Cognition, 11 (1982) 123–141

On the study of statistical intuitions*

DANIEL KAHNEMAN
University of British Columbia

AMOS TVERSKY
Stanford University

Abstract

The study of intuitions and errors in judgment under uncertainty is complicated by several factors: discrepancies between acceptance and application of normative rules; effects of content on the application of rules; Socratic hints that create intuitions while testing them; demand characteristics of within-subject experiments; subjects' interpretations of experimental messages according to standard conversational rules. The positive analysis of a judgmental error in terms of heuristics may be supplemented by a negative analysis, which seeks to explain why the correct rule is not intuitively compelling. A negative analysis of non-regressive prediction is outlined.

Much of the recent literature on judgment and inductive reasoning has been concerned with errors, biases and fallacies in a variety of mental tasks (see, e.g., Einhorn and Hogarth, 1981; Hammond, McClelland and Mumpower, 1980; Kahneman, Slovic, and Tversky, in press; Nisbett and Ross, 1980; Shweder, 1980; Slovic, Fishhoff and Lichtenstein, 1977; Tversky and Kahneman, 1974). The emphasis on the study of errors is characteristic of research in human judgment, but is not unique to this domain: we use illusions to understand the principles of normal perception and we learn about memory by studying forgetting. Errors of reasoning, however, are unique among cognitive failures in two significant respects: they are somewhat embarassing and they appear avoidable. We are not troubled by our susceptibility to the vertical-horizontal illusion or by our inability to remember a list of more than eight digits. In contrast, errors of reasoning are often disconcerting—either because the solution that we failed to find appears quite obvious in retrospect; or because the error that we made remains attractive although we know it to be an error. Many current studies of judgment are concerned with problems that have one or the other of these characteristics.

*This work was supported by the Office of Naval Research under Contract N00014-79-C-0077 to Stanford University. Reprint requests should be sent to Daniel Kahneman, Department of Psychology, University of British Columbia, 2075 Wesbrook Mall, Vancouver B.C., Canada V6T 1W5.

The presence of an error of judgment is demonstrated by comparing people's responses either to an established fact (e.g., that the two lines are equal in length) or to an accepted rule of arithmetic, logic or statistics. However, not every response that appears to contradict an established fact or an accepted rule is a judgmental error. The contradiction could also arise from the subject's misunderstanding of the question, or from the investigator's misinterpretation of the answer. The description of a particular response as an error of judgment therefore involves assumptions about the communication between the experimenter and the subject. (We shall return to this issue later in the paper.) The student of judgment should avoid overly strict interpretations, which treat reasonable answers as errors, as well as overly charitable interpretations, which attempt to rationalize every response.

Although errors of judgment are but a method by which some cognitive processes are studied, the method has become a significant part of the message. The accumulation of demonstrations in which intelligent people violate elementary rules of logic or statistics has raised doubts about the descriptive adequacy of rational models of judgment and decision making. In the two decades following World War II, several descriptive treatments of actual behavior were based on normative models: subjective expected utility theory in analyses of risky choice, the Bayesian calculus in investigations of changes of belief, and signal-detection theory in studies of psychophysical tasks. The theoretical analyses of these situations, and to a much lesser degree the experimental results, suggested an image of people as efficient, nearly optimal decision-makers. On this background, observations of elementary violations of logical or statistical reasoning appeared surprising, and the reaction may have encouraged a view of the human intellect that some authors have criticized as unfairly negative (Cohen, 1979, 1981; Edwards, 1975; Einhorn and Hogarth, 1981).

There are three related reasons for the focus on systematic errors and inferential biases in the study of reasoning. First, they expose some of our intellectual limitations and suggest ways of improving the quality of our thinking. Second, errors and biases often reveal the psychological processes and the heuristic procedures that govern judgment and inference. Third, mistakes and fallacies help the mapping of human intuitions by indicating which principles of statistics or logic are non-intuitive or counter-intuitive.

The terms 'intuition' and 'intuitive' are used in three different senses. First, a judgment is called intuitive if it is reached by an informal and unstructured mode of reasoning, without the use of analytic methods or deliberate calculation. For example, most psychologists follow an intuitive procedure in deciding the size of their samples but adopt analytic procedures to test the statistical significance of their results. Second, a formal rule or a fact of nature

is called intuitive if it is compatible with our lay model of the world. Thus, it is intuitively obvious that the probability of winning a lottery prize decreases with the number of tickets. but it is counter-intuitive that there is a better than even chance that a group of 23 people will include a pair of individuals with the same birthday. Third, a rule or a procedure is said to be part of our repertoire of intuitions when we apply the rule or follow the procedure in our normal conduct. The rules of grammar, for example, are part of the intuitions of a native speaker, and some (though not all) of the rules of plane geometry are incorporated into our spatial reasoning.

The present paper addresses several methodological and conceptual problems that arise in attempts to map people's intuitions about chance and uncertainty. We begin by discussing different tests of statistical intuitions, we then turn to a critique of the question-answering paradigm in judgment research, and we conclude with a discussion of the non-intuitive character of some statistical laws.

Tests of statistical intuitions

Errors and biases in judgment under uncertainty are the major source of data for the mapping of the boundaries of people's statistical intuitions. In this context it is instructive to distinguish between errors of application and errors of comprehension. A failure in a particular problem is called an error of application if there is evidence that people know and accept a rule that they did not apply. A failure is called an error of comprehension if people do not recognize the validity of the rule that they violated.

An error of application is most convincingly demonstrated when a person, spontaneously or with minimal prompting, clutches his head and exclaims: 'How could I have missed that?' Although many readers will recognize this experience, such displays of emotion cannot be counted on, and other procedures must be developed to demonstrate that people understand a rule that they have violated.

