ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 6, 649–744 (1971)

Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment[1]

PAUL SLOVIC AND SARAH LICHTENSTEIN
Oregon Research Institute

In recent years there have been several hundred studies within the rather narrowly-defined topic of information utilization in judgment and decision making. Much of this work has been accomplished within two basic schools of research, which we have labeled the "regression" and the "Bayesian" approaches. Each has its characteristic tasks and characteristic information that must be processed to accomplish these tasks. For the most part, researchers have tended to work strictly within a single approach and there has been minimal communication between the resultant subgroups of workers. Our objective here is to present a review and comparative analysis of these two approaches. Within each, we examine (a) the models that have been developed for describing and prescribing the use of information in decision making; (b) the major experimental paradigms, including the types of judgment, prediction, and decision tasks and the kinds of information that have been available to the decision maker in these tasks; (c) the key independent variables that have been manipulated in experimental studies; and (d) the major empirical results and conclusions. In comparing these approaches, we seek the answers to two basic questions. First, do the specific models and methods characteristic of different paradigms direct the researcher's attention to certain problems and cause him to neglect others that may be equally important? Second, can a researcher studying a particular substantive problem increase his understanding by employing diverse models and diverse experimental methods?

[1] This article will appear as a chapter in L. Rappoport and D. A. Summers (Eds.), *Human judgment and social interaction.* New York: Holt, Rinehart & Winston, in press. Sponsorship for this work comes from the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-68-C-0431, Contract Authority Ident. No. NR 153-311, and from Grants MH-15414 and MH-12972 from the United States Public Health Service. We are indebted to many individuals for their comments on an early draft of this article. In particular, we would like to thank Norman Anderson, Mats Björkman, Berndt Brehmer, Robyn Dawes, Lewis Goldberg, Kenneth Hammond, Leon Rappoport, James Shanteau, David Summers, and Amos Tversky for their careful and critical reading of the manuscript. Thanks are also due William Chaplin for his help in conceptualizing the relationships among the various paradigms.

ORGANIZATIONAL OUTLINE

INTRODUCTION

Our concern in this paper is with human judgment and decision making and, in particular, with the processing of information that

precedes and determines these activities. The distinction between judgments and decisions is a tenuous one and will not be maintained here; we shall use these terms interchangeably.

Regardless of terminology, one thing is certain: Judgment is a fundamental cognitive activity that vitally affects the well-being, or more accurately, the survival of us all. Decisions are frighteningly more important and more difficult than ever before. Ancient man's most important decisions concerned his personal survival, and only a limited number of alternatives were available to him. Technological innovation has placed modern man in a situation where his decisions now control the fate of large population masses, sometimes the whole earth, and his sights are now set on outer space. Even the personal decisions that direct an individual's daily life have become increasingly complicated. To cite but one example, consider the bewildering array of career choices that confront today's bright youth. And consider the extreme commitment of time, effort, and money necessary to obtain the specialized training most of these opportunities require. The result is a high-gain–high-risk decision, much more difficult and complex than that faced by his parents.

The difficulties attendant to decision making are usually blamed on the inadequacy of the available information, and, therefore, our technological expertise has been mobilized to remedy this problem. Devices proliferate to supply the professional decision maker with an abundance of data. The physician, for example, has access to sophisticated electronic sensors, and satellites now relay masses of strategic data for military intelligence. However, the problem of interpreting and integrating this information has received surprisingly little attention. At this point, the decision maker is typically left to his own devices. More likely than not he will proceed, as will the physician, businessman, or military commander, in much the same manner that has been relied upon since antiquity, and when you ask him what distinguishes a good judge from a poor one he will reply,

"It's a kind of locked in concentration, an intuition, a feel, nothing that can be schooled" (Smith, 1968, p. 20).

However, things have begun to change. Specialists from many disciplines have started to focus on the integration process itself. Their efforts center around two broad questions, "What is the decision maker doing with the information available to him?" and "What should he be doing with it?" The first is a psychological problem, that of understanding how man uses information. The second problem is a more practical one and involves the attempt to make decision making more effective and efficient.

The Focus of This Paper

Information processing occurs at several levels. Our concern here is not with events at the neural level but rather with cognitive operations performed on such grosser phenomena as symbols, signs, and facts. We shall focus on the processes and strategies that humans employ in order to integrate these discrete items of information into a decision. These are the deliberative processes commonly referred to by the terms "integrating," "weighing," "balancing," "trading off," or "combining" information.

Prior to 1960 there was relatively little research on information processing at this molar, judgmental level. However, the intellectual groundwork had already been laid by studies such as Brunswik's pioneering investigations of inference in uncertain environments (Brunswik, 1956; Hammond, 1955); the work on "probability learning" (Estes, 1959); investigations of gambling decisions, utility, and subjective probability (Edwards, 1954); Miller's (1956) elaboration of the limitations on the number of conceptual items that can be processed at one time; the concept formation studies by Bruner, Goodnow, and Austin (1956); and the research on computer simulation of thought by Newell, Shaw, and Simon (1958).

Since 1960, this early work has been supplemented by several hundred studies within the rather narrowly-defined topic of information utilization in judgment and decision making. The yearly volume of studies has been increasing exponentially, stimulated by a growing awareness of the significance of the problems and the aid of the ubiquitous computer. The importance of the latter cannot be overestimated. When Smedslund (1955) published the first multiple-cue probability learning study, he bemoaned having to compute 3200 correlations on a desk calculator. It's not surprising that the next study of its kind was not forthcoming for 5 more years.

Much of the recent work has been accomplished within two basic schools of research. We have chosen to call these the "regression" and the "Bayesian" approaches. Each has its characteristic tasks and characteristic information that must be processed to accomplish these tasks. For the most part, researchers have tended to work strictly within a single approach and there has been minimal communication between the resultant subgroups of workers.

Our objective in this chapter is to present a review and comparative analysis of these two approaches. Within each, we shall examine (a) the models that have been developed for describing and prescribing the use of information in decision making; (b) the major experimental

paradigms, including the types of judgment, prediction, and decision tasks and the kinds of information that have been available to the decision maker in these tasks; (c) the key independent variables that have been manipulated in experimental studies; and (d) the major empirical results and conclusions.

In comparing these approaches, we seek the answers to two basic questions. First, do the specific models and methods characteristic of different paradigms direct the researcher's attention to certain problems and cause him to neglect others that may be equally important? Second, can a researcher studying a particular substantive problem, such as the use people make of inconsistent or conflicting information, increase his understanding by employing diverse models and experimental methods?

Areas of Omission

Space limitations have forced us to omit several other paradigms that have made significant contributions to the study of human judgment. One of these is the process-tracing approach described by Hayes (1968) and exemplified by the work of Kleinmuntz (1968) and Clarkson (1962). Researchers following this approach attempt to build sequential branching models of the decision maker based upon his verbalizations as he works through actual decision problems. Yet another important approach to the study of judgment uses multidimensional scaling procedures to infer the cognitive structure of the judge. For coverage of this work the reader is referred to the chapter by Wiggins (in press). There have been several attempts to apply information theory to the study of human judgment. One of the most notable recent efforts along these lines is the work of Bieri, Atkins, Briar, Leaman, Miller, and Tripodi (1966) which examines the transmission of information in social judgment along the lines of Miller's (1956) well-known paradigm. Another area we shall omit here is that of probability learning. A recent and thorough review emphasizing the information processing implications of this work is presented by Jones (in press). Finally, we have not attempted to review signal detection theory, an approach that has produced a great deal of research concerning the integration of sensory information into decisions. The reader is referred to books by Swets (1964) and Green and Swets (1966) for detailed coverage of this area.

THE REGRESSION APPROACH

The regression approach is so named because of its characteristic use of multiple regression, and its close relative, analysis of variance (ANOVA), to study the use of information by a judge. Within this

broad approach we shall distinguish two different paradigms which we have labeled the "correlational" paradigm and the "ANOVA" paradigm.

The Correlational Paradigm

In the correlational paradigm, a judge's integration of information is described by means of correlational statistics. The basic approach requires the judge to make quantitative evaluations of a number of stimuli, each of which is defined by one or more quantified cue dimensions or characteristics. For example, a judge might be asked to predict the grade point average for each of a group of college students on the basis of high school grades and aptitude test scores. Sarbin and Bailey (1966) elaborate the aims of the correlational analyst in a study such as this:

> He correlates the information cues available to the inferring person with the judgments or inferences . . . What usually results is that the coefficients of correlation between cues and judgment make public the subtle, and often unreportable, inferential activities of the inferring person. That is, the coefficients reveal the relative degrees that the judgments depend on the various sources of information available to the judge (pp. 193–194).

The development of the correlational paradigm has followed two streams. One stream has focused on the judge; its goal is to describe the judge's idiosyncratic method of combining and weighting information by developing mathematical equations representative of his combinatorial processes (Hoffman, 1960).

The other stream developed out of the work of Egon Brunswik, whose philosophy of "probabilistic functionalism" led him to study the organism's successes and failures in an uncertain world. Brunswik's main emphasis was not on the organism itself, but on the adaptive interrelationship between the organism and its environment. Thus, in addition to studying the degree to which a judge used cues, he analyzed the manner in which the judge learned the characteristics of his environment. He developed the "lens model" to represent the probabilistic interrelations between organismic and environmental components of the judgment situation (Brunswik, 1952, 1956).

Because of his concern about the environmental determinants of judgment, Brunswik was also the foremost advocate of what he called "representative design." The essence of this principle is that the organism should be studied in realistic settings, in experiments that are representative of its usual ecology. The lens model provides a means for appropriately specifying the structure of the situational variables in such an experiment.

The lens model. The lens model has proved to be an extremely

valuable framework for conceptualizing the judgment process. Hammond (1955) described the relevance of the model for the study of clinical judgment, and recent work by Hursch, Hammond, and Hursch (1964), Tucker (1964), and Dudycha and Naylor (1966b) has detailed some important relationships among its components in terms of multiple-regression statistics. A diagrammatic outline of a recent version of the lens model (based on Dudycha & Naylor) is shown in Fig. 1. The variables X_1, X_2, . . . X_k are cues or information sources that define each stimulus object. For example, if the stimuli being evaluated are students whose grade point averages are to be predicted, the X_i can represent high school rank, aptitude scores, etc. The cue dimensions must be quantifiable, if only to the extent of a 0–1 (e.g., high *vs.* low or yes *vs.* no) coding. Each cue dimension has a specific degree of relevance to the true state of the world. This true state, also called the criterion value, is designated Y_e (e.g., the student's actual grade point average). The relevance of the ith information source in the environment is indicated by the correlation, $r_{i,e}$, across stimuli, between cue X_i and Y_e. This value, $r_{i,e}$, is called the *ecological validity* of the ith cue. The intercorrelations among cues, again across stimuli, are given by the $r_{i,j}$ values. On the subject's side, his response or judgment is Y_s (the judged grade point average), and the correlation of his judgments with the ith cue is $r_{i,s}$, also known as his *utilization coefficient* for the ith cue.

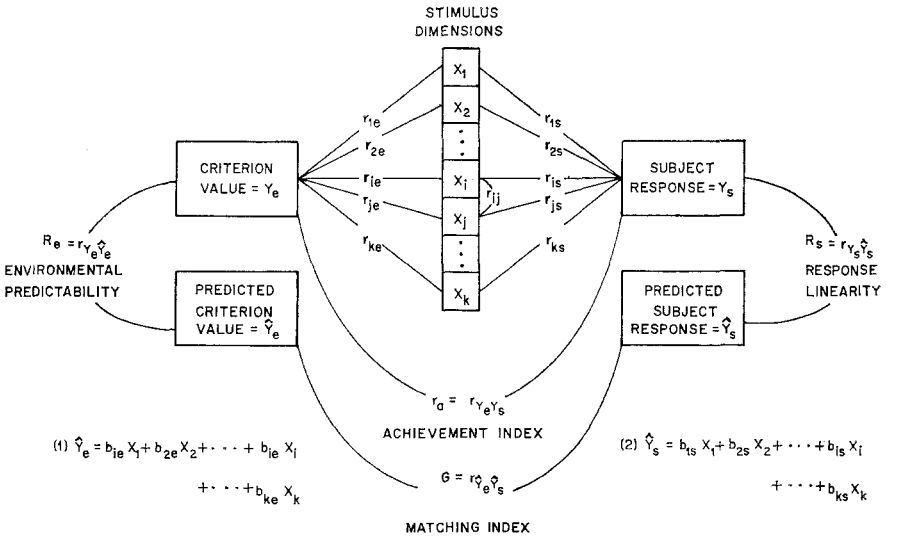

FIG. 1. Diagram of the lens model showing the relationship between cues, criterion values, and responses. (Based on Dudycha and Naylor, 1966b.)

Both the criterion and the judgment can be predicted from linear combinations of the cues as indicated by the following regression equations:

$$\hat{Y}_e = \sum_{i=1}^{k} b_{i,e} X_i, \tag{1}$$

$$\hat{Y}_s = \sum_{i=1}^{k} b_{i,s} X_i. \tag{2}$$

Equation (1) represents the prediction strategy that is optimal in the sense of minimizing the sum of squared deviations between $\hat{Y}_e$ and Y_e. The multiple correlation coefficient, $R_e = r_{Y_e \hat{Y}_e}$, indicates the degree to which the weighted combination of cues serves to predict the state of Y_e.

Equation (2) provides one possible model of the subject's decision-making strategy or *policy*. The multiple correlation coefficient, $R_s = r_{Y_s \hat{Y}_s}$, indicates how well his judgments can be predicted by a linear combination of cue values. It is also known as the subject's *response linearity*. The values of $b_{i,e}$ and $b_{i,s}$ provide measures of the importance of each cue in the environment and for the judge.

The two most important summary measures of the judge's performance are: $r_a = r_{Y_e Y_s}$, the achievement index, and $G = r_{\hat{Y}_e \hat{Y}_s}$, the matching index.

All of the above equations apply to linearly predictable relations and dependencies. The model has been further expanded by Hursch *et al.* to express nonlinear cue utilization by the introduction of the C coefficient. C is the correlation between the residual which cannot be linearly predicted in the criterion and the residual which cannot be linearly predicted in the judgment. If either of these residuals is random, C will be zero.

Tucker (1964) has shown that the indices of the lens model are related in a general equation for achievement:

$$r_a = R_e R_s G + C[(1 - R_e^2)(1 - R_s^2)]^{1/2}. \tag{3}$$

Equation (3) plays an extremely important role in many empirical studies and has come to be called the *lens model equation*. It demonstrates that achievement is a function of the statistical properties of the environment (R_e), as well as the statistical properties of the subject's response system (R_s), the extent to which the linear weightings of the two systems match one another (G), and the extent to which nonlinear variance of one system is correlated with nonlinear variance of the other (C). As Hammond (1966) has noted, the lens model permits an analysis of the relative contributions of environmental factors to a judge's achievement and thus serves as a valuable adjunct to research in the Brunswikian tradition.

Mathematical models of the judge. As we have seen, the lens model was developed to study the effects of the decision maker's environment on his performance. Because of this environmental emphasis, the focal components of the model are r_a, the achievement index, and the factors such as R_s, G, and C that determine achievement. Workers following the other stream of correlational research have had a different emphasis. They have been more interested in the judge's weighting process, his policy. In contrast with the Brunswikian tradition, they have placed less importance upon modeling the environment and, instead, have stressed the need to control the environmental situation. They tend to make the stimulus dimensions explicit and to vary their levels systematically, even though some degree of realism may be lost in the process.

A wide variety of mathematical models have been developed to capture judgmental policies. The first and most prominent of these is the *linear model* (Hoffman, 1960) which is exemplified by Eq. (2) of the lens model. Alternatively, when the judge is classifying stimuli into one of two categories, the *linear discriminant function*, rather than the multiple-regression equation, can be used to analyze the way that cues are weighted (Rodwan & Hake, 1964). In either form, the model captures the notion that the judge's predictions are a linear combination of each of the available cues. When judgment is represented by the linear model, the $b_{i,s}$ values of Eq. (2) and the utilization coefficients, $r_{i,s}$, are used to represent the relative importance given each cue. Hoffman (1960) proposed an alternative index, "relative weight," designed for this purpose. Relative weights are computed as follows:

$$RW_{i,s} = \frac{\beta_{i,s} r_{i,s}}{R_s^2}.$$

Since the sum of relative weights is 1.0, Hoffman's index describes the relative contribution of each of the predictors as a proportion of the predictable linear variance.

However, as Darlington (1968) has recently emphasized, all indices of relative weight become suspect when the factors are intercorrelated. This problem has led many judgment researchers to work with sets of stimuli in which the cues are made orthogonal to one another. One device used to insure orthogonality has been to construct stimuli by producing factorial combinations of the cues. Of course, this practice is anathema to Brunswikians, since it is the antithesis of representative design. Brunswik observed (1955; pp. 204–205) that factorial designs may produce certain combinations of values that are inherently incompatible or otherwise unrealistic and disruptive of the very process they were meant to disclose. This criticism cannot be taken lightly and some evidence does exist that judgment processes differ as a function of cue interrelationships

(Slovic, 1966; Dudycha & Naylor, 1966a). But to the researcher who is primarily interested in relative weights, rather than achievement, orthogonal designs often seem preferable to designs in which the cues are correlated in a representative fashion. Attempts are usually made, however, to mitigate potential disruptive effects by telling the judge that he will be dealing with a selected, rather than a random, sample of cases and by eliminating combinations of factors that are obviously unreal (see, for example, Hoffman, Slovic, & Rorer, 1968).

As we shall see, the linear model does a remarkably good job of predicting human judgments. However, judges' verbal introspections indicate that they believe they use cues in a variety of nonlinear ways, and researchers have attempted to capture these with more complex equations. One type of nonlinearity occurs when an individual cue relates to the judgments in a *curvilinear* manner. For example, the following quotation from a leading authority on the stock market suggests a curvilinear relation between the volume of trading on a stock and its future prospects:

> If you are driving a car you can get to your destination more quickly at 50 mph than at 10 mph. But you may wreck the car at 100 mph. In a similar way, increasing volume on an advance up to a point is bullish and decreasing volume on a rally is bearish, but in both cases only up to a point (Loeb, 1965, p. 287).

Such functions can be modeled by including exponential terms (i.e., X_i^2, X_i^3, etc.) as predictors in the judge's policy equation.

A second type of nonlinearity occurs when cues are combined in a *configural* manner. Configurality means that the judge's interpretation or weighting of an item of information varies according to the nature of other available information. An example of configural reasoning involving price changes, volume of trading, and market cycle was given by the same stock market expert:

> Outstanding strength or weakness can have precisely opposite meanings at different times in the market cycle. For example, consistent strength and volume in a particular issue, occurring after a long general decline, will usually turn out to be an extremely bullish indication. . . . On the other hand, after an extensive advance which finally spreads to issues neglected all through the bull market, belated individual strength and activity not only are likely to be shortlived but may actually suggest the end of the general recovery . . . (Loeb, 1965, p. 65).

When decision makers state that their judgments are associated with complex, sequential, and interrelated rules, it is likely that they are referring to some sort of configural process. It is important, therefore, that techniques used to describe judgment be sensitive to configurality. The C coefficient, described earlier, is unsatisfactory from a descriptive standpoint because of its lack of specificity.

The ANOVA Paradigm

One way of making the linear model sensitive to configural effects has been to incorporate cross-product terms into the policy equation of the judge. Thus, if the meaning of factor X_1 varies as a function of the level of factor X_2, the term $b_{12}X_1X_2$ can be added to the equation. When models become this complex, however, the proliferation of highly-intercorrelated terms in the equations becomes so great that estimation of the weighting coefficients is unreliable unless vast numbers of cases are available (Hoffman, 1968). For this reason, investigators such as Hoffman, Slovic, and Rorer (1968), Rorer, Hoffman, Dickman, and Slovic (1967), and Slovic (1969) have turned to the use of analysis of variance (ANOVA) to describe complex judgmental processes.

The structural model underlying ANOVA is quite similar to that of multiple regression, both being alternative formulations of a general linear model (Cohen, 1968). Although the factors that describe the cases can be either continuous or categorical, each must be partitioned into a relatively few discrete levels. In addition, the factors are typically made orthogonal to one another, although this is not a necessity. In return for these restrictions, the ANOVA technique provides a statistically efficient mechanism for detecting curvilinear and configural use of information.

When judgments are analyzed in terms of an ANOVA model, a significant main effect for cue X_1 implies that the judges' responses varied systematically with X_1 as the levels of the other cues were held constant. If sufficient levels of the factor were included in the design, and if these levels can be assigned interval scale values, the main effect may be divided into effects due to linear, quadratic, cubic, etc., trends. Similarly, a significant interaction between cues X_1 and X_2 implies that the judge was responding to particular patterns of those cues, i.e., the effect of variation of cue X_1 upon judgment differed as a function of the corresponding level taken by cue X_2.

