Psychological Review
1978, Vol. 85, No. 5, 395-416

Confidence in Judgment: Persistence of the Illusion of Validity

Hillel J. Einhorn
Graduate School of Business
University of Chicago

Robin M. Hogarth
London Graduate School of Business Studies
London, England
and Institut Européen d'Administration des
Affaires, Fontainebleau, France

An accumulating body of research on clinical judgment, decision making, and probability estimation has documented a substantial lack of ability of both experts and nonexperts. However, evidence shows that people have great confidence in their fallible judgment. This article examines how this contradiction can be resolved and, in so doing, discusses the relationship between learning and experience. The basic tasks that are considered involve judgments made for the purpose of choosing between actions. At some later time, outcome feedback is used for evaluating the accuracy of judgment. The manner in which judgments of the contingency between predictions and outcomes are made is discussed and is related to the difficulty people have in searching for disconfirming information to test hypotheses. A model for learning and maintaining confidence in one's own judgment is developed that includes the effects of experience and both the frequency and importance of positive and negative feedback.

Everyone complains of his memory and no one complains of his judgment. (La Rochefoucauld, 1959, p. 49)

Although the study and cataloguing of judgmental fallability has had a long history in psychology (see, e.g., Guilford, 1954, chap. 12; Johnson, 1972), an accumulating body of recent research on clinical judgment, decision making, and probability estimation has documented a substantial lack of ability across both individuals and situations (Slovic, Fischhoff, & Lichtenstein, 1977; Slovic & Lichtenstein, 1971). For example, predictive ability has been shown to have low (and even zero) validity in clinical settings (see, e.g., Einhorn, 1972; Goldberg, 1968, and his references). In addition, it is apparent that neither the extent of professional training and experience nor the amount of information available to clinicians necessarily increases predictive accuracy.

The fallibility of intuitive judgment has been further accentuated by the finding that simple statistical models for combining information consistently provide more accurate predictions than the judgments of clinicians (Meehl, 1954; Sawyer, 1966). Research on probability estimation (Hogarth, 1975) indicates several deficiencies: failure to appreciate the statistical notions of randomness, variance, and sampling variability; an inability to revise opinions to the extent prescribed by the normative rule of Bayes's theorem (Edwards, 1968); and reliance on judgmental heuristics (Tversky & Kahneman, 1974), which can lead to systematic biases in probability estimates. Further research indicates that judgment is not well calibrated in the sense that probabilities assigned to events are not of the same magnitude as the corresponding empirical relative frequencies (Lichtenstein, Fischhoff, & Phillips, 1977). Hindsight biases (Fischhoff, 1975) have also been documented.

This research was supported by a grant from the Illinois Department of Mental Health and Developmental Disabilities (Research and Development No. 740-01). We would like to thank Don Kleinmuntz and Steve Schacht for their programming assistance and Ken Hammond, Ed Joyce, Ben Kleinmuntz, Spyros Makridakis, and John Payne for comments on an earlier version of the article.

Requests for reprints should be sent to Hillel J. Einhorn, Graduate School of Business, University of Chicago, 5836 Greenwood Avenue, Chicago, Illinois 60637.

Copyright 1978 by the American Psychological Association, Inc. 0033-295X/78/8505-0395$00.75

Although intuitive predictions can be accurate (Murphy & Winkler, 1977), it might be thought that given the extensive evidence on the fallibility of human judgment, people would exhibit appropriate caution concerning their judgmental ability. However, both experimental evidence and casual empiricism do not support this view. For example, in a study by Oskamp (1965), self-confidence in judgments made by clinicians was found to increase as a function of the amount of information available to them but without any corresponding increase in judgmental accuracy (see also Ryback, 1967). Dawes (1976) has noted that after 20 years of research demonstrating the superiority of statistical over clinical prediction, clinicians continue to ignore the former and use the latter. Fischhoff, Slovic, and Lichtenstein (1977) have documented extreme overconfidence in probability judgments concerning answers for so-called "factual" questions (e.g., subjects expressed certainty in their answers to questions that were 18% to 27% incorrect). Moreover, in the calibration studies referred to above (Lichtenstein et. al., 1977), the major finding is that subjects are typically overconfident. Finally, Kahneman and Tversky (1973) have shown that people are most confident in judgment when information is consistent and/or extreme, even though these factors should induce them to decrease confidence in judgment. Indeed, Kahneman and Tversky (1973) state,

> The foregoing analysis shows that the factors which enhance confidence, for example, consistency and extremity, are often negatively correlated with predictive accuracy. Thus, people are prone to experience much confidence in highly fallible judgments, a phenomenon that may be termed the *illusion of validity*. Like other perceptual and judgmental errors, the illusion of validity persists even when its illusory character is recognized. (p. 249)

The question addressed in this article is the following: How can the contradiction between the considerable evidence on the fallibility of human judgment be reconciled with the seemingly unshakable confidence people exhibit in their judgmental ability? In other words, why does the illusion of validity persist? The importance of this question is that it concerns the relationship between learning and experience. That is, why does experience not teach people to doubt their fallible judgment?

It will be argued here that the answer to the above question is not primarily of a motivational nature, that is, that people selectively forget instances when their judgment is incorrect. Although motivated forgetting may occur, the concept does not account for the experimental results on overconfidence and is directly opposed to a functional analysis that emphasizes the development of coping mechanisms for survival. Furthermore, the use of cognitive theories to explain phenomena usually thought to be motivational in nature (e.g., conflict) has provided new insights into old problems (Brehmer, 1976; Hammond, 1965; Hammond & Brehmer, 1973). The approach followed here is to examine (a) the structure of judgmental tasks, (b) the extent to which people can observe the outcomes of judgments, and (c) how outcomes are coded and interpreted. That is, examination is made of outcomes that result from various configurations of task variables, the manner in which such feedback is coded in memory (Estes, 1976a, 1976b), and the way in which coded feedback is used for evaluating judgmental ability.

Task Structure and Psychological Process

The importance of understanding the environmental characteristics in which behavior occurs is difficult to overestimate. This point, originally made by Brunswik (1943), has been echoed recently by Simon and Newell (1971), Edwards (1971), Cronbach (1975), Dawes (1976), and Castellan (1977). This article therefore considers the structure of tasks in which judgments are made for the purpose of deciding between alternative courses of action. Note, however, that this is not the manner in which the validity of clinical judgment has usually been assessed. Most studies simply correlate judgments with criteria. How one then decides what action to take has been neglected (Einhorn & Schacht, 1977; Elstein, 1976). However, in real-world situations, judgments are made for the purpose of choosing between actions. This means that outcome information, which is available only after actions are taken, is frequently the only source of feedback with which to compare

judgment. Therefore, to understand how people learn about their judgmental ability, it is necessary to consider judgments, actions, and outcome feedback together.

Consider situations with two possible actions, A and B. Denote by x an overall evaluative judgment, which may itself be a function of various types and amounts of information. Furthermore, let x_c be a cutoff point, so that:

$$\begin{aligned} &\text{If } x \geq x_c, \text{ choose Action A;} \\ &\text{if } x < x_c, \text{ choose Action B.} \end{aligned} \tag{1}$$

Although simplistic, Equation 1 applies to many judgment and decision situations, for example, job hiring, promotion, admission to school, loan and credit granting, assignment to remedial programs, admission to social programs, journal article acceptance, grant awarding, and so on. In these cases, a judgment of the degree of deservedness typically determines which action is to be taken, since the preferred action cannot be given to all.

In order to compare judgment to a standard, the existence of a criterion, denoted y, is assumed to serve as the basis for evaluating the accuracy of judgment. The practical difficulties of finding and developing adequate criteria are enormous, but the focus here is theoretical: It is the concept of a criterion that is necessary for this analysis. To be consistent with the formulation of judgment, it is further assumed that the criterion has a cutoff point, y_c, so that $y \geq y_c$ and $y < y_c$ serve as the basis for evaluating the outcomes of judgment. Thus, as far as learning about judgment is concerned, representation of outcomes in memory is often of categorical form, that is, successes and failures (cf. Estes, 1976a).

Now consider the regression of y on x and the four quadrants that result from the intersection of x_c and y_c as illustrated in Figure 1. Denote the correct predictions as positive and negative hits and denote the two types of errors as false positives ($y < y_c | x \geq x_c$) and false negatives ($y \geq y_c | x < x_c$). To estimate the relationship between x and y, it is necessary to have information on each judgment–outcome combination. Assume first that such information becomes available over time (sequentially) and consider the experimental

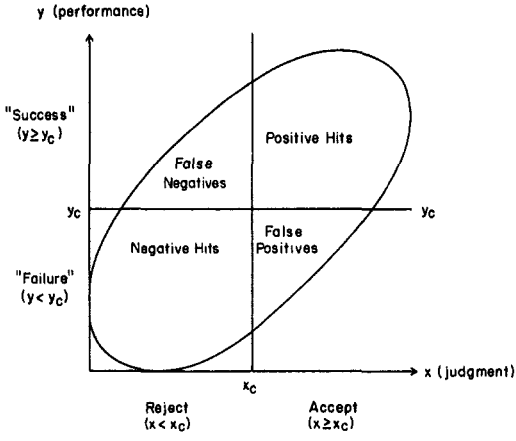


Figure 1. Action–outcome combinations that result from using judgment to make an accept or reject decision.

evidence concerned with learning the relationship between x and y in such circumstances. Research on the ability to judge the contingency between x and y from information in 2×2 tables (Jenkins & Ward, 1965; Smedslund, 1963, 1966; Ward & Jenkins, 1965) indicates that people judge the strength of relationship by the frequency of positive hits (in the terminology of Figure 1) while generally ignoring information in the three other cells. These results raise two important issues: First, why is information other than positive hits ignored? And second, why is frequency of positive hits the important variable rather than some measure of relative frequency or probability (either joint or conditional)?

It will be argued that the answer to both questions concerns the difficulty people have in making use of "disconfirming information," by which is meant the information that can be gained by the nonoccurrence of an action or prediction. Furthermore, a principal cause of this difficulty is the structure of judgmental tasks in the natural environment, since this determines the conditions in which inferential learning can occur. For example, consider Figure 1, where outcomes for judgments below x_c are usually not observable. Consequently, it is suggested that habits gained through experience of inferential learning in naturalistic settings prevent people from using information that is available in laboratory tasks (cf. Björkman, 1966).