The understanding of a rule can be tested by (1) eliciting from subjects or asking them to endorse, a statement of (2) a general rule or an argument for or against a particular conclusion. The combination of these features yields four procedures, which we now illustrate and discuss.

We begin with an informal example in which understanding of a rule is confirmed by the acceptance or endorsement of an argument. One of us has presented the following question to many squash players:

> 'As you know, a game of squash can be played either to 9 or to 15 points. Holding all other rules of the game constant, if A is a better player than B, which scoring system will give A a better chance of winning?'

Although all our informants had some knowledge of statistics, most of them said that the scoring system should not make any difference. They were then asked to consider the argument that the better player should prefer the longer game, because an atypical outcome is less likely to occur in a large sample than in a small one. With very few exceptions, the respondents immediately accepted the argument, and admitted that their initial response had been a mistake. Evidently, our informants had some appreciation of the effect of sample size on sampling errors, but they failed to code the length of a squash game as an instance of sample size. The fact that the correct conclusion becomes compelling as soon as this connection is made indicates that the initial response was an error of application, not of comprehension.

A more systematic attempt to diagnose the nature of an error was made in a study of a phenomenon labelled the conjunction effect (Tversky and Kahneman, in press). Perhaps the most elementary principle of probability theory is the conjunction rule, which states that the probability of a conjunction A&B cannot exceed either the probability of A or the probability of B. As the following example shows, however, it is possible to construct tests in which most judges—even highly sophisticated ones—state that a conjunction of events is more probable than one of its components.

To induce the conjunction effect, we presented subjects with personality sketches of the type illustrated below:

> 'Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.'

In one version of the problem, respondents were asked which of two statements about Linda was more probable: A. Linda is a bank teller; B. Linda is a bank teller who is active in the feminist movement. In a large sample of statistically naïve undergraduates, 86% judged the second statement to be more probable. In a sample of psychology graduate students, only 50% committed this error. However, the difference between statistically naïve and sophisticated respondents vanished when the two critical items were embedded in a list of eight comparable statements about Linda. Over 80% of both groups exhibited the conjunction effect. Similar results were obtained in a between-subject design, in which the critical categories were compared indirectly (Tversky and Kahneman, in press).

Tests of rule-endorsement and argument-endorsement were used in an effort to determine whether people understand and accept the conjunction rule. First, we presented a group of statistically naïve college students with several rule-like statements, which they were to classify as true or false. The statement: 'The probability of X is always greater than the probability of X

and Y' was endorsed by 81% of the respondents. In comparison, only 6% endorsed 'If A is more probable than B, then they cannot both occur'. These results indicate some understanding of the conjunction rule, although the endorsement is not unanimous, perhaps because of the abstract and unfamiliar formulation.

An argument-endorsement procedure was also employed in which respondents were given the description of Linda, followed by statements A and B above and were asked to check which of the following arguments they considered correct:

> (i) A is more probable than B because the probability that Linda is *both* a bank teller and an active feminist must be smaller than the probability that she is a bank teller.
> (ii) B is more probable than A because Linda resembles a bank teller who is active in the feminist movement more than she resembles a bank teller.

Argument (i) favoring the conjunction rule was endorsed by 83% of the psychology graduate students but only by 43% of the statistically naïve undergraduates. Extensive discussions with respondents confirmed this pattern. Statistically sophisticated respondents immediately recognized the validity of the conjunction rule. Naïve respondents, on the other hand, were much less impressed by normative arguments, and many remained committed to their initial responses, which had violated the conjunction rule.

Much to our surprise, naïve subjects did not have a solid grasp of the conjunction rule; they tended to endorse it in the abstract but not when it conflicted with a strong impression of representativeness. On the other hand, statistically trained subjects recognized the validity of the rule, and were able to apply it in an especially transparent problem. Statistical sophistication, however, did not prevent the conjunction effect in less transparent versions of the same problem. In terms of the present treatment, the conjunction effect appears to be an error of application, at least for the more sophisticated subjects. For further discussion of this issue see Tversky and Kahneman (in press).

In an attempt to describe the statistical intuitions of people at various levels of sophistication, Nisbett, Krantz, Jepson and Fong (in press) used an elicitation procedure, in which respondents were required to evaluate and justify certain conclusions and inferences attributed to characters in brief stories. The investigators observed large individual differences in the comprehension of basic statistical principles, which were highly correlated with the level of statistical training. Naturally, statistical intuitions vary with intelligence, experience, and education. As in other forms of knowledge, what is intuitive for the expert is often non-intuitive for the novice (see e.g.,

Larkin, McDermott, Simon and Simon, 1980). Nevertheless, some statistical results (e.g., the matching birthdays or the change of lead in a coin-tossing game) remain counter-intuitive even for students of probability theory (Feller, 1968, p. 85). Furthermore, there is some evidence that errors (e.g., the gambler's fallacy) that are commonly committed by naïve respondents can also be elicited from statistically sophisticated ones, with problems of greater subtlety (Tversky and Kahneman, 1971).

The elicitation method was also used by Wason and Evans (1975; Evans and Wason, 1976) in studies of logical intuitions in the well known four-card problem (Wason, 1966). In the standard version of this problem, the experimenter displays four cards showing A, T, 4 and 7, and asks subjects to identify the cards that should be turned over to test the rule 'if a card has a vowel on one side, it has an even number on the other'. The correct response is that the cards showing A and 7 should be examined, because the observation of an odd number on the first card or a vowel on the second would refute the rule. In a striking failure of logical reasoning, most subjects elect to look at the hidden side of the cards showing A and 4. Wason and Evans investigated different versions of this problem, and required their subjects to give reasons or arguments for their decisions of whether or not to look at the hidden side of each of the four cards. The investigators concluded that the arguments by which subjects justified their responses were mere rationalizations, rather than statements of rules that actually guided their decisions.