The ANOVA model thus has potential for describing the linear, curvilinear, and configural aspects of the judgmental process. Within the framework of the model, it is possible to calculate an index of the importance of individual or patterned use of a cue, relative to the importance of other cues. One such index is based upon the degree to which the mean judgment shifts as the levels of a factor are varied (see Slovic, 1969). Another is simply a transformation of these mean effects into an estimate of the proportion of the total variation in a person's judgments that can be predicted from knowledge of the particular levels of a given cue or pattern of cues (see Hays, 1963, p. 324, or Hoffman *et al.*, 1968). The latter index includes linear and nonlinear variance and, therefore,

it is analogous to, but more general than, Hoffman's index of relative weight.

In an ANOVA design, the usual way to produce orthogonal stimulus dimensions is to construct all possible combinations of the cue levels in a completely crossed factorial arrangement. Such an arrangement becomes unmanageable when the number of cues is large, or when it is desirable to include many levels of each cue. However, if one is willing to assume that some of the higher-order interactions are negligible, then it is possible to employ a fractional replication design and evaluate the importance of the main effects and lower-order interactions with a considerably reduced number of stimuli (Anderson, 1964; Cochran & Cox, 1957; Shanteau, 1970; Slovic, 1969).

Integration theory. Integration theory can be considered one extension of the regression approaches described above. As such, it has formed the basis of an intensive program of research in the areas of clinical and psychophysical judgment, personality impression formation, and decision making. The essential ideas stem from the work of Norman Anderson, and are summarized in Anderson (1968b, 1969, 1970, in press).

Integration theory is concerned simultaneously with two problems. The first is scaling the stimulus items and determining the weighting parameters. This component of the theory is called "valuation." The second concern, called "integration," tests theories about the specific composition rules used by the subjects. Particular attention has been given to tasks in which a simple algebraic model, involving adding, averaging, subtracting, or multiplying the informational inputs, serves as the substantive theory of judgment that is being tested.

Technically, integration theory relies upon factorial designs, due to the fact that the substantive theories studied thus far have almost always been reducible to an ANOVA model. Therefore, ANOVA has been the principal analytical tool, serving to represent the theoretical postulates and providing a goodness-of-fit test of the models. An invalid response scale could cause a valid model to fail the test of fit. Therefore, an important feature of Anderson's approach is the use of monotone rescaling procedures for the response variable. If the model is correct, it serves as a frame on which to rescale the response. Failure to obtain an adequate fit after rescaling argues against the model, and success argues for it. Once the model and response scale are established, the subjective values of the information items can be derived. Anderson has used the term "functional measurement" to describe this interplay between theory and scaling.

According to integration theory, each piece of information is represented by two properties: A subjective scale value, s, and a weight, w.

The weight represents the salience or importance of the information. The basic theoretical model is:

$$R = C + \sum_{k=0}^{n} w_k s_k, \tag{4}$$

where R is the subject's response or judgment. The first term in the sum, $w_0 s_0$, represents the weight and scale value of the initial opinion, prior to receiving any information. This basic model has been expanded in a variety of ways to encompass different substantive theories. The following description will focus on an additive model as applied to a two-way factorial design. Here, the subject is shown a stimulus containing two descriptive items, one from the row dimension and one from the column dimension. His response is considered to be the resultant of the two items and his initial opinion:

$$R_{ij} = w_0 s_0 + w_R s_{Ri} + w_C s_{Cj}, \tag{5}$$

where w_R and w_C are the weights associated with the row and column dimensions, s_{Ri} is the scale value of the information item in Row i of Dimension R, and s_{Cj} is the scale value of the item in Column j of Dimension C.

An example of the kind of task to which Eq. (5) is applicable is the study by Sidowski and Anderson (1967) in which subjects judged the attractiveness of working at a certain occupation (Doctor, Lawyer, Accountant, or Teacher) in a certain city (City A, B, C, or D). Each cell of the design corresponds to a pair of items (a city-occupation combination) that the judge is to integrate. Another example is an impression formation task similar to that used by Lampel and Anderson (1968), in which each cell is a person described by an adjective and a photograph. The adjective represents the row source and the photograph represents the column source. The salience of each of the two sources, represented by w_R and w_C in Eq. (5), is assumed to be constant. The values of the adjectives and photographs are captured by s_{Ri} and s_{Cj}.

The important properties of this model are that the weights are constant across levels of each dimension and that the model permits the scaling of subjective values for each item. Thus, Eq. (5) is similar to the linear model of Eq. (2) except that subjective scale values, rather than the physical or objective values, are employed in the linear equation. It is not assumed that the objective values of the stimulus dimensions are linearly related to the responses. If, for example, the judgment task involved the rating of occupations, and salary was one of the factors, the actual salary levels would enter into the linear model

of Eq. (2) as predictors of the judgments. But it is quite likely that the judge perceives salary in a nonlinear fashion. The subjective difference between \$20,000 and \$25,000 is probably less than the difference between \$5,000 and \$10,000. Integration theory attempts to discover these subjective scale values and to determine rules of composition based on *these* values, whereas the regression and ANOVA approaches described earlier attempt to discover the combination rule based on the objective dimensions.

Equation (5) implies that the row by column interaction should be zero in principle and nonsignificant in practice. Therefore, ANOVA serves to test the model's goodness of fit. If the model passes this test, it may be used to estimate the subjective values s_{Ri} and s_{Cj}. When Eq. (5) is averaged over columns, the mean response for row i is:

$$R_{i\cdot} = w_R s_{Ri} + \text{constant}, \tag{6}$$

where the dot subscript on R denotes the average over the column index. The constant expression represents the influence of the columns and is the same for all rows. Equation (6) says that these row means form a linear function of the subjective values of the row stimuli. Thus, the row means constitute an interval scale of the row stimuli. Similarly, the column means constitute an interval scale of the column stimuli.

The above results hold for an additive model. Similar analyses for subtractive and multiplicative models may be found in Shanteau and Anderson (1969) and Anderson and Shanteau (1970). Anderson (in press) summarized these results and presented averaging versions of these models in which the weights are constrained to sum to unity. In some of these models, w_k can be scaled but not s_k, or vice versa. In others, both parameters can be scaled.

Anderson (1969) noted that caution is required in interpreting the meaning of significant interactions when these occur. Interactions may result from cognitive configurality that is theoretically meaningful or from defects in the response scale, such as floor and ceiling effects. In some cases, a monotonic rescaling of the judgments can be used to eliminate the interaction (see, for example, Bogartz & Wackwitz, 1970). Whether or not to rescale the judgments is a delicate matter, one that depends upon the researcher's degree of confidence in the theoretical model that is being tested and his confidence in the validity of the scale on which the judgments are measured. For example, Lampel and Anderson (1968) found a significant interaction between visual and verbal information in an impression formation task. This interaction could have been eliminated by monotonic rescaling of the responses. However, as Anderson (1970) observed, they did not remove the interaction because

previous experimentation had given them confidence in the response scale. Thus the interaction was retained and given a psychological interpretation.

Anderson and colleagues have also applied integration theory to study the effects of serial position in information integration. In these studies the stimuli were presented successively and the serial positions corresponded to factors in the design. The weights indicated by the main effects thus produced a serial-position curve that was used to assess whether information was given more salience earlier (primacy) or later (recency) in the sequence (Anderson, 1965b, 1968a; Shanteau, 1970). Anderson (1965b) noted that when information is presented serially, the weighted average model can be reformulated in a manner that makes it particularly valuable for studying the step-by-step buildup of a judgment in response to each item. This form, called the proportional change model, asserts that the judgment, R_k, produced after receipt of the kth item of information, is given by:

$$R_k = R_{k-1} + w_k(s_k - R_{k-1}), \tag{7}$$

where R_{k-1} is the judgment prior, and R_k is the judgment posterior, to presentation of the kth item. The scale value of the kth item is denoted by s_k, and w_k is a change parameter that measures the influence of the kth item.

Conjoint measurement. The theory of conjoint measurement (Luce & Tukey, 1964; Tversky, 1967b; Krantz & Tversky, 1971) is analogous in many respects to integration theory. Both are concerned with discovering the psychological laws that govern the composition of several attributes (for our purposes, several items of information). However, integration theory deals with *quantitative* laws, whereas conjoint measurement is concerned with *qualitative* laws.

When the stimuli and the judgments can be measured independently on interval scales, the rule of combination can be tested directly. However, when the assumption of interval-scale measurement is of dubious validity, conjoint measurement is valuable, since it uses only *ordinal* properties of the judgments to test the proposed combination rule.

Krantz and Tversky delineated the testable ordinal relationships among judgments that can be used to diagnose which, if any, of several polynomial composition rules is appropriate. For example, one class of testable properties, called independence conditions, serves a valuable diagnostic function. The essence of ordinal independence is that the ordering of the judgments within any row of the factorial matrix is constant across rows.

Krantz and Tversky argued that the ordinal approach to the study

of composition rules, exemplified by conjoint measurement, should be regarded as complementary to numerical approaches such as the lens model or integration theory. From a practical standpoint, they noted that direct tests of ordinal properties are generally more powerful in their ability to discriminate alternative composition rules than are overall tests of goodness of fit. From a theoretical standpoint, they contended that qualitative properties may sometimes lead to a more fundamental understanding of psychological principles than do numerical analyses.

Conjoint measurement has been applied in only a few judgment studies thus far (see, for example, Coombs & Huang, 1970; Tversky, 1967a, 1967c; Tversky & Krantz, 1969; Wallsten, 1970) but the explication of its analysis techniques by Krantz and Tversky should stimulate greater use of this approach in the future.

THE BAYESIAN APPROACH

Brunswik proposed the use of correlations to assess relationships in a probabilistic environment. He could have used conditional probabilities instead; had he done so, he undoubtedly would have built his lens model around Bayes' theorem, an elementary fact about probabilities described in 1763 by the Reverend Thomas Bayes. The modern impetus for what we are calling the Bayesian paradigm can be traced to the work of von Neumann and Morgenstern (1947) who revived interest in maximization of expected utility as a core principle of rational decision making, and to L. J. Savage, whose book *The Foundations of Statistics* fused the concepts of utility and personal probability into an axiomatized theory of decision in the face of uncertainty, "a highly idealized theory of the behavior of a 'rational' person with respect to decisions" (Savage, 1954, p. 7). The Bayesian approach was communicated to businessmen by Schlaifer (1959) and to medical diagnosticians by Ledley and Lusted (1959). Psychologists were introduced to Bayesian notions by Ward Edwards (Edwards, 1962; Edwards, Lindman, & Savage, 1963) and much of the empirical work to be discussed was stimulated directly by the ideas in the latter two papers.

The Bayesian approach is thoroughly embedded within the framework of decision theory. Its basic tenets are that opinions should be expressed in terms of subjective or personal probabilities, and that the optimal revision of such opinions, in the light of relevant new information, should be accomplished via Bayes' theorem. Edwards (1966) noted that, although revision of opinion can be studied as a separate phenomenon, it is most interesting and important when it leads to decision making and action. Because of this concern with decision making, the output of

a Bayesian analysis is not a single prediction but rather a distribution of probabilities over a set of hypothesized states of the world. These probabilities can then be used, in combination with information about payoffs associated with various decision possibilities and states of the world, to implement any of a number of decision rules, including the maximization of expected value or expected utility.

Bayes' theorem is thus a normative model. It specifies certain internally consistent relationships among probabilistic opinions and serves to prescribe, in this sense, how men should think. Much of the psychological research has used Bayes' theorem as a standard against which to compare actual behavior and to search for systematic deviations from optimality.

The Bayesian model. Given several mutually exclusive and exhaustive hypotheses, H_i, and a datum,[2] D, Bayes' theorem states that:

$$P(H_i/D) = \frac{P(D/H_i)P(H_i)}{\sum_i P(D/H_i)P(H_i)}. \tag{8}$$

In Eq. (8), $P(H_i/D)$ is the posterior probability that H_i is true, taking into account the new datum, D, as well as all previous data. $P(D/H_i)$ is the conditional probability that the datum D would be observed if hypothesis H_i were true. For a set of mutually exclusive and exhaustive hypotheses H_i, the values of $P(D/H_i)$ represent the impact of the datum D on each of the hypotheses. The value $P(H_i)$ is the prior probability of hypothesis H_i. It, too, is a conditional probability, representing the probability of H_i conditional on all information available prior to the receipt of D. The denominator serves as a normalizing constant. Although Eq. (8) is appropriate for discrete hypotheses, it can be rewritten, using integrals, to handle a continuous set of hypotheses and continuously varying data.

If is often convenient to form the ratio of Eq. (8) taken with respect to two hypotheses, H_i and H_j:

$$\frac{P(H_i/D)}{P(H_j/D)} = \frac{P(D/H_i)}{P(D/H_j)} \cdot \frac{P(H_i)}{P(H_j)}.$$

For this ratio form, new symbols are introduced:

$$\Omega_1 = \mathrm{LR} \cdot \Omega_o,$$

where Ω_1 represents the *posterior odds*, LR is called the *likelihood ratio*, and Ω_o stands for the *prior odds*.

[2] Within the regression and ANOVA paradigms, a datum refers to a response made by a judge; for Bayesians, however, a datum is an item of information presented to the judge.

Bayes' theorem can be used sequentially to measure the impact of several data. The posterior probability computed for the first datum is used as the prior probability when processing the impact of the second datum, and so on. The order in which data are processed makes no difference to their impact on posterior opinion. The final posterior odds, given n items of data, are

$$\Omega_n = \prod_{k=1}^{n} \mathrm{LR}_k \cdot \Omega_0. \tag{9}$$

Equation (9) shows that data affect the final odds multiplicatively. If the $\log_{10}$ of this equation were taken, the log likelihood ratios would combine additively with the log prior odds. The degree to which the prior odds change, upon receipt of a new datum, is dependent upon the likelihood ratio for that datum. Thus the likelihood ratio is an index of data diagnosticity or importance analogous to the weights employed in regression models.

The use of Bayes' theorem assumes that data are conditionally independent, i.e.,

$$P(D_j/H_i) = P(D_j/H_i, D_k).$$

If this assumption is not met, then the combination rule has to be expanded. For two data, the expanded version is:

$$P(H_i/D_1, D_2) \propto P(D_2/H_i, D_1)P(D_1/H_i)P(H_i). \tag{10}$$

As more data are received, the equation requires further expansion and becomes difficult to implement.

The meaning of the conditional independence assumption might be clarified by an example. Height and hair length are negatively correlated, and thus non-independent, in the adult U. S. population (even these days), but within subgroups of males and females, height and hair length are, we might suppose, quite unrelated. Thus if the hypothesis of interest is the identification of a person as male or female, height and hair length data are *conditionally* independent, and the use of Bayes' theorem to combine these cues is appropriate. In contrast, height and weight are related both across sexes and within sexes, and are thus both unconditionally and conditionally non-independent. These cues could not be combined via Bayes' theorem without altering it as shown in Eq. (10). One way of thinking about the difference between these two examples is that in the first case the correlation between the cues is mediated by the hypothesis: the person is tall and has short hair *because* he is male. In the case of conditional non-independence, however, the correlation between the cues is mediated by something other

than the hypothesis: the taller person tends to weigh more because of the structural properties of human bodies.

Experimental paradigms. A hypothetical experiment, similar to one actually performed by Phillips and Edwards (1966), will illustrate a common use of the Bayesian paradigm. The subject is presented with the following situation: Two bookbags are filled with poker chips. One bookbag has 70 red chips and 30 blue chips, while the other bag holds 30 red chips and 70 blue chips. The subject does not know which bag is which. The experimenter flips a coin to choose one of the bags and then begins to draw chips from the chosen bag. After drawing a chip he shows it to the subject and then replaces it in the bag, stirring vigorously before drawing the next chip. The subject is asked to estimate the probability that the predominantly red bag is the one being sampled. At the start, before the first chip is drawn, the subject is required to give a probability estimate of .5, indicating that each bag is equally likely to have been chosen. Then, after each chip is drawn, the subject reflects the revision of his opinion by changing his probability estimate. The subject sees 10 successive chips drawn; the basic data for analysis are the 10 posterior probability estimates the subject made after each chip.

The optimal responses are computed from Bayes' theorem. The data (poker chips) are conditionally independent because each sampled chip is replaced before the next is drawn. The prior odds are on (the bookbags were equally likely to be chosen), and the likelihood ratios associated with red and blue chips are a function of the 70/30 proportions in each bag:

$$\mathrm{LR}_{\text{red chip}} = \frac{P(\text{red chip}/H_{70\%\ \text{red}})}{P(\text{red chip}/H_{70\%\ \text{blue}})} = \frac{.7}{.3};$$

$$\mathrm{LR}_{\text{blue chip}} = \frac{P(\text{blue chip}/H_{70\%\ \text{red}})}{P(\text{blue chip}/H_{70\%\ \text{blue}})} = \frac{.3}{.7}.$$

The posterior odds of the predominently red bag having been chosen, given a sample of, say, six red chips and four blue chips are calculated from Eq. (9):

$$\Omega_{10} = \left(\frac{7}{3}\right)^6 \cdot \left(\frac{3}{7}\right)^4 \cdot 1 \simeq 5.44.$$

The odds are greater than 5 to 1 that the predominantly red bag is the bag being sampled. This corresponds to a posterior probability for that bag of approximately .845.

The primary data analysis compares subjects' probability revisions upon receipt of each chip with those of Bayes' theorem. To supplement

direct comparisons of Bayesian probabilities and subjective estimates, Peterson, Schneider, and Miller (1965) introduced a measure of the degree to which performance is optimal, called the accuracy ratio:

$$AR = \frac{SLLR}{BLLR},$$

where SLLR is the log likelihood ratio inferred from the subjects' probability estimates (using Eq. (9)) and BLLR is the optimal (Bayesian) log likelihood ratio. The conversion to log likelihood ratios is made because the optimal responses then become linear with the amount of evidence favoring one hypothesis over the other. The accuracy ratio can be computed for each datum, or it can serve as a summary measure across many responses made to a variety of data. In the latter case, it is the slope of the regression line relating SLLR's to BLLR's, and is thus similar to a beta weight in the correlational model.

The task just described illustrates the use of a binomial data-generating model. The Bayesian paradigm, however, is capable of dealing with a great variety of different types of data, discrete or continuous, from the same or different sources, etc. For example, some Bayesian experiments have employed multinomial distributions to generate samples of data. Table 1 provides a hypothetical illustration. In this example three hypotheses concerning a college student's grade point average (GPA) are related to three data sources (e.g., verbal ability, achievement motivation, and credit hours attempted). Each data source is comprised of several subclasses of information (e.g., below average or above average achievement motivation). The entries in the cells of

TABLE 1
SOME MULTINOMIAL DATA GENERATING HYPOTHESES

		Hypotheses about a student's GPA		
		H_1 Lower 33%	H_2 Middle 33%	H_3 Upper 33%
D_1: Verbal ability	1. Below average	.55	.30	.15
	2. Average	.30	.40	.35
	3. Above average	.15	.30	.50
D_2: Achievement motivation	1. At or below average	.75	.50	.50
	2. Above average	.25	.50	.50
D_3: Credit hours attempted	1. Below 12	.15	.25	.20
	2. 12–15	.25	.30	.20
	3. 16–18	.30	.30	.30
	4. Above 18	.30	.15	.30

the resulting evidence-hypothesis matrix are conditional probabilities of the form $P(D_{jk}/H_i)$, i.e., the probability that the kth subclass of data class j would occur, given H_i.

If the data subclasses are mutually exclusive and exhaustive, as is the case here, the conditional probabilities within any data source and any one hypothesis must sum to 1.00 (e.g., a student must be either above, at, or below average on achievement motivation). Across hypotheses, the conditional probabilities need not sum to any constant (e.g., relatively few college students, regardless of GPA, take less than 12 credit hours of course work).

The critical measure of relatedness between a cue and a hypothesis is represented here by three conditional probabilities, $P(D_{jk}/H_1)$, $P(D_{jk}/H_2)$, and $P(D_{jk}/H_3)$, rather than by a single correlation. The diagnosticity of a particular datum, D_{jk}, rests on the ratios of the conditional probabilities across hypotheses. Thus below average verbal ability (D_{11}) is highly diagnostic, whereas 15–18 credit hours attempted (D_{33}) gives no information at all concerning GPA.

