see Legrenzi, Girotto, & Johnson-Laird, 1993). A recent study (Girotto, Evans, & Legrenzi, 1996) has shown that the difficulty of coping with multiple models is the main source of "pseudodiagnosticity," that is, the tendency to prefer information for the focal hypothesis over information about its alternative (see Mynatt, Doherty, & Dragan, 1993).

The model theory offers a unified account of reasoning leading to necessary conclusions and reasoning leading to conclusions about possibilities. A conclusion is necessary—it must be true—if it holds in all the models of the premises; a conclusion is possible—it may be true—if it holds in at least one model of the premises. Recent evidence confirms the theory's prediction of a key interaction (Bell & Johnson-Laird, 1998): It is easier to infer that a situation is possible (which demands only a single model of an example) than to infer that it is not possible (which demands its nonoccurrence in all models), whereas it is easier to infer that a situation is not necessary (which demands only a single model of a counterexample) than to infer that it is necessary (which demands its occurrence in all models).

The Model Theory of Naive Probability: Absolute Probabilities

The mental model theory applies in a natural way to naive probability. In particular, it accounts for extensional reasoning

Table 2
Models for the Sentential Connectives

Connective	Mental models	Fully explicit models
A and B	A B	A B
A or else B	A	A ¬B
	B	¬A B
A or B or both	A	A ¬B
	B	¬A B
	A B	A B
If A, then B	A B	A B
	...	¬A B
		¬A ¬B
If, and only if A, then B	A B	A B
	...	¬A ¬B

Note. See text for description. ¬ = negation; ... = wholly implicit model.

about probabilities, that is, reasoning that yields the probability of an event from the different possible ways in which it could occur. This account depends on three fundamental principles, the first of which we have already encountered.

1. The *truth* principle: People represent situations by constructing sets of mental models in which each model represents what is true in a true possibility. In contrast, a set of fully explicit models represents the mutually exclusive true possibilities completely.

Situations with a zero probability correspond to what is false, and so the theory implies that they are not normally represented in models.

2. The *equiprobability* principle: Each model represents an equiprobable alternative unless individuals have knowledge or beliefs to the contrary, in which case they will assign different probabilities to different models.

This principle is a variant on the "number of cases" intuition proposed by Shimojo and Ichikawa (1989). It differs from their subjective theorem in two ways: First, it applies to mental models of the partition; second, equiprobability is assumed only by default. An analogous principle of indifference or insufficient reason has a long history in classical probability theory (see Hacking, 1975), and we spell out below the difference between indifference and the equiprobability principle.

3. The *proportionality* principle: Granted equiprobability, the probability of an event, A, depends on the proportion of models in which the event occurs; that is, $p(A) = \frac{n_A}{n}$, where n_A is the number of models containing A, and n is the number of models.

A corollary is the *inclusion* principle: If one event, A, occurs in each model in which another event, B, occurs, then A is at least as probable as B, and if, in addition, A occurs in some models in which B does not occur, then A is more probable than B.

The equiprobability principle is similar to the classical principle of indifference, which Howson and Urbach (1993) express in terms of conditional probabilities, "If there are n mutually exclusive possibilities $h_1, \ldots, h_n$, and e gives no more reason to believe any one of these more likely to be true than any other, then $P(h_i|e)$ is the same for all i" (p. 52), where $P(h_1|e)$ is the conditional probability of h_1 given that the evidence, e, is the case. The principle of indifference lies at the heart of Laplace's (1820/1951) rule of succession—the notorious expression that if an event, A, occurs r times in s trials (or repeated observations), then A's probability on trial $s + 1$ is

$$\frac{r+1}{s+2}.$$

It was this formula that enabled Laplace to compute the odds that the sun would rise on the morrow as 1,826,214 to 1 (see Howson & Urbach, 1993, pp. 52 et seq.).

The problem with the principle of indifference, as these authors point out, is that it yields inconsistencies depending on how one chooses to partition possibilities. Suppose, for example, there are three marbles—a red, a green, and a blue marble—and a box contains the red marble, the green marble, or the blue marble, but the red marble is always with the green or the blue marble, but not both of them. One partition splits the possibilities into the red marble, the green marble, or the blue marble and then splits the cases of the red marble into one with the blue marble and one with the green marble. Indifference then yields a probability of 1/3 for selecting a box containing the red marble. Another partition splits the possibilities into the red and green marble, the red and blue marble, the green marble, and the blue marble; indifference now yields a probability of 1/2 for selecting a box with the red marble. The same difficulty applies to Shimojo and Ichikawa's (1989) and Falk's (1992) accounts of intuitive beliefs about probabilities. The principle of number of cases applies to alternatives, but the theories do not specify what counts as an alternative.

In contrast, the present principle of equiprobability applies to partitions corresponding to the mental models that reasoners construct. Table 2 embodies a theory of how models are based on sentential connectives, and so it is these models that yield the partition for a given description, which are then assumed to represent equiprobable alternatives. Given a box containing at least one of the three marbles, red, green, and blue, and the description,

If the red marble is in the box, then either the green or else the blue marble, but not both, is in the box too,

the theory postulates that reasoners should construct the following mental models:

red	green	
red		blue
	...	

though they are likely to forget the implicit model denoted by the ellipsis. In any case, they should assume equiprobability and use proportionality to infer a probability for a box with a red marble (p = 2/3 if they remember the implicit model and p = 1 if they forget it). The fully explicit set of models shows the real partition for the problem:

red	green	¬blue
red	¬green	blue
¬red	green	¬blue
¬red	¬green	blue
¬red	green	blue

Granted equiprobability over this partition, then proportionality yields a probability of 2/5 for a box containing at least the red marble.

The principle of equiprobability holds only by default. If individuals have knowledge or beliefs to the contrary, they will not assume equiprobability. For example, they tend not to assume that all the horses in a race are equally likely to win; they know that the favorite is more likely to win. Indeed, what we mean by a *default assumption* is one that can be abandoned in the face of evidence to the contrary. Default assumptions, which have a long history in cognitive science, play a critical role in the construction of mental models (see Johnson-Laird & Byrne, 1991, chap. 9). We return to the more subtle effects of knowledge and beliefs later in the article.

The three principles of naive probability combine to yield predictions about probabilistic reasoning. Individuals should construct models of the true possibilities on the basis of the premises (the truth principle). In the absence of contrary evidence, individuals should assume that the models represent equally probable alternatives (the equiprobability principle), and they should infer that the probability of an event, *A*, equals the proportion of models in which *A* occurs (the proportionality principle):

$$\mathrm{P}(A|premises) = \frac{n_A}{n},$$

where P(*A*|*premises*) is the conditional probability of *A* given the premises, n_A is the number of models of the premises in which *A* occurs, and *n* is the number of models of the premises. The twist in this prediction is that equiprobability applies to mental models, and mental models represent only what is true within the true possibilities.

Numerical Premises

The three principles provide an account of both basic extensional reasoning and inferences about relative probability. Relative probabilities, such as

> A is more likely than B

can be expressed without numerical values, and thus inferences about them can be drawn by individuals in nonnumerate cultures. Educated members of industrialized societies, however, are numerate, and so they are at least familiar with numerical values for probabilities even if they are naive about the calculus. With small integral values, such as

> The chances of a red marble in the box are 3 out of 4; otherwise, there is a green marble,

they can represent the probability by the relative frequency of the models,

red
red
red
green

and they can draw conclusions on the basis of proportionality. When the numerical values are larger, the theory postulates a more general procedure.

> 4. The *numerical* principle: If a premise refers to a numerical probability, the models can be tagged with their appropriate numerical values, and an unknown probability can be calculated by subtracting the sum of the $(n - 1)$ known probabilities from the overall probability of the *n* possibilities in the partition.

The procedure remains extensional but generalizes to any sort of numerical values, including probabilities expressed in terms of frequencies, fractions, decimals, or percentages (see also Stevenson & Over, 1995, who also postulated tagging models with probabilities). As an example, consider the problem

> There is a box in which there is one and only one of these marbles, a green marble, a blue marble, or a red marble. The probability that a green marble is in the box is .6, and the probability that a blue marble is in the box is .2. What is the probability that a red marble is in the box?

The premises are represented by the following annotated mental models

	Probabilities
green	0.6
blue	0.2
red	

and so the probability that the red marble is in the box equals .2.

If there is a stated probability for one possible event, but not for several alternative events, then is the equiprobability principle likely to apply to them? For example, suppose that we drop one of the stated probabilities from the previous problem:

> There is a box in which there is one and only one of these marbles, a green marble, or a blue marble, or a red marble. The probability that a green marble is in the box is .6. What is the probability that a red marble is in the box?

The theory commits us to the view that, by default, reasoners should assume that the blue and red marble are equiprobable. This tendency, however, is likely to be enhanced where the stated probability obeys the equiprobability assumption; for example, the probability of a green marble is .33 in the previous case.

The introduction of numbers into probabilistic problems complicates matters. Reasoners can no longer rely on the basic extensional principles alone (truth, equiprobability, and proportionality). Moreover, numbers call for calculations, and, most importantly, their difficulty depends in part on the particular numerical system. Consider, for example, the following three calculations:

$$\text{The probability of A} = \frac{(0.1)(0.8)}{0.533},$$

$$\text{the percentage probability of A} = \frac{(10\%)(80\%)}{(53.3\%)},$$

$$\text{the chances of A} = \frac{(1/10)(8/10)}{(8/15)}.$$

They are not equally easy; only the third calculation can be carried out mentally. Yet all three calculations express equivalent results. In general, extensional reasoning from numerical premises will be affected by the relative difficulty of the numerical calculations.

The probability calculus, as we have seen, distinguishes between absolute probabilities and conditional probabilities. We use

the same organization in what follows. We have shown how the model theory applies in principle to inferences concerning absolute probabilities, and in the next section, we present evidence in favor of this account. We then consider the model theory's explanation of biases in reasoning about absolute and relative probabilities. Finally, we consider conditional probabilities, Bayesian reasoning, and the theory's pedagogical implications.

Extensional Reasoning About Absolute Probabilities

Given the principles of basic extensional reasoning, the probability of an event should depend on the proportion of mental models in which it occurs. There were no relevant empirical results in the literature, and so we carried out two experiments to test this central prediction of the theory of naive probability. What we aimed to show in these studies was, first, that naive individuals who have no knowledge of a situation will adopt the equiprobability assumption and then use proportionality to infer the probabilities of events. As we saw earlier, Shimojo and Ichikawa (1989) and Falk (1992) postulated that naive individuals believe that given N alternatives the probability of any one of them is $1/N$. Our second aim was accordingly to show that this principle of equiprobability applies to mental models, not to the actual alternatives. The divergence should be most apparent in the case of conditionals, which have two mental models but which are compatible with three actual possibilities (see Table 2).