Other evidence for people's inadequate understanding of the rules of verification was reported by Wason (1969) and by Wason and Johnson-Laird (1970). In order to provide 'therapy', these investigators confronted subjects with the consequences of their judgments and called the subjects' attention to their inconsistent answers. This procedure had little effect on subsequent performance in the same task. Taken together, the results suggest that people's difficulties in the verification task reflect a failure of comprehension, not of application.

The examples that we have considered so far involved the endorsement of rules or arguments and the elicitation of arguments to justify a particular response. We have not discussed the procedure of asking respondents to state the relevant rule because such a test is often unreasonably demanding: we may want to credit people with understanding of rules that they cannot articulate properly.

The preferred procedures for establishing an error of application require a comparison of people's responses to a particular case with their judgment about a relevant rule or argument (McClelland and Rohrbaugh, 1978; Slovic and Tversky, 1974). It is also possible to confirm an error of application in other research designs. For example, Hamill, Wilson and Nisbett (1980)

showed subjects a videotaped interview allegedly conducted with a prison guard. Half the subjects were told that the opinions of the guard (very humane or quite brutal) were typical of prison personnel, while the other subjects were told that the guard's attitudes were atypical and that he was either much more or much less humane than most of his colleagues. The subjects then estimated the typical attitudes of prison personnel on a variety of issues. The surprising result of the study was that the opinions expressed by an atypical guard had almost as much impact on generalizations as did opinions attributed to a typical member of the group. Something is obviously wrong in this pattern of judgments, although it is impossible to describe any particular judgment as erroneous, and unlikely that many subjects would realize that they had effectively neglected the information about the guard's typicality (Nisbett and Wilson, 1977). In this case and in other between-subject studies, it appears reasonable to conclude that an error of application was made if the between-group comparison yields a result that most people would consider untenable.

We have defined an error of application as a response that violates a valid rule that the individual understands and accepts. However, it is often difficult to determine the nature of an error, because different tests of the understanding and acceptance of a rule may yield different results. Furthermore, the same rule may be violated in one problem context and not in another. The verification task provides a striking example: subjects who did not correctly verify the rule 'if a card has a vowel on one side, it has an even number on the other' had no difficulty in verifying a formally equivalent rule: 'if a letter is sealed it has a five cent stamp' (see Johnson-Laird, Legrenzi and Sonino-Legrenzi, 1972; Johnson-Laird and Wason, 1977; Wason and Shapiro, 1971).

These results illustrate a typical pattern in the study of reasoning. It appears that people do not possess a valid general rule for the verification of if-statements, or else they would solve the card problem. On the other hand, they are not blind to the correct rule or else they would also fail the stamp problem. The statement that people do not possess the correct intuition is, strictly speaking, correct—if possession of a rule is taken to mean that it is always followed. On the other hand, this statement may be misleading since it could suggest a more general deficit than is in fact observed.

Several conclusions of early studies of representativeness appear to have a similar status. It has been demonstrated that many adults do not have generally valid intuitions corresponding to the law of large numbers, the role of base rates in Bayesian inference, or the principles of regressive prediction. But it is simply not the case that every problem to which these rules are relevant will be answered incorrectly, or that the rules cannot appear compelling in particular contexts.

The properties that make formally equivalent problems easy or hard to solve appear to be related to the mental models, or schemas, that the problems evoke (Rumelhart, 1979). For example, it seems easier to see the relevance of 'not-q' to the implication 'p implies q' in a quality-control schema (did they forget to stamp the sealed letter?) than in a confirmation schema (does the negation of the conclusion imply the negation of the hypothesis?). It appears that the actual reasoning process is schema-bound or content-bound so that different operations or inferential rules are available in different contexts (Hayes and Simon, 1977). Consequently, human reasoning cannot be adequately described in terms of content-independent formal rules.

The problem of mapping statistical or logical intuitions is further complicated by the possibility of reaching highly unexpected conclusions by a series of highly intuitive steps. It was this method that Socrates employed with great success to convince his naïve disciples that they had always known truths, which he was only then making them discover. Should any conclusions that can be reached by a series of intuitive steps be considered intuitive? Braine (1978) discussed this question in the context of deductive reasoning, and he proposed immediacy as a test: A statement is intuitive only if its truth is immediately compelling, and if it is defended in a single step.

The issue of Socratic hints has not been explicitly treated in the context of judgment under uncertainty, and there are no rules that distinguish fair tests of intuitions, from contrived riddles on the one hand, and from Socratic instruction on the other. Imagine, for example, how Socrates might have taught a student to give the proper answer to the following question:

> 'Which hospital—a large or a small one—will more often record days on which over 60% of the babies born were boys?'.

This is a difficult question for Stanford undergraduates (Kahneman and Tversky, 1972, p. 441), but a correct answer can be elicited in a series of easy steps, perhaps as follows:

> 'Would you not agree that the babies born in a particular hospital on a particular day can be viewed as a sample?'
>
> 'Quite right. And now, would you have the same confidence in the results of a large sample, or of a small one?'
>
> 'Indeed. And would you not agree that your confidence is greater in a sample that is less likely to be in error?'
>
> 'Of course you had always known that. Would you now tell me what is the proportion of boys in a collection of babies which you consider the closest to an ideal of truth?'