A typical experiment, based on a multinomial task, proceeds as follows: The subject would be asked to assume some prior probability distribution across the hypothesis set (e.g., that each of the three GPA categories was equally likely). He then would receive a set of data, describing a student (e.g., verbal ability average, motivation above average, 16 credit hours). Following this he must revise his prior opinion about the likelihood that the various hypotheses had generated the evidence. To do this he needs some indication of the importance of each item of evidence. Either the $P(D_{jk}/H_i)$ values can be presented to him in the form of a table or he can be given feedback that will enable him to estimate these values on the basis of the relative frequency of occurrence of each item. The subject's estimates can be in the form of posterior probabilities or some analog such as posterior odds. These estimates are then compared with the optimal responses prescribed by Bayes' theorem.

Some Bayesian experiments require the subjects to estimate both $P(D/H)$ and $P(H/D)$. The goal of such a study is to see if the estimates are consistent; that is, to explore whether the $P(H/D)$ estimates can be predicted by aggregating the $P(D/H)$ estimates according to the optimal composition rule, Bayes' theorem.

Information-seeking experiments. The decision maker often has the option of deferring his decision while he gathers relevant information, usually at some additional cost. The information presumably will increase his certainty about the true state of the world and increase his chances of making a good decision. In seeking additional information, the decision maker must weight the relative advantage of the new information against

its cost. When the probabilistic characteristics of the task are well-defined, an optimal strategy can be specified that will, in conjunction with the reward for making a correct decision, the penalty for being wrong, and the cost of the information, determine a stopping point that is optimal in the sense of maximizing expected value (Edwards, 1965; Raiffa & Schlaifer, 1961; Wald, 1947). This task is a natural extension of the probabilistic inference tasks described above, since it requires the decision maker to link payoff considerations with his inferences in order to arrive at a decision. A large number of studies have investigated man's ability to make such decisions. For example, one commonly studied task uses the bookbag and the poker chip problem described earlier. As before, a sequence of chips is sampled, with replacement, from a bag with proportion of red chips equal to P_1 or P_2. Instead of estimating the posterior probabilities for each bag, the subject must decide from which bag the sample is coming. In some cases, he must decide, prior to seeing the first chip, how many chips he wishes to see (fixed stopping). In other cases, he samples one chip at a time and can stop at any point and announce his decision (optional stopping). Space limitations prohibit further analysis of this body of research here. The interested reader is referred to papers by Fried and Peterson (1969), Pitz (1969b, 1969c), Rapoport and Burkheimer (1970), and Wallsten (1968) for examples of this research.

COMPARISONS OF THE BAYESIAN AND REGRESSION APPROACHES

Having completed our overview of the basic elements of the regression and Bayesian approaches, it is appropriate to consider briefly some of the similarities and differences between them. At first glance, it would seem that the dissimilarities predominate. This impression is fostered, primarily, by the grossly different terminology used within each approach. However, closer examination reveals many points of isomorphism. In par-

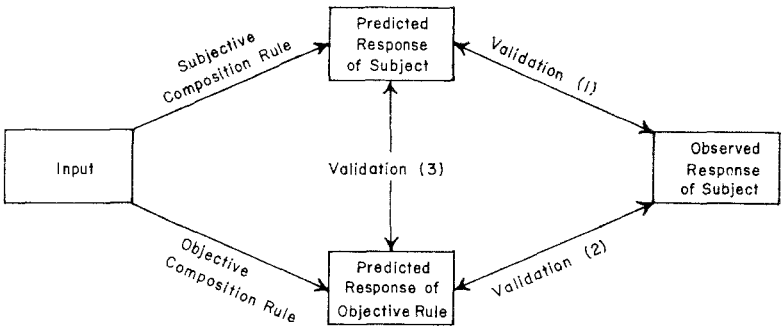

FIG. 2. General schema for comparing research paradigms.

ticular, each paradigm is based on a theoretical model of the composition rules whereby informational input is integrated into a judgmental response. The schematic diagram in Fig. 2 depicts some general relationships between input, process, and response, and will serve as the organizational framework for the present discussion.

Input

The information that serves as input to the decision maker varies somewhat both within and between each approach. The correlational paradigm typically involves dimensions of quantitative information. Data presented within the ANOVA and Bayesian studies, by contrast, are often discrete, categorical, or qualitative items, although these approaches can also process dimensional data. As to the relationships among data sources, ANOVA techniques require factorially combined information elements, and workers within the descriptive stream of correlational research also prefer orthogonal structure. Lens model research often uses items that are correlated in a fashion representative of the real world. Bayes' theorem, in its analytically convenient forms (Eqs. (8) and (9)), requires conditionally independent data. A rough translation of this requirement in correlational terms would be that the residual correlations between cues, with the criterion dimension partialled out, must be zero.

The Subject's Response

The response required of the subject also differs across paradigms. The correlational and ANOVA approaches usually deal with a single-valued prediction (point estimate) about some conceptually continuous hypothesis. Bayesians would say that there is a probability distribution over this continuous distribution and that the subject's single judgment must represent the output of some covert decision process in which some implicit decision rule is applied (e.g., the response may be interpreted as specifying the criterion value having the largest probability of occurrence), based on some implicit payoff matrix. Some Bayesian studies also require subjects to make predictions, usually concerning discrete hypotheses. When they do, the payoffs accompanying correct and incorrect predictions are usually made explicit to the decision maker. Most often, however, subjects in Bayesian studies estimate the posterior or conditional probability distribution (or some function thereof, such as odds) across various hypotheses. Although Bayes' theorem can, in principle, be applied to continuous hypotheses, the emphasis on probability distributions rather than point estimates makes such a task experimentally awkward (see, however, Peterson & Phillips, 1966).

Subjective Composition Rules

From the standpoint of psychological theory, the rule by which the subject combines or integrates the input information is the most important element being investigated. Each paradigm is based upon a theoretical model of this composition process. Furthermore, these various algebraic models are closely related. The simple additive model plays a key role in both correlational and ANOVA studies and more complex polynomial rules can also be studied within these paradigms. Bayes' theorem is multiplicative in form, much like the multiplicative models within integration theory and conjoint measurement. In addition, the proportional change model of integration theory (Eq. (7)) and the Bayesian model both conceptualize the step-by-step buildup of judgment in terms of a weighted combination of the present datum and prior opinion.

Each of these composition models contains descriptive parameters which assess the subjective impact or importance of information. Correlational studies rely on global measures which reflect importance across an entire dimension or data source, such as correlations ($r_{i,s}$), regression weights ($b_{i,s}$) or relative weights ($RW_{i,s}$). In contrast, Bayesians are usually interested in the subjective impact of each individual datum, as measured by its subjective likelihood ratio. In the Bayesian approach, the source or dimensionality of the datum is irrelevant. Integration theory makes a conceptual distinction between two kinds of impact which are confounded in the above measures: w_k reflects the importance of one dimension or source of data relative to other sources, and s_k measures the subjective value of a single datum relative to other data from the same source.

The Bayesian paradigm looks at fixed hypotheses and examines the manner in which the subjective probabilities of these hypotheses are revised in the light of new information. For this reason it has sometimes been called a "dynamic" paradigm. In contrast, most of the correlational research deals with "static" aspects of information processing: When a subjective weight is inferred from a subject's responses over 50 trials, it is assumed that the subject's view of the world remains unchanged over this period. However, the static *vs.* dynamic distinction is not inherent in the models. A good example of this point is found within integration theory, where the weighted average model takes both a static form (Eq. (4)) and a dynamic form, the proportional change model of Eq. (7). In like manner, a regression equation could process information sequentially, and the item-by-item revisions of the judgments could be compared to optimal revisions specified by the equation although such a study has not been reported.

Objective Composition Rules

Objective composition rules can be thought of as representing the environmental aspect of the judgment situation, and they encompass the same sorts of models as are applied to the subjects' judgments. In some cases, logical considerations suggest a normative rule (such as Bayes' theorem). In other situations, the rule is determined primarily on empirical grounds (e.g., the use of multiple regression in the lens model to predict some criterion variable).

Both the lens model and the Bayesian approach share a deep concern about the relationship between the decision maker and his environment. Both models compare what the decision maker does with what he should be doing. The optimality of multiple regression rests on the acceptance of a built-in payoff function: the least-squares criterion of goodness of fit. The posterior probabilities of the Bayesian model, however, can be combined with *any* payoff function to determine the best action. In certain circumstances, Bayesian and multiple regression models lead to identical solutions, as in the case of determining an optimal decision boundary between two hypotheses on the basis of normally distributed and standardized data (Koford & Groner, 1966).

Testing the Composition Rules

Validation of composition rules can take a number of different forms. Most important are comparisons between predicted responses (or other critical properties of subjective and objective rules) and the observed responses of the judge. These have been designated as validation types Nos. 1 and 2 in Fig. 2. The predictions of a particular subjective rule can also be compared with those of a particular objective rule. We have designated this as validation type 3. An example of this type of validation is the G index in the lens model.

With regard to validation types 1 and 2, researchers differ in their attitudes towards testing the models. Workers within correlational and Bayesian settings have typically been satisfied that high correlations between their model's predictions and the subject's responses provide adequate evidence for the validity of the model. For example, Beach (1966) observed correlations in the .90's between subjects' $P(H/D)$ estimates and Bayesian values that were calculated using subjects' earlier $P(D/H)$ estimates. He concluded that:

> . . . *S*s possess a rule for revising subjective probabilities that they apply to whatever subjective probabilities they have at the moment. . . .
>
> As has been amply demonstrated, the *S*s' revision rule is essentially Bayes' theorem. That is to say, *S*s' revisions can be predicted with a good deal of precision using Bayes' theorem as the model (Beach, 1966, p. 36).

However, Anderson, working within the framework of integration theory, has chided researchers for neglecting to test goodness of fit:

> Tests of quantitative predictions clearly require evaluation of the discrepancies from prediction. Much of the earlier work . . . is unsatisfactory in this regard since it is based on regression statistics and goes no further than reporting correlations between predicted and observed (Anderson, 1969, p. 64).

Anderson goes on to note that high correlations may occur despite a seriously incorrect model. As evidence he cites a study by Sidowski and Anderson (1967), which found a correlation of .986 between the data and a simple additive model, despite the fact that the ANOVA showed a statistically significant and substantively interpretable interaction. Finally, as we noted earlier, conjoint measurement provides qualitative, localized tests that may be quite powerful for discriminating among alternative composition models.

The paramorphic representation problem. Hoffman (1960, 1968) raised an issue particularly germane to the discrimination of alternative subjective composition functions. He observed that: (a) Two or more models of judgment may be algebraically equivalent yet suggestive of radically different underlying processes; and (b) two or more models may be algebraically different yet equally predictive, given fallible data. Drawing an analogy to problems of classification in mineralogy, Hoffman introduced the term "paramorphic representation" to remind psychologists that "the mathematical description of judgment is inevitably incomplete . . . , and it is not known how completely or how accurately the underlying process has been represented" (Hoffman, 1960, p. 125). Although Hoffman raised the paramorphic issue in connection with models based upon correlational techniques, all of the models we have discussed face the same problem when they are used descriptively.

EMPIRICAL RESEARCH

As the preceding discussion indicated, judgment researchers are studying similar phenomena but with somewhat different methods. In the remainder of this article we shall survey the empirical research spawned by the theory and methodology described above. Table 2 outlines the organization of our coverage. We have partitioned regression studies according to whether they were conducted within the correlational or ANOVA paradigms. We have further categorized the work according to five broad problem areas relating to the use of information by the decision maker.

The first category is devoted to a focal topic of research within each paradigm. For the correlational paradigm, this focal topic is the specification of the policy equation for the judge, including the closely related

TABLE 2
OVERVIEW OF TOPICS IN BAYESIAN AND REGRESSION STUDIES OF JUDGMENT

	Regression studies		Bayesian studies
	Correlational paradigm	ANOVA paradigm	Probability estimation paradigm
Typical dependent variables	R_s, $r_{i,s}$, $b_{i,s}$, r_a, G, C	Weights (w_k) and subjective scale values (s_k); significance tests for the models	Estimates of $P(H/D)$, $P(D/H)$, or ratios thereof; deviations from Bayes' theorem
Categories of Research			
I. The focal topic	Modeling a judge's policy	Models of impression formation	Conservatism
II. Task determinants of information use	Cue consistency Cue variability Cue format Number of cues Response mode	Set size Extremity of meaning Redundancy Inter-item consistency Contextual effects Sequential effects	Response mode Payoffs Data diagnosticity Prior probabilities Sequential effects
III. Learning to use information	Single-cue functional learning Positive *vs.* negative cues Multiple-cue learning Type of feedback Interpersonal learning Use of nonmetric cues	Little or no research	Effects of payoffs % and type of feedback
IV. Alternative subjective composition strategies	Strategies in correlational research	Little or no research	Strategies for estimating $P(H/D)$
V. Aiding the decision maker	Bootstrapping	Little or no research	Probabilistic information processing systems

problem of whether to include nonlinear terms in the policy equation.[3] A focal topic of ANOVA research is the distinction between two variants of

[3] This is not the only focal topic of correlational research. For workers in the Brunswik–Hammond tradition, the judge's learning or adaptation in an uncertain environment is a focal topic.

the linear model, the summation model and the averaging model, in impression formation. In Bayesian research, the focal topic is a particular form of suboptimal performance called conservatism. These topics are not closely interrelated. They are emphasized here simply because they have received a great deal of attention in the three areas of research.

The second category is devoted to the task determinants of information use. While many of these task variables are similar across differing paradigms, the dependent variables of such studies are less comparable, because they are so often closely related to the focal topics of the paradigms. For example, consider the number of items of information as a task variable. In the correlational paradigm, Einhorn (1971) has shown a decrease in R_s, subjects' linear consistency, as the number of cues increased. Anderson (1965a) used varying set sizes in an ANOVA paradigm to test predictions of an averaging model. In a Bayesian setting, Peterson, DuCharme, and Edwards (1968) have shown that larger sample sizes yield greater conservatism. Because task variables such as these are so closely linked to the focal topic in each area, we will report each group of studies directly after the relevant focal topic. We will restrict our coverage to what are primarily performance studies, i.e., studies in which the judge either has learned the relevant characteristics about the information he is to use prior to entering the experiment, or, alternatively, is given this information at the start. In other words, this particular research is concerned with evaluating how the judge uses the information he has and not with how he learns to use this information.

Additional research categories are devoted to learning to use information, subjective composition strategies other than those investigated as part of the focal topics, and techniques for helping the decision maker integrate information.

FOCAL TOPIC OF CORRELATIONAL RESEARCH: MODELING A JUDGE'S POLICY

The Linear Model

In a large number of studies, researchers have attempted to represent the judge's idiosyncratic weighting policy by means of the linear model (Eq. (2)). Examination of more than 30 of these studies illustrates the tremendous diversity of judgmental tasks to which the model has been applied. The tasks include judgments about personality characteristics (Hammond, Hursch, & Todd, 1964; Knox & Hoffman, 1962); performance in college (Dawes, 1970; Einhorn, 1971; Newton, 1965; Sarbin, 1942) or on the job (Madden, 1963; Naylor & Wherry, 1965); attractiveness of common stocks (Slovic, 1969) and other types of gambles (Slovic & Lichtenstein, 1968); physical and mental pathology (Goldberg, 1970;

Hoffman, Slovic, & Rorer, 1968; Oskamp, 1962; Wiggins & Hoffman, 1968); and legal matters (Kort, 1968; Ulmer, 1969).[4] In some cases, the stimuli were artificial and the judges were unfamiliar with the task. Typical of these is a study by Knox and Hoffman (1962), who asked college students to judge the intelligence of other students on the basis of grade point average, aptitude test scores, credit hours attempted, etc., and a study by Summers (1968), who asked students to rate the potential for achieving minority group equality as a function of legislated opportunities and educational opportunities. At the other extreme are studies of judgments made in complex but familiar situations by skilled decision makers who had other cues available besides those included in the prediction equation. For example, Kort (1968) modeled judicial decisions in workmen's compensation cases using various facts from the cases as cues. Brown (1970) modeled caseworkers' suicide probability estimates for persons phoning a metropolitan suicide prevention center; the cues were variables such as sex, age, suicide plan, etc., obtained from the telephone interview. And Dawes (1970) used a linear model to predict the ratings given applicants for graduate school by members of the admissions committee.

In all of these situations the linear model has done a fairly good job of predicting the judgments, as indicated by R_s values in the .80's and .90's for the artificial tasks and the .70's for the more complex real-world situations. Although most of these models were not cross-validated, the few studies that have applied the linear model derived from one sample of judgments to predict a second sample have found remarkably little shrinkage (Einhorn, 1971; Slovic & Lichtenstein, 1968; Summers & Stewart, 1968; and Wiggins & Hoffman, 1968).

Capturing a Judge's Policy

The various cues that define a stimulus are certainly not equally important and judges do not weight them equally. One of the purposes of using a linear model to represent the judgmental process is to make the judge's weighting policy explicit. Large individual differences among weighting policies have been found in almost every study that reports individual equations. For example, Rorer, Hoffman, Dickman, and Slovic

[4] Drs. Arthur Elstein and Lee Shulman have recently brought to our attention a study by Henry A. Wallace (who was later to become Secretary of Agriculture and Vice President) that used linear equations to model corn judges, and, in anticipation of the lens model, compared the judges' models with a model of the environment. Published in 1923, this study predates all other known work of this kind by several decades. The reference is:

WALLACE, H. A. What is in the corn judge's mind? *Journal of the American Society of Agronomy*, 1923, **15**, 300–304.

(1967) examined the policies whereby hospital personnel granted weekend passes to patients at a mental hospital. For five of the six cues there was at least one judge for whom that cue was the most important item and at least one judge for whom it was nonsignificant. A striking example of individual differences in a task demanding a high level of expertise comes from a study of nine radiologists by Hoffman, Slovic, and Rorer (1968). The stimuli were hypothetical ulcers, described by the presence or absence of seven roentgenological signs. Each ulcer was rated according to its likelihood of being malignant. There was considerable disagreement among radiologists' judgments as indicated by a median interjudge correlation, across stimuli, of only .38. A factor analysis of these correlations disclosed four different types of judges, each of which was associated with a particular kind of policy equation.

Even when expert judges don't disagree with one another, an attempt to model them can be enlightening. For example, seven of the nine radiologists studied by Hoffman *et al.* viewed small ulcer craters as more likely to be malignant than large craters. Yet a follow-up study by Slovic, Rorer, and Hoffman (1971) described statistical evidence obtained by other researchers indicating just the opposite, that large craters are more likely than small ones to be malignant.

The ability of regression equations to describe individual differences in judgment policies has led to the development of a number of techniques for grouping or clustering judges in terms of the homogeneity of their equations (Christal, 1963; Dudycha, 1970; Maguire & Glass, 1968; Naylor & Wherry, 1965; Wherry & Naylor, 1966; Williams, Harlow, Lindem, & Gab, 1970). Although a few of these studies have compared the methods, their relative utility remains to be demonstrated.

In summary, it is apparent that the linear model is a powerful device for predicting quantitative judgments made on the basis of specific cues. It is capable of highlighting individual differences and misuse of information as well as making explicit the causes of underlying disagreements among judges in both simple and complex tasks.

Nonlinear Cue Utilization

Despite the strong predictive ability of the linear model, a lively interest has been maintained in what Goldberg (1968) has referred to as "the search for configural judges." The impetus for this search comes from Meehl's (1954) classic inquiry into the relative validity of clinical *vs.* actuarial prediction. Meehl proposed that one possible advantage of the clinical approach might arise from the clinician's ability to make use of configural relationships between predictors and a criterion.

A clue to one outcome of the search was provided by Yntema and

Torgerson (1961) who hypothesized that, whenever predictor variables are monotonically related to a criterion variable, a simple linear combination of main effects will do a remarkably accurate job of predicting, even if interactions are known to exist. Yntema and Torgerson demonstrated their contention by presenting an example in which they showed that 94% of the variance of a truly configural function could be predicted from an additive combination of main effects.

Early work by Hoffman, some reported in Hoffman (1960) and some unpublished, indicated that configural terms based on the judge's verbalizations added little or no increment of predictable response variance to that contributed by the linear model. The R_s values were approximately as great as the retest reliabilities would permit, thus casting additional doubt about the existence of meaningful nonlinearities. Hursch, Hammond, and Hursch (1964); Hammond, Hursch, and Todd (1964); and Newton (1965) reported unsuccessful attempts to find evidence of nonlinearity using the C coefficient. However, these findings do not preclude the possibility of configural judgment processes, since either a lack of nonlinearity in the environment or a difference between the nonlinearity in the environment and in the judgmental systems is sufficient to produce low C values.