Support for the viewpoint that inferential

habits can lead to certain types of judgmental error has also been made in a somewhat different manner by Smedslund (1963) with reference to Piaget and Inhelder's (1951) theory of the development of probabilistic notions in children. Smedslund notes that the stage of concrete reasoning, which precedes the ability to apply "correlational reasoning," functions only on the basis of observable events. That is, disconfirming information cannot be used. Furthermore, Smedslund maintains that Piaget and Inhelder's (1951) stages of development represent different levels of cognitive functioning at which adults operate. Thus, although adults are capable of using disconfirming information for inferring relationships, they frequently fail to do so and operate at lower, more frequently used cognitive levels, for example, concrete reasoning. In other words, judgmental habits are based on experience with lower levels of cognitive functioning (see also Smedslund, 1966).

Consider the experimental evidence on the ability to use disconfirming information for making inferences. In an important series of papers, Wason (1960, 1966, 1968, 1969) has explored this issue in detail. In an early study (Wason, 1960), he presented subjects with a three-number sequence, for example, 2, 4, 6. Subjects were required to discover the rule to which the three numbers conformed (the rule being three ascending numbers). To discover the rule, they were permitted to generate sets of three numbers that the experimenter classified as conforming or not conforming to the rule. At any point, subjects could stop when they thought they had discovered the rule. The correct solution to this task should involve a search for disconfirming evidence rather than the accumulation of confirming evidence. For example, if someone believed that the rule had something to do with even numbers, this could only be tested by trying a sequence involving an odd number (i.e., accumulating vast amounts of confirming instances of even-number sequences would not lead to the rule). The fact that only 6 of 29 subjects found the correct rule the first time they thought they did illustrates the dangers of induction by simple enumeration. As Wason (1960) points out, the solution to this task must involve "a willingness to attempt to falsify hypotheses, and thus to *test those intuitive ideas which so often carry the feeling of certitude*" (p. 139, our emphasis).

It is important to emphasize that in Wason's experiment, where actions were not involved, a search for disconfirming evidence is possible. However, when actions are based on judgment, learning based on disconfirming evidence becomes more difficult to achieve. For example, consider how one might erroneously learn the rule "my judgment is highly predictive" and focus on the hypothetical case of a manager learning about his predictive ability concerning the potential of job candidates. The crucial factor here is that actions (e.g., accept or do not accept) are contingent on judgment. Therefore, at a subsequent date, the manager can only examine accepted candidates to see how many are successful. If there are many successes (which, as will be shown below, is likely), these instances all confirm the rule. Indeed, the important point here is that it would be difficult to disconfirm the rule, even though it might be erroneous.

One way in which the rule could be tested would be for the manager to accept a subset of those he judged to have low potential and then to observe their success rate. If their rate was as high as those judged to be of high potential, the rule would be disconfirmed. However, a systematic search for disconfirming evidence is rare and could be objected to on utilitarian or even ethical grounds, that is, one would have to withhold the preferred action from some of those judged most deserving and give it to some judged least deserving. Therefore, utilitarian or ethical considerations may prevent one from even considering the collection of possible disconfirming information. Note that the tendency not to test hypotheses by disconfirming instances is a direct consequence of the task structure in which actions are taken on the basis of judgment. Furthermore, as Wason (1960) points out, "In real life there is no authority to pronounce judgment on inferences: the inferences can only be checked against the evidence" (p. 139). Therefore, large amounts of positive feedback can lead to reinforcement of a nonvalid rule and hence to the illusion of validity.

A second series of experiments by Wason (1968, 1969) provides further insight into how people search for information to verify inferences (see also Wason & Johnson-Laird, 1972). Subjects were presented with four cards on which one letter or number appeared (a, b, 2, or 3). They were then told to verify the statement "All cards with a vowel on one side have an even number on the other" by indicating only those cards that would need to be turned over in order to determine whether the rule was true or false. Most subjects chose Cards a and 2 or Card a alone instead of the correct response of Cards a and 3.

The results of this experiment highlight two related points. The first is the lack of search for disconfirming evidence, namely, ignoring Card 3. The second point concerns the choice of Cards a *and* 2 and seems to follow from an assumed symmetry in the problem of the form: If P implies Q, then Q implies P. Although this assumed symmetry is clearly a logical fallacy, the choice of Card 2 in addition to Card a indicates that the subjects do not perceive it as such. The relevance of this observation for understanding how contingency judgments are made is obvious. In fact, Jenkins and Ward (1965) state that subjects in the contingency task reason as follows: If there were a contingency, favorable results would occur; since favorable results did occur, there was a contingency. Johnson (1972, chap. 6) also discusses the difficulties and errors involved in this kind of reasoning.

Although it may seem farfetched to those trained in scientific method and experimental design that people do not seek disconfirming evidence when testing hypotheses, the relative novelty of thinking in this manner should not be overlooked. For example, the concept of a control group, which is essential to scientific method and which illustrates the necessity of nonoccurrence (i.e., no treatment) for making valid inferences, came rather late in the history of thought (Boring, 1954). Moreover, the notion of equating experimental and control groups prior to treatment through randomization is a revolutionary twentieth-century notion attributed to R. A. Fisher. Finally Popper's (1959) views that hypotheses can only be disconfirmed by evidence but never confirmed have also only recently gained acceptance.[1]

Replication of Wason's Experiment

Since Wason's (1968, 1969) experimental results are important to the arguments presented in this article, we have collected additional evidence. Specifically, an attempt was made to find subjects known to have been trained in examining possible disconfirming evidence as well as an experimental stimulus related to checking predictive ability (and which was, as a consequence, less abstract than that used by Wason). The subjects were 23 statisticians (faculty members and graduate students of statistics departments of colleges of the University of London), who were attending a research seminar given by one of the authors. The importance of using high-level statisticians as subjects is that they are formally trained in testing statistical hypotheses, that is, null hypotheses are frequently formulated so that one can see whether they are rejected by the data. Consequently, if such subjects were to exhibit behavior similar to those of Wason, this would clearly be consistent with the notion that the habits of lower level cognitive functioning, for example, concrete reasoning, are strong.

The experimental stimulus concerned checking the claim of accurate predictive ability made by a consultant with respect to rises and falls in a particular market. The example had been developed by the experimenter when teaching elements of decision theory to experienced managers (who, incidentally, had extraordinary difficulty with the problem). The stimulus, which was also carefully explained verbally by the experimenter, was the following:

It is claimed that when a particular consultant says the market will rise (i.e., a favorable report), it always does rise.

---

[1] John Stuart Mill (1851) is generally credited with formally discussing the notion of control groups. However, Boring (1954) provides earlier (post-Renaissance) references to the recognition of the need for controls as a basis for comparison. On the other hand, he emphasizes the growth of the use of control groups in experimental work as a twentieth-century phenomenon strongly associated with the development of methods for statistically testing differences between groups.

You are required to check the consultant's claim and can observe any of the outcomes or predictions associated with the following:

1. favorable report.
2. unfavorable report.
3. rise in the market.
4. fall in the market.

Subjects were then asked the following question: What is the *minimum* evidence you would need to check the consultant's claim? Subjects were requested to respond by circling the appropriate statement number(s).

Results were as follows: Of the 23 statisticians, 12 requested a single piece of confirmatory information (11 asked for Response 1, and 1 for either Response 1 or 3); 1 person asked for any of the four possibilities; 2 people asked for either Number 1 or 4; 3 people asked for Number 4 alone; and a mere 5 people indicated the correct response of 1 and 4. Results were thus somewhat different from those of Wason. First, no one committed the logical fallacy implied by choosing Responses 1 and 3. Second, there is some evidence that scientific training may make people more aware of the need to seek disconfirming information in that almost half the responses did include Response 4. On the other hand, the fact remains that when checking a rule concerning predictive ability, a majority of analytically sophisticated subjects failed to make the appropriate response. In particular, half of the subjects chose to examine the same piece of confirmatory information, that is, Response 1.

What Is Learned From Outcomes?

In addition to the difficulty of learning from disconfirming information, the earlier work on correlational learning pointed to the frequency of positive hits as the determinant of the perceived relationship between x and y. If outcomes are coded in memory as frequencies rather than probabilities, this has major implications for explaining the persistence of the illusion of validity. Recent research on probability learning is relevant to this issue.

Despite the voluminous literature on probability learning (see, e.g., Estes, 1972; Jones, 1971; Myers, 1976), Estes (1976a) has recently pointed out that research into what is actually learned has been relatively scarce. In order to remedy this, he has performed an extensive series of experiments (Estes, 1976a, 1976b) concerned with the coding of outcomes in memory and how subjective probability and predictive behavior are based on such coding. Estes's experimental paradigm involved presenting subjects with three pairs of political candidates from a simulated public opinion poll together with outcome information concerning the number of wins obtained by each candidate. At the end of this observation stage, subjects were asked to predict the winners between candidates from different pairs, that is, those who had not been paired in the observation stage. The advantage of this design is that it allows for the separation of the effects of probability and frequency. For example, consider Pairs A versus B and C versus D. In the former, imagine 100 trials where A wins 75 times; thus, $p(\mathrm{A} > \mathrm{B}) = .75$. In the second pair, imagine 200 trials where C wins 100 times; thus, $p(\mathrm{C} > \mathrm{D}) = .50$. If A is now paired with C, and subjects are asked to predict the winner, responses based on probability would indicate A; whereas responses based on frequency of winning would indicate C.

Estes's (1976a) results indicated that subjects had a strong tendency to predict the more frequently winning candidate even when that candidate had a lower probability of winning. This result was sufficiently strong so that in another experiment (Estes, 1976b), the same losing candidate, previously paired with four winning candidates, was chosen over a new candidate that had appeared in no observation trials.