> 'We agree again. Does that not mean, then, that a day on which more than 60% of babies born are boys is a grave departure from that ideal?'
>
> 'And so, if you have great confidence in a sample, should you not expect that sample to reveal truth rather than error?'. Etc.

The Socratic procedure is a heavy-handed way of leading the respondent to a desired response, but there are subtler ways of achieving the same goal. Fischhoff, Slovic and Lichtenstein (1979) showed that subjects become sensitive to base rates and to the reliability of evidence, when they encounter successive problems that vary only in these critical variables. Although these investigators did not obtain an effect of sample size even in a within-subject design, such effects have been obtained by Evans and Dusoir (1977) and by Bar-Hillel (1979) with a more transparent formulation and more extreme sample outcomes.

The hint provided by parallel problems may lead subjects to assign weight to a variable that is actually irrelevant to the correct response: Fischhoff and Bar-Hillel (1980) demonstrated that respondents were sensitive to irrelevant base-rate information, if that was the only variable distinguishing a set of problems. Indeed, subjects are prone to believe that any feature of the data that is systematically varied is relevant to the correct response. Within-subject designs are associated with significant problems of interpretation in several areas of psychological research (Poulton, 1975). In studies of intuitions, they are liable to induce the effect which they are intended to test.

On the limitations of the question-answering paradigm

In the preceding section we raised the possibility that within-subject designs and Socratic hints could prompt the intuitions under study. The problem is actually much broader. Most research on judgment under uncertainty and on inductive inference has been conducted in a conversational paradigm in which the subject is exposed to information and is asked to answer questions or to estimate values, orally or in writing. In this section we discuss some difficulties and limitations associated with this question-answering paradigm.

The use of short questionnaires completed by casually motivated subjects is often criticized on the grounds that subjects would act differently if they took the situation more seriously. However, the evidence indicates that errors of reasoning and choice that were originally established with hypothetical questions are not eliminated by the introduction of substantial incentives (Grether, 1979; Grether and Plott, 1979; Lichtenstein and Slovic, 1971, 1973; Tversky and Kahneman, 1981). Hypothetical questions are ap-

propriate when people are able to predict how they would respond in a more realistic setting, and when they have no incentive to lie about their responses. That is not to say that payoffs and incentives do not affect judgment. Rather, we maintain that errors of reasoning and choice do not disappear in the presence of payoffs. Neither the daily newspaper nor the study of past political and military decisions support the optimistic view that rationality prevails when the stakes are high (Janis, 1972; Janis and Mann, 1977; Jervis, 1975).

Perhaps a more serious concern regarding the question-answering paradigm is that we cannot safely assume that 'experimental conversations' in which subjects receive messages and answer questions will simulate the inferences that people make in their normal interaction with the environment. Although some judgments in everyday life are made in response to explicit questions, many are not. Furthermore, conversational experiments differ in many ways from normal social interaction.

In interpreting the subjects' answers, experimenters are tempted to assume (i) that the questions merely elicit from subjects an overt expression of thoughts that would have occurred to them spontaneously, and (ii) that all the information given to the subject is included in the experimental message. The situation is quite different from the subject's point of view. First, the question that the experimenter asks might not spontaneously arise in the situation that the experiment is meant to simulate. Second, the subject is normally concerned with many questions that the experimenter never thought of asking, such as: 'Is there a correct answer to this question? Does the experimenter expect me to find it? Is an obvious answer at all likely to be correct? Does the question provide any hints about the expected answer? What determined the selection of the information that I was given? Is some of it irrelevant and included just to mislead, or is it all relevant?' The single overt answer that the experimenter observes is determined in part by the subject's answers to this cluster of tacit questions. And the experimental message is only one of the sources of information that subjects use to generate both the covert and the overt answers (Orne, 1973).

Following Grice's William James lectures in 1967 (Grice, 1975), a large body of literature in philosophy, linguistics and psycholinguistics has dealt with the contribution of the cooperativeness principle to the meaning of utterances (for references, see Clark and Clark, 1977). By this principle, the listener in a conversation is entitled to assume that the speaker is trying to be 'informative, truthful, relevant and clear' (Clark and Clark, 1977, p. 560). Grice listed several maxims that a cooperative speaker will normally follow. For example, the maxim of quantity prohibits the speaker from saying things that the listener already knows, or could readily infer from the context or from the rest of the message. It is by this maxim that the state-

ment 'John tried to clean the house' conveys that the attempt was unsuccessful: the listener can assume that a successful attempt would have been described by the simpler sentence: 'John cleaned the house'.

Subjects come to the experiment with lifelong experience of cooperativeness in conversation. They will generally expect to encounter a cooperative experimenter, although this expectation is often wrong. The assumption of cooperativeness has many subtle effects on the subjects' interpretation of the information to which they are exposed. In particular, it makes it exceptionally difficult for the experimenter to study the effects of 'irrelevant' information. Because the presentation of irrelevant information violates rules of conversation, subjects are likely to seek relevance in any experimental message. For example, Taylor and Crocker (1979) commented on the fact that subjects' impressions of a person are affected by statements that are true of everybody, e.g., 'Mark is shy with his professors'. But the subjects' inference that Mark is unusually shy could be justified by the belief that a cooperative experimenter would not include a wholly redundant statement in a personality description. Similar issues arise in other studies (e.g., Kahneman and Tversky, 1973; Nisbett, Zukier and Lemley, 1981) which investigated the impact of irrelevant or worthless information.

The role of presuppositions embedded in a question was illustrated in a study by Loftus and Palmer (1974), who showed that eye-witnesses give a higher estimate of the speed of a car when asked 'how fast was the car going when it smashed the other car?' than when the question is 'how fast was the car going when it hit the other car?'. The use of the word 'smash' in the question implies that the questioner, if sincere and cooperative, believes that the car was going fast.