In light of the simple but compelling arithmetic underlying Yntema and Torgerson's "main effect hypothesis" the results of this early research should not have been too surprising. Yet the search continued, buoyed by (a) the repeated assertions of judges to the effect that their processes really were complex and configural; (b) the possibility that previous experimenters had not yet studied the right kinds of tasks, tasks that were "truly configural," and (c) the possibility that the experimental designs and statistical procedures used in previous studies were not optimally suited for uncovering the existing configural effects.

For example, Wiggins and Hoffman (1968) used a more sophisticated approach in their study of the diagnosis of neuroticism *vs.* psychoticism from the MMPI. These 861 MMPI profiles were selected because MMPI lore considered this task to be highly configural with respect to this type of diagnosis. In addition to criterion diagnoses, the judgments of 29 clinical psychologists were available for each profile. Besides using the linear model, Wiggins and Hoffman employed a "quadratic model," which included the 11 MMPI scale scores (X_i), as in the linear model, along with all 11 squared values of these scales (X_i^2) and the 55 cross-product terms (X_iX_j). The third model tested was a "sign model" which included 70 diagnostic signs from the MMPI literature, many of which were nonlinear. The coefficients for each model and each judge were derived using a stepwise regression procedure. Cross validation of the models in a new

sample indicated that 13 subjects were best described by the sign model, three by the quadratic model, and 12 by the linear model. But even for the most nonlinear judge the superiority of his best model over the linear model was slight (.04 increase in R_s).

Studies in which judges predicted the effects of various foreign policies (Summers & Stewart, 1968), rated the attractiveness of gambles (Slovic & Lichtenstein, 1968), evaluated the quality of patient care in hospital wards (Huber, Sahney, & Ford, 1969), and made decisions about workmen's compensation cases in a court of law (Kort, 1968) also found only minimal improvements in predictability as a consequence of including curvilinear and configural terms.

In an attempt to demonstrate the existence of configural effects, a number of investigators dropped the regression approach in favor of ANOVA designs applied to systematically constructed stimuli in tasks ranging from medical diagnosis to stock market forecasting (Hoffman, Slovic, & Rorer, 1968; Rorer, Hoffman, Dickman, & Slovic, 1967; Slovic, 1969). These studies did succeed in uncovering numerous instances of interaction among cues but the increment in predictive power contributed by these configural effects was again found to be small.

This line of research, employing both correlational and ANOVA techniques, can be summarized simply and conclusively. The hypothesis of Yntema and Torgerson has clearly been substantiated. The linear model accounts for all but a small fraction of predictable variance in judgments across a remarkably diverse spectrum of tasks.

However, the ANOVA research and other recent studies aimed at assessing the predictive power of nonlinear effects have exposed a different view of the problem, one that accepts the limited predictive benefits of nonlinear models but, simultaneously, asserts the definite, indeed widespread, existence of nonlinear judgment processes, and emphasizes their substantive meaning. To illustrate the complexity inherent in judgments that are quite predictable with a linear model, consider the data from the study of ulcer diagnosis conducted by Hoffman *et al.* An ANOVA technique showed that each of the nine radiologists who served as subjects exhibited at least two statistically significant interactions. One showed 13. Across radiologists there were 24 significant two-way, 17 three-way, and 14 four-way interactions. A subset of only 17 cue configurations, out of a possible 57, accounted for 43 of the 57 significant interactions. Thus numerous instances of configurality were evidenced and a subset of specific interactions occurred repeatedly across radiologists. Hoffman *et al.* did not attempt to probe into the content of the interactions they observed but Slovic (1969), in his study of stockbrokers, and Kort (1968), in his study of workmen's compensation decisions, did and both

uncovered information about the rationale behind their judges' nonlinear use of the cues.

Anderson has also paid careful attention to interactions obtained in his ANOVA studies of impression formation and has found several of substantive interest. For example, Anderson and Jacobson (1965) asked subjects to judge the likableness of persons described by sets of three adjectives. They found an interaction which implied that the weight given a particular adjective was less for sets where that adjective was inconsistent with the implications of the other adjectives than for sets in which it was consistent. Sidowski and Anderson (1967) asked subjects to judge the attractiveness of working at a certain job in a certain city. While the judgments of each city-job combination were found to be a weighted sum of the values for the two components, the attractiveness of being a teacher was more dependent upon the attractiveness of the city than were the other occupations, perhaps because teachers are in more direct contact with the cities' socioeconomic milieu. Other interesting examples of interactive cue utilization have been found by Lampel and Anderson (1968) and Gollob (1968).

Hoffman (1968) has observed that an undirected search for configural relations within a finite set of data is fraught with statistical difficulties. Green (1968) concurred and criticized standard regression and ANOVA techniques for being essentially fishing expeditions. A better strategy, he suggested, is to form some specific hypothesis about configurality and seek support for it. In this vein, Slovic (1966) hypothesized and found differences in subjects' strategies for combining information as a function of whether cues were in conflict. When the implications of important cues were congruent, subjects seemed to use both. When they were inconsistent, subjects focused on one of the cues or turned to other cues for resolution of the conflict. This study, and related experiments reported in Hoffman (1968) and in Anderson and Jacobson (1965), indicate that the linear model may need to be amended to include a term sensitive to the level of incompatibility among important cues.

Tversky (1969) and Einhorn (1971) also hypothesized, and found, specific nonlinear uses of information. Tversky's subjects sometimes chose among a pair of two-dimensional gambles by a lexicographic procedure in which they selected the gamble with the greater probability of winning, provided that the difference between gambles on this dimension exceeded some small value ϵ. If the probability difference was less than or equal to ϵ, these subjects selected the gamble with the higher payoff. In contrast to the linear model, this sort of strategy is noncompensatory, inasmuch as no amount of superiority with regard to payoff can overcome a deficiency greater than ϵ on the probability dimension.

Einhorn developed mathematical functions that could be incorporated into a prediction equation to approximate conjunctive and disjunctive judgmental processes as postulated by Coombs (1964) and Dawes (1964). Dawes described the evaluation of a potential inductee by a draft board physician as one example of a conjunctive process. The physician requires that the inductee meet an entire set of minimal criteria in order to be judged physically fit. A disjunctive evaluation, on the other hand, depends solely on the attribute with highest value. For example, a scout for a professional football team may evaluate a player purely in terms of his best specialty, be it passing, running, or kicking. Neither the conjunctive nor the disjunctive models balances one attribute against another as does the linear model. Einhorn pitted his conjunctive and disjunctive models against the linear model in two tasks, one where faculty members ranked applicants for graduate school, the other where students ranked jobs according to their preferences. He found that many subjects were fit better by the conjunctive model than by the linear model, particularly in the job preference task. Einhorn concluded by criticizing the notion that cognitive complexity and mathematical complexity go hand in hand. He argued that nonlinear, noncompensatory strategies may be more simple, cognitively, than the linear model, despite their greater mathematical complexity.

At this point, it seems appropriate to conclude that notions about nonlinear processes are likely to play an increasing role in our understanding of judgment despite their limited ability to outpredict linear models.

Subjective Policies and Self-Insight

Thus far we have been discussing weighting policies that have been assessed by fitting a regression model to the judge's responses. We think of these as "computed" or "objective" policies. Judges in a number of studies were asked to describe the relative weights they had used in the task. The correspondence between these "subjective weights" and the computed weights serves as an indicant of the judge's insight into his own policy. Martin (1957) and Hoffman (1960) proposed a technique that has been used to examine these "after the fact" opinions—that of asking the judge to distribute 100 points according to the relative importance of each attribute. Martin found that the linear model based on subjective weights produced a mean R_s of .77 in predicting evaluations of a student's sociability. A linear model computed in the usual way, but not cross-validated, did better, producing a mean R_s of .89. Hoepfl and Huber (1970), Hoffman (1960), Oskamp (1962), Pollack (1964), Slovic (1969), and Slovic, Fleissner, and Bauman (in press)

all found serious discrepancies between subjective and computed relative weights for their judges.

One type of error in self-insight has emerged in all of these studies. Judges strongly overestimate the importance they place on minor cues (i.e., their subjective weights greatly exceed the computed weights for these cues) and they underestimate their reliance on a few major variables. Subjects apparently are quite unaware of the extent to which their judgments can be predicted by only a few cues. Across a number of studies, varying in the number of cues that were available, three cues usually sufficed to account for more than 80% of the predictable variance in the judge's responses. The most important cue usually accounted for more than 40% of this variance.

Shepard (1964, p. 266) presented an interesting explanation of the subjective underweighting of important cues and overweighting of minor cues. He hypothesized:

> Possibly our feeling that we can take into account a host of different factors comes about because, although we remember that at some time or other we have attended to each of the different factors, we fail to notice that it is seldom more than one or two that we consider at any one time.

Slovic, Fleissner, and Bauman (in press), studying the policies of stockbrokers, examined the relationship between number of years as a broker and accuracy of self-insight. The latter was measured by correlating a broker's subjective weights with his computed weights across eight cue factors. Across 13 brokers, the Spearman rank correlation between the insight index and experience was —.43. Why should greater experience lead to less valid self-insight? Perhaps the recent training of the young brokers necessitated an explicit awareness of the mechanics of the skill that they were attempting to learn. Skills generally demand a great deal of conscious attention as they are being acquired. With increasing experience, behaviors become more automatic and require much less attention. Because of this they may also be harder to describe. It may be that the most experienced judges produce verbal rationales for their evaluations that are less trustworthy than those of their inexperienced colleagues. This hypothesis is an intriguing one and needs to be investigated with more precision than was done in this study. In fact, all of the studies described in this section need to be replicated using a variety of psychophysical methods to assess subjective weights. Some methods that could be used in such studies are suggested by Eckenrode (1965), Klahr (1969), Yntema and Klem (1964), and Yntema and Torgerson (1961).

TASK DETERMINANTS IN CORRELATIONAL RESEARCH

Cue Interrelationships

Several studies have examined the role of intercorrelational structure and conflict among cues in determining the weighting of those cues. Slovic (1966; see also Hoffman, 1968) found that, when important cues agreed in their implications, subjects typically used both and weighted them equally. When they disagreed, subjects focused on one of the cues or turned to other cues for resolution of the conflict. Also, in situations of higher cue conflict, linear consistency (R_s) was much lower. These effects were found both when cue conflict and cue intercorrelations varied together and when conflict was varied while keeping the cues orthogonal (uncorrelated). Dudycha and Naylor (1966a) have also studied the effect of varying the intercorrelations among cues upon policy equations. They found that profiles of cue-utilization correlations ($r_{i,s}$) showed more variance and R_s decreased as the correlation between cues decreased. However, the regression weights ($b_{i,s}$ values) remained relatively stable across sets varying in cue correlation and, therefore, much of the decrease in R_s may have been a statistical artifact (see Schenck & Naylor, 1968).

Cue Variability and Cue Utilization

Uhl and Hoffman (1958) hypothesized that an increase in the variability of a salient cue, across a set of stimuli, would lead to greater weighting of this cue by a judge. This increased weight would persist, they proposed, even among subsets of stimuli for which this cue was not unusually variable. Underlying this hypothesis was the assumption that the judge is motivated to make differential predictions and that cues which increase his ability to differentiate will reinforce his use of those attributes. The variability of a salient cue is one such feature that correlates with differentiability. Presumably judges will focus their attention on the more highly variable cues, other things being equal.

Uhl and Hoffman tested this hypothesis in a task where subjects judged IQ on the basis of profiles made up of nine cues. Each subject judged several sets of profiles on different days. The variability of a particular cue was increased on 1 day by providing a greater number of extreme levels. On a following day, the cue was returned to normal and its relative weight was compared with the weight it received prior to the manipulation of variability. The hypothesized effect was found in seven out of ten subjects when a strong cue was manipulated. Increasing the variability of a minor cue had no effect upon its subsequent

use. The authors concluded that the judge may alter his system of judgment because of the characteristics of samples he sees.

Morrison and Slovic (1962) independently tested a version of the variability hypothesis in a different type of setting. Each of their stimuli consisted of a circle (Dimension 1) paired with a square (Dimension 2). Subjects had to rank order a set of these stimuli on the basis of their total area (circle and square combined). The results indicated that if the variability of circle area was greater than the variability of square area in the set of stimuli, then circle area would be assigned much heavier weight in the judgment. If the variability of square area was higher in the set, then square area became the dominant dimension.

Cue Format

Knox and Hoffman (1962) examined the effect of profile format on judgments of a person's intelligence and sociability. Each subject based his judgments on profiles of cues. In one condition, the cues were presented as T scores with a mean of 50 and SD of 10. A second condition presented the information in percentile scores. The percentile profiles showed considerably more scatter and had more extreme values than did the T scores, which tended to appear rather "squashed." Judgments were made on a stanine scale with a normal distribution suggested but not forced. Judgments made to percentile scores were found to be much more variable. It appeared that judges were responding not only to the underlying meaning of the scores but to the graphical position of the points on the profile in an absolute sense. Subjects were reluctant to make ratings on the judgment scale that were more extreme than the stimulus scores. Being statistically naive, they were unable to gauge the true extremeness of certain T scores. Judgments made to percentile scores were also more reliable and produced higher values of R_s when linear models were fitted to them. Regression weights did not differ significantly between formats.

Number of Cues

There has been surprisingly little correlational research on the effects of varying the number of cues. A pilot study by Hoffman and Blanchard (1961) obtained some interesting results but was limited by a small number of subjects. Hoffman and Blanchard had subjects predict a person's weight on the basis of either two, five, or seven physical characteristics. Increased numbers of cues led to lower R_s values, decreased accuracy, lower test-retest reliability, and lower response variance. This latter finding may be the cause of some of the other results and may

itself be due to an increased number of conflicts among cues in the larger data sets.

Hayes (1964) and Einhorn (1971) also found that linear consistency decreased as the number of cues increased. Einhorn interpreted this decrease in R_s as indicating that his subjects were using more complex models in the high information conditions—models whose variance was not predictable from the linear and nonlinear models he tested. However, the reliability of his subjects' judgments was not assessed and it is possible that greater information merely produced more unreliable rather than more complex judgments. Hayes found that increased numbers of cues also led to a reduction in decision quality when subjects were working under a time limit.

Oskamp (1965) had 32 judges, including eight experienced clinical psychologists, read background information about a case. The information was divided into four sections. After reading each section of the case the judge answered 25 questions about the attitudes and behaviors of the subject and gave a confidence rating with each answer. The correct answers were known to the investigator. Oskamp found that, as the amount of information about the case increased, accuracy remained at about the same level while confidence increased dramatically and became entirely out of proportion to actual correctness.

In summary, there is a small amount of evidence that increasing the amount of information available to the decision maker increases his confidence without increasing the quality of his decisions and makes his decisions more difficult to predict.

Cue-Response Compatibility

Fitts and Deininger (1954) introduced the concept of stimulus–response compatibility to explain the results of several studies of paired-associates learning and reaction time. Compatibility was defined as a function of the similarity between the spatial position of the stimulus in a circular array and the position of the correct response in the same sort of array. High compatibility produced the quickest learning and fastest reaction time. In an experiment concerned with risk-taking judgments, Slovic and Lichtenstein (1968) observed a related type of compatibility effect that influenced cue utilization. They found that when subjects rated the attractiveness of a gamble, probability of winning was the most important factor in their policy equations. In a second condition, subjects were required to indicate the attractiveness of a gamble by an alternative method, viz., equating the gamble with an amount of money such that they would be indifferent between playing

the gamble and receiving the stated amount. Here it was found that attractiveness was determined more by a gamble's outcomes than by its probabilities. The outcomes, being expressed in units of dollars, were readily commensurable with the units of the responses, also dollars. On the other hand, the probability cues had to be transformed by the subject into values commensurable with dollars before they could be integrated with these other cues. It seems plausible that the cognitive effort involved in making this sort of transformation greatly detracted from the influence of the probability cues in the second task.

This finding suggests the general hypothesis that greater compatibility between a cue and the required response should enhance the importance of that cue in determining the response. Presumably, the more complex the transformation needed to make a cue commensurable with other important cues and with the response, the less that cue will be used.

FOCAL TOPIC OF THE ANOVA PARADIGM: MODELS OF IMPRESSION FORMATION

There is a substantial body of literature concerned with the problem of understanding how component items of information are integrated into impressions of people. Much of this research can be traced to the work of Asch (1946) who asked subjects to evaluate a person described by various trait adjectives. In one of Asch's studies, the adjective "warm" was added to the set of traits. Another group saw the trait "cold." All other adjectives were identical. The subjects wrote a brief description of the person and completed an adjective checklist. Asch found that substitution of the word "warm" for "cold" produced a decided change in the overall characterization of the person being evaluated. He interpreted this as being due to a shift in meaning of the traits associated with the key adjectives "warm" and "cold." This view has much in common with the notions of configurality and interaction we have been discussing.

More recent endeavors have centered around the search for quantitative models of the integration process. That is, they attempt to develop a mathematical function of the scale values of the individual items to predict the overall impression. Although Asch explicitly denied that an impression could be derived from a simple additive combination of stimulus items, the additive model and its variations have received the most attention (Rosenberg, 1968). Most of these studies have used rigorous experimental control, factorial designs, and statistical techniques such as ANOVA as the basis for their analyses.

One of the first studies to test the additive model was carried out by Anderson (1962). His subjects rated a number of hypothetical persons,

each described by three adjectives, on a 20-point scale of likableness. Within each set there was one item each of high, medium, and low scale value as determined in a separate normative study. An additive model gave an excellent fit to the data.

The additive model serves as a more general case for two derivative models—one based on the principle of summation of information, the other an averaging formulation. In the summation model, the values of the stimulus items are added to arrive at an impression. The averaging model asserts that an impression is the mean, rather than the sum, of the separate item values. The adding *vs.* averaging question has fundamental importance in integration theory. It governs the theoretical interpretation of the results and it affects the scaling of the stimuli (Anderson, in press).

The study by Anderson (1962) did not attempt to distinguish between the averaging and summation formulations. To do so requires careful attention to subtle facets of stimulus construction and experimental design. Recent research has taken on this challenge, varying task and design characteristics in an attempt to determine the validity of these and other competing models. The following section will review briefly the types of situational manipulations that have been brought to bear on this problem.

TASK DETERMINANTS OF INFORMATION USE IN IMPRESSION FORMATION

Set Size

The number of items of information in a set is one factor that has been varied in attempts to distinguish summation and averaging models. Fishbein and Hunter (1964) provided four groups of subjects with different amounts of positively evaluated information about a fictitious person. The information was presented sequentially in such a way that the total summation of effect increased as a function of the number of items while the mean decreased, i.e., the more highly evaluated items came first. The subjects used a series of bipolar adjective scales to evaluate the stimulus persons. The judgments became more favorable as the amount of information increased, presumably supporting the summation model. The Fishbein and Hunter study has been criticized by Rosenberg (1968) who argued that presenting the most favorable adjective first permitted possible sequential effects to influence the results. Also, the intended decrease in mean evaluation of the individual adjectives as a function of set size was not statistically significant. As Anderson (1971 in press) observed, increased extremity of judgment

with sequential presentation of essentially equal adjectives can be accounted for by a version of the averaging model.

Anderson (1965a) also used set size to contrast the two models. He had subjects rate the likableness of persons described by either two or four traits. He found that sets consisting of two moderately valued traits and two extremely valued traits produced a less extreme judgment than sets consisting of the two extreme traits alone. This result was taken as support for the averaging model and was later replicated by Hendrick (1968). Another result of Anderson's study, that sets of four extreme adjectives were rated more extreme than sets with two extreme adjectives, confirmed earlier findings by Anderson (1959), Podell (1962) and Stewart (1965) to the effect that increased set size produces more extreme ratings. While this result seems to support a summation model, Anderson showed how it could be accommodated using an averaging model that incorporates an initial impression with nonzero weight and scale value s_0. Anderson (1967b) provided further support for this model.

Extremity of Information

The adjectives in Anderson's (1965a) study were presumed to be of equal weight. Thus the averaging model predicted that the judgment of a stimulus set containing four items having extreme scale values averaged with the judgment of a set containing four items of moderate value would equal the judgment of a set containing two extreme and two moderate items. Anderson found that this prediction did not hold for negatively evaluated items. The discrepancy suggested that the extreme negative items carried more weight than did moderately negative information.

Studies by Himmelfarb (1970), Kerrick (1958), Manis, Gleason, and Dawes (1966), Oden and Anderson (in press), Osgood and Tannenbaum (1955), Podell and Podell (1963), Weiss (1963), and Willis (1960) also found indications that the weight of an information item is associated with the extremity of its scale value. Manis *et al.* found that two positive or two negative items of information of different value would lead to a judgment less extreme than the most extreme item but more extreme than that predicted by a simple averaging of the items. At the same time these judgments were not extreme enough to be produced by the summation model. To account for these results, the authors suggested a version of the averaging model that weights items in proportion to their extremity. Himmelfarb found support for an averaging model in which neutral information received less weight than more polarized information. Oden and Anderson found that the importance of a stimulus dimension depended on its value but the direction of dependence varied across different types of judgment tasks.