Experiment 1: A Preliminary Study of Extensional Reasoning

Our preliminary study tested whether naive reasoners spontaneously adopt the equiprobability principle. Because we wanted to isolate the problems from the participants' expectations based on general knowledge or beliefs, we used the same sort of contents as the bookbag and poker chips of many other studies of probabilistic inference (see, e.g., von Winterfeldt & Edwards, 1986). The premises were in the form of inclusive disjunctions, and a typical problem was:

> There is a box in which there is a black marble, or a red marble, or both.
>
> Given the preceding assertion, according to you, what is the probability of the following situation?
>
> In the box there is a black marble with or without another marble.
>
> Probability: __%

Expert reasoners may refuse to answer on the grounds that the problem is ill-posed; it is difficult, if not impossible, to assess the probability of a single unique event, and even if the problem were based on repeated events, the probability could be anywhere between 0% to 100% because there is no information about the respective probabilities of the three possibilities. Naive reasoners, if the model theory is right, should have no such scruples. The premise yields three models, which represent the true components of the true possibilities:

```
black
      red
black red
```

These models are the components of the partition with nonzero probabilities. Granted their equiprobability, the probability of a black marble with or without another marble (e.g., the red one) depends on the proportion of models in which the black marble occurs, namely, 67%. So, if the model theory is correct, then reasoners should tend to infer the following probabilities about the contents of the box: probability of a black marble with or without another = 67%, probability of a black marble and a red marble = 33%, probability of a black marble without a red marble = 33%, probability of neither a black marble nor a red marble = 0%. It seemed possible, however, that naive reasoners would either balk at the task—much as expert reasoners should—or else make inferences in some other, perhaps haphazard, way.

Method. The participants acted as their own controls and carried out each inference on the basis of a separate premise with a different content. The order of the problems was as they are stated in the preceding section. This experiment and the subsequent ones were carried out in French with native speakers. The four problems all concerned objects in a box, but each problem was about different objects in different colors. The participants were tested individually, and the instructions at the top of the first sheet described how they were to make their estimates in terms of percentages:

> With each of the following questions, we would like you to give estimates of the probability of certain events. Your estimates must be given in percentages (from 0% [*impossible*] to 100% [*certain*]) on the basis of the following scale.

```
0%        50%        100%
|----------|----------|
Impossible            Certain
```

Each problem was printed on a separate sheet, and the participants worked through the problems in the fixed order and were not allowed to go back to a previous problem. We tested 13 volunteers, who were students at the University of Aix-en-Provence, native speakers of French, and unfamiliar with the probability calculus.

Results and discussion. Table 3 presents the results. In order to examine the data of all participants, and to allow for numerical inexactitude, we analyzed how many of each participants' inferences were within ± 5% of the predicted values. We adopted this analysis in all of our experiments. The a priori probability of answering a question by chance within ± 5% of the predicted value is 1/10, and so a participant who makes at least one such inference out of the four trials is performing better than chance. In fact, 2 out of the 13 participants inferred probabilities that matched exactly the numerical predictions, 3 participants were within ± 5% of the predictions on three of their inferences, and all but 1 of the 13 participants performed better than chance (Sign test, $p < .002$).

The probability of A should be greater than, or equal to, the probability of A *and* B, otherwise the inferences commit an extensional version of the "conjunction fallacy"; that is, the participants infer that the probability of a conjunction is greater than the probability of one of its conjuncts (see Tversky & Kahneman, 1983). This fallacy often occurs in nonextensional reasoning about a conditional probability. Thus, its typical demonstrations call for estimates of two probabilities, such as $p(A|C)$ and $p(A \text{ and } B|C)$, where C is a description of an individual, say, Charles, A is a category such as "accountant," and B is a category such as "plays baseball as a hobby." The

Table 3
The Predicted Percentage Inferences and the Means of the Observed Percentage Inferences in Experiment 1 for Inclusive Disjunctions of the Form A or B, or Both

	Types of conclusions			
Results	p(*A*)	p(*A* and *B*)	p(*A* and not-*B*)	p(not-*A* and not-*B*)
Predicted percentages	66	33	33	0
Observed percentages	56	27	33	7

evidence shows that the conjunction fallacy derives from a difference in the representativeness of *C* as an instance of *A* and as an instance of *B*. In extensional reasoning, the model theory predicts the conjunction fallacy if reasoners omit, or forget, certain models of the premises. Thus, given a conditional *if A, then B,* the fallacy would occur if a participant forgot the implicit model in assessing $p(A \text{ and } B)$, which would yield a probability of 100%, but did not forget the implicit model in assessing $p(A)$, which would yield a probability of 50% (see the mental models in Table 2). In our experiment, comparable oversights are unlikely with simple disjunctions, and indeed 12 out of 13 participants fitted the correct inequality (Sign test, $p < .002$). The probability *A* should also be greater than, or equal to, the probability of *A and not-B,* and it was for all 13 participants ($p = .5^{13}$, i.e., less than 1 in 8,000). Finally, 7 participants inferred the probability of *A* (e.g., a black marble) as 50% instead of the predicted 67%. One reason may be that they thought that the two possibilities, a black marble or red marble, were equiprobable. In other words, they may have neglected the description "with or without another marble," or they may have thought that this description did not include the red marble. We clarified the statement of the question in the next experiment.

In general, the participants were happy to carry out the task and did not balk at the inferences. They did indeed appear to assume that each model was roughly equiprobable, as shown by the fact that both the probability of *A and B* and of *A and not-B* were inferred as around 33% by more than half the participants. These results corroborated the principles of equiprobability and proportionality, but they concerned only a single connective. Our next task was to test the model theory's prediction for a broader range of sentential connectives.

Experiment 2: Extensional Probabilities for the Main Connectives

This experiment examined three connectives: exclusive *or,* inclusive *or,* and *if.* Because *and* does not allow any inferences that are not definite, we used it only in two practice trials.

For exclusive disjunction, *A or B but not both,* the model theory (see Table 2) predicts that individuals should construct two models,

A
B

and so they should infer probabilities of 50% for *at least A* and for *A and not-B,* and probabilities of 0%, for *A and B* and for *neither A nor B.* For inclusive disjunction, *A or B or both,* the model theory makes the predictions that we described for Experiment 1; that is, individuals should infer a probability of 67% for *at least A,* probabilities of 33% for *A and B* and for *A and not B,* and a probability of 0% for *neither A nor B.*

The conditional premises are more critical for a comparison between the mental model theory and the accounts of Shimojo and Ichikawa (1989) and Falk (1992). A conditional *if A, then B* is compatible with three distinct possibilities (see the fully explicit models in Table 2), whereas it has only two mental models:

A B
. . .

Hence, according to the model theory, individuals should infer probabilities of 50% for *at least A* and for *A and B;* however, they are likely to construct other models when they are asked to infer the probability of the corresponding state of affairs. That is, individuals are most unlikely to hold in mind all the possible models of the conditional. According to the theory, they will make a mental footnote that A occurs only in the explicit model, and so they should infer a probability of 0% for *A and not-B.* They should be able to flesh out the implicit model to represent the state of affairs:

¬A ¬B

They will contrast this model with the initial explicit model and accordingly infer the probability of this new model as 50%. If they omit or forget the implicit model—a common failing when there is another connective in the same assertion (see, e.g., Johnson-Laird & Savary, 1996)—then they will infer a probability of 100% for *A* and for *A and B;* that is, they will treat the conditional as though it were a conjunction. A conditional *if A, then B* can be, and often is, interpreted as a biconditional. In this case, the two models exhaust the set of possibilities,

A B
¬A ¬B

and so participants who make the biconditional interpretation will infer a probability of 0% for *B and not A.* Otherwise, they will construct the model

¬A B

and infer a probability of 50% (comparing this model with only the initial explicit model). In summary, the model theory's predictions for the conditional are that participants should infer probabilities of 50% for *at least A,* for *A and B,* and for *neither A nor B* and a probability of 0% for *A and not-B.* The probability that they infer for *B and not-A* will be either 0% (the biconditional interpretation)

or 50% (the conditional interpretation). Because participants should construct certain models only when they are asked certain questions, they are likely to overestimate probabilities so that they sum to greater than 1, particularly in the case of a conditional interpretation.

Neither Shimojo and Ichikawa (1989) nor Falk (1992) described how people represent alternatives, but they did not deny that individuals have access to the actual possibilities. In the case of a conditional, individuals should therefore infer probabilities of 33% for *at least A,* for *A and B,* and for *not-A and B* and a probability of 0% for *A and not-B.* In the case of a biconditional interpretation, they should infer probabilities of 50% for *A and B* and *not-A and not-B* and 0% for *A and not-B* and *not-A and B.* The possible omission of the implicit model and the conditional rather than the biconditional interpretation discriminate between the two theories.

Method. There were 13 experimental problems (4 for exclusive *or,* 4 for inclusive *or,* and 5 for *if*). The participants acted as their own controls, and after two practice trials with *and,* they carried out all the experimental problems. These were presented in a different random order to each participant, with the constraint that no problem with a particular connective was immediately followed by another problem with the same connective. The format of the trials was the same as those in Experiment 1. The statements of the four connectives were as follows:

A and B: There is a box in which there is a yellow card and a brown card.

A or B, not both: There is a box in which there is a yellow card, or a brown card, but not both.

A or B, or both: There is a box in which there is a yellow card, or a brown card, or both.

If A then B: There is a box in which if there is a yellow card then there is a brown card.

The five sorts of questions were as follows:

At least A: In the box there is at least a yellow card.

A and B: In the box there is a yellow card and a brown card.

A and not B: In the box there is a yellow card and there is not a brown card.

B and not A: In the box there is a brown card and there is not a yellow card.

Neither A nor B: In the box there is neither a yellow card nor a brown card.

The formulation of the question about the probability of *A* was modified from the version in Experiment 1 to *at least A* in order to simplify it.

The content of all 15 problems was distinct. Once again, we needed to isolate it from the participants' knowledge, and so it was based on five objects—marble (*bille*), cube (*cube*), ball (*boule*), card (*carte*), and chalk (*craie*)—and 15 distinct pairs of colors from the following list: blue (*bleu*), yellow (*jaune*), brown (*marron*), black (*noir*), green (*vert*), red (*rouge*). Two separate allocations of the resulting contents were made to the 15 problems; half of the participants were tested with one allocation and half were tested with the other allocation.

The participants were tested individually, and the instructions and procedure were identical to those of Experiment 1, except that each participant worked through the problems in a different random order. We tested 22 volunteers from the same population as before.

Results and discussion. We deal with the results for each connective separately, and, as before, we analyze how many inferences were within ± 5% of the predicted values, so a participant who makes at least one such inference out of four trials is performing better than chance.

Table 4 summarizes the results for exclusive disjunction, *A or B but not both.* Of the 22 participants, 8 inferred exactly the predicted numerical values, a further 9 made one departure from the four predictions, and all participants performed better than chance (Sign test, $p = .5^{22}$, i.e., less than 1 in 4 million).

Table 5 summarizes the results for inclusive disjunction: *A or B or both.* In this case, 8 participants inferred exactly the predicted numerical values, a further 5 made just one departure from the predicted pattern, and only 1 out of the 22 participants failed to perform better than chance (Sign test, $p \ll .0005$). There was no apparent difference between p(*A* and *B*) and p(*A* and not-*B*): 14 ties, and the remaining 8 participants were split equally. Hence, the results corroborated the pattern that we observed in Experiment 1.