The normative analysis of such an inference can be divided into two separate problems: (i) should the witness be affected by the question in forming a private opinion of the speed of the car? (ii) Should the witness be affected by the question in formulating a public estimate? The answer to (i) must be positive if the question conveys new information. The answer to (ii) is less clear. On the one hand, it appears inappropriate for the reply to a question to echo information contained in the question. On the other hand, the cooperative witness is expected to give the best possible estimate in responding to a question about a quantity. What is the witness to do if that estimate has just been influenced by the question? Should the reply be: 'Before you asked me, I would have thought ...'? Whatever the normative merits of the case, the evidence indicates that people are often unable to isolate past opinions from current ones, or to estimate the weight of factors that affected their views (Fischhoff, 1977; Goethals and Reckman, 1973; Nisbett and Wilson, 1977; Ross and Lepper, 1980).

Our research on anchoring (Tversky and Kahneman, 1974) further illustrates the potency of subtle suggestions. In one study we asked a group of subjects to assess the probability that the population of Turkey was greater than 5 million, and we asked another group to assess the probability that the population of Turkey was less than 65 million. Following this task, the two groups recorded their best guesses about the population of Turkey; the median estimates were 17 million and 35 million, respectively, for the groups exposed to the low and to the high anchors. These answers can also be rationalized by the assumption that the values that appear in the probability questions are not very far from the correct one.

We have argued that suggestion effects can sometimes be justified because there is no clear demarcation between suggestion and information. It is important to note, however, that people do not accept suggestions *because* it is appropriate to do so. In the first place, they usually do not know that they have been affected by a suggestion (Loftus, 1979; Nisbett and Wilson, 1977). Second, similar suggestion effects are observed even when respondents cannot reasonably believe that an anchor they are given conveys information. Subjects who were required to produce estimates of quantities by adjusting up or down from a randomly generated value showed strong evidence of anchoring effects (Tversky and Kahneman, 1974). It is not suggestibility as such that is troublesome, but the apparent inability to discard uninformative messages.

When subjects are required to indicate their response by choosing an answer from a list, or by constructing a probability distribution over a given set of alternatives, the experimenter's choice of categories could be informative. Loftus (1979) has shown that respondents report many more headaches per week when the response scale is expressed as 1–5, 5–10, 10–15, etc., than when the scale is expressed as 1–3, 3–5, 5–7, etc. In this case, the scale could legitimately affect the boundaries of what is to be called a headache. Even when such reinterpretations are not possible, subjects may be expected to favor the middle of the range in their estimates of quantities, and to produce subjective probability distributions in which each category is assigned a non-negligible probability (Olson, 1976; Parducci, 1965).

Suggestions implied by the questionnaire could also contribute to a result observed by Fischhoff, Slovic and Lichtenstein (1978) who asked naïve subjects and experienced garage mechanics to evaluate the probability of different malfunctions that could cause failure in starting a car. They found that the estimated probability of the category 'all other problems' was quite insensitive to the completeness of the list, and was hardly increased when a major factor (e.g., the entire electrical system) was deleted from that list.

Even subtle and indirect clues can be effective. In a recent study we gave subjects the following information: 'Mr. A is Caucasian, age 33. He weighs 190 pounds'. One group of subjects were asked to guess his height. Other subjects also guessed his height, after first guessing his waist size. The average estimate was significantly higher in the first group, by about one inch. We surmise that subjects who first guessed waist size attributed more of Mr. A's weight to his girth than did subjects who only guessed his height.

We conclude that the conversational aspect of judgment studies deserves more careful consideration than it has received in past research, our own included. We cannot always assume that people will or should make the same inferences from observing a fact and from being told the same fact, because the conversational rules that regulate communication between people do not apply to the information that is obtained by observing nature. It is often difficult to ask questions without giving (useful or misleading) clues regarding the correct answer, and without conveying information about the expected response. A discussion of a related normative issue concerning the interpretation of evidence is included in Bar-Hillel and Falk (1982).

Naturally, the biasing factors that we have mentioned are likely to have most impact in situations of high uncertainty. Subjects' interpretations of the experimenter's conversational attitude will not be given much weight if they conflict with confident knowledge of the correct answer to a question. In the grey area where most judgment research is carried out, however, variations of conversational context can affect the reasoning process as well as the observed response.

Judgmental errors: positive and negative analyses

It is often useful to distinguish between positive and negative accounts of judgmental errors. A positive analysis focuses on the factors that produced a particular incorrect response; a negative analysis explains why the correct response was not made. For example, the positive analysis of a child's failure in a Piagetian conservation task attempts to specify the factors that determine the child's response, e.g., the relative height or surface area of the two containers. A negative analysis of the same behavior would focus on the obstacles that make it difficult for the child to acquire and to understand the conservation of volume. In the investigation of judgment under uncertainty, positive analyses are concerned with the heuristics that people use to make judgments, estimates and predictions. Negative analyses are concerned with the difficulties of understanding and applying elementary rules of reasoning. In the case of an error of comprehension, the negative analysis focuses on

the obstacles that prevent people from discovering the relevant rule on their own, or from accepting simple explanations of it. The negative analysis of an error of application seeks to identify the ways in which the coding of problems may mask the relevance of a rule that is known and accepted.