Redundancy

Both summation and averaging models assume that the values of the stimulus items are independent of the other items in the set. This assumption has been the focus of concern for a number of studies. Dustin and Baldwin (1966), for example, had subjects evaluate persons described by single adjectives, A and B, and by the combined pair AB. Ratings of AB pairs tended to be more extreme than the mean of the individual items; this tendency was dependent upon the degree of redundancy or implication between A and B as measured by their intercorrelation in a normative sample. Schmidt (1969) did a similar study but varied the relatedness of the items differently. He combined trait sentences (Mr. A is kind) with instance sentences (Mr. A is kind to B). The two sentences just given are obviously highly redundant. By changing the trait adjectives this redundancy can be greatly reduced. Schmidt found that judgments based on less redundant sets were consistently more extreme than those based on more redundant information. Wyer reported similar findings in studies where redundancy was measured by the conditional probability of A given B (Wyer, 1968) and by the degree to which the joint probability of occurrence (P_{AB}) exceeded the product of the two unconditional probabilities (P_A and P_B) (Wyer, 1970). It seems apparent that models in this area will need further revision to handle the effects of redundancy.

Inter-Item Consistency

The data just described indicate that highly redundant information has less impact. But information with too great a "surprise value" shares a similar fate. Anderson and Jacobson (1965) found that an item whose scale value is highly inconsistent with its accompanying items (as is the trait "gloomy" in the set "honest-considerate-gloomy") was likely to be discounted, i.e., given less weight. The discounting was slight when subjects were told that all three traits were accurate and equally important, but increased when subjects were cautioned that one of the items might be less valid than the others. Anderson and Jacobson argued that the averaging model might have to include differential weights to accommodate the reduced impact of inconsistent information.

Wyer (1970) defined inconsistency among two adjectives as the degree to which their joint probability (P_{AB}) was less than the product of their unconditional probabilities (P_A and P_B). Note that this places high inconsistency at the negative end of a continuum defined by P_{AB}–P_AP_B, with maximum redundancy at the positive end. After constructing stimuli according to this definition, Wyer found that inconsistency produced a discounting of the less polarized of a pair of adjectives, leading to a

more extreme evaluation. However, when inconsistency became too great, both adjectives appeared to be discounted, producing a less extreme evaluation.

Himmelfarb and Senn (1969) studied the effects of stimulus inconsistency in experiments concerned with judgments of a person's social class. The stimulus persons were described by dimensional attributes, occupation, income, and education. Surprisingly, discounting of inconsistent information was not found. The authors speculated that their failure to find discounting might have been due to the lack of directly contradictory information or to the possibility that social class stimuli, being objective aspects of an individual, might be less easily discounted than personality traits.

Other Contextual Effects

Anderson and Lampel (1965) and Anderson (1966) had subjects form an impression based on three adjectives and then rate one of the component traits alone. Both studies produced context effects, judgments of the single trait being displaced towards the values of the other traits. A natural interpretation of this effect is that the value or meaning of the test word has changed as a function of the impression formation process, much as Asch originally suggested. Wyer and Watson (1969) argued in favor of this change-of-meaning interpretation over several competing hypotheses but their data are equivocal. Anderson and Lampel (1965) hypothesized that positive context effects are due to a generalized halo effect rather than a change of meaning. According to this theory, once a component has been integrated into the whole, it no longer has a distinct individual meaning. The subject then rates the individual component according to a weighted average of its context-free meaning and the overall impression. Anderson (1971) describes two experiments supporting this position.

Primacy and Recency Effects

Several dozen studies of impression formation have attempted to determine whether information presented early in a sequence is more or less influential than information presented later, other things being equal. Greater influence of early information is called primacy. Its opposite effect is called recency.

These studies can be categorized according to whether the subject's task was to average verbal items of information such as adjectives, or quantitative and perceptual items such as numbers, weights, lines, and loudness of sounds. A typical experimental design in such studies goes as follows: First the items are scaled individually with respect to the

criterion. These items are then sorted into homogeneous subsets having high (H) or low (L) scale values. Then blocks of H and L items are presented in varying order. For example, primacy would lead the final judgment for an HHLL sequence to be higher than that for an LLHH sequence, while recency would produce the opposite effect.

The results of these investigations indicate that order effects are highly pervasive phenomena, appearing in studies that employ quite diverse stimuli and response modes. However, whether primacy or recency effects occur is dependent upon task characteristics. When adjectives are used as stimuli and the subject responds only at the end of the information sequence, primacy is usually found (Asch, 1946; Anderson, 1965b; Anderson & Barrios, 1961; Anderson & Norman, 1964; Luchins, 1957). Primacy effects in these studies have been attributed to decreased attention being given the later adjectives and to discounting of the inconsistent information provided by the later adjectives (Anderson, 1968a). When attentional demands were changed by having subjects recall the adjectives after making their rating, pronounce each adjective, or make a rating after each new item was presented, recency predominated (Anderson, 1959, 1968a; Levin & Schmidt, 1969; Luchins, 1958; Anderson & Hubert, 1963; Hendrick & Costantini, 1970a; Rhine, 1968; Rosenkrantz & Crockett, 1965; Stewart, 1965). When numbers, weights, lines, or sounds are used as stimuli, the information items are homogeneous and not likely to create feelings of incongruity. In these studies recency is observed, regardless of whether the judgments are made during the sequence or only at the end (Anderson, 1964, 1967a; Anderson & Jacobson, 1968; Hendrick & Costantini, 1970b; Parducci, Thaler, & Anderson, 1968; Weiss & Anderson, 1969). Although many hypotheses have been proposed to account for these effects, their causes remain to be determined precisely.

FOCAL TOPIC OF BAYESIAN RESEARCH: CONSERVATISM

The most common Bayesian study deals with probability estimation, often in some variant of the bookbag and poker chip experiment described earlier. The primary finding has been labeled *conservatism:* Upon receipt of new information, subjects revise their posterior probability estimates in the same direction as the optimal model, but the revision is typically *too small;* subjects act as if the data are less diagnostic than they truly are. Subjects in some studies (Peterson, Schneider, & Miller, 1965; Phillips & Edwards, 1966) have been found to require from two to nine data observations to revise their opinions as much as Bayes' theorem would prescribe for one observation.

Much of the Bayesian research has been motivated by a desire to

discover the determinants of conservatism in order that its effects might be minimized in practical diagnostic settings. A spirited debate has been raging among Bayesians about which part of the judgment process leads subjects astray. The principal competing explanations as to the "locus of conservatism" are the misperception, misaggregation, and artifact hypotheses (Edwards, 1968).

Misperception

In order to perform optimally, subjects must have some understanding of the data generator, the model, device, equations or other assumptions used by the experimenter to generate the stimuli shown to the subject. If the subject misunderstands the data generator, he may misperceive the conditional probability of the data given the hypothesis, $P(D/H)$; this misperception is hypothesized by some to be the source of conservatism. For example, Lichtenstein and Feeney (1968) showed that subjects performed very poorly when dealing with a circular normal data generator despite 150 training trials with feedback. But subjects' data and comments suggested an entirely different (and incorrect) model regarding the meaning of each datum, and reanalyses of their responses showed them to be quite consistent with this simpler yet incorrect view of the data generator. Does such a simple and popular data generator as the binomial distribution also lead to misperceptions about the meaning of data? Vlek and Beintema (1967) and Vlek and van der Heijden (1967) showed that it does. Vlek and Beintema presented subjects with samples (e.g., five black and four white) drawn from an urn whose constituent proportions were known to the subject, and asked them how often such a sample might be expected to occur in 100,000 samples of the same size. Vlek and van der Heijden asked for the probability that such a sample would occur in 100 trials. Both studies showed that subjects had poor understanding of the likelihood of data.

If such misperceptions are the cause of conservatism, then one would expect estimates of posterior probabilities to be consistent with, and predictable from, estimates about the data generator, that is, estimates of $P(D/H)$. Peterson, DuCharme, and Edwards (1968) had subjects in a binomial task estimate $P(H/D)$, then $P(D/H)$. Then they were instructed in $P(D/H)$ by being shown several theoretical sampling distributions and discussing them with the experimenter. For example, they observed how the distribution became more peaked as the number of draws and the dominant proportion increased. Finally they were again asked to estimate $P(H/D)$. Peterson *et al.* found that subjects' conservative $P(H/D)$ estimates could be explained by the deviations of

their $P(D/H)$ estimates from the optimal values. They also found that instruction about the sampling distributions reduced conservatism in the final stage, but the reduction was small in relation to the amount of conservatism.

Subjects in the study by Peterson *et al.* did not have the theoretical sampling distribution available at the time they made postinstruction $P(H/D)$ estimates. Pitz and Downing (1967) gave subjects similar instruction and, in addition, allowed them to refer to histogram displays of the theoretical sampling distributions as they made predictions about which of two populations was generating the data. However, their predictions were not improved by this instruction. Wheeler and Beach (1968) trained subjects by having them observe samples of eight draws, make a bet on which of two populations generated the data, and then observe the correct answer. Prior to training the subjects' sampling distributions were too flat, their betting responses were conservative, and these two errors were consistent with one another. After training, the subjects' sampling distributions were more veridical, their betting responses were less conservative, and again the two sets of responses were consistent.

A particular kind of misperception error relates to the impact of rare events. Vlek (1965) suggested that unlikely events, when they occur, are seen as uninformative. He argued for the compelling nature of this error by giving an exaggerated example,

> The posterior probability that a sample of 2004 chips, 1004 of which are red, is taken from bag A ($P_r = .70$), and not from B ($P_r = .30$), is equal to .967. But who will accept hypothesis A as a possible generator of these data, and, if forced to do so, who dares to base an important decision of such a small difference in the—seemingly biased—sample? (p. 15).

The answer to his question is, of course, that Bayes' theorem dares. In the optimal model, it matters not at all that a datum may be highly unlikely under both hypotheses. The only determinant of its impact is the relative possibility of its occurrence: the likelihood ratio. The violation of this likelihood principle has been demonstrated by Vlek (1965) and Vlek and van der Heijden (1967), who showed a systematic increase in conservatism as a function of the rarity of the data, and this violation can serve as an explanation for Lichtenstein and Feeney's (1968) results. Beach (1968) directly tested Vlek's hypothesis. Beach constructed decks of cards, each with a letter, from A to F, written on it in green or red ink. The task of the subjects was to estimate the posterior probability that the letters sampled were drawn from the green deck rather than the red deck, given complete information about the frequency of each letter

in each deck. Two groups of subjects used different decks of cards; the likelihood ratios were the same between groups, but the relative frequencies of the letters differed between groups. This permitted a test of whether the impact of rare events was misperceived, with likelihood ratio held constant. The results verified Vlek's hypothesis; subjects were more conservative when responding to less likely events.

Misaggregation

Another explanation of conservatism is that subjects have great difficulty in aggregating or putting together various pieces of information to produce a single response. Proponents of this view draw support from several sources (Edwards, 1968). First, they point out that in the studies just reported as supporting the misperception hypothesis, subjects were shown samples of several data at once. When shown a sample of, say, six red and three blue chips, and asked to state the probability that such a sample might occur, the subject must, in a sense, aggregate the separate impact of each chip, even though the sample is presented simultaneously. Viewed in this light, both estimation of $P(H/D)$ and of $P(D/H)$ in studies like Wheeler and Beach (1968) are aggregation tasks; thus the consistency between the two tasks does not provide a discrimination between the misperception and misaggregation hypotheses. Beach (1968), testing the rare event hypothesis, did present subjects with only one datum at a time, but he presented three data per sequence. Gettys and Manley (1968) reported two experiments in which five levels of frequency of data and five levels of likelihood ratios were factorially combined in 100 binomial problems. For each problem the subject was shown the contents of two urns and the result of a single sampling of one datum. In this situation with no aggregation required, the rare event effect was not found. The subjects were sensitive to changes in likelihood ratio but not to differing event frequencies. The authors argued that the rare event effect found in other studies is attributable to aggregation difficulties.

A related source of support for the misaggregation hypothesis comes from the finding that subjects perform best on the first trial of a sequence. DuCharme and Peterson (1968) reported this finding based on a task using normal data generators. The subjects were shown samples of heights and asked the posterior odds that the population being sampled was of men or women. They were virtually optimal for single-datum sequences and for the first trial of four-data sequences, but conservative on subsequent trials. Similar results were obtained by Peterson and Swensson (1968).

It might be noted that Peterson and Miller (1965) found conservatism with just one datum per problem, but this presents no special problem for

the misaggregation hypothesis, since in that study the one datum had to be aggregated with a varying value for the prior probability of the hypothesis. Peterson and Phillips (1966) also found first-trial conservatism. However, they, like Peterson and Miller, used a probability response mode. This mode, as will be discussed later, is highly susceptible to a nonoptimal but simple strategy which produces artifactual results. DuCharme and Peterson (1968) and Peterson and Swensson (1968) avoided this criticism by asking for responses in terms of posterior odds rather than probabilities, and found first-trial optimality.

Both the misperception and misaggregation hypotheses received support in a study by Phillips (1966; also reported in Edwards *et al.*, 1968). His subjects misperceived the impact of each datum, and, in addition, were not consistent with that misperception in a subsequent aggregation task.

Additional evidence for the misaggregation hypothesis was given by Hammond, Kelley, Schneider, and Vancini (1967) who found that the log likelihood ratios inferred from subjects' $P(D/H)$ estimates did not match the log likelihood ratios inferred from these subjects' $P(H/D)$ estimates. Similar results were recently obtained by Grinnell, Keeley, and Doherty (1970).

Finally, man's difficulties in aggregating data have been demonstrated in a series of man-machine systems studies. A system where men estimate $P(D/H)$ separately for each datum and the machine combines these into posterior probabilities via Bayes' theorem has consistently been found superior to a system where the man, himself, must aggregate the data into a $P(H/D)$ estimate (Edwards, Phillips, Hays, & Goodman, 1968; Kaplan & Newman, 1966; and Schum, Southard, & Wombolt, 1969).

Artifact

The third explanation of conservatism, that conservatism is artifactual, was originally suggested by Peterson (see Edwards, 1968). A similar explanation, called *response bias*, was advanced by DuCharme (1970). DuCharme hypothesized that subjects are capable—and optimal—when dealing with responses in the odds range from 1:10 to 10:1, but are conservative when forced, either by the accumulation of many data or by the occurrence of one enormously diagnostic datum, to go outside that range. He pointed out that such a response bias would explain many of the conservatism effects reported in the literature, including increased conservatism attributed to increasing diagnosticity and the superiority of first-trial performance. DuCharme tested his hypothesis directly in a task where subjects had to determine whether observed samples of heights came from a male or female population. His subjects gave

sequential posterior odds estimates to sequences varying in length from one to seven data. The results supported the response bias hypothesis. First-trial estimates and later-trial estimates in the same probability range were similarly optimal. Second- and third-trial estimates were more conservative following a highly diagnostic first datum (LR = 99) than were estimates to those same data following an undiagnostic first trial (LR = 1.3). Optimality of response was exhibited within a central range of posterior odds, while conservatism occurred outside this range.

TASK DETERMINANTS IN BAYESIAN RESEARCH

The Effects of Response Mode

Direct estimation methods. Direct estimation of posterior probabilities has several drawbacks. First, the amount of change in $P(H/D)$ induced by a single datum decreases as the probabilities prior to the receipt of that datum become more extreme. Subjects may have difficulty coping with this nonlinear relationship between stimulus and response. In addition, there is a potential problem with floor and ceiling effects because of the boundedness of the probability scale at zero and one. The subject may be reluctant to give an extreme response early in the sequence, for fear of "using up" the scale before the last data arrive. If subjects estimate odds or log odds (odds spaced logarithmically on a scale), the above difficulties are avoided. Phillips and Edwards (1966) compared these various response modes in a binomial task and found, as expected, that "odds" and "log odds" responses showed less conservatism than did estimates of probabilities.

Indirect methods. Instead of asking the subject for probabilities, indirect methods infer his probabilities from some other response. Sanders (reported in Edwards, 1966) used bookbag and poker chip situations to compare a direct response, verbal odds, with two different indirect responses, choice among bets and bidding for bets. He found substantial agreement, as measured by similarity of accuracy ratios, between the direct, verbal estimates of odds and the estimates inferred from choices among bets. The bidding mode produced considerably more optimal behavior than the other two modes.

Beach and Phillips (1967) compared direct probability estimates with probabilities inferred from choices among bets, and found high correlations between the two responses. Strong agreement between probability estimates and probabilities inferred from bids has also been found in two studies by Beach and Wise (1969a, 1969b). However, Beach and Olson (1967) have shown that probabilities inferred from choices among bets were highly susceptible to the gambler's fallacy (e.g., subjects overestimated the probability of a red after four greens were sampled, and

underestimated it after four reds occurred), while direct estimates of probabilities were much more optimal. Geller and Pitz (1968) have explored the use of decision speed, measured without the subject's knowledge, as an indirect measure of probability in a bookbag and poker chip task. A high correlation was found between the speed of decision and the Bayesian probability that the decision was correct. In addition, relative changes in decision speed approximated optimal changes in probability more closely than did changes in subjects' confidence estimates.

Effects of intermittent responding. Perhaps the very act of making repeated responses, once after each datum is presented, affects the final response of the subject. This hypothesis was tested by Beach and Wise (1969b), who compared verbal estimates of posterior probabilities made only at the end of a sequence of three data with estimates made after each datum. They found satisfactory correspondence between the two estimate methods. Pitz (1969a), however, did find differences attributable to repeated responses. When subjects responded after each new datum and their previous responses were displayed, confidence increased with increased sample size, holding diagnosticity constant. However, when the responses made after each datum were not continuously displayed, or when the subject made confidence responses only at the end of the sequence, such nonoptimal tendencies were not observed. Halpern and Ulehla (1970), using a signal detection task, also found differences between repeated responses and a single, final response.

Nominal vs. probability responses. Is there any difference between a nominal response (yes–no; predominantly red–predominantly blue) and a probability response which is later converted to a nominal response by the experimenter? Swets and Birdsall (1967), using an auditory detection task, found that the probability-response data provided a better fit to the signal detection model than the nominal-response data. Similar results were found by Ulehla, Canges, and Wackwitz (1967). However, Halpern and Ulehla (1970) found exactly the opposite results in a visual discrimination task.

Using a Bayesian task with three hypotheses, Martin and Gettys (1969) found better performance using a nominal response than using a probability response. Attaching probabilities to *two* less likely hypotheses as well as to the favored hypothesis was apparently difficult enough to degrade subjects' performance.

The Effects of Payoffs

The use of payoffs in probability estimation tasks may have a motivational effect, persuading the subjects to try harder, and an instructional effect, helping subjects to understand what the experimenter wants from

them (Winkler & Murphy, 1968). These effects were explored by Phillips and Edwards (1966), who used three different payoff schemes and a control group in a bookbag and poker chip task. The subjects estimated the posterior probability of each bag for 20 sequences of 20 draws each. The control group received no payoff but were told which hypothesis was correct after each sequence. The three payoff groups were paid $v(p)$ points, later converted to money, where p was the subject's estimate for the correct hypothesis, and $v(p)$ was calculated as follows:

$$\begin{aligned} &\text{Quadratic:} && v(p) = 10{,}000 - 10{,}000\,(1-p)^2. \\ &\text{Logarithmic:} && v(p) = 10{,}000 + 5{,}000\,\log_{10} p. \\ &\text{Linear:} && v(p) = 10{,}000p. \end{aligned}$$

The quadratic and log payoffs share the characteristic that the only way the subject can maximize his expected winnings is by reporting his true subjective probability (Toda, 1963). For the linear payoff, the subject should always estimate 1.0 for the more likely hypothesis. The results indicated that payoffs help to decrease conservatism, but do not eliminate it. The instructional value of payoffs was reflected in more learning by the payoff groups than the control groups, and by the lower between-subject variance for the payoff groups.

These findings were amplified in a study by Schum, Goldstein, Howell, and Southard (1967) using a complex multinomial task with six hypotheses and 4, 8, or 12 data, of varying diagnosticity, in each of 324 sequences. A log payoff group was found to be conservative. A linear payoff group was not conservative, but their responses were highly variable: their posterior odds were as likely to be 50 times too great or too small as they were to be accurate. When the responses were simply scored as "correct," meaning that the true hypothesis received the largest estimated posterior probability, or "incorrect," differences among the payoff groups were eliminated.