Again, it should be asked why outcomes appear to be coded as frequencies rather than probabilities. While no definitive answer can be given, the following observations are pertinent: (a) Probability differs from frequency in that frequency must be divided by all elementary events in the sample space (assuming, of course, that all events have the same probability of occurrence). However, to take all elementary events into account, one must also pay attention to instances of the nonoccurrence of the event of interest. Therefore, inability to deal adequately with nonoccurrence of events would favor the coding of outcomes as frequencies rather than

probabilities. Parenthetically, the statistician Lindley (1965, p. 5) points out that it is easy to produce paradoxes in probability theory by failure to mention the conditioning event (i.e., the sample space against which frequency should be compared). (b) It should also be recalled that the notion of probability itself has developed relatively late, a major stumbling block being the division of frequency by the appropriate sample space (David, 1955; Hacking, 1975; Kendall, 1956). This late development is even more remarkable when one considers that notions of gambling and games of chance existed in antiquity (Cohen, 1972; David, 1962).[2]

Although outcomes appear more likely to be coded in memory as frequencies than as probabilities, the manner in which predictions and subjective probability judgments are made on the basis of such coding is more complicated. Estes (1976a) suggests that there is no general rule, that is, task characteristics and experience may have specific effects. However, in the contingency tasks considered here, the experimental evidence is certainly consistent with the notion that frequency is more salient in memory than probability. In fact, in a number of experiments (Estes, 1976b), the total number of winning outcomes for a given alternative was the major determinant of subjects' choice behavior.

Model for Learning Confidence in Judgment

A model is now developed that relates positive hits and false positives to confidence in one's judgment. The basic assumptions of the model relate to the previous discussion: (a) Evidence about outcomes contingent on the action not taken is missing, or if outcomes are available, attention is not paid to them (as in the contingency studies). (b) Action–outcome combinations are coded as frequencies rather than probabilities.

The variables in the model are denoted as follows: N = number of decisions made; $\phi = p(x \geq x_c)$ = selection ratio, that is, unconditional probability of giving Action A (see Equation 1); N_p = number of positive hits; N_f = number of false positives; $ph = p(y \geq y_c | x \geq x_c)$ = positive hit rate; and $fp = p(y < y_c | x \geq x_c)$ = false positive rate $= 1 - ph$. From these definitions, it is easily shown that

$$N_p = N\phi ph$$
$$\text{and} \qquad (2)$$
$$N_f = N\phi(1 - ph).$$

Now consider that confidence in judgment is defined as the strength of the learned concept "my judgment is accurate." The significance of this conceptualization is that it places "confidence" clearly within concept learning. Therefore, in principle, the learning of confidence in one's judgment should be no different from the learning of other concepts, and the role of the reinforcing values of positive and negative feedback will be particularly important.

Let β_1 and β_2 denote the relative reinforcing values of positive and negative feedback, respectively; and, since β_1 and β_2 are relative weights, define their sum as

$$\beta_1 + \beta_2 = 1.0 \quad (\beta_1, \beta_2 \geq 0). \qquad (3)$$

Therefore, when $\beta_1 > .5$, positive feedback has greater reinforcement value than negative feedback. Furthermore, to combine the relative importance of the type of feedback with the amount of feedback (i.e., N_p and N_f), let the total feedback effect (F) on confidence (C) be written as

$$F = \beta_1 N_p - \beta_2 N_f. \qquad (4)$$

Equation 4 defines the total feedback effect as the difference between the amounts of positive and negative feedback weighted by their relative reinforcement values. It is further assumed that confidence in judgment (C) is an increasing monotonic function of F, that is,

$$C = f(F),$$

---

[2] By "late" is meant the sixteenth and seventeenth centuries, Cardano, Galileo, Huygens, Pascal, and Fermat. In commenting on the development of probabilistic ideas, David (1955) asks, "The question which constantly recurs to one while studying the games of the past is 'Why did not someone notice the equiproportionality property of the fall of the die?' " (p. 6). In addition, Kendall (1956) states, "It might have been supposed that during the several thousand years of dice playing preceding, say, the year A.D. 1400, some idea of the permanence of statistical ratios and the rudiments of a frequency theory of probability would have appeared. I know no evidence to suggest that this was so" (p. 3).

where f is an increasing monotonic function. The model expressed in Equation 4 is quite general. For example, the implication of people only considering positive hits would be: $\beta_2 = 0$, $\beta_1 = 1$, and $F = N_p$. Exactly what affects the relative sizes of β_1 and β_2 is problematic and is considered below.

Now consider the following implication of the model. Substitute Equation 2 into Equation 4 to obtain

$$\begin{aligned} F &= \beta_1 N\phi ph - \beta_2 N\phi(1 - ph) \\ &= N\phi(\beta_1 ph - \beta_2 + \beta_2 ph) \\ &= N\phi[(\beta_1 + \beta_2)ph - \beta_2]. \end{aligned} \tag{5}$$

Since from Equation 3, $\beta_1 + \beta_2 = 1.0$, Equation 5 can be rewritten as

$$F = N\phi(ph - \beta_2). \tag{6}$$

Three important points are illustrated by Equation 6:

1. The sign of the total feedback effect (F) is determined by the difference $(ph - \beta_2)$. It is shown in the next section that the positive hit rate (ph) is greater than .5 in many situations. Therefore, in order to have $F < 0$, β_2 must be greater than .5, which means that negative feedback must have greater reinforcement value than positive feedback (i.e., $\beta_2 > \beta_1$). While this is possible and would lead to a decrease in confidence, the evidence on the relative effects of positive versus negative reinforcement is clearly consistent with the notion that $\beta_1 > \beta_2$. Indeed, in some studies, evidence is consistent with $\beta_2 = 0$ (Estes, 1976a, 1976b). However, although F can be expected to be positive in most cases, the model in Equation 6 has the advantage of showing when F will be negative, a point that is discussed further below.

2. When $F > 0$, note that the size of the total feedback effect is directly related to N, the number of decisions made (or the number of learning trials). Therefore, F increases with N, as does confidence in one's judgment. This aspect of the model is particularly illuminating, since it helps to explain why judges with greater experience (larger N) may feel considerable confidence in judgment that is no more valid than those who have little experience. While empirical evidence on the relationship between confidence and total feedback effect is lacking, it is interesting to speculate as to the form of the functional relationship. One possibility that has considerable appeal is an **S**-shaped function, since this would imply that confidence in judgment is built up slowly with experience, rises rapidly with moderate amounts of experience, and then levels off (and reaches asymptote) with large amounts of experience. Clearly, further work is necessary to test this conjecture.

3. When $F > 0$, the total feedback effect is a function of positive and negative reinforcement that occurs on a partial-reinforcement schedule (Skinner, 1953). Therefore, although acquisition of the concept "my judgment is accurate" should be slow (lending some credence to the idea of an **S**-shaped learning function), it should be highly resistant to extinction. The implications of this prediction are quite disturbing, since it suggests that once confidence in judgment is learned, even negative evidence will not quickly extinguish the concept. Therefore, when confidence in judgment is unwarranted, the illusion of validity can be maintained in the face of contradictory evidence.

It is clear that the major implications of the model depend on the size of the positive hit rate. Therefore, the factors affecting the positive hit rate and the values it is likely to assume are now considered.

Factors Affecting the Positive Hit Rate

From Figure 1, it can be seen that three factors affect the positive hit rate: (a) judgmental ability as measured by ρ_{xy}, that is, the correlation between x and y; (b) the unconditional probability of being judged above the cutoff, that is, the so-called selection ratio; and (b) the base rate or unconditional probability of a success. The selection ratio was defined previously and denoted ϕ, while the base rate is $p(y \geq y_c)$ and is denoted here by br.

The effects of these three factors on the positive hit rate are well known. Taylor and Russell (1939), for example, have shown that one can increase the positive hit rate for any given ρ_{xy} and base rate by reducing the selection ratio (ϕ), that is, by raising the cutoff point for the preferred action (assuming

$\rho_{xy} \neq 0$). Therefore, even if ρ_{xy} is low, it is possible to have a high positive hit rate, depending on the values of ϕ and br. Taylor and Russell (1939) provide tables of positive hit rates for a wide range of values of ρ_{xy}, ϕ, and br. Examination of these tables shows that low correlations between judgments and criteria are not incompatible with large positive hit rates.

In addition to the three factors already mentioned, a fourth factor must be considered, which can be illustrated by imagining the following experiment. Assume that a series of judgments is made about some persons. Of those judged to be above x_o, randomly assign half to Action A and half to Action B. Similarly, do the same for those judged below x_o. At some later time, measure performance and calculate the proportion of those with $y \geq y_o$ in each cell (each person is assigned a 0 or 1 to indicate whether he or she is below or above the cutoff on y, the proportion above y_o being simply the mean of that cell). This is a 2×2 factorial design, with one factor being "judgment" and the other "type of action." Note that because the criterion cannot be measured immediately before the decision (indeed, if it could, there would be no need for judgment), people receiving Actions A and B have also received different experimental treatments.

If this experiment were done, one could test for the main effect of judgment (which measures its accuracy); the main effect for the action, that is, whether receiving Action A or B in itself causes differences in performance; and the interaction between judgment and action. Observe that the advantage of the experiment is that it allows one to untangle the accuracy of judgment from the treatment effects of the action. However, such an experiment is rarely done, even conceptually and especially not by people without extensive training in experimental design. Therefore, judgmental accuracy will almost always be confounded with possible treatment effects due to actions. Furthermore, and with reference to the earlier discussion, this experiment allows one to examine disconfirming information. Therefore, in contrast to most real judgmental tasks, it would permit one to disconfirm the hypothesis of judgmental accuracy as well as to estimate any treatment effects due to the action.

To illustrate how treatment effects may increase confidence in judgment, consider the decision to award or not to award grants to researchers.[3] Assume that grant applications are judged on some basis of potential, where those judged above x_o receive awards and those judged below x_o are denied. Assume also that the granting agency wishes to determine whether its judging procedures produce satisfactory results. To this end, it develops a criterion that reflects both quantity and quality of completed research. It then examines funded projects and calculates the proportion considered successes. (If the agency were wise, it might also try to discover the proportion of successful projects it had refused to fund. The difficulty of doing this, however, illustrates the earlier point about the rarity of having complete information to evaluate judgment.)