The analysis of the results for the conditionals of the form *if A, then B* is more complicated because of their potential ambiguity, which concerns p(*B* and not-*A*). Table 6 therefore summarizes the results for the other four inferences. In this case, 9 of the participants inferred exactly the predicted numerical values, a further 3 made just one departure from the predicted pattern, and all of them performed better than chance (Sign test, $p = .5^{22}$, i.e., less than 1 in 4 million). As the theory predicted, some participants appeared to forget the implicit model, and thus 4 participants inferred a 100% probability for *at least A,* and 8 participants inferred a 100% probability for *A and B.* The model theory also predicted that participants should tend to infer higher probabilities for *A and B* than for *neither A nor B;* both are possible for conditional and biconditional interpretations, but only the former corresponds to an

Table 4
The Predicted Percentage Inferences and the Means of the Observed Percentage Inferences in Experiment 2 for Exclusive Disjunctions of the Form A or B but not Both

	Types of conclusions			
Results	p(A)	p(A and B)	p(A and not-B)	p(not-A and not-B)
Predicted percentages	50	0	50	0
Observed percentages	45	6	53	16
No. of participants[a] within ±5%	16	19	18	16

[a] $n = 22$.

Table 5
The Predicted Percentage Inferences and the Means of the Observed Percentage Inferences in Experiment 2 for Inclusive Disjunctions of the Form A or B or Both

Results	Types of conclusions			
	p(A)	p(A and B)	p(A and not-B)	p(not-A and not-B)
Predicted percentages	67	33	33	0
Observed percentages	60	45	44	7
No. of participants[a] within ±5%	9	15	14	20

[a] $n = 22$.

initially explicit model. The predicted difference occurred for 8 participants, and the remaining 14 participants were ties (Sign test, $p < .004$). Every participant inferred a probability of 0% for *A and not-B,* the only condition in the experiment where everyone made the same response. The inferences for *B and not-A* reflect the interpretation of the conditional: 12 participants inferred a probability of 0% (the biconditional interpretation), 4 participants inferred a probability of 50% (the conditional interpretation), and the remaining 6 participants inferred some other probability.

The inferences for *if A, then B* corroborated another predicted pattern. Responses can be either for a conditional or biconditional interpretation. In either case, we have inferences for all four possibilities, and they should sum to 100%. A biconditional has fewer explicit models than a conditional, and those participants who made the biconditional interpretation tended to infer probabilities that summed correctly to 100%, whereas those participants who made a conditional interpretation tended to infer probabilities that summed to more than 100%. If we classify as a biconditional interpretation those patterns of responses that assign positive probabilities to *A and B* and to *neither A nor B* and 0% probabilities to the other two cases, then 7 out of 9 participants inferred probabilities that summed to 100%. Likewise, if we classify as a conditional interpretation those patterns of inferences that assign 0% probability only to *A and not-B,* then 7 out of 8 participants inferred probabilities that summed to more than 100%. Hence, as predicted, the participants were more likely to make overestimates when they had to contend with three models as opposed to only two models (Fisher–Yates exact test, $p < .005$). They fail to bring to mind all the models of the premises and so overestimate the probability of the model that corresponds to the event for which they are trying to infer a probability (cf. the subadditivity predicted by Tversky & Koehler's, 1994, "support" theory of nonextensional reasoning).

The relative difficulty of deductive reasoning with different connectives has the following trend: *if* is easier than exclusive *or,* which in turn is easier than inclusive *or* (Johnson-Laird & Byrne, 1991). This pattern reflects the number of explicit models that are needed to make deductions. The present task is different in that it calls for participants to consider possibilities that are not normally represented explicitly, and, as we have seen, the participants had some difficulty with them. The ambiguity of *if* also complicates matters. However, we can make a straightforward comparison between inclusive and exclusive disjunction: The mean number of inferences (out of 4) according to the predictions was 3.1 for exclusive disjunction and 2.6 for inclusive disjunction; 9 participants performed better with exclusive disjunctions, 4 participants performed better with inclusive disjunctions, and there were 9 ties (Wilcoxon's $T = 22$, $n = 13$, $z = 1.7$, $p < .05$, one-tailed).

Overall, the results corroborated the theory of naive probability. Participants appear to infer probabilities by constructing models of the premises, adopting the equiprobability principle, and assessing the proportion of models in which the events occur. Their inferences from conditional premises of the form *if A, then B* were particularly relevant. As the model theory predicted, the participants tended to infer that the probability of *A and B* was 50%. Because 12 of the participants interpreted the premise as a biconditional, it is hard to decide between the model theory and the theory of Shimojo and Ichikawa (1989) and Falk (1992); both theories predict a probability of 50% for *A and B* in the case of a biconditional. A more telling datum in favor of the model theory is that 8 participants inferred the probability of *A and B* as 100%, which is predicted from the omission of the implicit model of the conditional. The participants could construct other models when the questions called for them, but they then tended to compare each new model with just the single prior explicit model.

Could it be that the participants in the experiment were merely

Table 6
The Predicted Percentage Inferences and the Means of the Observed Percentage Inferences in Experiment 2 for Conditionals of the Form If A, Then B

Results	Types of conclusions			
	p(A)	p(A and B)	p(A and not-B)	p(not-A and not-B)
Predicted percentages	50	50	0	50
Observed percentages	58	68	0	38
No. of participants[a] within ±5%	14	12	22	12

[a] $n = 22$.

seeking to oblige us, that the situation led them to make responses in which they did not believe but that they thought the experimenter expected? We doubt it, because the participants clearly put thought into their answers, which were often not the precise proportions predicted by the model theory. Moreover, when we have put similar questions to experts in the probability calculus, they typically balk at the question or describe the range of possible probabilities. However, because the participants knew nothing about the circumstances in which the marbles, or other objects, were placed in the box or withheld from it, how could they experience intuitions about the requested probability? The fact is, of course, that people do have intuitions about probabilities even when they know nothing about the relevant circumstances. They judge the probability of heads as roughly 1/2 even when they know nothing about whether the toss or coin was fair. Another worry is whether our sample of participants is representative of naive individuals. As a reviewer of this article wrote, "It is difficult for me to believe that a random sample of Chicagoans, for example, would have responded to such an ill-defined question with much more than a blank stare." However, Shimojo and Ichikawa (1989) also observed that their participants from the University of Tokyo believed that if there were *N* possible events, then the probability of any one of them was 1/*N*. Falk (1992) also reported that her participants were strongly committed to it. The principle of equiprobability is thus not unique to our studies. The reviewer may well be right about Chicagoans. The moral that we draw is that they, unlike students in Aix-en-Provence, Tokyo, and the Hebrew University of Jerusalem, are not naive about the probability calculus.

Another worry arises about problems that are based on a premise of the form *A or B or both.* It describes three possibilities in three constituents of the sentence, and so the participants might have estimated *A* as having a probability of 1/3 purely on the basis of a superficial linguistic analysis. We examine the feasibility of this explanation next, and we return to the contrast between actual alternatives and the mental model theory's prediction of biases.

Biases in Extensional Reasoning

The results of our previous studies can be explained by the model theory and perhaps by the theory proposed by Shimojo and Ichikawa (1989) and endorsed by Falk (1992). The results might even be explained by extending a rule theory of the sort that we described earlier. So, how can we test whether individuals rely on mental models? The answer is that mental models predict the existence of systematic biases in extensional reasoning because models represent only what is true, not what is false. As readers will recall, this principle of truth applies at two levels: Individuals construct models that make explicit only true possibilities, and they make explicit only those propositions that are true within them. It is important to emphasize that what is omitted concerns falsity, not negation. For example, given an inclusive disjunction, *not-A or B,* reasoners should construct the following three models:

¬A	
	B
¬A	B

where each model represents only those literal propositions (affirmative or negative) in the premise that are true within each true possibility. Thus, B is false in the first model, and not-A is false in the second model, but this information is represented in mental footnotes that are easily forgotten. For certain premises, the loss of these footnotes should produce biased extensional inferences. The mental models yield different partitions from those based on fully explicit models. By definition, premises that yield different answers in these two cases are potential experimental problems in which the participants' inferences should follow the mental models, whereas premises that yield the same answers in the two cases are potential control problems in which the participants' inferences should be based on the actual partition corresponding to the fully explicit models. We wrote a computer program implementing the construction of both mental models and fully explicit models, and it searched systematically for both sorts of premises in the vast space of possible premises.

Here is an example of an experimental problem that should create a bias:

> There is a box in which there is at least a red marble, or else there is a green marble and there is a blue marble, but not all three marbles.
>
> Given the preceding assertion, what is the probability of the following situation?
>
> In the box there is a red marble and a blue marble.

The mental models of the premise are as follows:

red		
	green	blue

Reasoners should therefore infer that the probability of a red marble and a blue marble is 0%. The fully explicit models of the premises, however, take into account that where it is true that there is a red marble, there are three distinct ways in which it can be false that there is both a green marble and a blue marble:

red	green	¬blue
red	¬green	blue
red	¬green	¬blue
¬red	green	blue

The unbiased inference based on the actual partition is that a red marble and a blue marble, granted equiprobability, has a probability of 25%.

In general, we define an *unbiased inference* as one that applies the equiprobability principle to the actual alternatives, which correspond to the fully explicit models. The following control problem, for example, should elicit an unbiased inference, because its mental models yield the same inference as its fully explicit models:

> There is a box in which there is a grey marble and either a white marble or else a mauve marble, but not all three marbles are in the box.
>
> Given the preceding assertion, what is the probability of the following situation?
>
> In the box there is a grey marble and there is a mauve marble.

The premise elicits the mental models

grey	white	
grey		mauve

and so participants should respond 50%. The fully explicit models of the premise are as follows:

grey	white	¬mauve
grey	¬white	mauve

They support exactly the same inference, and so it is unbiased, granted the equiprobability principle. Experiment 3 was designed to compare the model theory's predictions about biases with those of the other theories of extensional reasoning.

Experiment 3: A Test of Bias in Extensional Reasoning

The experiment compared two sorts of problems: experimental problems, in which the mental models predict systematic biases, and matched control problems of a comparable complexity, in which the mental models of the premises support the same inferences as the fully explicit models. The model theory should predict the inferences for both sorts of problems. However, if reasoners base their inference on actual alternatives (cf. Falk, 1992; Shimojo & Ichikawa, 1989), then the experimental and control problems alike should yield unbiased inferences, that is, conclusions that correspond to the fully explicit models.

Method. The experiment was based on 3 experimental premises and 3 matched control premises, and each of these 6 premises was used to elicit 3 different inferences, that is, a total of 18 problems preceded by 2 practice problems. (In fact, the experiment compared 4 pairs of premises, but we describe only 3 pairs because a design error vitiated the 4th pair.) The 3 inferences with each premise were made on separate trials with a different content, and each trial consisted of the presentation of a single premise and a single question about a probability (as in Experiment 2). The problems were presented in a different random order to each participant, with the constraint that no problem from a set of the 6 matching experimental and control problems was ever immediately followed by a problem from the same set.