Pitz and Downing (1967) manipulated payoffs in a binary prediction task, to test subjects' sensitivity to changes in strategy required by the optimal model. Subjects were asked to guess which of two specially constructed dice was being rolled, after five data were presented. Five different payoff matrices were used. The first matrix was symmetric, in that rewards and penalties were the same for both dice. The other matrices were biased. In order to maximize their expected winnings, subjects should alter their strategies when payoff matrices are biased. For example, they should guess the less likely die when the reward for being correct is great and the cost for being wrong is small. The subjects were highly optimal when using the symmetric matrix. With the biased payoff matrices, they altered their predictions as a function of varying payoffs, but they did not change nearly enough; they were unwilling to make

responses which had a smaller probability of being correct, even though, because of the biased payoffs, these responses would have increased their expected gains. Pitz and Downing suggested that subjects have a high utility for making a correct guess. A similar suggestion was made by Ulehla (1966), who found essentially the same result in a study of perceptual discrimination of lines tilted left or right. With a symmetric payoff scheme, subjects closely fit the signal detection model, but biased payoffs led to insufficient change in strategy.

The Effects of Diagnosticity

One of the simplest ways of varying the diagnosticity of the data in a probability estimation task is to change the data generator. In a bookbag and poker chip experiment, the diagnostic impact of a sample of one red chip is greater when the bag being sampled contains 80 red and 20 green *vs.* 20 red and 80 green than when the possible contents of the bag are more similar, say 60 red and 40 green *vs.* 40 red and 60 green. In several experiments (Peterson, DuCharme, & Edwards, 1968; Peterson & Miller, 1965; Phillips & Edwards, 1966; Pitz, Downing, & Reinhold, 1967; and Vlek, 1965) diagnosticity has been manipulated in this way and all have shown greater conservatism with more diagnostic data. Very low levels of diagnosticity sometimes produce the opposite of conservatism: subjects' responses are more extreme than Bayes' theorem specifies (Peterson & Miller, 1965).

When the data generator is a complex multinomial system, different samples can differ greatly in total diagnosticity, i.e., in the certainty with which the sample points to one of several hypotheses. Studies by Martin and Gettys (1969), Phillips, Hays, and Edwards (1966), and Schum, Southard, and Wombolt (1969) all showed that samples of higher overall diagnosticity lead to greater conservatism. Martin and Gettys found that their least diagnostic samples produced the same extremeness of response (opposite of conservatism) as found by Peterson and Miller (1965) in a binomial task.

Another way of varying diagnosticity is to vary sample size. In general, the larger the sample, the more diagnostic it is. Vlek (1965), Pitz, Downing, and Reinhold (1967), and Peterson, DuCharme, and Edwards (1968), using binomial tasks, and Schum (1966b, also Schum, Southard, & Wombolt, 1969), using a multinomial task, have shown that larger sample sizes yield greater conservatism. Diagnosticity can be held constant across different sample sizes, however. In any symmetric binomial task, diagnosticity is solely a function of the difference between the number of occurrences of one type and of the other type. Thus the occurrence of four reds and two blues in a sample of six chips has the same diagnosticity as the occurrence of 12 reds and 10 blues in a sample

of 22 chips. Studies by Vlek (1965) and Pitz (1967) showed that when this difference was held constant, the larger sample sizes yielded lower posterior estimates, hence greater conservatism. However, when Schum, Southard, and Wombolt (1969) held diagnosticity constant in a multinomial task, variations in sample length had no effect upon the size of subjects' final posterior probability estimates. The method used for holding sample diagnosticity constant as sample size increases differed in the binomial and multinomial task. In the binomial task, diagnosticity is held constant by holding constant the difference between red and blue chips, as the sample size increases. But in the multinomial task used by Schum *et al.*, the total diagnosticity of a large sample was equated with that of a small sample by using data in the large sample each of which was, on the average, less diagnostic than the average datum used in the small sample. Since data of low diagnosticity have been shown to produce less conservatism, this may account for the discrepancy between Schum's findings of no sample-size effect and the finding of large effects by Vlek (1965) and Pitz (1967).

Sample size and diagnosticity can also be varied by holding the total number of data constant and varying the number of data presented to the subject at any one time. Peterson, Schneider, and Miller (1965) presented subjects with 48 trials of one datum each, with 12 trials of four data each, with four trials of 12 data each, and with a single trial containing 48 data. Conservatism was large when subjects responded after each single datum, but was even larger when the number of data (and hence the average diagnosticity) per trial increased. Vlek (1965) also found poorer performance with larger blocks of data presented simultaneously.

All these studies tell the same story: Increased diagnosticity, no matter how produced, increases conservatism. The sole exception to this statement is reported by Schum and Martin (1968), who used a multinomial task–six hypotheses and six data per sample. They used two different data-generating models, Model A and Model B. Diagnosticity was varied both within and between the two models. The results for Model A were typical of diagnosticity studies—subjects were sensitive to changes in diagnosticity within the model, but as diagnosticity increased, subjects became increasingly conservative. The results from Model B represented a unique finding; as diagnosticity increased in Model B samples, the responses deviated further from Bayes' theorem, but in a nonconservative direction (i.e., they became increasingly more extreme). This finding is unexplained by Schum and Martin. One possible explanation is that subjects completely disregarded the difference between Model A and Model B, responding solely to the number of items favoring the most likely hypothesis. The subsequent comparison of such responses with the optimal responses derived from the two different models would make

similar responses look conservative in one case and extreme in the other case.

The Effects of Manipulating Prior Probabilities

A decision makers' beliefs about the state of the world before he receives relevant information are expressed in the prior probabilities he attaches to the hypotheses. Prior probabilities may be manipulated in a bookbag and poker chip experiment, e.g., by showing the subject not two, but ten bookbags, n of which are predominantly red and $(10 - n)$ predominantly blue. When the experimenter chooses one bag at random, the subject has reason to believe, before the first chip is drawn, that the prior probability of the bag being predominantly red in $n/10$. Variation of n thus constitutes variation of prior probabilities.

When subjects' responses are analyzed in terms of inferred likelihood ratios or Accuracy Ratios, such measures should not change when prior probabilities are varied. Phillips and Edwards (1966) and Schum (1966b) reported such invariance. However, Peterson and Miller (1965) did find a systematic relationship between prior probability and Accuracy Ratio in a binomial task. Across nine levels of prior probability, from .1 to .9, subjects' Accuracy Ratios increased (the subjects became less conservative) as the priors became more extreme (departed from .5). This finding, however, may be an artifactual result of the response mode—probabilities expressed with a sliding pointer on an equal-interval scale. If subjects simply moved the slider a constant amount, up for a black datum, down for a white datum, regardless of its initial setting, the reported relationship between the Accuracy Ratio and prior probabilities would occur.

The one general characteristic of the Bayesian research summarized so far is that subjects are never as sensitive to the experimental conditions as they ought to be. This statement characterizes conservatism itself, as well as the effects of payoffs and diagnosticity. Partial sensitivity to variations in prior probabilities has been found using signal detection models by Ulehla (1966) and Galanter and Holman (1967). Wendt (1969) found partial sensitivity to prior odds. He asked his subjects to bid for each datum; this bid was interpreted as the value of the datum for the subject. Wendt found that the bids were closer to optimal when the prior odds were 1:1 than when the prior odds were extreme.

The Effects of Sequence Length

Several studies have found that subjects are more hesitant to commit themselves fully to a probability revision when they know that there will be opportunity for additional revision on later trials than when they know any revision taking place must be made immediately. Vlek (1965) compared $P(H/D)$ estimates made after the ninth trial in a 19-trial

sequence with estimates made after the simultaneous presentation of nine data items (no more were to be presented). The probability estimates were less extreme in the former condition where subjects knew they had ten additional opportunities for revisions. This effect might be attributed to the difference between simultaneous *vs.* serial presentation in the above study. However, Pitz, Downing, and Reinhold (1967) used serial presentation with responding after each item and found the average revision of $P(H/D)$ to be greater for shorter sequences than for longer ones. Similarly, Shanteau (1970) found that shorter sequences produced more extreme $P(H/D)$ responses at any serial position, holding the evidence constant. Although none of the above studies put any pressure on subjects to make their intermediate responses maximally accurate, Roby (1967) used a payoff system to motivate subjects to be accurate at every response point and he, too, found that they tended to delay for several trials before modifying their estimates. These results should be viewed with caution, however, because each of these studies employed a response mode highly susceptible to ceiling effects. Replication of these findings using odds or log odds response modes is needed to discover if subjects are affected by sequence length even when they feel comfortable that they cannot "use up" the response scale.

Primacy and Recency Effects

Studies investigating sequential use of probabilistic information have generally required subjects to make judgments after each new datum was presented. Three of these studies have reported primacy effects (Peterson & DuCharme, 1967; Dale, 1968; Roby, 1967) and two have obtained recency (Pitz & Reinhold, 1968; Shanteau, 1970). Both primacy and recency violate the Bayesian model. The occurrence of primacy here contrasts with the recency effects generally found when subjects make intermittent judgments upon receipt of verbal or perceptual information. One possible explanation for this is the fact that studies showing primacy effects each presented subjects with a long sequence of items of information that first pointed strongly to one hypothesis and then suddenly changed in character so that the less favored hypothesis became at least as probable as the first. The resulting inconsistency of the latter data is extremely implausible in a stationary environment, and it is not surprising that subjects tended to discount those data. Neither of the two studies obtaining recency effects used such inconsistent sequences of data.

An Inertia Effect in Bayesian Research

Anderson (1959) invoked the concept of inertia in discussing the "basal component" of an opinion—that part which becomes increasingly resist-

ant to change as information accumulates. More recently Pitz and his associates have conducted a series of studies demonstrating the existence of inertia in opinions that are formed and revised on the basis of probabilistic evidence. Pitz, Downing, and Reinhold (1967) found that subjects revised their $P(H/D)$ estimates much less following evidence contradictory to their currently favored hypotheses than they did after confirming evidence. Revision should have been equal in either direction. Especially interesting was the finding that probability estimates sometimes moved towards greater certainty after a single disconfirming datum was observed. This phenomenon, labeled an "inertia effect," was also found by Geller and Pitz (1968).

Geller and Pitz investigated two possible explanations of the inertia effect. The first was that inertia stems from strong commitment to a hypothesis whereby subjects become unwilling to change their stated level of confidence even though their opinions might change. This hypothesis was suggested by findings in studies by Gibson and Nichol (1964), Brody (1965), and Pruitt (1961). Pruitt found that subjects required more information to change their minds about a previous decision than to arrive at that decision in the first place. Brody found that initial commitment to an incorrect decision slowed down the rate of increase in confidence for the correct choice. Geller and Pitz obtained data indicating that subjects' speed of decision decreased markedly following disconfirming evidence even though the stated confidence in that decision had not decreased. They argued that this supported the commitment hypothesis and also concluded that stated confidence may not indicate the subject's true opinions. A second hypothesis tested by Geller and Pitz was that subjects may expect an occasional disconfirming event to occur when information is probabilistic. For example, if the task is to determine whether the samples of marbles are coming from an urn that is 60% red and 40% blue or vice versa and the first nine draws produce six red and three blue marbles, the drawing of a blue on the next trial may not be upsetting to subjects who believe the urn to contain 40% blue marbles. When subjects were asked to predict the next event in the sample, Geller and Pitz found that the inertia effect was greater following predicted disconfirming events than nonpredicted disconfirming events, and this was taken as support for the second hypothesis.

Further evidence for the commitment hypothesis comes from a study by Pitz (1967). His subjects stated their confidence in their opinions only after an entire sample was presented. When confidence was plotted as a function of increasing sample size, with Bayesian probabilities held constant, mean confidence judgments decreased, rather than increasing as would be predicted from the inertia effect. This lack of inertia was at-

tributed to the fact that there was no prior judgment to which subjects would have been committed. A later study (Pitz, 1969a) found that when subjects were not allowed to keep track of their trial-by-trial responses, inertia was eliminated.

Pitz (1966) had subjects make sequential judgments of the proportion of particular events in a sample. When subjects' previous judgments were displayed to them or could be recalled, their estimates showed a delay in revision towards .5 that seems analogous to the inertia effect found in studies of confidence or subjective probability. Here, too, a group whose previous judgments were not displayed showed no such effect.

LEARNING TO USE INFORMATION

There has been considerable investigation into the learning of information processing and judgmental skills. Our focus here will be on studies in which the subject has to learn to use information to make a prediction or judgment. We shall neglect a rather sizable literature that explores whether subjects can learn to detect correlational or probabilistic contingencies among events but does not require that this knowledge be used in decisions.

Regression Studies of Learning

Researchers working within the regression framework, and in particular with the lens model, have been quite interested in learning. In fact, learning could be categorized, along with the problem of modeling, as a focal topic within the correlational paradigm. One way to partition the studies that have been conducted is according to whether subjects had available only one cue or multiple cues in the learning task.

Single-cue learning. Research with single cues has focused upon what Carroll (1963) has called "functional learning." Carroll attempted to discover whether subjects could learn the functional relationships between a scaled cue or stimulus variable, X, and a scaled criterion, Y. The environment was deterministic; i.e., there was a perfect 1–1 correspondence between all values of X and Y. Across tasks, Carroll varied the mathematical complexity of the functions as determined by the number of parameters needed to describe them. He found that subjects' responses seemed to follow continuous subjective functions, even when the stimuli and criterion feedback were randomly ordered. Not surprisingly, simple functions were learned more accurately. Later work by Björkman (1965) and Naylor and Clark (1968) centered around the relative ease of learning positive *vs.* negative linear functions both in deterministic and probabilistic settings. The results of these studies indicated that positive relationships between cue and criterion are learned much more readily than negative ones.

Björkman (1968) defined "correlation learning" as functional learning where error ($R_e < 1.00$) was involved. He observed that correlational learning requires a subject to learn both the function relating stimulus and response, and the probability distributions around this function. In one experiment he found that the variance of a subject's responses about his own regression curve decreased as a consequence of training. A second experiment varied the extent to which there was a definite function to learn. Conditions with less pronounced cue-criterion trends resulted in larger ratios of subjects' response variance to criterion variance. From these results, Björkman concluded that correlational tasks are learned through a two-stage process involving both functional learning and probability learning, with the former occurring temporally prior to the latter.

Conservatism in single-cue learning. Do subjects in single-cue learning experiments exhibit conservatism such as occurs in Bayesian studies of performance? The results of several studies have been brought to bear upon this question but they must be viewed cautiously because of the problems in assessing conservatism in correlational tasks. For example, Naylor and Clark (1968) measured conservatism by dividing the stimulus distribution into thirds and computing the variance of each subject's responses within each third of the range. These variances were compared with the variances of the criterion values computed over the same subranges. The assumptions underlying this measure are (a) that the criterion distribution reflects the true probabilities of the various hypothesis states within each subrange of cue values and (b) that a subject's distribution of point responses represents an adequate picture of his perceived subjective probabilities for each of these hypothesis states. Given these assumptions, Naylor and Clark's subjects were conservative, inasmuch as the average dispersion of their judgments was found to exceed the dispersion of the criterion values in the upper and lower thirds of the cue distribution.

Naylor and Clark also proposed that the standard error of estimate ($\sqrt{1 - R_s^2}$) could be taken as an index of conservatism. Conservatism was presumed to increase this index, leading subjects to scatter their responses rather than consistently predicting the same criterion value, given a particular cue value. By this measure, Naylor and Clark's subjects, as well as subjects in studies by Björkman (1965), Gray (1968), Gray, Barnes, and Wilkinson (1965), were not conservative. In these studies, R_s typically exceeded R_e and the discrepancy ($R_s - R_e$) was inversely related to R_e. Thus the two measures proposed by Naylor and Clark lead to opposite conclusions about conservatism.

Brehmer and Lindberg (1970) have criticized the above conclusions, arguing that conservatism really means that subjects do not change their inferences as much as they should when the cue values change. They

argued that the indices used by Naylor and Clark confound two sources of variance—the consistency of the subjects and their conservatism or extremeness. Therefore, Brehmer and Lindberg proposed that conservatism be assessed by the relationship between b_e and b_s, the slopes of the regression lines relating the criterion values and judgments to the cue dimension.[5]

The experiments by Gray (1968), Gray *et al.* (1965), and Naylor and Clark (1968) found that b_s exceeded b_e for low values of R_e (and b_e) but not for high values. Since R_e and b_e were confounded in these studies, Brehmer and Lindberg decided to vary R_e, holding b_e constant. Lower values of R_e simply had greater deviation about a regression line that was the same for each condition. They found that subjects' judgments were consistently more extreme than the criterion values; i.e., b_s was greater than b_e. This was especially true when R_e was low. This result, along with similar findings by Gray, and Naylor and Clark, was interpreted as indicating that subjects are not conservative in this type of task.

Multiple-cue learning. Multiple-cue research is assumed to have relevance for a variety of "real-world" situations in which an individual must integrate information from several sources. Most of the studies rely upon the lens model for conceptual and analytical guidance. Independent variables are the number of cues, their $r_{i,e}$ values and the multiple correlation, R_e, the forms of the functional relationships between cues and criterion, and the intercorrelation between cues. Typically, the subject is presented with a set of cues, he makes a quantitative judgment on the basis of these cues, and then receives the criterion value as feedback. Among the major results are (a) subjects can learn to use linear cues appropriately (Lee & Tucker, 1962; Smedslund, 1955; Summers, 1962, and Uhl, 1963); (b) learning of nonlinear functions occurs but is slower and less effective than learning of linear relationships (Brehmer, 1969a; Hammond & Summers, 1965; Sheets & Miller, in press; Summers, 1967; and Summers, Summers, & Karkau, 1969) and is especially difficult if subjects are not properly forewarned that the relations may be nonlinear (Earle, 1970; Hammond & Summers, 1965; and Summers & Hammond, 1966); (c) subjects can learn to detect changes in relative cue weights over time although they do so slowly (Peterson, Hammond, & Summers, 1965a; Summers, 1969); (d) it is easier for subjects to learn which cue to use than to discover which functional rule relates a known valid cue to the criterion; learning both of these simultaneously is

[5] In single-cue studies, $r_{i,e}$ is equivalent to R_e; similarly, $r_{i,s}$ equals R_s, while $b_{i,e}$ and $b_{i,s}$ equal b_e and b_s, respectively.

especially difficult (Summers, 1967, 1969); (e) in a two-cue task, pairing a cue of low or medium validity with one of high validity is detrimental to performance (a distraction effect), while pairing a cue of low validity with another of medium or low validity is facilitative (Dudycha & Naylor, 1966b); and (f) subjects can learn to use valid cues even when they are not perceived with perfect reliability (Brehmer, 1970).

Cue intercorrelations (redundancies) have been varied in several learning studies (Armelius & Lenntoft, 1970; Miller & Sarafino, in press; Naylor & Schenck, 1968). The major result is that subjects' beta weights match the cue-criterion correlations rather than the cue-criterion beta weights. Thus subjects fail to take appropriate account of redundancies.

Conservatism has not been an explicit concern in many multiple-cue learning studies. However, Peterson, Hammond, and Summers (1965b) found that subjects failed to weight the most valid of three cues heavily enough and slightly overweighted the cue with lowest validity. Peterson *et al.* noted the similarity of these results to those of Bayesian performance tasks in which conservatism and data diagnosticity are positively related.

A few studies have investigated the effects of different modes of feedback upon correlational learning. When subjects are given the correct answer on every trial ("outcome feedback"), learning is relatively slow. Lens model feedback, indicating how a subject's cue utilization coefficients compare with the ecological validities, is far more effective (Hammond & Boyle, 1970; Newton, 1965; Todd & Hammond, 1965). Magnusson and Nystedt (1969) found that providing subjects with the ecological validities was more effective than providing feedback about their cue-utilization coefficients.

The lens model paradigm has also been extended to the problem of analyzing interpersonal learning and conflict between pairs of individuals (Hammond, 1965; Hammond & Brehmer, in press; Hammond, Todd, Wilkins, & Mitchell, 1966; Hammond, Wilkins, & Todd, 1966; Rappoport, 1965). A typical experiment trains pairs of subjects to use one of two cues in either linear or nonlinear fashion. Each subject learns to use a different cue, perhaps in different ways as well. After training, subjects are brought together to learn to predict a new criterion, using the same cues. Typically both cues must be used in this second task, and the subjects' training leads them initially to disagree with one another and with the outcome feedback they receive from the task. Lens model analysis of each subject's individual judgments and the pair's joint judgments provides a great deal of information about the mechanisms whereby subjects learn from the task and from one another. A study by Brehmer (1969b) found that the differences between subjects' weighting policies are rapidly

reduced in the joint task but this reduction is accompanied by increased inconsistency such that overt discrepancies are not very much diminished by the end of the conflict period. Because of such findings, Hammond and Boyle (1970) and Hammond and Brehmer (in press) argued that it is necessary to invent methods to display to the subjects the real sources of their disagreement and both of these studies describe a computer system which does this. Another interesting result from this area is that persons initially trained to have nonlinear policies are more likely to change than are persons with more linear policies (Brehmer, 1969b; Earle, 1970).