In general, a positive analysis of an error is most useful when the same heuristic explains judgments in a varied set of problems where different normative rules are violated. Correspondingly, a negative analysis is most illuminating when people consistently violate a rule in different problems, but make errors that cannot be attributed to a single heuristic. It then becomes appropriate to ask why people failed to learn the rule, if routine observations of everyday events offer sufficient opportunities for such learning. It also becomes appropriate to ask why people resist the rule, if they are not convinced by simple valid arguments. The difficulties of learning statistical principles from everyday experience have been discussed by several authors, notably Goldberg (1968), Einhorn and Hogarth (1978), and Nisbett and Ross (1980). Failures of learning are commonly traced to the inaccessibility of the necessary coding of relevant instances, or to the absence of corrective feedback for erroneous judgments. The resistance to the acceptance of a rule is normally attributed to its counter-intuitive nature. As an example, we turn now to the analysis of the reasons for the resistance to the principle of regressive prediction.

Studies of intuitive prediction have provided much evidence for the prevalence of the tendency to make predictions that are radical, or insufficiently regressive. (For a recent review of this literature see Jennings, Amabile and Ross, in press.) In earlier articles we offered a positive analysis of this effect as a manifestation of the representativeness heuristic (Kahneman and Tversky, 1973, 1979). However, as we shall see below, there are reasons to turn to a negative analysis for a more comprehensive treatment.

A negative analysis is of special interest for errors of comprehension, in which people find the correct rule non-intuitive, or even counter-intuitive. As most teachers of elementary statistics will attest, students find the concept of regression very difficult to understand and apply despite a lifetime of experience in which extreme predictions were most often too extreme. Sportcasters and teachers, for example, are familiar with manifestations of regression to mediocrity: exceptional achievements are followed more often than not by disappointment, and failures by improvement.

Furthermore, when the regression of a criterion variable on a predictor is actually linear, and when the conditional distributions of the criterion (for fixed values of the predictor) are symmetric, the rule of regressive prediction can be defended by a compelling argument: it is sensible to make the same prediction for all cases that share the same value of the predictor variable,

and it is sensible to choose that prediction so that the mean and the median of the criterion value, for all cases that share the same predicted value Y, will be equal to Y. This rule, however, conflicts with other intuitions, some of which are discussed below.

(i) 'An optimal rule of prediction should at least permit, if not guarantee, perfectly accurate predictions for the entire ensemble of cases'. The principle of regressive prediction violates this seemingly reasonable requirement. It yields a set of predicted values which has less variance than the corresponding set of actual criterion values, and thereby excludes the possibility of a set of precisely accurate predictions. Indeed, the regression rule guarantees that an error will be made on each pair of correlated observations: we can never find a son whose height was correctly predicted from his father's height, and whose height also allowed an accurate prediction of the father's height, except when both values are at the mean of the height distribution. It appears odd that a prediction rule that guarantees error should turn out to be optimal.

(ii) 'The relation between an observation and a prediction based on it should be symmetric'. It seems reasonable to expect that, if B is predicted from knowledge of A, then A should be the appropriate prediction when B is known. Regressive predictions violate this symmetry, of course, since the predictions of the two variables from each other are not governed by the same regression equation. A related asymmetry is encountered in comparing regressive predictions to the actual values of the criterion variable. Regressive predictions are unbiased, in the sense that the mean criterion value, over all cases for which a particular value Y was predicted, is expected to be Y. However, if we consider all the cases for which the criterion value was Y, it will be found that the mean of their predicted scores lies between Y and the group average. These asymmetries are puzzling and counter-intuitive for intelligent but statistically naïve persons.

The asymmetries of regressive prediction are especially troubling when the initial observation and the criterion are generated by the same process and are not distinguishable *a priori*, as in the case of repeated sampling from the same population, or in the case of parallel forms of the same test. The main mode of prediction that satisfies symmetry in such situations is an identity rule, where the score on the second form is predicted to be the same as the initial observation. The principle of regressive prediction introduces a distinction for which there is no obvious reason: how is it possible to predict the sign of the difference between two values drawn from the same population, as soon as one of these values is known?

(iii) 'Any systematic effect must have a cause'. The difference between initial observations and the corresponding criterion values is a fact, which can be observed in any scatter plot. However, it appears to be an effect with-

out a cause. In a test-retest situation, for example, the knowledge that the first score was high entails the prediction that the second will be lower, but the first observation does not cause the second to be low. The appearance of an uncaused effect violates a powerful intuition. Indeed, the understanding of regression is severely hindered by the fact that any instance of regression on which one stumbles by accident is likely to be given a causal explanation. In the context of skilled performance, for example, regression from an initial test to a subsequent one is commonly attributed to intense striving after an initial failure and to overconfidence following an initial success. It is often difficult to realize that performers would regress even without knowledge of results, merely because of irreducible unreliability in their performance. The regression of the first performance on the second is also surprising because it cannot be given a simple causal explanation.

We have sketched above a negative analysis of people's difficulties to understand and apply the concept of regressive prediction. We propose that people have strong intuitions about statistical prediction, and that some normatively correct principles are counter-intuitive precisely because they violate existing intuitions. In this view, the 'principles' that people adopt represent significant beliefs, not mere rationalizations, and they play a substantial role in retarding the learning of the correct rules. These beliefs, however, are often contradictory and hence unrealizable. For example, it is impossible to construct a non-degenerate joint distribution of the height of fathers and (first) sons so that the mean height of a father will be an unbiased predictor of the height of his son and the height of a son will be an unbiased predictor of the height of his father.

In conclusion, we have proposed that some errors and biases in judgment under uncertainty call for a dual analysis: a positive account that explains the choice of a particular erroneous response in terms of heuristics, and a negative account that explains why the correct rule has not been learned. Although the two analyses are not incompatible, they tend to highlight different aspects of the phenomenon under study. The attempt to integrate the positive and the negative accounts is likely to enrich the theoretical analysis of inductive reasoning.