The forms of the problems are summarized in Appendix A, together with their mental models and the fully explicit models. The appendix table also shows the predicted and the unbiased values of the probabilities. Because the premises containing conditionals also include a disjunction, the implicit model of a conditional should tend to be forgotten (Johnson-Laird & Savary, 1996), and so, as the table shows, the predictions do not take implicit models into account. One experimental problem had a predicted value that was unbiased (see the last problem for Premise 3 in Appendix A), but the rest had predicted values that were biased, whereas all the predicted values for the control problems were unbiased. Each problem was about different colored objects in a box. They were based on six different objects: a flower (*une fleur*), a card (*une carte*), a cup (*une tasse*), a ball (*une boule*), a chalk (*une craie*), and a marble (*une bille*). We devised two separate sets of distinct triples of colors drawn from the following set: black (*noire*), white (*blanche*), red (*rouge*), blue (*bleue*), green (*verte*), and brown (*marron*). The six objects were assigned at random four times to the triples of colors. The resulting contents were assigned in two separate random allocations to the problems. Half of the participants were tested with one of the resulting sets of contents, and half were tested with the other of the resulting sets of contents.

The participants were tested individually. The instructions were identical to those of Experiment 1, and at the top of the first sheet the instructions described how the participants were to make their estimates in terms of percentages from 0% (*impossible*) to 100% (*certain*). As before, each problem was printed on a separate sheet, and the participants worked through the problems in random order. We tested 25 new participants from the same population as before.

Results and discussion. Appendix B presents the results of the experiment for each of the nine pairs of matched experimental and control problems. It shows for each problem the number of participants (from $n = 25$) who made the predicted inference (within ± 5%) and the number of participants who made the unbiased inference (within ± 5%, on the assumption of equiprobability over the actual partition corresponding to the set of fully explicit models). Overall, the participants made a mean of 5.6 unbiased inferences for the nine control problems but only a mean of 1.1 unbiased inferences for the eight experimental problems (discounting the problem in which the predicted value was unbiased), and every single participant made more inferences that were unbiased with the control problems than with the experimental problems ($p = .5^{25}$, i.e., a probability of less than 1 in 33 million). Likewise, all of the control problems elicited more unbiased inferences than their corresponding experimental problems ($p = .5^{8}$, i.e., $p < .005$).

In general, the participants' inferences matched the predictions of the model theory reliably better than chance; that is, they corresponded to the mental models of the premises. The mean numbers of matches were 5.8 for the nine experimental problems and 5.6 for the nine control problems. The chance probability of a match is 1/10, and so any participant who makes two or more matches for a set of nine problems is, conservatively, responding at a better than chance level. All participants performed at a better than chance level for both the experimental and the control problems ($p = .5^{25}$ in both cases).

The model theory's predictions were borne out for all 18 problems except for 1 control problem. For the control premise 1′

> A and either B or else C, but not all three

only 6 participants inferred the unbiased probability of *at least A* as 100%, and 10 participants made inferences of 50%. One possible explanation is that the particular question led the majority to mis-parse the structure of the premise, treating it as equivalent to

> A and B, or else C, but not all three,

which yields the models

A B

C

Is it possible that the participants' inferences depended on a reinterpretation of the premises according to the pragmatics of daily life? In particular, given a premise for Problem 1, such as

> There is a box in which there is at least a red marble, or else there is a green marble and there is a blue marble, but not all three marbles,

the participants may have assumed that if there was a red marble in the box, then there was no other marble there. This factor may have played a part in performance with this particular problem. But, consider a premise for Problem 3, such as

> There is a box in which if there is a red marble in the box, then there is either a green marble or else a blue marble, but not all three marbles.

As the model theory predicts, the participants tended to overlook the possibilities in which it was false that there was a red marble in the box. The phenomenon may be compatible with the prag-

matics of everyday usage, but pragmatic theories do not, at present, predict that individuals represent only what is true.

The unbiased values differed between the control problems and the experimental problems. However, if we consider the subset of problems with the same unbiased values of 0% or 33%, of which there are two experimental problems and five control problems, there is still a reliable difference in performance. The participants made a total of 6 unbiased inferences for the two experimental problems (a mean of 3.0 out of 25 for each problem) but a total of 86 unbiased inferences for the five control problems (a mean of 17.2 out of 25 for each problem), and the difference was highly reliable (Sign test, $p < .0005$). Hence, the imbalance in the distributions of unbiased values does not explain why the control problems elicited more unbiased inferences than the experimental problems.

The experiment's purpose was to establish that reasoning relies on mental models. It showed that mental models lead predictably to biases for the experimental problems, but not for the control problems. Are the results attributable merely to the complexity of the problems—that is, are the biases simply a result of an increase in the number of real alternatives? Certainly, the experimental problems have more real alternatives than the control problems. Mere complexity does not predict the participants' inferred probabilities, however. The model theory predicts the difficulty of inferences—the greater the number of models the harder the task—but it also predicts the inferences drawn by the participants. Hence, the results show that naive individuals reason extensionally, not from the actual partition of a problem (the fully explicit models) but from the partition corresponding to its mental models. The results corroborate the model theory and show that Shimojo and Ichikawa's (1989) and Falk's (1992) accounts are viable only if they assume that what is equiprobable are mental models. Likewise, any extension of rule theories to naive probability must similarly allow for the equiprobability of mental models.

Models and Illusory Inferences About Relative Probabilities

The model theory predicts that naive reasoners infer the probabilities of events from the proportions of mental models in which these events occur, and the previous experiments have confirmed this prediction. The theory also makes predictions about inferences of relative probability, that is, that one event is more probable than another. In these inferences, however, reasoners might rely on an alternative to the principle of proportionality. They could use the less risky principle of inclusion:

> If A occurs in each model in which B occurs, but A occurs in some model in which B does not occur, then A is more likely than B.

In other words, if the models in which A occurs contain the models in which B occurs as a proper subset, then A is more likely than B. This principle is valid provided that reasoners take into account all the premise models with nonzero probabilities. It is also a special case of the principle of proportionality, because if the models of A include the models of B as a proper subset, then A occurs in a greater proportion of the models than B does.

So, do people use inclusion or proportionality? The way to address this question depends on an unexpected prediction. For certain premises, mental models yield grossly erroneous conclusions. The model theory accordingly predicts that naive reasoners will infer from some premises that A is more probable than B when, in fact, A is impossible, but B is not. It predicts in other cases that naive reasoners will infer that A is equally probable with B when, in fact, A is certain, but B is not. These premises should give rise to *illusory* inferences; that is, most people should draw the same conclusion, which should seem obvious and yet which is wrong. Such illusory inferences, if they occur, may also provide decisive support for the model theory because they are contrary to current rule theories. These theories use only valid rules of inference (see, e.g., Braine & O'Brien, 1991; Rips, 1994), and so they do not readily account for a phenomenon in which most people draw the same invalid conclusion.

As an example of a potential illusion, consider the following problem:

> Suppose that *only one* of the following assertions is true about a specific hand of cards:
>
> There is a king in the hand or there is an ace in the hand, or both.
>
> There is a queen in the hand or there is an ace in the hand, or both.
>
> Which is more likely to be in the hand: the king or the ace?

Readers may care to answer the question and to make a note of their answer for future reference. The mental models of the first premise are

king
ace
king ace

and the mental models of the second premise are

queen
ace
queen ace

The rubric that only one of the two premises is true means that either one premise is true or the other premise is true, but not both of them. That is, the rubric calls for an exclusive disjunction of the premises, and the mental models for an exclusive disjunction of the form *A or else B* are (see Table 2)

A
B

and so the disjunction calls merely for all the models of the two alternatives. Hence, the problem as a whole calls for the following models:

king
ace
king ace
queen
ace
queen ace

Thus, a consequence of the principle of truth is that individuals think about what it means for one premise to be true and what it means for the other premise to be true, but they fail to take into account that when one premise is true the other premise is false.

They treat the other premise as though it has ceased to exist rather than that it is false. If reasoners infer relative probabilities using proportionality, then they should conclude that the ace is more probable than the king, because the ace occurs in a greater proportion of models than the king. However, if they infer relative probabilities using inclusion, they will respond that the problem is indeterminate, because the models containing kings neither include nor are included by the models containing aces. Of course, reasoners may use proportionality but reject equiprobability, and so in this case they would also respond that the problem is indeterminate, that, for example, the probability of the king alone could be greater than the probabilities of all the other models summed together.

In fact, it is an egregious error to respond that the ace is more probable than the king, or to respond that the problem is indeterminate. When the first disjunction is true, the second disjunction is false; that is, there is not a queen and not an ace. Likewise, when the second disjunction is true, the first disjunction is false, and there is not a king and not an ace. The fully explicit models of the premises are accordingly

king	¬queen	¬ace	(first disjunction true, second disjunction false)
¬king	queen	¬ace	(first disjunction false, second disjunction true).

The ace cannot occur in the hand, but the king can occur in the hand, and so the king is more probable than the ace! This conclusion follows logically from the premises, but the model theory predicts that reasoners will fail to draw it because they are unable to cope with falsity.

Here is another example of a potential illusion:

> Suppose that *only one* of the following assertions is true about a specific hand of cards:
>
> If there is a jack in the hand, then there is a queen in the hand.
>
> If there is a ten in the hand, then there is a queen in the hand.
>
> Which is more likely to be in the hand: the queen or the jack?

The mental models of the disjunction of the two conditionals are

jack	queen
ten	queen
. . .	

and so reasoners should tend to respond that the queen is more probable than the jack, on the basis of either inclusion or proportionality or both. Once again, however, this conclusion is wrong. According to the rubric that only one premise is true, if the first conditional is true, the second conditional is false, and so there is a ten and not a queen. Likewise, if the second conditional is true, the first conditional is false, and so there is a jack and not a queen. (When people are asked to falsify conditionals, they make these responses; see, e.g., Oaksford & Stenning, 1992.) The fully explicit models of the premises are accordingly

¬jack	ten	¬queen	(first conditional true, second conditional false)
jack	¬ten	¬queen	(first conditional false, second conditional true).

Hence, the queen cannot occur in the hand, whereas the jack can, and so the jack is more probable than the queen! An inclusive interpretation of the disjunction or a biconditional interpretation of the conditionals or both yield a tautology, and so nothing follows about the relative probabilities of the two cards. As with the first problem, the correct answer depends on taking into account the false cases.

These cognitive illusions do occur, as Johnson-Laird and Savary (1996) have shown experimentally. Overall, 21 out of the 24 participants in one experiment chose as more probable a card that could not occur in the hand for one or both of the illusory problems. In contrast, the participants performed competently with control problems in which mental models supported the same conclusion as fully explicit models. The results also implied that individuals use proportionality rather than inclusion to assess relative probabilities; they infer that one event is more probable than another if it occurs in a greater proportion of models than does the other event. A second experiment corroborated a variety of other illusions, including those that depend on biconditionals as the major connective. The procedure was also more sensitive because the participants made independent estimates of the probabilities of each pair of cards, responding by clicking a mouse to mark their estimates on separate scales presented on a computer screen. The experiment showed that illusions could be created in a minimal way by assertions containing only two sentential connectives, as in the following scenario:

> If one of the following assertions is true about a specific hand of cards, then so is the other assertion:
>
> There is a jack in the hand if, and only if, there is a queen in the hand.
>
> There is a jack in the hand.