Bayesian Studies of Learning

Bayesian researchers have been notably uninterested in the topic of learning. Many Bayesian studies have used situations like bookbags and poker chips, with which the experimenters assume the subject is already familiar. Others (e.g., Lichtenstein & Feeney, 1968) have given initial training trials, with feedback; however, this training data is usually not analyzed. The epitome of indifference to learning is illustrated in a study by Peterson (1968). Though his subjects responded to more than 8000 four-data sequences, Peterson did not mention whether feedback was given (presumably it was not), and all analyses were based on all the data, without any attention paid to changes over time. Peterson, like most other Bayesian researchers, was interested in how subjects *behave*—not how they learn. Nonetheless, a few Bayesian studies do consider learning and thus merit attention.

The effects of feedback. Edwards, Phillips, Hays, and Goodman (1968) reported a study which compared two groups of subjects who gave likelihood ratio responses; these responses were then cumulated, that is, converted into posterior odds estimates, by the experimenters, using Bayes' theorem. One group received feedback of these cumulated posterior odds after each estimate; the other group received no feedback. This type of feedback was found to degrade the cumulated posterior odds—making them more conservative—although changes over time were not reported.

Martin and Gettys (1969) gave subjects either nominal feedback (e.g., H_1 generated the data) or probabilistic feedback (e.g., the posterior probabilities that each hypothesis generated the data are .769 for H_1, .108 for H_2, and .123 for H_3) in a multinomial task. These authors found that probabilistic feedback produced more optimal responses than nominal feedback, but they found no evidence that learning had occurred, either across four blocks of 50 trials, within the first 50 trials, or in a 20-trial replication. However, learning may have occurred in the five-pre-experimental practice trials.

The effects of payoff. Phillips and Edwards (1966) presented 20 sequences of binomial data to three groups, each with different payoff schemes, and to one group which received no payoff. They found that the no-payoff group showed a small amount of learning (decreasing discrepancy from optimal responses); all payoff groups showed more learning, with no evidence of asymptote by the end of the experiment. Performance showed greater improvement in the latter half of these 20-item sequences than in the first half, suggesting that the subjects learned to use large probabilities as the evidence for one hypothesis mounted.

Learning specific aspects of a probabilistic setting. Schum (1966a) showed that subjects can learn and utilize existing conditional nonindependence in multinomial data. The subjects were warned which data sources might be nonindependent, but they were not told the form of the relationship, nor which of the hypotheses mediated the relationship. They were taught to tabulate the frequencies with which the data occurred in such a way that the nonindependence could be seen. Thus the outstanding achievement of the subjects was not that they could learn what interdependencies existed, but that they could utilize this information appropriately in their posterior probability estimates; their responses more closely matched a model utilizing the nonindependence than a model in which independence was falsely assumed.

Two additional learning studies were oriented to the misperception explanation of conservatism. In order to strengthen the point that subjects' conservatism resulted from their misunderstanding of the kinds of samples to expect from a given population, Peterson, DuCharme, and Edwards (1968) found that subjects were less conservative after they had been shown 100 illustrative samples of data from a binomial population. Wheeler and Beach (1968) not only showed their subjects 200 binomial samples, but they asked the subjects to make a bet on which population generated the data, for each sample. Outcome feedback was given immediately after each bet. The effects of such training were seen in increased accuracy of subjects' estimated sampling distributions and decreased conservatism.

DESCRIPTIVE STRATEGIES: A SEARCH FOR ALTERNATIVE SUBJECTIVE COMPOSITION RULES

Thus far we have tied our presentation of theoretical notions and empirical results rather closely to the Bayesian and regression paradigms. In doing so, we have accepted the validity of their models rather uncritically as descriptive indicators of cognitive processes. However, despite the fairly adequate global fit provided by these models, close examination of judgmental data often reveals discrepancies that may

carry important theoretical implications. In this section we shall discuss some of the alternative subjective composition strategies suggested by these discrepancies.

Strategies in Correlational Research

Starting-point and adjustment strategies. The present authors have recently conducted several experiments that seem to provide insight into the cognitive operations performed by decision makers as they attempt to integrate information into an evaluative judgment. In a study by Slovic and Lichtenstein (1968), the stimuli were gambles, described by four risk dimensions: probability of winning (P_W), amount to win ($\$_W$), probability of losing (P_L), and amount to lose ($\$_L$). One group of subjects was asked to indicate their strength of preference for playing each bet on a bipolar rating scale. Subjects in a second group indicated their opinion about a gamble's attractiveness by equating it with an amount of money such that they would be indifferent between playing the gamble or receiving the stated amount. This type of response is referred to as a "bid." The primary data analysis consisted of correlating each subject's responses with each of the risk dimensions across a set of gambles. These correlations indicated that the subjects did not weight the risk dimensions in the same manner when bidding as when rating a gamble in monetary units. Ratings correlated most highly with P_W, while bids were influenced most by $\$_W$ and $\$_L$.

Both bids and ratings presumably reflect the same underlying characteristic of a bet, viz., its worth or attractiveness. Why should subjects employ probabilities and payoffs differently when making these related responses? The introspections of one individual in the bidding group are especially helpful in providing insight into the type of cognitive process that could lead bidding responses to be overwhelmingly determined by just one payoff factor. This subject said,

> If the odds were . . . heavier in favor of winning . . . rather than losing . . . , I would pay about ¾ of the amount I would expect to win. If the reverse were true, I would ask the experimenter to pay me about . . . ½ of the amount I could lose.

Note this subject's initial dependence on probabilities followed by a complete disregard for any factor other than the winning payoff for attractive bets or the losing payoff for unattractive bets. After deciding he liked a bet, he used the amount to win, the upper limit of the amount he could bid, as a starting point for his response. He then reduced this amount by a fixed proportion in an attempt to integrate the other dimensions into the response. Likewise, for unattractive bets, he used

the amount to lose as a starting point and adjusted it proportionally in an attempt to use the information given by the other risk dimensions. Such adjustments, neglecting to consider the exact levels of the other dimensions, would make the final response correlate primarily with the starting point–one of the payoffs in this case.

It is interesting to note that this starting point and adjustment process is quite similar to the fixed-percent markup rule that businessmen often use when setting prices (Katona, 1951). This type of process can be viewed as a cognitive shortcut employed to reduce the strain of mentally weighting and averaging several dimensions at once.

The observation of simple starting point and adjustment procedures in bidding and pricing judgments has led the first author to examine the strategies by which subjects average two numerical cues into an evaluative judgment. Preliminary analysis of the data indicates that, even in this relatively simple task, subjects tend to use a single cue as a starting point for their judgment. Next, they adjust this starting judgment rather imprecisely in an attempt to take the other cue into account. These data suggest that the subjects, although college students of above average intelligence, resorted to simple strategies in order to combine the two cue values. They were not skilled arithmeticians, able to apply regression equations or produce weighted averages without computational aids.

Strategies in multiple-cue learning. Close examination of multiple-cue learning studies provides further evidence for the occurrence of simple strategies. For example, Azuma and Cronbach (1966) studied the manner in which subjects learned to predict a criterion value on the basis of several cues. When subjects' responses were correlated with the cue values over blocks of trials, the results indicated an orderly progression towards proper weighting of the cues. However, when successful learners were asked to give introspective accounts of the process by which they made their judgments, these reports bore little resemblance to the weighting function employed by the experimenters. Instead they typically described a sequence of rather straightforward mechanical operations. Azuma and Cronbach observed that, although the experimenter regards the universe of stimuli as an undifferentiated whole, their subjects isolated subuniverses and employed different rules within each of these. The imposition, by the experimenter, of a correlational composition model may obscure the more local rules used by the subjects.

Strategies for Estimating $P(H/D)$

Students of the theory of probability have been continually amazed at its subtlety and the extent to which results derived from it conflict

with their intuitive expectations. Nevertheless, a recent review by Peterson and Beach (1967) concerning man's capabilities as an "intuitive statistician" came to an optimistic conclusion. Peterson and Beach asserted that:

> Experiments that have compared human inferences with those of statistical man show that the normative model provides a good first approximation for a psychological theory of inference. Inferences made by subjects are influenced by appropriate variables and in appropriate directions [Pp. 42–43].

Even the spectre of conservatism has failed to dampen the optimism of many Bayesian researchers who have attributed conservatism to erroneous subjective probabilities rather than an inadequate (i.e., non-Bayesian) processing of this information.

Our own examination of the experimental literature suggests that the Peterson and Beach view of man's capabilities as an intuitive statistician is too generous. Instead, the intuitive statistician appears to be quite confused by the conceptual demands of probabilistic inference tasks. He seems capable of little more than revising his response in the right direction upon receipt of a new item of information (and the inertia effect is evidence that he is not always successful in doing even this). After that, the success he obtains may be purely a matter of coincidence—a fortuitous interaction between the optimal strategy and whatever simple rule he arrives at in his groping attempts to ease cognitive strain and to pull a number "out of the air."

Constant Δp strategy. There are several simple strategies that seem to highlight subjects' difficulties in conceptualizing the requirements of probabilistic inference tasks and, at the same time, explain many of the ethereal phenomena that comprise the "conservatism" effect. The first such strategy is to revise one's $P(H/D)$ response by a constant, Δp, regardless of the prior probability of the hypothesis or the diagnosticity of the data. The strongest evidence for this strategy comes from Pitz, Downing, and Reinhold (1967). Subjects saw sequences of either 5, 10, or 20 data items and made a probability revision after each datum. Three different levels of data diagnosticity were employed, using a binomial task. The results indicated the usual inverse relationship between diagnosticity and conservatism, with some subjects overreacting to data of low validity. Longer sequences produced greater conservatism. Pitz *et al.* noted that events which confirmed the favored hypothesis resulted in approximately equal changes in subjective probability, regardless of a subject's prior probability. There was little difference between changes for sequences of lengths 5 and 10, but the average change for sequences of length 20 was considerably lower, as

if subjects were holding back in anticipation of a greater amount of future information. The experimenters also reported the "remarkable fact" that the average change was not a function of the nature of the two hypotheses but, instead, was approximately the same across the three levels of diagnosticity. They concluded with the observation that:

> The fact that changes in subjective probability were a constant function of prior probabilities, were independent of the nature of the hypotheses, yet were not independent of the length of the sequence of data, implies that a subject's performance in a probability revision task is nonoptimal in a more fundamental way than is implied by discussions of conservatism. Performance is determined in large part by task characteristics which are irrelevant to the normative model. . . . It may not be unreasonable to assume that . . . the probability estimation task is too unfamiliar and complex to be meaningful (Pitz, Downing, & Reinhold, 1967; p. 392).

This same sort of insensitivity to gross variations in diagnosticity is evident in studies by Martin (1969), Peterson and Miller (1965), Peterson, Schneider, and Miller (1965), and Schum and Martin (1968) and serves to explain many of their results.

Similarity strategies. The second type of strategy for making probability estimates appears in several studies. The subjects base their responses on the similarity of the sample data with whatever representative feature of the hypothesis seems most salient. This strategy was observed by Dale (1968) in a pseudo-military task. The values of $P(D_j/H_i)$ were displayed as histograms, one for each of the four hypotheses. Data categories were represented on the X-axis of these histograms and probabilities on the Y-axis. As the subjects received the data reports, they often physically arranged these reports to form a frequency histogram which they then compared with the four conditional probability displays. The relative magnitudes of their responses appeared to be based upon the similarity between the pattern formed by the data and the pattern formed by each of the conditional distributions. Dale noted that the subjects were at a loss to know what magnitude of probability to assign a given level of similarity. One subject, when he had assessed the probability of the correct hypothesis at .38 (the Bayesian probability was .98) remarked: "Getting mighty high!"

Lichtenstein and Feeney (1968) also observed a kind of similarity strategy. Their subjects were shown the locations of bomb blasts and had to estimate the probability that the intended target was City A or City B. The subjects were told that the errors were unbiased, in that a bomb was just as likely to miss its target in any direction. They were also told that a bomb was more likely to fall near its target than far from it. The subjects' responses were clearly discrepant from the

optimal responses derived from the circular normal data generator. Several subjects reported that they compared the distances of the bomb site from the two cities and based their estimates on this comparison, that is, on the similarity between the location of the datum and the locations of the cities. A model assuming that probability estimates were simply a function of the ratio of the two distances did a much better job of predicting the responses of most subjects than did the "correct" circular normal model.

Use of sample proportions. The results of several independent studies using binomial tasks suggest that the subjects were matching their $P(H/D)$ responses to the sample proportions. For example, Beach, Wise, and Barclay (1970), using a task with a simultaneous sample of items, found a remarkably close relationship between the sample proportion and the mean posterior probability estimates. Several of their subjects remarked that sample proportions were very compelling because they were available (and somehow relevant) numbers in a difficult and foreign task. A study by Kriz (1967) obtained similar results. Shanteau (1970) found that subjects gave essentially the same responses regardless of whether they were inferring $P(H/D)$ or estimating the population proportion. Subjects who confuse proportion estimation with inference will fail to take into account the likelihood of the data, as specified by the population proportions. Their inferences thus would not change across tasks that varied in population proportion (diagnosticity). This lack of sensitivity has been reported by Beach, Wise, and Barclay (1970) and by Vlek (1965), who suggested that ". . . subjects do not look further than the sample presented to them" (p. 22).

For the usual levels of diagnosticity found in binomial tasks, a strategy of using the sample proportion as an estimate of $P(H/D)$ will produce very conservative performance. Beach *et al.* (1970) concluded that this strategy is a spurious one that invalidates the bookbag and poker chip task as an indicant of subjective probability revision. It seems to us that this may be too harsh a judgment in light of the ubiquity of simple strategies for inference across a variety of laboratory and real-life judgment situations.

AIDING THE DECISION MAKER

Experimental work such as we have just described documents man's difficulties in processing multidimensional and probabilistic information. Unfortunately, there is abundant evidence indicating that these difficulties persist when the subject leaves the artificial confines of the laboratory and resumes the task of using familiar sources of information to make decisions that are important to himself and to others. Examples

of overly simplistic use of information have been found in business decision making (Katona, 1951), military decision making (Wohlstetter, 1962), governmental policy (Lindblom, 1964), design of scientific experiments (Tversky & Kahneman, in press), and management of our natural resources (Kates, 1962; Russell, 1969; White, 1966). Agnew and Pyke (1969, p. 39) note that a decision maker left to his own devices ". . . uses, out of desperation, or habit, or boredom, or exhaustion, whatever decision aids he can—anything that prepackages information." Among the vast assortment of decision aids described by Agnew and Pyke are rumors, cultural biases and self-evident truths, common sense, appeals to authority, and appeals to experts who, themselves, are all too fallible.

The need for effective decision aids has not gone unnoticed, however. This is an age of technological advancement that creates more difficult and more important decision problems as it provides man with ever more power to manipulate his environment. It is not surprising, therefore, that this same technological bent has been focused upon the decision-making process itself. One interesting new development is the cognograph, a computer system that provides lens model feedback to the judge (Hammond & Boyle, 1970; Hammond, in press).

The aim of this section is to describe two other recent and distinctive contributions to the regression and Bayesian approaches to the improvement of decision making.

Probabilistic Information Processing Systems

A great deal of Bayesian research has centered about the use of probability assessments in applied diagnostic systems. Edwards (1962) introduced the notion of a probabilistic information processing (PIP) system because of his concern about the optimal use of information in military and business settings. He distinguished two types of probabilistic outputs for such a system. The first was diagnosis (what is the probability that this activity indicates an enemy attack?), and the second was parameter estimation (how rapidly is that convoy moving and in what direction?). Edwards proposed the following design for a PIP system: let men estimate $P(D/H)$, the probability that a particular datum would be observed given a specified hypothesis, and let machines integrate these $P(D/H)$ estimates across data and across hypotheses by means of Bayes' theorem. After all the relevant data have been processed, the resulting output is a posterior probability, $P(H/D)$, for each hypothesis. Edwards originally designed the PIP system with the intention of using Bayes' theorem as a labor-saving device. However, when research subsequently indicated that difficulties in aggregating

data led subjects' unaided posterior probability estimates to be markedly conservative, the need to develop an antidote for conservatism added considerable impetus to the development of such systems.

Edwards and Phillips (1964) promoted the PIP system as a promising alternative to traditional command and control systems. They hypothesized that this system would produce faster and more accurate diagnoses for several reasons. First, Bayes' theorem is an optimal procedure for extracting all the certainty available in data. It automatically screens information for relevance, filters noise, and weights each item appropriately. In addition, PIP systems promise to permit men and machines to complement one another, using the talents of each to best advantage.

Sometimes $P(D/H)$ values are readily calculable from historical information or from some explicit model of the data-generating device. However, in many cases, no such probabilities exist. For example, what is the probability that Russia would have launched 25 reconnaissance satellites in the last 3 days if she planned a missle attack on the United States? As Edwards and Phillips observed, only human judgment can evaluate this type of information; PIP systems attempt to obtain and use such judgments systematically.

The basic idea of a PIP system suggested a number of questions for research. Edwards and Phillips discussed the need to verify the basic premise that men can be taught to be good estimators for probabilities. One question concerned the most effective method for making such estimates. For example, should men estimate $P(D/H)$ values directly or estimate other quantities from which $P(D/H)$ can be inferred? Subsequent research indicated that it is easier to estimate likelihood ratios than to estimate $P(D/H)$ values themselves, because the latter are influenced by many irrelevant factors such as the level of detail with which the datum is specified (Edwards, Lindman, & Phillips, 1965).

Perhaps the most important research need was to evaluate the effectiveness of PIP systems in realistically complex environments. A number of such studies have been completed in recent years. One of the most extensive studies was by Edwards, Phillips, Hays, and Goodman (1968). They constructed an artificial future world (complete with "history" up to 1975). The subjects related sequences of data to six hypotheses concerning war within the next 30 days, e.g., H_1 was "Russia and China are about to attack North America," while H_6 was "Peace will continue to prevail." Four groups of subjects received intensive training in the characteristics of the "world," and then each group was trained in a particular response task. The *PIP group's* responses were likelihood ratios. To each datum, five ratios were given, comparing in turn the likelihood of the datum given each of the war hypotheses against

the likelihood of the datum given the peace hypothesis. The responses were registered on log-odds scales. The *POP group* responded with posterior odds, estimated upon receipt of each new datum. Again, each of the war hypotheses was compared in turn to the peace hypothesis. The *PEP group* responded by naming, for each war hypothesis, the fair price for an insurance policy that would pay 100 points in the event of that particular war, and nothing in the event of peace. The *PUP group* gave probability estimates comparable to the PEP group's price estimates. Thus, of the four groups, only the PIP group, who gave likelihood ratios, were relieved of the task of cumulating evidence across the data in each sequence. In this group, the aggregation was done by machine to compute posterior odds.

No optimal model could be devised for this simulation. The "true" hypothesis for any sequence of data was not known. Results showed that the PIP group arrived at larger final odds than other groups. When the PIP system showed final odds of 99:1, other groups showed final odds from 2:1 to 12:1. Because of this greater efficiency, the authors concluded that PIP was superior to the other systems.

The problem of finding a task complex enough to warrant the comparison of $P(D/H)$ responses (PIP) with $P(H/D)$ responses (POP), while still providing an optimal model against which to evaluate both methods, was tackled by Phillips (1966; also reported in Edwards, 1966). The data were thirty bigrams, combinations of two letters such as "th" or "ed." The hypotheses were that the bigrams were drawn either from the first two letters of words, or from the last two letters of words. The bigram "ed" might thus be viewed as beginning a word (like *ed*itor) or ending a word (like look*ed*). Phillips' subjects were six University newspaper editorial writers; data came from their own editorials. Frequency counts using the subjects' editorials (not shown to them) provided the veridical probabilities against which their responses could be compared. For the PIP task, all subjects estimated the likelihood ratio $(P(D/H_1)/P(D/H_2))$ for each bigram. Then, for the POP task, they were asked to imagine that the bigrams had been placed in two bookbags according to their frequencies of use, i.e., if "my" had occurred 20 times at the beginning of words and 40 times at the end of words, the 20 "my" bigrams were placed in bag B, and 40 in bag E. One bag was chosen by the flip of a coin, and 10 bigrams were successively sampled. The subjects gave posterior odds estimates after each draw. Following this POP task, they repeated the PIP task. Results showed that in the PIP task subjects were modestly successful at estimating the relative frequencies of their own use of bigrams, but five of the six subjects were conservative. In the POP task they were much more conservative;

they treated all but two of the bigrams as if they provided little or no diagnostic information.