If the proportion of successes for those given grants is high, the agency might feel that its judgmental procedures are quite accurate. However, note that the treatment effect of receiving a grant is completely confounded with judgmental accuracy, for example, obtaining a grant can give a researcher time and resources to do more and better work. If there were a main effect for the action (in the direction assumed here), one might still experience a high positive hit rate, even if the accuracy of the judgment were low (or perhaps zero). Note that the proper experiment would be difficult to do, since it would require withholding grants from some deserving cases while awarding grants to some who do not deserve them. Consequently, the assumed validity of judgment can be continually reinforced by experience.

Model for Determining Positive Hit Rates

To assess the effects on positive hit rates of the above four factors, a simulation experiment

[3] It should be emphasized that this example is used only for illustrative purposes. We are aware of no data on this particular issue. However, similar treatment effects have been documented in the literature and are discussed further.

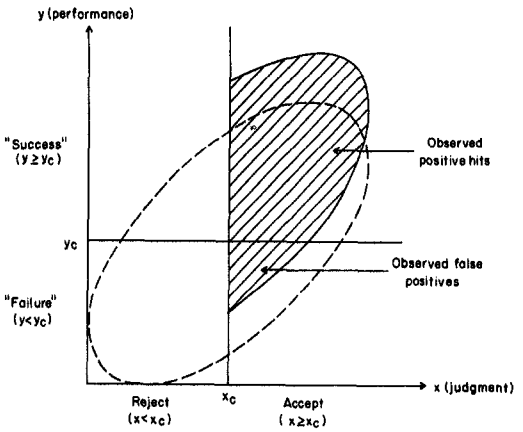


Figure 2. Effects of treatment on the observed positive hit rate.

was performed based on the following model. Assume that in the absence of any treatment effects, both x and y are standardized, and that they are distributed as bivariate normal. Furthermore, let true judgmental ability be defined as ρ_{xy}, the correlation between x and y. By "true" judgmental ability is meant the correlation that would occur if there were no treatment effects. It is important to think of ρ_{xy} in this manner, since empirical correlations between x and y may be contaminated by treatment effects. Attention will be limited to a possible additive treatment effect, to be denoted t, which occurs for those persons judged to exceed x_c. Under these assumptions, the relationship between y and x can be expressed as

$$y = \rho_{xy}x + zt + \epsilon, \tag{7}$$

where z is a dummy variable with the specification

$$z = \begin{cases} 1 & \text{if } x \geq x_c \\ 0 & \text{if } x < x_c, \end{cases}$$

t represents a treatment effect measured in units of the standard deviation of performance (e.g., $t = .5$ means that for those judged above x_c, the treatment increases y by .5.), and ϵ denotes a random disturbance with mean of 0.

Note that the model could also incorporate a negative treatment effect (i.e., people below x_c receive an action that reduces their y scores) by changing the specification of the dummy variable when $x < x_c$ from 0 to -1. It follows from Equation 7 that the conditional expectation of y is

$$E(y|x, \rho_{xy}, z, t) = \rho_{xy}x + zt. \tag{8}$$

Therefore, the conditional probability of observing a success, that is, an outcome above y_c, for any $x \geq x_c$, can be found by making use of the conditional distribution of y given x; that is,

$$p(y \geq y_c|x, \rho_{xy}, z, t) = \int_{y_c}^{\infty} [2\Pi(1 - \rho_{xy}^2)]^{-\frac{1}{2}} \times \exp\left\{\frac{-(y - \rho_{xy}x - zt)^2}{2(1 - \rho_{xy}^2)}\right\} dy. \tag{9}$$

From Equation 9, it can be seen that the probability of observing a successful outcome depends on (a) y_c, and thus the base rate, br; (b) x_c, and thus the selection ratio, ϕ—since z is a function of x_c; (c) ρ_{xy}, true judgmental ability; and (d) t, the size of the treatment effect.

Treatment effects are illustrated in Figure 2. The dotted ellipse is that shown in Figure 1 and represents the true relationship between judgments and outcomes. The shaded portion indicates those outcomes that can be observed; hence, only values for which $x \geq x_c$ are shown. The treatment effect occurs in that the outcomes (i.e., performance) of all those given Action A are increased by a constant amount, so that the number of positive hits is greater than would have been observed in the absence of treatment effects. From a psychological viewpoint, the key aspect of Figure 2 is that the nature of the feedback to the judge is contaminated; the number of positive hits is inflated.

To quantify the extent to which the factors discussed above affect the positive hit rate, the simulation study involved combinations of the following levels of the four factors: four levels of treatment effects ($t = 0, .5, 1.0$, and 1.5), five levels of base rate ($br = .1, .3, .5, .7$, and .9), five levels of selection ratio ($\phi = .1, .3, .5, .7$, and .9), and five levels of judgmental ability ($\rho_{xy} = 0, .2, .4, .6$, and .8). This design resulted in $4 \times 5 \times 5 \times 5 = 500$ combinations of all factor levels. For each of these

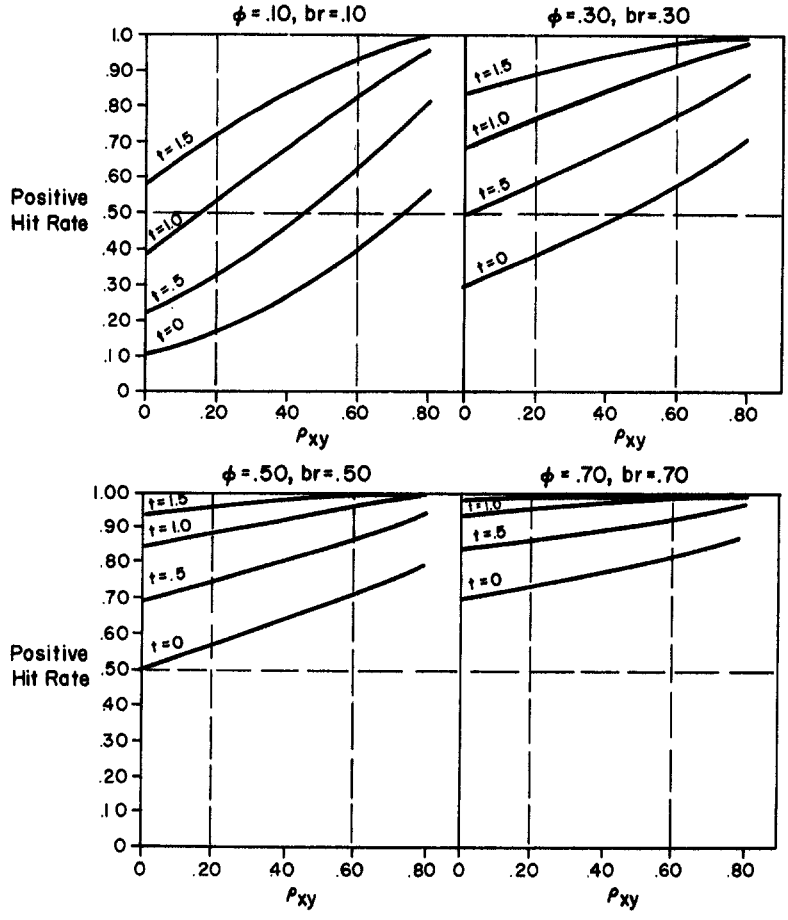


Figure 3. Positive hit rate as a function of correlation for differing treatment effects. ($\phi = br$.)

combinations, the positive hit rate was computed using the following formula:

$$p(y \geq y_c | x \geq x_c, \rho_{xy}, t, z)$$

$$= \frac{\int_{x_0}^{\infty} \int_{y_0}^{\infty} f[x, (y+t)]dx\, d(y+t)}{\int_{x_0}^{\infty} f(x)dx}, \quad (10)$$

where $f[x, (y + t)]$ denotes the joint normal distribution of x and $(y + t)$ and $f(x)$ the marginal distribution of x.

The positive hit rates for each of the 500 combinations of t, br, ϕ, and ρ_{xy} were calculated via Equation 10 using a computer program that generates joint and marginal normal distributions. To present the results of such a large amount of data, the positive hit rate was plotted as a function of ρ_{xy} for varying levels of the treatment (t). Moreover, these functions are shown for three conditions of the base rate and selection ratio: $\phi = br$, $\phi < br$, and $\phi > br$.

To illustrate, consider Figure 3. Each panel in the figure shows the positive hit rate as a function of ρ_{xy} and t for specific values of ϕ and br, under the condition that $\phi = br$. The horizontal dotted line in each figure indicates where the positive hit rate is .5. The two vertical dotted lines indicate the positive hit rates (for varying values of t) when ρ_{xy} is in the range .2 to .6. Since the results of research on the accuracy of clinical judgment strongly suggest that the validity of judgment is low to moderate, the results between the two vertical dotted lines are most likely to represent actual situations.