Most of the participants erroneously inferred that the two cards had the same probability of being in the hand, though in fact the queen is certain to be in the hand, but the jack is not.

According to the model theory, the illusions arise because reasoners cope with truth but not with falsity. For the control problems, this failure does not matter because reasoners will still reach the right conclusion even if they do not consider falsity. For the experimental problems, however, the failure leads to systematic errors. The experiment also ruled out some alternative explanations, including the hypothesis that the illusions are based on the frequency of mention of cards in the premises or that they arise from misinterpretations of the sentential connectives. These alternatives can account for a few illusions, whereas the model theory accounts for all of them. Of course, rule theorists could invoke a different invalid formal rule to deal with each of the different illusions, but such an account would be post hoc. And, if reasoners were guided by such rules, their capacities for rational thinking would be inexplicable. In contrast, the model theory assumes that reasoners are rational in principle but that they err in practice because their working memories are limited and they can usually cope only with truth.

Models and Conditional Probabilities

Basic extensional reasoning concerns inferences about the absolute or relative probabilities of events. These inferences are

within the competence of naive reasoners, but many other inferences are beyond them. For example, they do not answer the following question correctly:

> An unbiased coin is tossed 20,000 times, and a cumulative record is kept of the outcomes. At any point in the sequence, there will be more heads than tails, more tails than heads, or an equal number of heads and tails. How many times is the lead likely to change from heads to tails, or vice versa?

The answer is that, no matter how long the series, the most probable number of changes of lead is zero (see Feller, 1957, p. 68). Such technical aspects of probability are bound to transcend common knowledge. What lies on the border of naive ability is reasoning about conditional probabilities. The reason is twofold. First, reasoners need to understand that a problem calls for an inference about a conditional probability. Many problems, as Nickerson (1996) has emphasized, are ambiguous, conceptually obscure, and rest on unstated assumptions. Second, reasoners need to establish the appropriate relation in models, which often calls for fleshing them out into a fully explicit form. We examine these two factors in turn.

The first source of difficulty is understanding that a conditional probability is needed in order to solve a problem. Consider, for example, the following puzzle (Bar-Hillel & Falk, 1982):

> The Smiths have two children. One of them is a girl. What's the probability that the other is a girl?

Naive reasoners—and some not so naive reasoners—respond that the probability is about 1/2. They assume the problem calls for models of the absolute probability of a girl (i.e., the other child is either a girl or a boy). They may qualify their answers with "approximately" if they know that the frequencies of the two sexes are not quite identical. The puzzle appears to call for the probability that a child is a girl, but it really calls for a conditional probability, p(other child is girl|one child is a girl). The partition of possibilities for two children is as follows:

Firstborn	Secondborn
girl	girl
girl	boy
boy	girl
boy	boy

Granted that either the firstborn or the secondborn is a girl, we can eliminate the last of these four possibilities. It follows that the probability that the other child is a girl is 1/3 (see Nickerson, 1996, for the subtleties in assumption that can influence this analysis).

The second source of difficulty is the need to represent the relation corresponding to a conditional probability. This problem is a special case of the general difficulty of constructing fully explicit models, which overtaxes the processing capacity of working memory. The same problem occurs in representing conditional assertions that make no reference to probabilities, such as

> If the DNA matches, then the suspect is guilty.

Such a conditional is represented by one explicit and one implicit model (see Table 2)

DNA matches	guilty
. . .	

and reasoners need to make a mental footnote that the antecedent (the DNA matches) is false in those cases represented by the implicit model. The converse conditional,

> If the suspect is guilty, then the DNA matches,

yields the models

guilty	DNA matches
. . .	

Because the two sets of models have the same explicit content, naive reasoners often take the two conditionals to be equivalent to one another. They also make the analogous inference that *all A are B* is equivalent to *all B are A* (see Evans, Newstead, & Byrne, 1993, for a review of illicit conversions). Of course, if a conditional is interpreted as a biconditional, then its conversion is valid. Otherwise, its conversion is invalid, and the key to blocking the inference is to flesh out the models in a fully explicit way, using negations that are true to represent the false cases. The conditional

> If the DNA matches, then the suspect is guilty

has the fully explicit models,

DNA matches	guilty
¬DNA matches	guilty
¬DNA matches	¬guilty

These models do not support the converse conditional,

> If the suspect is guilty, then the DNA matches,

because the second model is inconsistent with this assertion. Three fully explicit models, however, place a considerable load on working memory, and so individuals have difficulty in resisting the inference.

The same difficulty occurs with conditional probabilities, as exemplified by the following problem, which may have stumped the British Court of Appeal (see the epigraph to this article):

> The suspect's DNA matches the crime sample. If the suspect is not guilty, then the probability of such a DNA match is 1 in a million. Is the suspect likely to be guilty?

People tend to say "Yes." The numerical principle of the theory (see the earlier section on *Numerical Premises*) postulates that they can represent the conditional probability p(DNA matches|not guilty) = 1 in a million in the following way:

		Frequencies
¬guilty	DNA matches	1
. . .		999,999

where the given assumption is stated first in the explicit model. They will confuse this probability with its converse, p(not guilty|DNA matches) = 1 in a million. Hence, they will infer that the probability that the suspect is not guilty, given that the DNA matches, is very small. Once again, the true partition depends on fleshing out the models in a fully explicit way. The premise

p(DNA matches|not guilty) = 1 in a million establishes only the following frequencies:

		Frequencies
¬guilty	DNA matches	1
¬guilty	¬DNA matches	999,999

A conditional probability depends on a subset relation in models, and these models establish the subset within the "not guilty" models in which the DNA matches. This subset relation does not tell us anything about the probability that the suspect is guilty given a DNA match, p(guilty|DNA matches). This conditional probability depends on the numerical values of the following subset relation:

DNA matches guilty
DNA matches ¬guilty

Thus, suppose that the frequencies for all the true possibilities are as follows:

		Frequencies
¬guilty	DNA matches	1
¬guilty	¬DNA matches	999,999
guilty	DNA matches	9

where the frequency of cases of *guilty and ¬DNA matches* is zero. In this case, the probability that the suspect is *not* guilty given a DNA match is quite high, that is, 1/10. In other words, the data support the conclusion

> If the DNA doesn't match, then the suspect is almost certainly not guilty. (There are no counterexamples in the data.)

But, they do not support the conclusion

> If the DNA matches, then the suspect is almost certainly guilty. (There is 1 chance in 10 of innocence.)

The confusion between one conditional probability and another is thus inherited from the general problem of distinguishing one subset relation from another in models that normally make explicit only what is true.

The DNA problem is just one example of many that confuse naive reasoners (see, e.g., Eddy, 1982). Ironically, one case concerns statistical tests of significance. Suppose you carry out, say, a binomial test that yields the following conditional probability of the data given the "null" hypothesis of chance: p(data|null hypothesis) = .005. This result is deductively valid granted the assumptions of the test. However, one can make no valid inference from this result about the probability that the null hypothesis is false. In particular, it does not imply p(null hypothesis|data) = .005, contrary to the claims of some researchers (see Gigerenzer & Hoffrage, 1995, who point out the error).

Certain problems combine both the need to grasp that a problem calls for a conditional probability and the need to flesh out models explicitly. These problems typically concern Bayesian inference, and so we consider them in the next section.

Models and Bayesian Inference

Naive individuals, as we have argued, have difficulty with conditional probabilities. Previous theories of extensional reasoning have explained why naive individuals go wrong in trying to solve difficult Bayesian problems. In particular, such reasoners do not appear to use Bayes's theorem, as expressed by Equation 1 or its cognates, but rather they rely on various beliefs (see the earlier section Intuitive Beliefs About Probabilities, which discusses Shimojo & Ichikawa, 1989, and Falk, 1992). These accounts are consistent with the model theory, which specifies the nature of the mental representations that reasoners are likely to rely on. Yet, naive reasoners are able to infer posterior probabilities for certain sorts of problems. The model theory transcends these earlier accounts because it offers an explanation of how, in these cases, naive individuals can infer posterior probabilities without relying on Bayes's theorem. We describe the key principle and then examine the implications for two contentious matters: the alleged neglect of base rates and the doctrine of frequentism that we outlined earlier. In the subsequent section, we consider the implications of the model theory for the pedagogy of Bayesian reasoning.

The logical skeleton of Bayesian reasoning is as follows:

> There is a set of alternatives that have different consequences. Given information about the occurrence of one of these consequences, reasoners draw a conclusion about one of the alternatives.

The simplest version of this skeleton occurs when the given information entails a categorical conclusion:

> Pat has either the disease or a benign condition. If she has the disease, then she is likely to have a certain symptom. If she has a benign condition, then she will definitely not have the symptom. In fact, she has the symptom. So, does she have the disease?

The initial premises yield the following models of the possibilities:

disease symptom
disease ¬symptom
benign ¬symptom

The categorical premise that Pat has the symptom eliminates the latter models to leave only a single model,

disease symptom

from which it follows that Pat has the disease.

Probabilities often enter into such inferences where the premises concern frequencies. For example,

> According to a population screening, 4 out of 10 people have the disease, 3 out of 4 people with the disease have the symptom, and 2 out of the 6 people without the disease have the symptom. A person selected at random has the symptom. What's the probability that this person has the disease?

Naive individuals can build equiprobable models:

disease symptom
disease symptom
disease symptom
disease ¬symptom
¬disease symptom
¬disease symptom

. . .

where the implicit model represents individuals who have neither the disease nor the symptom. Alternatively, reasoners can build numerical models:

		Frequencies
disease	symptom	3
disease	¬symptom	1
¬disease	symptom	2
. . .		4

Either set of models establishes that the probability that a person has the symptom is 5/10 and that the probability that a person has the disease given the presence of the symptom is 3/5. The posterior probability can be computed from either set of models without having to use Bayes's theorem. A simpler algorithm suffices, which we capture in the final principle of the model theory.

> 5. The *subset* principle: Granted equiprobability, a conditional probability, p(*A*|*B*), depends on the subset of *B* that is *A*, and the proportionality of *A* to *B* yields the numerical value. Otherwise, if the models are tagged with their absolute frequencies (or chances), then the conditional probability equals the frequency (chance) of the model of *A and B* divided by the sum of all the frequencies (chances) of models containing *B*. In computing the ratio of the subset relation, reasoners can err in assessing either the numerator or, more likely, the denominator.

Naive reasoners, and experts who treat probabilities as degrees of belief, are happy to assign probabilities to unique events (see, e.g., Kahneman & Tversky, 1996), even though such probabilities raise deep philosophical problems. For example, the truth conditions of the assertion

> The chances that Pat has the disease are 4/10

are mysterious because the assertion is consistent with either Pat having the disease or not. Psychologically speaking, however, the assertion is about the chances of various possible scenarios. The model theory allows that such probabilities can be represented in models, and that the assertion can be represented by either a model of equiprobable possibilities or a model with numerical tags on the possibilities.

Bayesian inference about either unique events or repeated events should be easier when models directly represent probabilities according to the three principles of basic extensional reasoning (truth, equiprobability, and proportionality). For example,

> The chances that Pat has the disease are 4/10. If she has the disease, then the chances are 3/4 that she has the symptom. If she does not have the disease, then the chances are 2/6 that she has the symptom.