Summary

We addressed in this essay three clusters of methodological and conceptual problems in the domain of judgment under uncertainty. First, we distinguished between errors of application and errors of comprehension and discussed different methods for studying statistical intuitions. Second, we reviewed

some limitations of the question-answering paradigm of judgment research and explored the effects of tacit suggestions, Socratic hints and rules of conversation. Third, we discussed the roles of positive and negative explanations of judgmental errors.

The considerations raised in this paper complicate the empirical and the theoretical analysis of judgment under uncertainty; they also suggest new directions for future research. We hope that a deeper appreciation of the conceptual and the methodological problems associated with the study of statistical intuitions will lead to a better understanding of the complexities, the subtleties, and the limitations of human inductive reasoning.

References

Bar-Hillel, M. (1979) The role of sample size in sample evaluation. *Organiz. Behav. Hum. Perf.*, *24*, 245–257.

Bar-Hillel, M. and Falk, R. (1982) Some teasers concerning conditional probability. *Cog.*, *11*, 109–122.

Braine, M. D. S. (1978) On the relation between the natural logic of reasoning and standard logic. *Psychol. Rev.*, *85*, 1–21.

Clark, H. H. and Clark, E. V. (1977) *Psychology and language*. New York: Harcourt Brace Jovanovich.

Cohen, L. J. (1979) On the psychology of prediction: Whose is the fallacy? *Cog.*, *7*, 385–407.

Cohen, L. J. (1981) Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences*, *4*, 317–331.

Edwards, W. (1975) Comment. *J. Amer. Statist. Ass.*, *70*, 291–293.

Einhorn, H. J. and Hogarth, R. M. (1978) Confidence in judgment: Persistence of the illusion of validity. *Psychol. Rev.*, *85*, 395–416.

Einhorn, H. J. and Hogarth, R. M. (1981) Behavioral decision theory: Processes of judgment and choice. *An. Rev. Psychol.*, *32*, 53–88.

Evans, J. St. B. T. and Dusoir, A. E. (1977) Proportionality and sample size as factors in intuitive statistical judgment. *Acta Psychol.*, *41*, 129–137.

Evans, J. St. B. T. and Wason, P. C. (1976) Rationalization in a reasoning task. *Brit. J. Psychol.*, *67*, 486–497.

Feller, W. (1968) *An introduction to probability theory and its applications*. New York, Wiley.

Fischhoff, B. (1977) Perceived informativeness of facts. *J. exper. Psychol.. Hum. Percept. Perf.*, *3*, 349–358.

Fischhoff, B. and Bar-Hillel, M. (1980) Focusing techniques as aids to inference. *Decision Research Report, 80-9*, Decision Research, Eugene, Oregon.

Fischhoff, B., Slovic, P. and Lichtenstein, S. (1978) Fault trees: Sensitivity of estimated failure probabilities to problem representation. *J. exper. Psychol.: Hum. Percept. Perf.*, *4*, 330–344.

Fischhoff, B., Slovic, P. and Lichtenstein, S. (1979) Subjective sensitivity analysis. *Organiz. Behav. Hum. Perf.*, *23*, 339–359.

Goethals, G. R. and Reckman, R. F. (1973) The perception of consistency in attitudes. *J. exper. Soc. Psychol.*, *9*, 491–501.

Goldberg, L. R. (1968) Simple models or simple processes? Some research on clinical judgments. *Amer. Psychol.*, *23*, 483–496.

Grether, D. M. (1979) Bayes rule as a descriptive model: The representativeness heuristic. *Social Science Working Paper* no. 245, California Institute of Technology.

Grether, D. M. and Plott, C. R. (1979) Economic theory of choice and the preference reversal phenomenon. *Amer. Econ. Rev., 59*, 623–638.

Grice, H. P. Logic and conversation. In D. Davidson and G. Harmon (eds.), *The logic of grammar*. Encino, Dickenson.

Hamill, R., Wilson, T. D. and Nisbett, R. E. (1980) Insensitivity to sample bias: Generalizing from atypical cases. *J. Personal. Soc. Psychol., 39*, 578–589.

Hammond, K. R., McClelland, G. H. and Mumpower, J. (1980) *Human judgment and decision making*. New York, Praeger.

Hayes, J. R. and Simon, H. A. (1977) Psychological differences among problem isomorphs. In N. J. Castellan, Jr., D. B. Pisoni, and G. R. Potts (eds.) *Cognitive theory*. Hillsdale, NJ, Erlbaum.

Janis, I. L. (1972) *Victims of groupthink*. Boston, Houghton-Mifflin.

Janis, I. L. and Mann, L. (1977) *Decision making*. New York: The Free Press.

Jennings, D., Amabile, T. and Ross, L. (in press) Informal covariation assessment. In D. Kahneman, P. Slovic, and A. Tversky (eds.) *Judgment under uncertainty: Heuristics and biases*. New York, Cambridge University Press.

Jervis, R. (1975) *Perception and misperception in international relations*. Princeton, Princeton University Press.

Johnson-Laird, P. N., Legrenzi, P. and Sonino-Legrenzi, M. (1972) Reasoning and a sense of reality. *Brit. J. Psychol., 63*, 395–400.

Johnson-Laird, P. N. and Wason, P. C. (1977) A theoretical analysis of insight into a reasoning task. In P. N. Johnson-Laird and P. C. Wason (eds.), *Thinking*. Cambridge, Cambridge University Press.

Kahneman, D., Slovic, P. and Tversky, A. (eds.) (in press) *Judgment under uncertainty: Heuristics and biases*. New York, Cambridge University Press.