The most important finding is that the

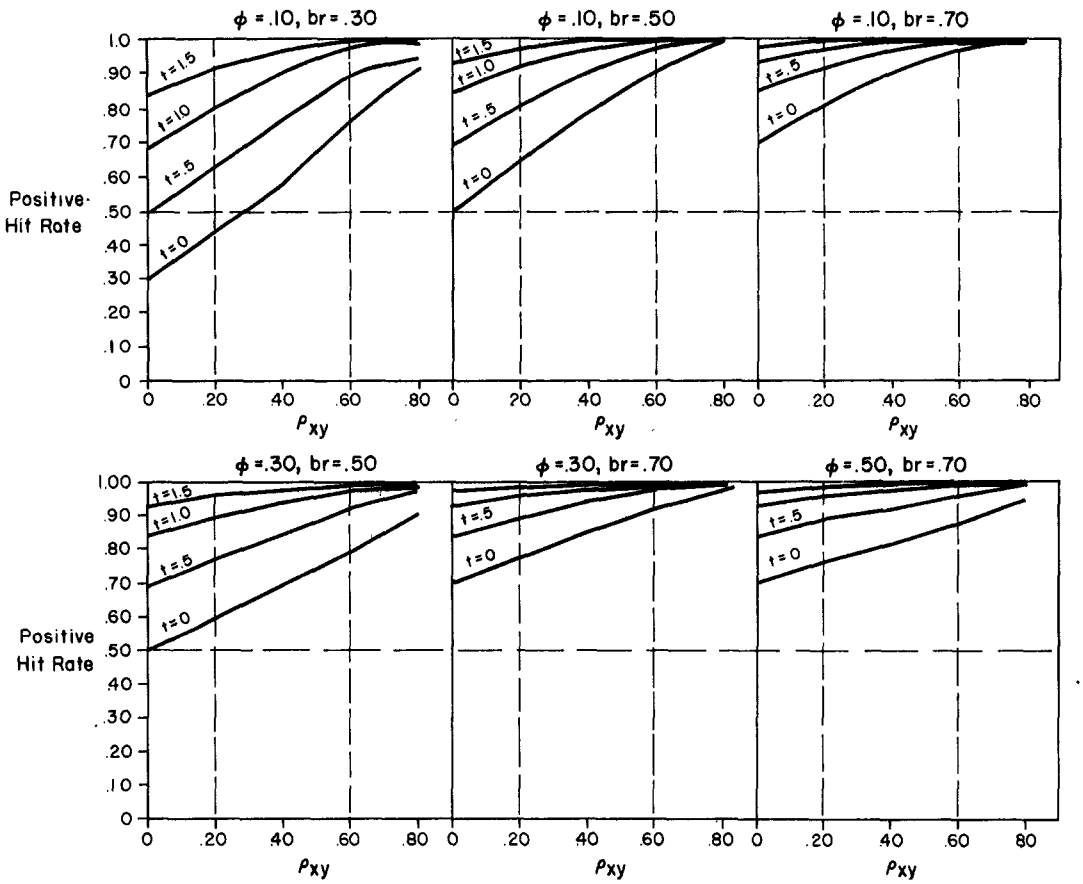


Figure 4. Positive hit rate as a function of correlation for differing treatment effects. ($\phi < br$.)

positive hit rate is generally quite high (compare the solid lines in Figure 3 to the dotted line at .5). As ϕ and br increase, all positive hit rates go up. However, of particular interest is the fact that the lines representing treatment effects seem to get compressed, that is, closer together, and to flatten out (have a lower slope) as ϕ and br increase. This result has two important implications: First, treatment effects will affect the positive hit rate most strongly when ϕ and br are low. Therefore, when one is highly selective (and the base rate is close to ϕ), treatment effects have substantial influence on the positive hit rate. Conversely, treatment effects have a smaller influence when ϕ is high. Second, the slope of any line for a given treatment effect reflects how sensitive the positive hit rate is to ρ_{xy} (steeper slopes indicate greater changes in this probability for changes in ρ_{xy}). The results show that as ϕ and br increase, judgmental ability plays less of a role in determining the positive hit rate; therefore, when $\phi = br$, positive hit rates are most affected by treatment effects *and* judgmental ability when ϕ is low.

The second set of results concerns situations where $\phi < br$. A representative sample of these cases is shown in Figure 4. The major result in Figure 4 is again the high levels of positive hit rates. Furthermore, treatment effects influence the positive hit rate most when br is low (as in Figure 3). In addition, for any given ρ_{xy} ($\rho_{xy} \neq 0$), br, and t, one can always increase the positive hit rate by decreasing ϕ (as discussed earlier). This relation can be seen graphically by comparing the positive hit rates for Figures 3 and 4 that have the same base rates, for example, the bottom panel in Figure 3 with the top panel

in Figure 4. The latter panel shows a steeper slope for any given treatment effect.

The final results involve situations where $\phi > br$. A representative sample of these cases is shown in Figure 5. When $\phi > br$, we have the condition where the positive hit rate will be lowest (holding ρ_{xy} and t constant). However, note that in the cases where the positive hit rate is low, treatment effects are quite influential. Therefore, even in the worst case, if there are treatment effects, one may still experience positive hit rates that are substantial.

Although the above results have been discussed in detail, the detail should not obscure the basic point, namely, that when treatment effects exist, many judgmental situations are so structured that even poor judgment will result in high positive hit rates.

General Discussion

The simulation results and the implications of the learning model are now discussed with respect to four issues: (a) conditions for learning from experience; (b) the nature of outcome feedback, especially environmental effects on outcomes and outcome coding; (c) determination of when confidence in judgment will be low; and (d) improving one's ability to learn from experience.

Conditions for Learning Probabilities and Clinical Inference

In discussing probability learning, Estes (1976a) states two general conditions for veridical estimation of probabilities: "(a) The alternative events involved in a situation

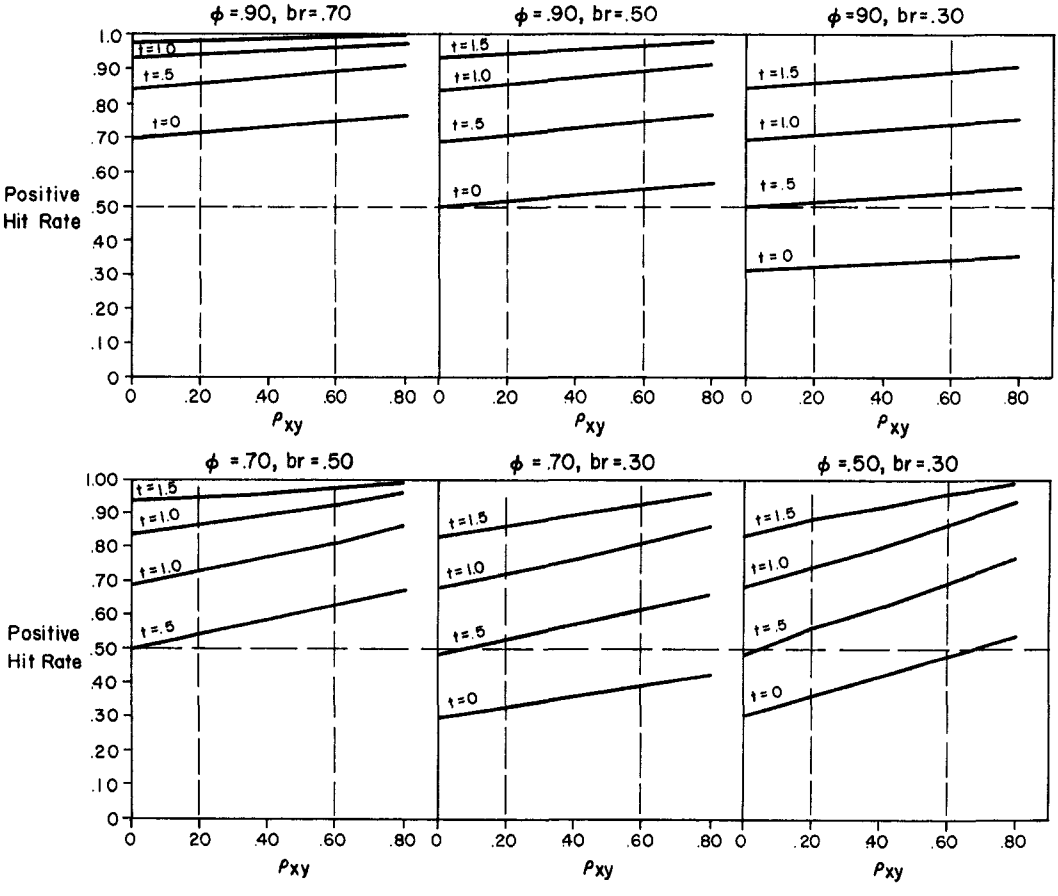


Figure 5. Positive hit rate as a function of correlation for differing treatment effects. ($\phi > br$.)

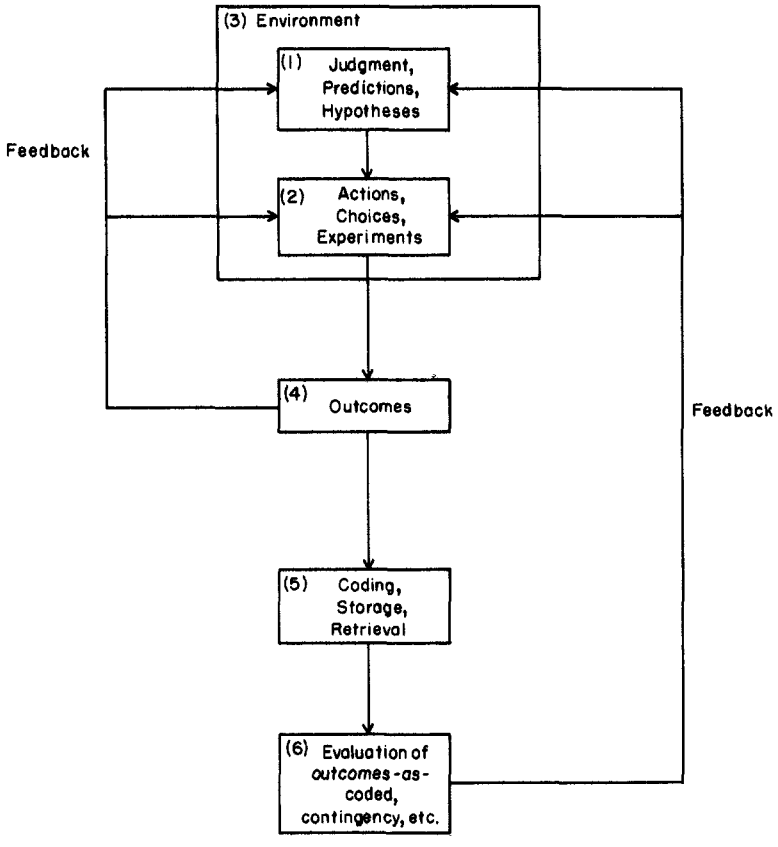


Figure 6. Schematic representation of a judgment–action situation.

must have equal opportunities of occurrence and (b) the learner must attend to and encode occurrences of all the alternative events with equal uniformity or efficiency" (p. 53). While, as Estes points out, such conditions are probably met in situations like weather forecasting (which, incidentally, may explain why weather forecasters are better calibrated than other experts; Slovic et al., 1977), they are clearly violated in many other situations. In particular, this must be the case when judgments lead to actions, since the choice of one alternative excludes the others, leading to violations of both Conditions a and b. That is, outcomes contingent on actions taken and not taken will typically have unequal opportunities of occurrence. Moreover, even when equality obtains, attention is unlikely to be focused on cases bearing disconfirming information. This latter point is, of course, supported by the considerable evidence discussed above on the failure to seek disconfirming evidence and the reliance on positive instances to make judgments of contingency. Indeed, as Golding and Rorer (1972) point out with regard to clinical psychology, a clinician who entertains the idea that Symptom X implies Disease Y is not likely to consider Y when X is absent.