Reasoners can build a simple set of models of the possibilities as above. If they are asked what is the probability that Pat has the symptom, they should be able to respond, "5/10." However, a problem may arise if they are asked for the posterior conditional probability:

> Pat has the symptom. So, what are the chances that she has the disease?

Some reasoners may be able to grasp the appropriate subset relation and compute that the chances are 3/5, but others may focus on the model

disease symptom 3

and then compute the probability that Pat has a symptom *and* the disease, that is, 3/10. Girotto and Gonzalez (in press) observed exactly this sort of error. Its antidote is to force reasoners to consider separately the denominator and the numerator of the posterior probability. The researchers asked the participants to complete the following sort of sentence by adding the two missing numbers:

> Pat is tested now. Out of the entire 10 chances, Pat has __ chances of having the symptom; among these chances, __ chances will be associated with the disease.

This instruction elicited the correct answer (among the five chances of having the symptom, three chances are associated with the disease) from 53% of the participants in comparison with only 8% correct in a control problem.

The problem about Pat calls for the conditional probability, p(diseaselsymptom), and the proportionality of one set (disease) to the other (symptom) in the models themselves is 3/5. The set of models is easy to build because it concerns chances expressed in simple integer ratios, such as "4 out of 10," the numerator of this base rate equals the denominator of the given conditional probability, and the integral values yield trivial numerical inferences. In contrast, the following problem violates all of these principles:

> According to a population screening, a person has a 40% probability of having the disease, a 75% probability of having the symptom if she has the disease, and a 33% probability of having the symptom if she does not have the disease. Pat is tested now. She has the symptom. What is the probability that she has the disease?

A diagram of the problem shows the difficulty of envisaging the correct mental models:

	Probability		Conditional probability
disease	40%	symptom	75%
		¬symptom	25%
¬disease	60%	symptom	33%
		¬symptom	67%

Even if individuals envisage such models, it is not obvious how to calculate the posterior probability. The numerical version of the subset principle applies only to absolute frequencies or chances, not to conditional probabilities, and so the only route to the answer appears to be to use Bayes's theorem (see equation 3). Naive reasoners are neither familiar with the theorem nor easily able to carry out its computations.

The Neglect of Base Rates

What light does the model theory throw on the alleged neglect of base rates, which we discussed earlier in the paper? When a problem concerns two binary variables, *A* and *B*, the partition calls for the construction of four models:

A	*B*
A	¬*B*
¬*A*	*B*
¬*A*	¬*B*

Once reasoners know the probabilities for each possibility in a partition, then they know everything that is to be known from a

probabilistic standpoint; that is, they can work out any conditional probability and the probability of any assertion about the domain based on truth-functional connectives, for example, the probability of *A or B or both.* Unfortunately, naive reasoners do not appreciate the need to fix such probabilities, and, as we have seen, four models is at the limit of their abilities. Consider, for example, the following problem (translated from the original French):

> A test has been discovered. All people affected by a given disease have a positive reaction to this test. However, 10% of people who are not affected by this disease have a positive reaction too. Marie has tested positive. What is the probability that she is actually affected by the disease?

In a pilot study (conducted in collaboration with Michel Gonzalez), we observed that naive participants tend to infer a probability ranging from 10% to 90%. They appeared to construct the following models:

		Percentage conditional probability
disease	test positive	100%
¬disease	test positive	10%

Some individuals merely assume that the two models are equiprobable (and infer an answer of 50%), others take the mean of the two conditional probabilities (and infer an answer of 55%), and still others subtract one conditional probability from the other (and infer an answer of 90%). Few individuals, however, realize that the problem is ill-posed. We may have been oversold on the base rate fallacy, but naive reasoners can fail to notice that a problem does not even state a base rate (see also Hammerton, 1997).

The "Frequentist" Hypothesis

The frequentist hypothesis postulates that human beings possess an inferential module embodying the probability calculus and operating on frequency data (Cosmides & Tooby, 1996; cf. Gigerenzer & Hoffrage, 1995). Cosmides and Tooby carried out a series of experiments in which they compared various statements of a Bayesian problem. Some versions stated the problem in terms of absolute frequencies, and other versions stated it in terms of percentage probabilities (cf. Casscells, Schoenberger, & Graboys, 1978). There was a much higher success rate for problems stated with frequencies than for problems stated with percentage probabilities, and Cosmides and Tooby concluded that base-rate neglect disappears and good Bayesian reasoning emerges when problems are posed in frequentist terms. Gigerenzer and Hoffrage (1995) drew the same conclusion from their experiments.

Cosmides and Tooby (1996), wrote

> Frequentist mechanisms could not elicit Bayesian reasoning unless our minds contained mechanisms that embody at least some aspects of a calculus of probability. This means that the more general conclusion of the literature on judgment under uncertainty—that the human mind does not embody a calculus of probability, but has instead only crude rules-of-thumb—must also be re-examined. This conclusion was based largely on subjects' responses to single-event probability problems. But if those inductive reasoning procedures that do embody a calculus of probability take frequency representations as input and produce frequency representations as output, then single-event probability problems cannot, in principle, reveal the nature of these mechanisms. (p. 62)

In response, Howson and Urbach (1993) commented,

> A more accurate conclusion is that their respondents are competent at whole number arithmetic, which is anyway hardly surprising in view of the fact that they are often university students. But with probabilistic reasoning, and especially with reasoning about frequency probabilities, Cosmides and Tooby's results have very little to do at all, despite their dramatic claims. (p. 422)

It is indeed crucial to show that the difference between frequencies and probabilities transcends mere difficulties in calculation and that difficulties with problems about unique events are not attributable merely to difficulties in numerical calculation. In fact, data in the form of frequencies by no means guarantee good Bayesian reasoning. Girotto and Gonzalez (in press) reported that not a single participant inferred the right response to the following Bayesian problem (translated from the original French):

> According to a recent epidemiological survey:
>
> Out of 100 tested people, there are 10 infected people.
>
> Out of 100 infected people, 90 people have a positive reaction to the test.
>
> Out of 100 non-infected people, 30 have a positive reaction to the test.
>
> Imagine that the test is given to a new group of people. Among those who have a positive reaction, how many will actually have the disease? __ out of __ .

Intuitions about evolution are an interesting heuristic for generating hypotheses about how the mind solves ill-posed problems, that is, problems that would be insoluble without innate constraints on the process, such as the stereoptic recovery of depth information from disparate visual images (Marr, 1982). However, it is hard, if not impossible, to test intuitions about the mental processes of our evolutionary ancestors (Lewontin, 1990). Hence, the claim that frequencies trigger an inductive module needs to be examined on its own merits. The real burden of the findings of Gigerenzer and Hoffrage (1995) is that the mere use of frequencies does not constitute what they call a "natural sample." Whatever its provenance, as they hint, a natural sample is one in which the subset relation can be used to infer the posterior probability, and so reasoners do not have to use Bayes's theorem. A separate question is whether problems that concern unique events guarantee bad Bayesian reasoning. In fact, as the problem about Pat showed, reasoners *can* infer posterior probabilities, for unique events provided that the probabilities are stated in simple numerical terms, such as "3 chances out of 5," and that they allow easy numerical calculations in the use of the subset principle.

Mental Models and the Pedagogy of Bayesian Reasoning

Shimojo and Ichikawa (1989) report that even when people are taught Bayesian principles they still find them counterintuitive. Their main pedagogical recommendation is to try to integrate Bayes's theorem into the intuitive system by way of Equation 2 (see the section *Bayes's Theorem and Studies of Bayesian Reasoning*). Likewise, Falk (1992) argues that the expedient way to grapple with Bayesian puzzles is to carry out these steps:

1. Uncover the hidden assumptions and check whether they are warranted.
2. Explicate the random process that generated the data.
3. Apply Bayes's theorem.

Indeed, it is remarkable that many pedagogical discussions of Bayesian problems, such as the three prisoners puzzle, treat them as though the only road to success is by way of Bayes's theorem (see Shimojo & Ichikawa, 1989, and Falk, 1992, for relevant references). As the model theory shows, however, naive reasoners can infer posterior probabilities without having to carry out the computations required by Bayes's theorem. It follows that they should be able to solve puzzles, such as the three prisoners problem, provided that they are framed appropriately. It should also be possible to develop a pedagogical method to make the correct posterior probabilities seem intuitive. Our aim in this section is to explore these possibilities.

A famous example of a Bayesian brainteaser is the eponymous "Monty Hall" problem, which has the same structure as the three prisoners puzzle (see, e.g., Falk, 1992; Nickerson, 1996). A TV quiz show host (Monty Hall) asks a guest to choose one of three doors, only one of which hides a prize. The guest makes a choice of, say, the door on the left. Before the door is opened, the host, who knows the location of the prize, opens one of the two remaining doors, say, the middle door, to reveal, as always, that it does not hide the prize, and then offers the guest the chance of switching to the other unopened door. Most guests stick with their first choice. According to the model theory, they begin with the following models of the possibilities:

left door	middle door	right door
prize		
	prize	
		prize

They choose the left door. The host opens the middle door to reveal that the prize is not there, and so they eliminate the relevant model. They are left with the following models of possibilities:

chosen door	right door
prize	
	prize

The prize is either behind the door they have chosen or behind the unopened door, and so there appears to be no advantage in switching choices. This account accommodates the view that naive individuals believe that when one alternative is eliminated, the remaining possibilities are equiprobable or have the same ratio as the ratio of their prior probabilities (Falk, 1992; Shimojo & Ichikawa, 1989). Insofar as a naive individual holds such beliefs, they could be emergent properties of this manipulation of models.

The model theory suggests a way to reach the right answer without using Bayes's theorem. This method is as follows:

1. Represent the initial situation.
2. Make explicit the relevant conditional relations on the basis of the assumptions and processes that generated the data.
3. Construct a diagram of the *equiprobable* possibilities (a partition akin to a set of mental models).
4. Use the subset principle.

Let us apply this method to the Monty Hall problem. The first step is to represent the initial possible locations of the prize. They are shown in the left door—middle door—right door diagram earlier in this section. Hence, the guest has a probability of 1/3 of choosing the door with the prize, assuming that the location of the prize is chosen at random. Suppose that the guest chooses the left-hand door. The second step is to represent the conditional relations:

> If the prize is behind the chosen left-hand door, then the host opens either the middle door or the right-hand door, presumably at random with equal probabilities.
>
> If the prize is behind the middle door, then the host must open the right-hand door.
>
> If the prize is behind the right-hand door, then the host must open the middle door.