Kahneman, D. and Tversky, A. (1972) Subjective probability: A judgment of representativeness. *Cog. Psychol., 3*, 430–454.

Kahneman, D. and Tversky, A. (1973) On the psychology of prediction. *Psychol. Rev., 80*, 237–251.

Kahneman, D. and Tversky, A. (1979) Intuitive prediction: Biases and corrective procedures. *TIMS Studies in Management Science, 12*, 313–327.

Larkin, J., McDermott, J., Simon, D. P. and Simon, H. A. (1980) Expert and novice performance in solving physics problems. *Science, 208*, 1335–1342.

Lichtenstein, S. and Slovic, P. (1971) Reversals of preference between bids and choices in gambling decisions. *J. exper. Psychol., 89*, 46–55.

Lichtenstein, S. and Slovic, P. (1973) Response-induced reversals of preference in gambling: An extended replication in Las Vegas. *J. exper. Psychol., 101*, 16–20.

Loftus, E. F. (1979) *Eyewitness testimony*. Cambridge, Harvard University Press.

Loftus, E. F. and Palmer, J. C. (1974) Reconstruction of automobile destruction: An example of the interaction between language and memory. *J. verb. Learn. verb. Behav., 16*, 585–589.

McClelland, G. and Rohrbaugh, J. (1978) Who accepts the Pareto axiom? The role of utility and equity in arbitration decisions. *Behav. Sci., 23*, 446–456.

Nisbett, R. E., Krantz, D. H., Jepson, C. and Fong, G. T. (in press) Improving inductive inference. In D. Kahneman, P. Slovic and A. Tversky (eds.), *Judgment under uncertainty: Heuristics and biases*. New York, Cambridge University Press.

Nisbett, R. E. and Ross, L. (1980) *Human inference: Strategies and shortcomings*. Englewood Cliffs, NJ, Prentice-Hall.

Nisbett, R. E. and Wilson, T. D. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychol. Rev., 84*, 231–259.

Nisbett, R. E., Zukier, H. and Lemley, R. E. (1981) The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cog. Psychol., 13*, 248–277.

Olson, C. L. (1976) Some apparent violations of the representativeness heuristic in human judgment. *J. exper. Psychol.: Hum. Percept. Perf., 2*, 599–608.

Orne, M. T. (1973) Communication by the total experiment situation: Why it is important, how it is evaluated, and its significance for the ecological validity of findings. In P. Pliner, L. Krames and T. Alloway (eds.) *Communication and Affect.* New York, Academic Press.

Parducci, A. (1965) Category judgment: A range-frequency model. *Psychol. Rev., 72*, 407–418.

Poulton, E. C. (1975) Range effects in experiments with people. *Amer. J. Psychol., 88*, 3–32.

Ross, L. and Lepper, M. R. (1980) The perseverance of beliefs: Empirical and normative considerations. In R. A. Shweder (ed.), *New directions for methodology of behavioral sciences: Fallible judgment in behavioral research.* San Francisco, Jossey-Bass.

Rumelhart, D. E. (1979) Schemata: The building blocks of cognition. In R. Spiro, B. Bruce and W. Brewer (eds.), *Theoretical issues in reading comprehension.* Hillsdale, NJ, Lawrence Earlbaum Associates.

Shweder, R. A. (ed.) (1980) *New directions for methodology of behavioral sciences: Fallible judgment in behavioral research.* San Francisco, Jossey-Bass.

Slovic, P., Fischhoff, B. and Lichtenstein, S. (1977) Behavioral decision theory. *An. Rev. Psychol., 28*, 1–39.

Slovic, P. and Tversky, A. (1974) Who accepts Savage's axiom? *Behav. Sci., 19*, 368–373.

Taylor, S. E. and Crocker, J. (1979) The processing of context information in person perception. Unpublished manuscript. Harvard University.

Tversky, A. and Kahneman, D. (1971) The belief in the law of small numbers. *Psychol. Bul., 76*, 105–110.

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

Tversky, A. and Kahneman, D. (1981) The framing of decisions and the psychology of choice. *Science, 211*, 453–458.

Tversky, A. and Kahneman, D. (in press) Judgments of and by representativeness. In D. Kahneman, P. Slovic and A. Tversky (eds.), *Judgment under uncertainty: Heuristics and biases.* New York, Cambridge University Press.

Wason, P. C. (1966) Reasoning. In B. Foss (ed.), *New horizons in psychology.* Harmondsworth, Middlesex, Penguin.

Wason, P. C. (1969) Regression in reasoning? *Brit. J. Psychol., 60*, 471–480.

Wason, P. C. and Evans, J. St. B. T. (1975) Dual processes in reasoning? *Cog., 3*, 141–154.

Wason, P. C. and Johnson-Laird, P. N. (1970) A conflict between selecting and evaluating information in an inferential task. *Brit. J. Psychol., 61*, 509–515.

Wason, P. C. and Shapiro, D. (1971) Natural and contrived experience in a reasoning problem. *Q. J. exper. Psychol., 23*, 63–71.

Résumé

Plusieurs facteurs rendent complexe l'étude des intuitions et des erreurs de jugement dans des conditions d'incertitude: l'étude des écarts entre l'acceptation et l'application des règles normatives, les effets du contenu sur l'application des règles, les allusions Socratiques créatrices d'illusions pendant qu'on les teste, les contraintes spécifiques des expériences inter-sujets, les interprétations par les sujets des messages expérimentaux selon les règles conversationnelles standards. L'analyse positive d'une erreur de jugement en terme d'heuristiques peut être complétée par une analyse négative pour expliquer pourquoi une règle correcte n'est pas intuitivement contraignante. Une analyse négative de prédiction non-regressive est proposée.