In addition to the fact that Estes's (1976a) Condition a may not arise frequently in the natural environment, two reasons suggest that people will not actively seek to make it occur. First, it was pointed out that ethical/utilitarian considerations may prevent one from performing the type of experiment needed to disentangle judgment from treatment effects. However, such experiments are necessary to meet Condition a in that they allow all events the opportunity to occur. Second, the simulation results indicated high positive hit rates in many types of situations. Therefore, in the presence of such positive feedback,

there would be little motivation to seek additional and possibly disconfirming evidence about judgmental ability.

Finally, one further point should be made concerning Estes's (1976a) Condition b. Although research on the learning of contingencies indicates that people ignore relevant information, Jenkins and Ward (1965) have shown that this tendency can be reduced when subjects are given intact 2×2 tables rather than being presented with the information sequentially. One can presume that the former method focuses greater attention on the disconfirming instances, since memory for nonoccurrences is not required. However, information in the natural environment is typically acquired sequentially (cf. Lathrop, 1967). Thus, unless memory for disconfirming instances is aided, Condition b is likely to be violated.

In the area of clinical judgment, Goldberg (1968) lists three conditions for learning: (1) feedback, which is necessary but not sufficient; (2) the ability to rearrange cases so that hypotheses can be verified or discounted; and (3) the ability to tally the accuracy of one's hypotheses (keep a "box score"). If Conditions 2 and 3 were met, an important issue would clearly be to determine when feedback would be necessary and sufficient for learning. However, any attempt to answer this question must first discuss the nature of outcome feedback.

Nature of Outcome Feedback

Consider Figure 6, which shows a schematic diagram of the judgment–action situation, and concentrate initially on Boxes 1 to 4. Boxes 5 and 6 will be considered subsequently. The crucial aspect of the diagram is that judgment and actions are taken within particular task environments. Indeed, this point is so obvious that it tends to be overlooked. By "task environment" is meant such factors as base rates, selection ratios, treatment effects, uncertainty of the task, sequential versus simultaneous presentation of information, completeness of judgment–action combinations, and so on. It is the combination of judgments, actions, and environments that produces outcomes. Consequently, outcome feedback to either or both of Boxes 1 and 2 must pass through Box 3. However, if awareness of environmental variables and their effects is lacking, outcome feedback will be ineffective. In fact, much research on multiple-cue probability learning shows just that (Castellan, 1977; Slovic & Lichtenstein, 1971). Furthermore, in the absence of adequate control or understanding of environmental factors, inference regarding causal relationships between judgment–actions and outcomes is problematic.

For example, in a recent paper, Hammond (1978) has discussed six "modes of inquiry" that people use to learn about the world. These modes vary from formal experimentation to quasi-experiments to unaided judgment. An important dimension that these modes vary on is the degree to which manipulation of variables is possible. As one leaves the experimental modes,

> Inability to hold certain variables constant, and to manipulate other variables leaves the question of causal directions ambiguous. As a result, interdependent variables must be disentangled sheerly by cognitive activity, that is, by reaching a judgment about what the results of disentanglement might be . . . for the disentanglement of causal relations by (passive) cognition instead of (active) experimentation is subject to a variety of psychological factors, such as memory loss, information overload, and recency and primacy effects, to mention only a few. (Hammond, 1978, p. 16)

While Hammond's point concerning the difficulty of disentangling variables by unaided judgment is important, of equal importance is the fact that intuitive judgment frequently lacks awareness of environmental effects. Thus, when judgments or actions are evaluated by comparison with outcomes, environmental factors influencing these outcomes may not even be considered. Evidence on this point is now discussed with regard to regression effects, base rates, selection ratios, and treatment effects.

Environmental Effects on Outcomes

Regression effects occur when there is an imperfect relationship between judgments and criterion values ($\rho_{xy} \neq 1.0$). However, despite the prevalence of regression effects in the environment, people exhibit great difficulty in acquiring a proper understanding of the phenomenon (Kahneman & Tversky, 1973).

For example, when actions are given to extreme groups (as measured by some x), outcomes (y) will be regressive with respect to x. However, unless one understands the regressive nature of the environment, it is easy to incorrectly attribute outcomes to actions. Furthermore, regression effects are symmetric, that is, x is also regressive with respect to y. When actions are based on judgments, this has several nonintuitive implications (Einhorn & Schacht, 1977). It should be emphasized that inadequate understanding of regression is not restricted to nonscientists. Tversky and Kahneman (1974) point out that regression effects seem to have been ignored in the debate as to the relative importance of reward and punishment in learning, although research on the topic has spanned some 50 years!

Next, consider the effects of base rates. First, there is a growing amount of research indicating that people ignore base rate information in making probability judgments (e.g., Lyon & Slovic, 1976; Nisbett, Borgida, Crandall, & Reed, 1976; Tversky & Kahneman, 1974). According to normative statistical theory, this tendency can result in large mistakes. The implications of ignoring base rate information for the persistence of the illusion of validity are equally serious. For example, consider a person who experiences an 80% positive hit rate. Without knowledge of the base rate, this hit rate may seem to be indicative of accurate judgment. However, if the base rate were 70%, the 80% hit rate would not look as impressive. Therefore, accuracy of judgment should be evaluated as the marginal increase in the hit rate over the base rate. If people do not use the marginal hit rate to evaluate their judgment, they are likely to overestimate their judgmental ability. An extreme example might be one in which the base rate is .75, judgmental ability and treatment effects are both zero, yet the positive hit rate is .75 (equal to the base rate).

Even if the base rate is not ignored, the positive hit rate is greatly affected by the relationship between the base rate and selection ratio (for any given correlation and treatment effect). The simulation results indicate that the positive hit rate will be highest when the selection ratio is less than the base rate. It is, therefore, instructive to consider the kinds of situations in which this condition holds. First, consider situations involving budgetary or physical constraints, for example, limited access to expensive medical treatment, research grants, or admission to certain schools or jobs. Here the cutoff is set at a point where the judge tries to choose the most deserving cases. Many other deserving cases are not judged greater than x_o, since the cutoff is a function of resources rather than a judgment of deservedness in some absolute sense. In these kinds of situations, which must be frequent, positive feedback should occur even if judgmental ability is low.

Second, there will be many situations where the real selection ratio is in fact smaller than might appear to the judge. Specifically, consider what happens when there are systematic self-selection biases, so that persons who feel they have a good chance of being accepted are more liable to subject themselves to evaluative judgment. In these cases, the population of judgments is skewed so that if the judge decides to fix, say, a 10% selection ratio, he or she is in fact operating with a much smaller selection ratio.[4]

Third, there are many instances where the cost of observable errors resulting from the judge's decision is high; for example, personnel officers can be penalized for engaging the wrong person for their organizations but are rarely questioned about false negatives, since there are no follow-ups of rejected candidates. In such circumstances, it is clear that control is exercised over observable errors by maintaining a low selection ratio. Furthermore, results of the simulation (see Figure 4) show that when judgmental ability is low, high selectivity (low selection ratio) not only increases the positive hit rate for a given level of treatment effect but also increases the sensitivity of the positive hit rate to treatment effects. Therefore, for several reasons, high selectivity is likely to result in large positive hit rates.

When selection ratios are greater than base rates, positive hit rates are lowest. However,

[4] Self-selection should cause the distributions of x and y to be skewed. However, little is known about the robustness of normality assumptions to self-selection effects.

examination of Figure 5 shows that the positive hit rate is quite sensitive to treatment effects in these situations. One can speculate as to the types of situations that are likely to fall into this category. Such situations may be those where a false negative is costly, that is, one does not want to reject a deserving person. Remedial action situations probably fall into this class. If this is the case, treatment effects are likely to exist, thereby causing increases in the positive hit rates. Of course, if there are no treatment effects, large selection ratios relative to base rates will result in low positive hit rates (unless the correlation is extremely high). Such situations may exist in government programs to help the needy. Under the condition that the selection ratio is greater than the base rate, larger numbers of false positives (e.g., welfare cheaters) will exist (cf. Einhorn, 1978). However, actual judgmental ability involved in determining deservedness may be just as valid as when judgment is exercised in situations where the selection ratio is smaller than the base rate. Unfortunately, those in the latter position cannot understand why the government does not run programs more effectively, since their own experience tells them that it is possible to make small numbers of mistakes. Of course, they are rarely aware that the task structure is causing the difference and that their good results may occur in spite of, rather than because of, their own ability.

Finally, consider treatment effects due to actions. It is important to note that the magnitude of treatment effects will be determined to some degree by the nature of the task. For example, if one judges that rain is likely and then bases action on that judgment by carrying an umbrella, it seems absurd to consider that carrying the umbrella can have any effect on the occurrence of rain ($t = 0$). However, in other situations, treatment effects due to actions can be substantial without awareness of their influence. The most compelling evidence of this occurs in medicine and is commonly known as the "placebo effect" (Shapiro, 1960). However, the discovery that any action, no matter how worthless from a pharmacological point of view, can improve patients was very slow in developing. In fact, the invention of placebo control groups is a twentieth-century idea. In his fascinating history of the placebo effect, Shapiro (1960) relates a story that illustrates why it took so long to understand the phenomenon:

> In 1794, Dr. Ranieri Gerbi, a professor at Pisa, published a manuscript describing a miraculous cure for toothache due to any cause which lasted for a whole year. A worm species, called curculio antiodontaligious, was crushed between the thumb and forefingers of the right hand. The fingers then touched the affected parts. An investigatory commission found that 431 of 629 toothaches were stopped immediately. (p. 112)

Although unintended treatment effects due to actions are well known in medical (and psychiatric) science, such effects are also known in psychology. Most notable is the Hawthorne effect. However, other evidence compiled by Rosenthal (1966) and Rosenthal and Jacobson (1968) is of relevance. In the former, experimenter effects were documented, that is, experimenters holding a particular hypothesis have a greater chance of confirming their hypotheses than those holding contrary positions. This clearly points to the difficulty of separating judgments from actions in interpreting outcomes. In the latter (and controversial) study, evidence on the nature of self-fulfilling prophecy is clearly consonant with treatment effects that result from actions based on judgment.