The third step is to construct a diagram of equiprobable possibilities. What complicates matters is the first of these conditionals. It can be represented as follows:

chosen door	middle door	right door
prize	open	
prize		open

In order to ensure equiprobability, each of the other two conditionals must be represented by two models to match the two models for the first conditional. The second conditional calls for these two identical models in the diagram:

chosen door	middle door	right door
	prize	open
	prize	open

The third conditional calls for these two identical models in the diagram:

chosen door	middle door	right door
	open	prize
	open	prize

The set of equiprobable models as a whole is therefore

chosen door	middle door	right door
prize	open	
prize		open
	prize	open
	prize	open
	open	prize
	open	prize

The fourth step is to apply the subset principle to the models in the diagram. There are four cases out of the six models in which switching from the chosen door to the closed door yields the prize, and there are two cases out of the six in which switching to the closed door loses the prize. In other words, switching choices yields the prize with a probability of 2/3, whereas sticking with the original choice yields the prize with a probability of 1/3. Hence, it pays to switch. The same reasoning applies mutatis mutandis for whichever door hides the prize. The problem, however, is that naive reasoners are unlikely to realize that in order to ensure equiprobability they need to construct two identical models for each of the cases where the prize is not behind the chosen door. The cause of the difficulty in our view is the load on working

memory, which leads naive reasoners not to construct fully explicit models. They think, "If the prize is behind the door I have chosen, then the host chooses another door." They need to think, "If the prize is behind the door I have chosen, say, the left-hand door, then the host chooses either the right-hand door or the middle door."

A study by Macchi and Girotto (1994) showed that merely spelling out the conditional relations does not help naive individuals to infer the correct probabilities. Naive individuals are still inclined to start with the models of the three possibilities, then to eliminate the case where the opened door hid the prize, and finally to treat the two remaining cases as equiprobable. However, Macchi and Girotto found that performance was reliably better with the following variant of the puzzle (translated from the original Italian):

> There are three boxes, A, B, and C. Only one of them contains a prize.
>
> Depending on its location, either a red or green light comes on.
>
> If the prize is in Box A, either the red or the green light comes on.
>
> If the prize is in Box B, the red light never comes on.
>
> If the prize is in Box C, the red light always comes on.
>
> The red light comes on. Which is more likely—that the prize is in Box A or in Box C?

In this case, naive reasoners more readily grasp the need to construct two models for the case in which the prize is in Box A and thus realize that the probability of the red light for Box A is half the probability of the red light for Box C. This improvement in performance is difficult to explain if individuals are bound to reason using a module that is based on faulty intuitions about probabilities.

We return finally to the problem of the three prisoners, which we discussed earlier. We analyze a very difficult variant introduced by Shimojo and Ichikawa (1989) so that we can demonstrate the power of diagrams that represent equiprobable possibilities:

> Three men, A, B, and C, were in jail. One of them is to be set free, and the other two to be executed. A had reason to believe that their chances of being freed were, A = ¼, B = ¼, and C = ½. After their fates had been decided, A, who didn't know the outcome of the decision, asked the jailer, who did, "Since two of the three will be executed, it is certain that either B or C will be, at least. You will give me no information about my own chances if you give me the name of one man, B or C, who is going to be executed." Accepting this argument, the jailer said, "B will be executed." Thereupon A felt happier because now either he or C would go free, so his chance had increased from ¼ to ½. The prisoner's happiness may or may not be reasonable. What do you think?

The participants in the experiment generated a variety of estimates of the probability that A will go free given the jailer's statement that B will be executed, and few of their estimates were correct. Contrary to their intuitions, the jailer's news is bad for A, because the probability that he will go free reduces to ⅓.

To show how the pedagogical method based on the model theory works, we use it to solve the puzzle. The first step is to represent the initial situation of the probabilities of each prisoner going free (A = ¼, B = ¼, and C = ½) in equiprobable models:

A	B	C
free		
free		
	free	
	free	
		free
		free
		free
		free

The second step is to formulate the conditional relations:

> If A is to be freed, then, assuming a random choice, the jailer will choose at random between B or C as the victim to name.
>
> If B is to be freed, then the jailer must name C as the victim.
>
> If C is to be freed, then the jailer must name B as the victim.

The third step is to use this information to complete the construction of the equiprobable models

A	B	C	jailer names the person to be executed as
free			B
free			C
	free		C
	free		C
		free	B
		free	B
		free	B
		free	B

The fourth step is to use the subset principle. The probability that A goes free is the one chance out of the five possibilities in the diagram in which the jailer named B as the person to be executed.

The use of equiprobable models and the subset principle is simpler than the tree diagrams that are so common in textbooks because the models contain no explicit numerical information and the subset principle can be applied directly to them. In theory, the method can be used for inferring posterior probabilities for many problems. In practice, such diagrams can become too large to be tractable. However, once naive individuals have grasped the underlying principles, they can be taught to use numerically tagged diagrams. The problem just described, for example, can be represented in the following way, using integers as frequencies that sum to the common denominator (in this case, 8):

A	B	C	jailer names the person to be executed as	frequency
free			B	1
free			C	1
	free		C	2
		free	B	4

Once again, the final calculation is a straightforward application of the subset principle. The probability that A goes free, given that the jailer named B to be executed, is 1/(1 + 4). The numerical method is similar to the more standard treatments in texts.

General Discussion

Our major goal has been to advance a new theory of naive extensional reasoning about probability. This theory depends on five principles:

1. Truth: People represent situations by constructing a set of mental models in which each model represents what is true in a true possibility.

2. Equiprobability: Each model represents an equiprobable alternative unless individuals have knowledge or beliefs to the contrary, in which case they will assign different probabilities to different models.

3. Proportionality: Granted equiprobability, the probability of an event depends on the proportion of the models in which it occurs.

4. Numerical representation: If a premise refers to a numerical probability, models can be tagged with the appropriate numerical values, and an unknown probability can be calculated by subtracting the sum of the $(n - 1)$ known probabilities from the overall probability of the n possibilities in the partition.

5. Subsets: Granted equiprobability, a conditional probability, $p(A|B)$, depends on the subset of B that is A, and the proportionality of A to B yields the numerical value. Otherwise, if the models are tagged with their frequencies (or chances), then the conditional probability equals the frequency (or chance) of the model of *A and B* divided by the sum of all the frequencies (or chances) of models containing B.

In short, as our experiments confirmed, individuals construct models of true possibilities, assume equiprobability by default, and infer probabilities from proportionality. Laplace's (1820/1951) principle of indifference, as we saw, ran into insuperable difficulties because when one is totally ignorant there is more than one partition of events. In our experiments, however, the partitions were fixed by the premises and the mental models that individuals constructed from them. When naive reasoners are told, for example,

> If there is a red marble in the box, then there is either a green marble or else a blue marble in the box, but not all three marbles,

they are likely to envisage the following partition (see Problem 3 in Appendix A):

red	green	
red		blue

It is possible that reasoners represent the partition using Euler circles or Venn diagrams, if they have been taught them, but they rapidly forget the implicit model representing the possibilities in which the antecedent of the conditional is false. As our results show, reasoners assume that the two models above represent equiprobable alternatives, and so they tend to infer that the probability of at least a red marble in the box is 100%. Likewise, reasoners infer that it is impossible for there to be a green and a blue marble in the box. Thus, the results bear out the principle of truth: Mental models represent only what is true. In contrast, the actual alternatives (see the fully explicit models for Problem 3 in Appendix A) show that the probability of a red marble is 33% (assuming equiprobability over the actual alternatives), and they also show that it is possible for there to be both a green and a blue marble. Hence, where mental models and the real alternatives diverge, then naive reasoners tend to follow their mental models. In this sense, the present theory accommodates the earlier accounts of Shimojo and Ichikawa (1989) and Falk (1992) but shows that these theories should be based on mental models rather than actual alternatives.

In our experiments, we used problems about marbles in boxes. Such materials are often used in studies of probabilistic reasoning to insulate problems from the participants' general knowledge or beliefs (see von Winterfeldt & Edwards, 1986). The trouble is that general knowledge is readily triggered by any materials to which it seems relevant. Consider, for example, the following conditional (from an anonymous reviewer):

> If dirixil acid is present on Europa, then there is life on Europa.

If people have no knowledge of these matters, then, according to the model theory, they should infer a probability of 50% for dirixil acid and life on Europa (see the mental models of the conditional in Table 2). However, people who know that Europa is one of Jupiter's moons probably have some beliefs about the likelihood of life there. We, for instance, would assign a low probability to life on Europa. The following conditional (from the same reviewer) concerns two sorts of dirixil acid:

> If either dirixil-1 or dirixil-2 acid is present on Europa, then there is life on Europa.

What the model theory predicts in this case is that naive reasoners should infer that the following two assertions have the same probability:

> There is dirixil-1 acid on Europa and there is life on Europa.

> There is dirixil-2 acid on Europa and there is life on Europa.

Nonextensional reasoning from beliefs—perhaps about the lack of life on the Earth's moon—is likely to yield a low probability for each of these two propositions; however, extensional reasoning from the models of the conditional are likely to yield equal probabilities for both propositions. This mixture of extensional and nonextensional reasoning is typical in daily life.

What are probabilities? Are they degrees of belief, partial entailments, limits on relative frequencies, or some other entity? As we pointed out at the beginning of this article, there is no consensus about such matters among probabilists. We deliberately avoided the debate about the proper interpretation of probabilities. Do our participants have an internal conviction that, say, the probability of a blue marble in the box is 50%? We are confident that they are not just saying so to oblige the experimenter, or that they made responses in which they did not believe because they thought the experimenter expected them. The participants in Experiments 1 and 2 were clearly thinking hard, and their inferences were often not the precise round numbers predicted by the model theory. The problems in Experiment 3 were more complex, and it would be odd for individuals to have two sets of intellectual machinery for probabilities: one for generating the inferences expected by the experimenter and the other for generating inferences in which they believe. Moreover, when naive reasoners encounter such tasks in real life, such as in the Monty Hall puzzle (see the previous section), they do indeed assign equal probabilities to their mental models of the partition. Marilyn vos Savant, the author of the "Ask Marilyn" column in *Parade* magazine, published a correct analysis of the problem. Such is the power of equiprobability, as Falk (1992, p. 203) noted, that vos Savant received thousands of letters from readers, many of them from universities and research institutes, and about 90% of them insist-

ing that she was wrong. Bar-Hillel and Falk (1982) corroborated these readers' judgments experimentally. Yet, when naive reasoners have knowledge to the contrary, they do abandon equiprobability. They do not believe that the presence of life on Europa is as equally probable as its absence. Obviously, they abandon equiprobability when the premises stipulate different explicit probabilities. In such cases, proportionality is also suspended, and mental models can be tagged with numerical values. Naive reasoners will still be able to infer probabilities provided that the arithmetical calculations are not too taxing.

The model theory dispels five misconceptions about probabilistic reasoning. The first misconception is that it is necessarily an inductive process. As we have shown, many inferences about probabilities are deductively valid, particularly if background knowledge is taken into account as an additional set of premises. The second misconception is that naive reasoning is seldom, if ever, extensional. However, following D. Kahneman (personal communication, January 26, 1994), we have argued that naive reasoners make both extensional and nonextensional inferences about probabilities. The third misconception is that extensional reasoning depends on a tacit knowledge of the probability calculus, perhaps embodied in an innate inferential module. The results of the present experiments provide strong support for our alternative theory of naive probability. The fourth misconception is that extensional reasoning occurs only when premises concern the natural frequencies of events. Our studies show that individuals can reason extensionally when premises concern neither frequencies nor probabilities. Reasoning about conditional probabilities is a borderline ability for naive individuals, because it calls for models to be fleshed out in a fully explicit way. One way to overcome this problem, especially in Bayesian reasoning, is to allow reasoners to make extensional inferences from simple sets of models in which the numerical results emerge from the subset algorithm. The fifth misconception is that cognitive illusions occur only in nonextensional reasoning and disappear in extensional reasoning. "Subadditivity" is a well-known phenomenon of nonextensional reasoning in which estimates of the probability of an implicit disjunctive category, such as "accidents," is less than the sum of its explicit disjuncts, such as "accidents in the home" and "accidents outside the home." According to Tversky and Koehler's (1994) support theory, the description of an event in greater detail recruits more evidence in favor of it and thus leads to a higher judged probability (see also Miyamoto, Gonzalez, & Tu, 1995). Extensional reasoning, too, can yield subadditivity. Given an inclusive disjunction of the form *A or B, or both,* for instance, participants in Experiment 2 inferred that the probability of *at least A* is 60% (see Table 5), which is much less than the sum of their inferences for its two components, *A and B* (45%), and *A and not B* (44%). The model theory predicted subadditivity on the grounds that reasoners have difficulty in calling to mind all the possible models of premises.

The most striking cognitive illusions in extensional reasoning arise from the failure of reasoners to cope with falsity. As the model theory predicts, they succumb to gross illusions about relative probabilities. From certain premises, they infer as the more probable of two events one that is, in fact, impossible; from other premises, they infer as the less probable of two events one that is, in fact, certain (Johnson-Laird & Savary, 1996). These erroneous conclusions corroborate the model theory but count against any account that embodies the probability calculus in any current psychological theory that is based on formal rules, because these theories rely only on valid rules (see, e.g., Braine & O'Brien, 1991; Rips, 1994).

A naive grasp of probability provides the mental foundations for expertise on the topic. Such expertise depends on knowledge of a variety of matters. One important component is a knowledge of combinations and permutations. For example, if a coin is tossed twice, which is more likely to occur: Two heads or one head and one tail? Naive individuals reason that there are three possibilities: two heads, two tails, or one head and one tail, and so conclude that the two outcomes are equally likely. In contrast, experts know that it is necessary to take permutations into account, and that there are accordingly four possible outcomes—two heads, two tails, a head followed by a tail, and a tail followed by a head—and so the correct answer is that one head and one tail is more likely, because it can be obtained in two distinct ways, whereas two heads can be obtained in only one way. Whether expertise of this sort is conceptual or a result of observation is a moot point (cf. Hacking, 1975). Another component of expertise is a knowledge of numbers and arithmetic, a factor that is easily overlooked, as we have seen, in studies of probabilistic reasoning. Still another component is the explicit acquisition of the laws of probability, such as Bayes's theorem. Underlying the ability to acquire these technical matters, in our view, are the simple principles of extensional reasoning based on mental models.

The import of our results is clear: They substantiate the model theory of naive probability. This theory is based on a small number of simple, but powerful, principles. Reasoners make extensional inferences about probabilities from mental models representing what is true. They assume by default that each model represents an equiprobable alternative. They also infer the probability of an event from the proportion of models in which it occurs. In cases where the premises include numerical statements of probability, reasoners build the same sorts of models, tag them with numerical probabilities, and, if possible, use simple arithmetic to calculate probabilities. Extensional problems that cannot be solved in these ways are likely to be beyond the ability of naive reasoners.

References

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

Bell, V. A., & Johnson-Laird, P. N. (1998). A model theory of modal reasoning. *Cognitive Science, 22,* 25–51.

Braine, M. D. S., & O'Brien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98,* 182–203.

Braine, M. D. S., Reiser, B. J., & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. *The psychology of learning and motivation (Vol. 18, pp. 313–371).* New York: Academic Press.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299,* 999–1001.

Collins, A., & Michalski, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science, 13,* 1–49.

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

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems

and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases (pp. 249–267).* Cambridge, England: Cambridge University Press.

Edwards, W., & Tversky, A. (Eds.). (1967). *Decision making.* Harmondsworth, England: Penguin.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Hillsdale, NJ: Erlbaum.

Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition, 43,* 197–223.

Feller, W. (1957). *An introduction to probability theory and its applications* (2nd ed.). New York: Wiley.

Fisher, R. A. (1932). *The design of experiments* (3rd ed.). Edinburgh, Scotland: Oliver and Boyd.

Garnham, A. (1987). *Mental models as representations of discourse and text.* Chichester, England: Ellis Harwood.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103,* 592–596.

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

Girotto, V., Evans, J. St. B. T., & Legrenzi, P. (1996, October). Relevance of information and consideration of alternatives: Pseudodiagnosticity as a focusing phenomenon. Paper presented at the Third International Conference on Thinking, University College, London.

Girotto, V., & Gonzalez, M. (in press). Mental models and statistical reasoning. In W. Schaeken (Ed.), *Strategies in deductive reasoning.* Mahwah, NJ: Erlbaum.

Hacking, I. (1975). *The emergence of probability.* Cambridge, England: Cambridge University Press.

Hammerton, M. (1997, March). Ignoring the base-rate—a robust phenomenon. Paper presented at the March meeting of the Experimental Psychology Society, University of Oxford, England.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago: La Salle.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, England: Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction.* Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N., & Savary, F. (1996). Illusory inferences about probabilities. *Acta Psychologica, 93,* 69–90.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions: A reply to Gigerenzer's critique. *Psychological Review, 103,* 582–591.

Keynes, J. M. (1943). *A treatise on probability.* London: Macmillan.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19,* 1–53.

Laplace, P. S. de. (1951). *Philosophical Essay on Probabilities.* New York: Dover. (Original work published 1820).

Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (1993). Focussing in reasoning and decision making. *Cognition, 49,* 37–66.

Lewontin, R. C. (1990). The evolution of cognition. In D. N. Osherson, & E. E. Smith (Eds.), *An invitation to cognitive science: Vol. 3. Thinking (pp. 229–246).* Cambridge, MA: MIT Press.

Macchi, L., & Girotto, V. (1994). Probabilistic reasoning with conditional probabilities: The three boxes paradox. Paper presented at the Annual Meeting of the Society for Judgement and Decision Making, St. Louis, MO.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman.

Miyamoto, J., Gonzalez, R., & Tu, S. (1995). Compositional anomalies in the semantics of evidence. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Decision making from the perspective of cognitive psychology: Vol. 32. Psychology of learning and motivation (pp. 319–383).* New York: Academic Press.

Mynatt, C. R., Doherty, M. E., & Dragan, W. (1993). Information relevance, working memory, and the consideration of alternatives. *Quarterly Journal of Experimental Psychology, 46A,* 759–778.

Nickerson, R. S. (1996). Ambiguities and unstated assumptions in probabilistic reasoning. *Psychological Bulletin, 120,* 410–433.

Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 835–854.

Osherson, D. N. (1976). *Logical abilities in children: Vol. 4. Reasoning and concepts.* Hillsdale, NJ: Erlbaum.

Osherson, D. N. (1996). Probability judgment. In D. N. Osherson, & E. E. Smith (Eds.), *An invitation to cognitive science: Vol. 3. Thinking* (2nd ed., pp. 35–75). Cambridge, MA: MIT Press.

Phillips, L., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72,* 346–354.

Rips, L. J. (1994). *The psychology of proof.* Cambridge, MA: MIT Press.

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

Shimojo, S., & Ichikawa, S. (1989). Intuitive reasoning about probability: Theoretical and experimental analyses of the "problem of three prisoners." *Cognition, 32,* 1–24.

Stevenson, R. J. (1993). *Language, thought and representation.* New York: Wiley.

Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology, 48A,* 613–643.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232.

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

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547–567.

von Mises, R. (1957). *Probability, statistics and truth.* London: Allen & Unwin.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research.* Cambridge, England: Cambridge University Press.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology.* Harmondsworth, England: Penguin.

Appendix A

The 18 Problems of Experiment 3, Their Mental Models and Fully Explicit Models, Predicted Values, and Unbiased Values

Experimental problems

1. A or both B and C, not all three

Mental models	Fully explicit models
A	A ¬B ¬C
B C	A B ¬C
	A ¬B C
	¬A B C

Conclusions	Predicted value	Unbiased value
p(A and C)	0%	25%
p(B and C)	50%	25%
p(at least A)	50%	75%

2. A or else if B then C, not all three

Mental models	Fully explicit models
A	A B ¬C
B C	¬A B C
	¬A ¬B C
	¬A ¬B ¬C

Conclusions	Predicted value	Unbiased value
p(A and B)	0%	25%
p(empty box)	0%	25%
p(A alone)	50%	0%

3. If A then either B or C, not all three

Mental models	Fully explicit models
A B	A B ¬C
A C	A ¬B C
	¬A B ¬C
	¬A ¬B C
	¬A B C
	¬A ¬B ¬C

Conclusions	Predicted value	Unbiased value
p(B and C)	0%	17%
p(at least A)	100%	33%
p(at least B)	50%	50%

Control problems

1.′ A and either B or else C, not all three

Mental models	Fully explicit models
A B	A B ¬C
A C	A ¬B C

Conclusions	Predicted unbiased value
p(A and C)	50%
p(B and C)	0%
p(at least A)	100%

2.′ A or else B or else C, but no more than one

Mental models	Fully explicit models
A	A ¬B ¬C
B	¬A B ¬C
C	¬A ¬B C

Conclusions	Predicted unbiased value
p(A and B)	0%
p(empty box)	0%
p(A alone)	33%

3.′ A or else B, and C, not all three

Mental models	Fully explicit models
A C	A ¬B C
B C	¬A B C

Conclusions	Predicted unbiased value
p(A and B)	0%
p(at least A)	50%
p(at least B)	50%

Note. The mental models predict unbiased values for the control problems. The predicted value was also the unbiased one for the third experimental problem that was based on Premise 3.

Appendixes continue

Appendix B

Results of Problems in Experiment 3

Experimental problems			Control problems	
1. A or both B and C, not all three			1.′ A and either B or else C, not all three	
	No. predicted	No. unbiased		No. predicted and unbiased
A and C:	18	2	A and C:	14
B and C:	17	0	B and C:	11
at least A:	17	0	at least A:	6
2. A or else if B then C, not all three			2.′ A or else B or else C, but no more than one	
	No. predicted	No. unbiased		No. predicted and unbiased
A and B:	19	0	A and B:	24
empty box:	17	1	empty box:	20
A alone:	14	5	A alone:	11
3. If A then either B or C, not all three			3.′ A or else B, and C, not all three	
	No. predicted	No. unbiased		No. predicted and unbiased
B and C:	14	0	A and B:	20
at least A:	12	1	at least A:	17
at least B:	16	16[a]	at least B:	17
Overall means:	16.0	2.8		15.6

Note. This Appendix shows the numbers of participants who made inferences within ±5% of the predicted values, and within ±5% of the unbiased values for the experimental problems. $N = 25$.

[a] The unbiased inference was also the predicted one for this problem (see Appendix A).

Received February 10, 1997
Revision received October 15, 1997
Accepted April 2, 1998 ■