Although treatment effects due to actions have been documented, the extent and magnitude of such effects are difficult to ascertain, since the kinds of experiments needed to study them are difficult and often impossible to perform. However, they are probably large, otherwise regression effects would surely teach people to have less confidence in judgment than they do. For example, if something is judged to be three standard deviations above the mean ($x = 3.0$), and judgmental ability is moderate (e.g., $\rho_{xy} = .5$), the best estimate of performance is $y = 1.5$, which is considerably less than the original judgment. Moreover, the probability of performance being as high as (or higher than) the corresponding judgment is only .04.[5] Kahneman and Tversky (1973)

---

[5] The probability of any y being greater than or equal to x is given by

$$p(y \geq x \mid x, \rho_{xy}) = 1 - F[(x - \rho_{xy}x)/(1 - \rho_{xy}^2)^{\frac{1}{2}}],$$

where F is the cumulative normal distribution.

have nonetheless found that people have great confidence in extreme judgments, even though these should be the most regressive. If treatment effects are present, however, they will tend to cancel the regression effects and again lead one to judge predictions as being quite accurate. Treatment effects clearly play an important role in maintaining the illusion of validity.

Finally, the concept of treatment effects provides a possible explanation for why people perform poorly in certain laboratory studies. When treatment effects exist, outcomes are not conditionally independent of judgment, that is,

$$p(y|x, t) \neq p(y|x).$$

However, consider the effects of introducing such conditional nonindependence into the normative models that have been used as standards for evaluating judgment: "Conservatism" effects are negligible with conditionally dependent data (Winkler & Murphy, 1973); in a regression context, the presence of treatment effects will mean that the performance (y) of someone who is judged high on potential (x) cannot be expected to regress in the manner the normative model would indicate. Furthermore, if there is lack of independence, sampling variability will be considerably reduced, so that small samples can be on occasion highly representative of populations (Tversky & Kahneman, 1974).

The point being made here is not that the systematic examination of people's limited judgmental ability has been misguided. Rather, it is emphasized that the kinds of failings uncovered are consistent with the experience of processes generating data that are not conditionally independent; treatment effects are one important form of nonindependence. The intended contribution is to suggest why experience with one's own judgmental ability does not concur with the documented lack of ability. People simply do not live in a world characterized by conditionally independent data (Brunswik, 1943, 1952; Winkler & Murphy, 1973).

Coding, Memory, and Feedback

The discussion to this point has focused on outcome feedback. However, it has been suggested that process feedback (i.e., feedback that explicates the relationships between cues in the environment and the event to be predicted) should induce more effective learning. Empirical evidence, on the other hand, indicates that this is not necessarily the case (for a review, see Castellan, 1977). Although process feedback has been found to be superior to outcome feedback on occasion (see, e.g., Hammond, Summers, & Deane, 1973), insufficient emphasis has been given to how outcomes are coded, stored, and retrieved in memory and how such transformed outcomes are evaluated. Consider Figure 6 again. Note that outcomes are first coded in memory and then evaluated. This evaluation is then fed back to Boxes 1 and 2 through Box 3. Thus, to understand the effects of feedback on learning, one must know something about the links between Boxes 4, 5, and 6.

To illustrate, consider that outcomes are coded as frequencies rather than probabilities and that evaluation of judgment–outcome contingency is based on the frequency of positive hits. Such an encoding process will inevitably lead to an "availability" bias (Tversky & Kahneman, 1973) in recall and corresponding "illusory correlation" (Chapman & Chapman, 1969). However, from the perspective of this article, the more important issue is that attempts to overcome this bias through feedback are not likely to succeed unless the feedback is designed to focus attention on disconfirming cases (Golding & Rorer, 1972). Therefore, presenting people with feedback, where, for example, a symptom and disease classification are uncorrelated in a statistical sense, is unlikely to alleviate illusory correlation for the simple reason that people do not make contingency judgments on the basis of statistical theory. Finally, although it is necessary to understand the links between Boxes 4, 5, and 6, it should be stressed that the whole process (Boxes 1–6) is involved in understanding feedback and why the illusion of validity persists.

When Confidence Will Be Low

Since the results of the simulation experiment show that the positive hit rate will be greater than .5 in many situations, it might

be thought that substantial confidence in judgment is an inevitable result of engaging in predictive behavior. However, there is evidence to the contrary. For example, Oskamp (1962) reports an experiment where inexperienced judges expressed more confidence in judgment than trained clinicians. It is therefore important to consider when confidence is likely to be low. The learning model discussed previously provides a way for discerning the relevant conditions.

For convenience, Equation 6 is rewritten here to aid the discussion:

$$F = N\phi(ph - \beta_2).$$

Note that when $ph = \beta_2, F = 0$ and experience (N) should have no effect on confidence. Under what conditions is this likely to happen? Assuming that $\beta_2 < \beta_1$, ph will have to be small, which according to the simulation results will occur with low ρ_{xy}, $br < \phi$, and $t \simeq 0$. Such a combination of conditions suggests remedial actions where there is much uncertainty in the environment (hence, low ρ_{xy}) and negligible treatment effects. Furthermore, under these conditions, it might be the case that $ph < \beta_2$ if the cost of false positives is high, for example, giving expensive therapy that does not work. In these circumstances, $F < 0$, and greater experience should lead to decreasing amounts of confidence. Therefore, note that the structure of the task again plays an important role as to the likely effects of experience on confidence.

Two further comments are necessary. First, the direction of treatment effects due to actions is not known a priori. Boomerang effects can occur. For example, imagine the case of a person accepted for a fellowship who, because of the financial security provided by the award, works less hard and so performs less effectively. Such boomerang effects clearly lower the positive hit rate and influence the total feedback effect. However, although boomerang effects can be expected to decrease confidence in judgment, this issue is by no means settled. Consider the case of Benjamin Rush, a highly respected physician and professor at the first medical school in America. Believing in the theory that febrile illnesses resulted from an excess stimulation and excitement of the blood, he advocated and practiced bloodletting as a cure. When he fell ill with yellow fever, he instructed that he be bled plentifully. As reported by Eisenberg (1977),

> From illness and treatment combined, he almost died; his convalescence was prolonged. That he did recover persuaded him that his methods were correct. Neither dedication so great that he risked his life to minister to others, nor willingness to treat himself as he treated others, nor yet the best education to be had in his day was sufficient to prevent Rush from committing grievous harm in the name of good. Convinced of the correctness of his theory of medicine and lacking a means for the systematic study of treatment outcome, he attributed each new instance of improvement to the efficacy of his treatment and each new death that occurred despite it to the severity of the disease. (p. 1106)

Finally, in some situations, outcomes contingent on the action not taken may be so salient that they cannot be ignored. For example, Hirsch (1969, 1972) has discussed the perception of executives in the pop record business that predicting successful records is extremely difficult even though records that are initially selected as promising are given elaborate and expensive publicity (a large treatment effect). A major reason for the lack of confidence in prediction in this case is that records that were rejected initially can (and do) become popular successes. This false negative error cannot be ignored, and indeed, attention is focused on why such records were not originally chosen. Therefore, under these kinds of conditions, confidence in prediction is more problematic.

Improving One's Ability to Learn From Experience

The difficulty of learning from experience has been traced to three main factors: (a) lack of search for and use of disconfirming evidence, (b) lack of awareness of environmental effects on outcomes, and (c) the use of unaided memory for coding, storing, and retrieving outcome information. What, if anything, can be done to alleviate these problems? With regard to the use of disconfirming evidence, formal training in experimental design, teaching the logic of control groups and baseline predictions, and so on would seem to be a necessary but not sufficient condition. If sophisticated subjects, who are trained in these matters, make similar mistakes

to those without training, the prospects for overcoming such tendencies is certainly disheartening. One must hope that future research will be aimed at providing methods for overcoming the cognitive difficulties associated with disconfirming information.

In order to gain awareness of the environmental effects on outcomes, the use of a model of the environment, as advocated by Hammond and his colleagues (Hammond, 1971; Hammond, Mumpower, & Smith, 1977), has much to recommend it. Such models draw attention to the structure of the environment and the manner in which structure affects outcomes. Furthermore, the use of process rather than outcome feedback should also focus attention on environmental variables. However, it should be emphasized that process feedback, in the absence of an understanding of the task structure, may not be effective.

Finally, is there any way of improving one's memory for outcomes? Since the time between outcomes and judgments may be large, memory can be considerably aided by simply keeping a box score (Goldberg, 1968). This also has the advantage of keeping a record of disconfirming instances. Moreover, memory can be aided so that it is not only categoric. Tversky and Kahneman (1974) have suggested that people should attempt to encode events not by their substantive content but by judged probability. When events are grouped in this manner, it is possible to keep a tally of the extent to which judged probabilities match subsequent empirical relative frequencies. Otherwise, as Tversky and Kahneman (1974) point out, dichotomous feedback indicating whether an event did or did not occur provides inadequate feedback concerning one's ability to make probabilistic judgments. In a recent article, Kahneman and Tversky (in press) make further suggestions to overcome judgmental biases in predictive activity.

Conclusion

This investigation began by asking why a substantial discrepancy exists between the findings of research on judgmental fallibility and people's confidence in their own judgment. In trying to answer this question, the focus has been on the structure of judgmental tasks as it affects outcomes and the manner in which outcomes are interpreted and used. It has been shown that good outcomes are quite likely even when judgmental ability is low. Furthermore, the learning model helps to explain how the concept, "my judgment is accurate," is both learned and maintained even though judgment may be invalid.

The results and indeed the whole paper pose an interesting paradox. If, as Rogers (1961) believes, experience in the form of "self-discovered" or "self-appropriated learning" is the only form of learning that significantly affects behavior, then perhaps this and other studies are exercises in futility in that they provide insufficient stimulus for people to question their judgmental ability. It is, of course, hoped that this is not the case, since the issues are too important. Just how important can be gleaned from the following question (adapted from Hammond, 1978): If we believe we can learn from experience, is it possible to learn that we can't?

References

Björkman, M. Predictive behavior, some aspects based on an ecological orientation. *Scandinavian Journal of Psychology*, 1966, *7*, 43–57.

Boring, E. G. The nature and history of experimental control. *American Journal of Psychology*, 1954, *67*, 573–589.

Brehmer, B. Social judgment theory and the analysis of interpersonal conflict. *Psychological Bulletin*, 1976, *83*, 985–1003.

Brunswik, E. Organismic achievement and environmental probability. *Psychological Review*, 1943, *50*, 255–272.

Brunswik, E. *Conceptual framework of psychology.* Chicago: University of Chicago Press, 1952.

Castellan, N. J., Jr. Decision making with multiple probabilistic cues. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2). Hillsdale, N.J.: Erlbaum, 1977.

Chapman, L. J., & Chapman, J. P. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 1969, *74*, 271–280.

Cohen, J. *Psychological probability: Or the art of doubt.* London, England: George Allen & Unwin, 1972.

Cronbach, L. J. Beyond the two disciplines of scientific psychology. *American Psychologist*, 1975, *30*, 116–127.

David, F. N. Studies in the history of probability and statistics. I. Dicing and gaming (A note on the history of probability). *Biometrika*, 1955, *42*, 1–15.

David, F. N. *Games, gods and gambling.* London, England: Charles Griffin, 1962.

Dawes, R. M. Shallow psychology. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior.* Hillsdale, N. J.: Erlbaum, 1976.

Edwards, W. Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

Edwards, W. Bayesian and regression models of human information processing—A myopic perspective. *Organizational Behavior and Human Performance,* 1971, *6,* 639–648.

Einhorn, H. J. Expert measurement and mechanical combination. *Organizational Behavior and Human Performance,* 1972, *7,* 86–106.

Einhorn, H. J. Decision errors and fallible judgment: Implications for social policy. In K. R. Hammond (Ed.), *Judgment and decision in public policy formation.* Denver, Colo.: Westview Press, 1978.

Einhorn, H. J., & Schacht, S. Decisions based on fallible clinical judgment. In M. Kaplan & S. Schwartz (Eds.), *Judgment and decision processes in applied settings.* New York: Academic Press, 1977.

Eisenberg, L. The social imperatives of medical research. *Science,* 1977, *198,* 1105–1110.

Elstein, A. S. Clinical judgment: Psychological research and medical practice. *Science,* 1976, *194,* 696–700.

Estes, W. K. Research and theory in the learning of probabilities. *Journal of the American Statistical Association,* 1972, *67,* 81–102.

Estes, W. K. The cognitive side of probability learning. *Psychological Review,* 1976, *83,* 37–64. (a)

Estes, W. K. Some functions of memory in probability learning and choice behavior. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 10). New York: Academic Press, 1976. (b)

Fischhoff, B. Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance,* 1975, *1,* 288–299.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance,* 1977, *3,* 552–564.

Goldberg, L. R. Simple models or simple processes? Some research on clinical judgments. *American Psychologist,* 1968, *23,* 483–496.

Golding, S. L., & Rorer, L. G. Illusory correlation and subjective judgment. *Journal of Abnormal Psychology,* 1972, *80,* 249–260.

Guilford, J. P. *Psychometric methods* (2nd ed.). New York: McGraw-Hill, 1954.

Hacking, I. *The emergence of probability.* New York: Cambridge University Press, 1975.

Hammond, K. R. New directions in research in conflict resolution. *Journal of Social Issues,* 1965, *21,* 44–66.

Hammond, K. R. Computer graphics as an aid to learning. *Science,* 1971, *172,* 903–908.

Hammond, K. R. Toward increasing competence of thought in public policy formation. In K. R. Hammond (Ed.), *Judgment and decision in public policy formation.* Denver, Colo.: Westview Press, 1978.

Hammond, K. R., & Brehmer, B. Quasi-rationality and distrust: Implications for international conflict. In L. Rappoport & D. A. Summers (Eds.), *Human judgment and social interaction.* New York: Holt, Rinehart & Winston, 1973.

Hammond, K. R., Mumpower, J. L., & Smith, T. H. Linking environmental models with models of human judgment: A symmetrical decision aid. *IEEE Transactions on Systems, Man, and Cybernetics,* 1977, *SMC-7*(5), 358–367.

Hammond, K. R., Summers, D. A., & Deane, D. H. Negative effects of outcome-feedback on multiple-cue probability learning. *Organizational Behavior and Human Performance,* 1973, *9,* 30–34.

Hirsch, P. *The structure of the popular music industry.* Ann Arbor: University of Michigan, Institute for Social Research, 1969.

Hirsch, P. Processing fads and fashions: An organization-set analysis of cultural industry systems. *American Journal of Sociology,* 1972, *77,* 639–659.

Hogarth, R. M. Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association,* 1975, *70,* 271–289.

Jenkins, H. M., & Ward, W. C. Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied,* 1965, *79*(1, Whole No. 594).

Johnson, D. M. *A systematic introduction to the psychology of thinking.* New York: Harper & Row, 1972.

Jones, M. R. From probability learning to sequential processing: A critical review. *Psychological Bulletin,* 1971, *76,* 153–185.

Kahneman, D., & Tversky, A. On the psychology of prediction. *Psychological Review,* 1973, *80,* 251–273.

Kahneman, D., & Tversky, A. Intuitive prediction: Biases and corrective procedures. *Management Science,* in press.

Kendall, M. G. Studies in the history of probability and statistics. II. The beginning of a probability calculus. *Biometrika,* 1956, *43,* 1–14.

La Rochefoucauld, F. *The maxims of La Rochefoucauld* (L. Kronenberger, Trans.). New York: Random House, 1959.

Lathrop, R. G. Perceived variability. *Journal of Experimental Psychology,* 1967, *73,* 498–502.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art. In H. Jungermann & G. de Zeeuw (Eds.), *Decision making and change in human affairs.* Dordrecht, The Netherlands: Reidel, 1977.

Lindley, D. V. *Introduction to probability and statistics from a Bayesian viewpoint. Part I. Probability.* Cambridge, England: Cambridge University Press, 1965.

Lyon, D., & Slovic, P. Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica,* 1976, *40,* 287–298.

Meehl, P. E. *Clinical versus statistical prediction.* Minneapolis: University of Minnesota Press, 1954.

Mill, J. S. *A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence, and the methods of scientific investigation.* London, England: J. W. Parker, 1851.

Murphy, A. H., & Winkler, R. L. The use of credible intervals in temperature forecasting: Some experimental results. In H. Jungermann & G. de Zeeuw

(Eds.), *Decision making and change in human affairs*. Dordrecht, The Netherlands: Reidel, 1977.

Myers, J. L. Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Volume 3). Hillsdale, N.J.: Erlbaum, 1976.

Nisbett, R. E., Borgida, E., Crandall, R., & Reed, H. Popular induction: Information is not necessarily informative. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior*. Hillsdale, N.J.: Erlbaum, 1976.

Oskamp, S. The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs: General and Applied*, 1962, *76* (28, Whole No. 547).

Oskamp, S. Overconfidence in case-study judgments. *Journal of Consulting Psychology*, 1965, *29*, 261–265.

Piaget, J., & Inhelder, B. *La genèse de l'idée de hasard chez l'enfant*. Paris, France: Presses Universitaires de France, 1951.

Popper, K. R. *The logic of scientific discovery*. London, England: Hutchinson, 1959.

Rogers, C. R. *On becoming a person*. Boston: Houghton Mifflin, 1961.

Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts, 1966.

Rosenthal, R., & Jacobson, L. *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston, 1968.

Ryback, D. Confidence and accuracy as a function of experience in judgment-making in the absence of systematic feedback. *Perceptual and Motor Skills*, 1967, *24*, 331–334.

Sawyer, J. Measurement *and* prediction: Clinical *and* statistical. *Psychological Bulletin*, 1966, *66*, 178–200.

Shapiro, A. K. A contribution to a history of the placebo effect. *Behavioral Science*, 1960, *5*, 109–135.

Simon, H. A., & Newell, A. Human problem solving: The state of the theory in 1970. *American Psychologist*, 1971, *26*, 145–159.

Skinner, B. F. *Science and human behavior*. New York: Macmillan, 1953.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 1971, *6*, 649–744.

Slovic, P., Fischhoff, B., & Lichtenstein, S. Behavioral decision theory. *Annual Review of Psychology*, 1977, *28*, 1–39.

Smedslund, J. The concept of correlation in adults. *Scandinavian Journal of Psychology*, 1963, *4*, 165–173.

Smedslund, J. Note on learning, contingency, and clinical experience. *Scandinavian Journal of Psychology*, 1966, *7*, 265–266.

Taylor, H. C., & Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, 1939, *23*, 565–578.

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

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

Ward, W. C., & Jenkins, H. M. The display of information and the judgment of contingency. *Canadian Journal of Psychology*, 1965, *19*, 231–241.

Wason, P. C. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 1960, *12*, 129–140.

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

Wason, P. C. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 1968, *20*, 273–281.

Wason, P. C. Regression in reasoning? *British Journal of Psychology*, 1969, *60*, 471–480.

Wason, P. C., & Johnson-Laird, P. N. *Psychology of reasoning. Structure and content*. London, England: Batsford, 1972.

Winkler, R. L., & Murphy, A. H. Experiments in the laboratory and the real world. *Organizational Behavior and Human Performance*, 1973, *10*, 252–270.

Received September 2, 1977
Revision received March 6, 1978 ■