Kaplan and Newman (1966) reported the results of three experiments designed to evaluate PIP in a military setting. In two of these studies the PIP technique showed a definite superiority over a POP condition. This superiority was particularly evident early in the data sequence. The authors speculated that the relatively poor performance of the PIP system in the other experiment may have been due to the fact that subjects there were provided with the output of Bayes' theorem after each datum was presented, making it difficult to evaluate each item of information on its own merit. Edwards, Phillips, Hays, and Goodman (1968) and Schum, Southard, and Wombolt (1969) also found a detrimental effect from showing $P(D/H)$ estimators the current state of the system.

A major effort to evaluate the idea of a PIP system within the context of threat evaluation has been carried out at Ohio State University under the direction of David Schum and his colleagues. The results are described in Briggs and Schum (1965), Howell (1967), Schum (1967, 1968, 1969), Schum, Goldstein, and Southard (1966), and Schum, Southard, and Wombolt (1969). The Ohio State research employed a situation in which the experimenters specified an arbitrarily constructed $P(D/H)$ matrix that governed the sampling of data. Subjects had to learn the import of various data items by accumulating relative frequencies linking data and hypotheses. The subjects were intensively trained in making probabilistic judgments and were quite familiar with the characteristics of the information with which they were dealing. Howell (1967) has summarized the first 6 years of research at Ohio State, concluding that automation of the aggregation process (i.e., PIP) can be expected to improve the quality of decisions in a wide variety of diagnostic conditions. He also observed that the superiority of a PIP system is most pronounced under degraded, stressful, or otherwise difficult task conditions.

In contrived or simulated diagnostic situations, the PIP system seems to be a promising device for improving estimates of posterior probabilities. Future work will undoubtedly see the extension of the system to nonmilitary settings along with greater attention to the practical details of implementing such systems in applied contexts. PIP systems have already been proposed for medicine (Lusted, 1968; Gustafson, 1969; Gustafson, Edwards, Phillips, & Slack, 1969) and probation decision making (McEachern & Newman, 1969), and applications to weather forecasting, law, and business seem imminent.

As promising as the PIP idea seems to be, however, a number of

serious problems have yet to be faced. Some problems of particular importance noted by Schum, Southard, and Wombolt (1969) include hypothesis definition, source unreliability (uncertainty about which datum is being observed), nonstationarities of the environment, and nonindependence of data. Schum (1969) observed that in systems where data accumulates rapidly, experts who assess $P(D/H)$ may have to aggregate their judgments over a series of data (i.e., judge $P(D_1, D_2, D_3, \ldots D_n/H_i)$. When data items are nonindependent, these conditional probabilities can become quite complex. Three experiments reported by Schum, Southard, and Wombolt (1969) found that highly trained subjects could adequately aggregate diagnostic import across small samples of conditionally nonindependent data. However, the subjects were given access to the relative frequencies of the lower order conditional probabilities and, as the experimenters' noted, the results gave little indication about "how one uses his educated intuition in assigning import to evidence in the absence of a relative frequency 'crutch' " (Schum *et al.*, 1969, p. 44).

Tversky and Kahneman have recently scrutinized subjects' "educated intuitions" about probabilities in a series of experiments (Tversky & Kahneman, 1970, in press; Kahneman & Tversky, 1970) and their results imply that man may be as poor at estimating $P(D/H)$ values as he is at estimating posterior probabilities. One of their hypotheses is that the number of instances of an event that are readily retrieved from memory or the ease with which they come to mind are major clues used for estimating the probability of that event. The "availability" of instances is affected by many subtle factors such as recency, salience, and imaginability, all of which may be unrelated to the correct probability. Tversky and Kahneman point out the implications of these biases for PIP systems and they suggest that informing the decision maker of his susceptibility to these influences might be valuable, although by doing so one risks imparting new biases.

Bootstrapping

Can a system be designed to aid the decision maker that is based on his own judgments of complex stimuli? One possibility is based on the finding that regression models, such as the linear model, can do a remarkably good job of simulating such judgments. An intriguing hypothesis about cooperative interaction between man and machine is that these simulated judgments may be better, in the sense of predicting some criterion or implementing the judge's personal values, than were the actual judgments themselves. Dawes (1970) has termed this phenomenon "bootstrapping."

The rationale behind the bootstrapping hypothesis is quite simple. Although the human judge possesses his full share of human learning and hypothesis generating skills, he lacks the reliability of a machine. As Goldberg (1970, p. 423) has noted:

> He 'has his days': Boredom, fatigue, illness, situational and interpersonal distractions all plague him, with the result that his repeated judgments of the exact same stimulus configuration are not identical. He is subject to all these human frailties which lower the reliability of his judgments below unity. And, if the judge's reliability is less than unity, there must be error in his judgments—error which can serve no other purpose than to attenuate his accuracy. If we could . . . [eliminate] the random error in his judgments, we should thereby increase the validity of the resulting predictions.

Of course, the bootstrapping procedure, by foregoing the usual process of criterion validation, is vulnerable to any misconceptions or biases that the judge may have. Implicit in the use of bootstrapping is the assumption that these biases will be less detrimental to performance than the inconsistency of unaided human judgment.

Bootstrapping seems to have been explored independently by at least four groups of investigators. Yntema and Torgerson (1961) reported a study that suggested its feasibility. Their subjects were taught, via outcome feedback, to predict a criterion that was nonlinearly related to the cues. After 12 days of practice, their average correlation with the criterion was found to be .84. Then a linear regression model was computed for each subject on the basis of his responses during the final practice day. When these models were used to predict the criterion, the average correlation rose to .89. Thus consistent application of the linear model improved the predictions, even though the subjects had presumably been taking account of nonlinearities in making their own judgments. Yntema and Torgerson saw in these results the possibility that "artificial, precomputed judgments may in some cases be better than those the man could make himself if he dealt with each situation as it arose" (p. 24). More recently, Dudycha and Naylor (1966b) have reached a similar conclusion on the basis of their observation that subjects in a multiple-cue learning task were employing the cues with appropriate relative weights but were being inaccurate due to the inconsistency of their judgments. They concluded that although humans may be used to generate strategies, they should then be removed from the system and replaced by their strategies.

Bowman (1963) outlined a bootstrapping approach within the context of managerial decision making that has stimulated considerable empirical research (see Gordon, 1966; Hurst & McNamara, 1967; Jones, 1967; and Kunreuther, 1969). Kunreuther, for example, developed a linear model

of production scheduling decisions in an electronics firm. Coefficients were estimated to represent the relative importance of sales and inventory variables across a set of decisions made by the production manager. Under certain conditions, substitution of the model for the manager was seen to produce decisions superior to those the manager made on his own.

At about the time that Bowman was proposing his version of bootstrapping, Ward and Davis (1963) were advocating the same kind of approach to man–computer cooperation. Although they presented no data, Ward and Davis outlined several applications of the method in tasks such as estimating the time it would take to retrain 500 people, who now hold 500 existing jobs, to 500 new, possibly different jobs. Here a model would be built to capture an expert judge's policy on the basis of a relatively small number of cases. The model could then be substituted for the expert on the remaining cases. Ward and Davis also outlined an application of bootstrapping for the purpose of assigning personnel to jobs so as to maximize the payoff of the assignments.

Goldberg (1970) evaluated the merits of bootstrapping in a task where 29 clinical psychologists had to predict the psychiatric diagnoses of 861 patients on the basis of their MMPI profiles. A linear model was built to capture the weighting policy of each clinician. When models of each clinician were constructed on the basis of all 861 cases, 86% of these models were more accurate predictors of the actual criterion diagnoses than the clinicians from whom the models were derived. There was no instance of a man being greatly superior to his model. When a model was constructed on only one-seventh of the cases and used to predict the remaining cases, it was still superior to its human counterpart 79% of the time. While the average incremental validity of model over man was not large, the consistent superiority of the model suggested considerable promise for the bootstrapping approach.

Another demonstration of bootstrapping comes from a study of a graduate student admissions committee by Dawes (1971). Dawes built a regression equation to model the average judgment of the four-man committee. The predictors in the equation were overall undergraduate grade point average, quality of the undergraduate school, and a score from the Graduate Record Examination. To evaluate the validity of the model and the possibility of bootstrapping, Dawes used it to predict the average committee rating for his sample of 384 applicants. The R_s value for predicting the new committee ratings was .78. Most important, however, was the finding that it was possible to find a cutting point on the distribution of predicted scores such that no one who scored below it was invited by the admissions committee. Fifty-five percent of the applicants scored below this point, and thus could have been elim-

inated by a preliminary screening without doing any injustice to the committee's actual judgments. Furthermore, the weights used to predict the committee's behavior were better than the committee itself in predicting later faculty ratings of the selected students. In an interesting cost-benefit analysis, Dawes estimated that the use of such a linear model to screen applicants to the nation's graduate schools could result in an annual savings of about $18,000,000 worth of professional time.

A recent paper by Dawes and Diller (1970) derives a formula based on observable properties of judgments that indicates when bootstrapping with linear models may be expected to occur. In addition, a procedure is presented for amalgamating the judge's predictions with the predictions of his linear model in such a way that the amalgamation is superior to both the judge and model.

CONCLUDING REMARKS

Some Generalizations about the State of our Knowledge

What have we learned about human judgment as a result of the efforts detailed on the preceding pages? Several generalizations seem appropriate. First, it is evident that the judge responds in a highly predictable way to the information available to him. Furthermore, much of what we call "intuition" can be explicated in a precise and quantitative manner. When this is done, the judge's insight into his own cognitive processes is often found to be inaccurate.

Second, we find that judges have a very difficult time weighting and combining information, be it probabilistic or deterministic in nature. To reduce cognitive strain, they resort to simplified decision strategies, many of which lead them to ignore or misuse relevant information.

The order in which information is received affects its use and integration but the specific form of sequential effects that occur is dependent upon particular circumstances of the decision task. Similarly, the manner in which information is displayed and the nature of the required response greatly influence the use of that information. In other words, the structure of the judgment situation is an important determinant of information use.

Finally, despite the great deal of research already completed, it is obvious that we know very little about many aspects of information use in judgment. Few variables have been explored in much depth–even such fundamental ones such as the number of cues, cue redundancy, or the effects of various kinds of stress. And the enormous task of integrating this area with the mainstream of cognitive psychology–work on concept formation, problem solving, memory, learning, attention, etc.–remains to be undertaken.

Does the Paradigm Dictate the Research?

One of the objectives of this chapter was to determine whether the specific models and methods characteristic of each research paradigm tend to focus the researcher's attention on certain problem areas while causing him to neglect others. Such focusing has obviously occurred. For example, the Bayesians have been least concerned with developing descriptive models of subjective composition rules, concentrating instead on comparing subjects' performance with that of an optimal model, Bayes' theorem. They have paid little attention to the learning of optimality, however. Researchers within the correlational paradigm have spent a great deal of effort using correlational methods to describe a judge's idiosyncratic weighting process and researchers using ANOVA designs to study impression formation have concentrated on distinguishing various additive and averaging models and delineating sequential effects at the group level.

These different emphases are further illustrated by the fact that experimental manipulations which are similar from one paradigm to the other have been undertaken for quite different purposes. For example, the Bayesians have studied sequence length to gauge its effects on conservatism; set size was studied in impression formation in order to distinguish additive and averaging models; and the number of cues was varied by correlational researchers to study the effects upon consistency and complexity of subjects' strategies.

Can these differences in focus be attributed to the influence of the model used? Is a researcher inevitably steered in a particular direction by his chosen model? To some small extent, this is certainly true. A correlationalist would find it difficult to use, as his cues, intelligence reports such as: "General Tsing was seen last Monday lunching with Ambassador Hsieh." Instead, he will feel more comfortable with cues such as MMPI scores, or grade point averages. Similarly, a Bayesian is most comfortable working with a small number of hypotheses, while the correlationalist can work conveniently with many, provided they are unidimensionally scaled.

In general, however, we believe that the major differences in research emphasis cannot be traced to differences between the models. On the one hand, we see neglected problems for which a model is well suited. For example, the Bayesians neglect learning, although they have a numerical response, which can easily be compared to a numerical optimal response, for every trial; they need not partition the data into blocks (as correlationalists must in order to compute a beta weight). On the other hand, we see persistent, even stubborn, pursuit of topics for which the model is awkward. Correlationalists have devoted much effort to

the search for configural cue utilization, yet the linear model is extraordinarily powerful in suppressing such relationships, and interactions in ANOVA must be viewed with suspicion because the technique lacks invariance properties under believable data transformations.

Several research paradigms have been wound up around common points of interest and are chugging rapidly down diverging roads. Since any study almost always raises additional questions for investigation, there has been no dearth of interesting problems to fuel these research vehicles. Unfortunately, these vehicles lack side windows, and few investigators are looking far enough to the left or right. Of several hundred studies, only a handful indicate any awareness of the existence of comparable research under another paradigm. The fact remains, however, that all these investigators are interested in the same general problem, that of understanding how humans integrate fallible information to produce a judgment or decision. Singleminded dedication to one paradigm is disturbing since it suggests a lack of concern with basic, substantive issues. As Platt (1964) put the matter:

> To paraphrase an old saying, Beware of the man of one method or one instrument, either experimental or theoretical. He tends to become method oriented rather than problem oriented. The method oriented man is shackled; the problem oriented man is at least reaching freely toward what is most important. (Platt, 1964, p. 351).

Applied vs. Theoretical Objectives

If one is problem oriented, the distinction between applied and theoretical objectives becomes relevant to the selection of an experimental design. Methods suited for one of these general aims may be inadequate for the other. Thus correlational research, with its emphasis on predictability, may be quite useful for certain applied work but less adequate for theoretical endeavors which require sharper hypotheses and tests of fit. Similarly, in an applied man–machine system, the response scale is often valid by definition. But when the research emphasis shifts from practical problems and normative models to theoretical issues, the choice of a response scale may become critical. A clear awareness of these distinctions would seem to be important for both theoretical and applied research.

Towards an Integration of Research Efforts

We suggest that researchers should employ a multiparadigm approach, searching for the most appropriate tasks and models to attack the substantive problems of interest to them. We will try to show, for several such problems, how such a broader perspective might be advantageous.

Sequential effects. The potential value of a diverse approach is illustrated by research on primacy and recency effects. Hendrick and Costantini (1970a) found no effect of varying information inconsistency in an impression formation task, where adjectives served as cues. They argued that attention decrement, not inconsistency, accounts for the primacy commonly found in studies of impression formation. Yet a number of Bayesian studies did obtain primacy when early and late data were, to varying degrees, inconsistent (Dale, 1968; Peterson & DuCharme, 1967; Roby, 1967) and recency when later data were not inconsistent (Pitz & Reinhold, 1968; Shanteau, 1970). The discrepancy between the Hendrick and Costantini data and the Bayesian results would seem to be worth investigating.

The study by Shanteau (1970) provides a nice example of the utility of applying methods and tasks from different paradigms when studying sequential effects. Shanteau used an ANOVA design with a Bayesian task. He presented subjects with sequences of data constructed according to factorial combinations of binary events. Their task was to estimate $P(H/D)$ after each datum was received. Sequential effects appear as main effects of serial position in such a design. Two experiments clearly showed that recency was operating throughout all stages of sequences as long as 15 items.

The inertia effect in Bayesian research can be viewed as a type of primacy effect. The fact that inertia is so dependent upon the degree to which subjects' previous judgments are displayed or otherwise highlighted suggests that this same factor might also be operating in studies of primacy and recency. It is perhaps relevant that most of the studies of impression formation that employed intermittent responding and obtained recency effects used spoken ratings, slash marks, or required subjects to fill out detailed questionnaires. None of these formats gives particular salience to previous judgments. The one study that exhibited primacy effects (Anderson, 1959) employed a more standard written response, although subjects did have to turn the page for each new item of information. In addition, in each of the Bayesian studies that obtained primacy (Dale, 1968; Peterson & DuCharme, 1967; Roby, 1967), subjects were asked to make estimates on some mechanical device that preserved the previous response and required it to be physically manipulated when changes were made. While all this is obviously *post hoc* analysis, future research on primacy and recency effects should take a close look at the manner in which the previous response in the sequence is made and stored.

Novelty. How do subjects handle data that are rare or novel? Wyer (1970) examined the effects of novelty, defined in terms of the un-

conditional probability of an adjective, upon impression formation. Novel adjectives were seen to carry greater weight, making impressions more polarized. This increased weight attached to rare data appears to be in contradiction with findings from Bayesian research on rare events (Beach, 1968; Vlek, 1965; Vlek & van der Heijden, 1967). These studies have presented evidence that rare events are viewed as uninformative, i.e., they are not given enough weight in the decision process.

Learning. Hammond and his colleagues (e.g., Hammond & Brehmer, in press; Todd & Hammond, 1965) have long contended that specific feedback derived from the lens model (i.e., feedback about the weight the subject gives to each cue, and the weight the environment gives to each cue) is more effective than nonspecific feedback (i.e., the "correct" answer). How does this result relate to the finding by Martin and Gettys (1969) that probabilistic feedback is better than nominal feedback, or to the evidence from Wheeler and Beach (1968) and Peterson, DuCharme, and Edwards (1968) that subjects give more optimal $P(H/D)$ estimates after they have received training in $P(D/H)$? If specific feedback enhances performance, why then did Pitz and Downing (1967) find that subjects' binary predictions were not improved by detailed information about the sampling distributions?

Diagnosticity and conservatism. Both the Bayesian and the correlational models have well-defined measures of the diagnosticity of data– $P(D/H)$ and $b_{i,e}$, respectively. A unified approach to this topic seems natural. In the past, correlationalists have done little exploration in performance (nonlearning) studies where diagnosticity was varied. Bayesian research on this topic has been extensive and has demonstrated the difficulties subjects have in integrating probabilistic information. The different data and response formats possible within the correlational paradigm would seem to provide an excellent opportunity to investigate the generality of these difficulties. Subjects could be taught how to use various linear and curvilinear cues individually and could then be asked to integrate them into their judgments.

From a theoretical standpoint the question of conservatism rests upon the validity of the response scale. We have already seen that different response modes produce different degrees of conservatism, but these responses have been selected on the basis of practical, not theoretical considerations. Response rescaling and model testing, as practiced within integration theory and conjoint measurement, could be quite useful in evaluating conservatism.

Self-insight. The analysis of the judge's insight into his own subjective composition function is a fascinating and important area that has been investigated only within the correlational paradigm. And in

these studies, only one rather arbitrary technique has been used to scale the judge's perceptions of a cue's importance–the "distribute-100-points" technique. Besides the obvious step of employing new scaling techniques in correlational studies, it would seem valuable to test the accuracy of self-insight in a variety of tasks, using Bayesian and ANOVA models as well.

Decision aids. The idea of bootstrapping, which was developed in the context of regression equations, has some interesting relationships with the PIP system designed by Bayesians to improve human judgment. Both view human judgments as essential and attempt to blend them optimally (see Pankoff & Roberts, 1968, for an elaboration of this point). However, advocates of the PIP system assume that the aggregation process is faulty and attempt to circumvent this by having subjects estimate $P(D/H)$ values and letting a machine combine them. They decompose the judgment task into a number of presumably simpler estimation tasks. Bootstrapping assumes that subjects can aggregate information appropriately except for unreliability that must be filtered out. The success of bootstrapping and PIP systems suggests that the assumptions of both have some validity; judges are biased *and* unreliable in their weighting of information. Perhaps a system can be designed to minimize both these sources of error or, at least, to differentiate situations where PIP might excel bootstrapping or vice versa.

New Directions

Although a diverse program, integrating Bayesian and regression paradigms, might be quite valuable in increasing our understanding of information processing, some new approaches might be even more illuminating. Our own inclination is to move towards more molecular analyses of the heuristic strategies that subjects employ when they integrate information, along the lines we discussed earlier in this article. At this level, the evidence to date seems to indicate that subjects are processing information in ways fundamentally different from Bayesian and regression models. Thus, if we are to pursue this line of research we will have to develop new models and different methods of experimentation. Use of eye movements, introspection, and tasks where subjects have to search for or request information as they need it may help provide insights into molecular strategies. The search for new models should bring judgment research into closer contact with the more traditional areas of psychology. For example, Bruner, Goodman, and Austin's (1956) work on the role of cognitive strain in determining concept formation strategies, and Simon's analyses of cognitive limitations and their influence on problem solving (Simon, 1956, 1969), seem

likely to have much relevance for our understanding of composition strategies in judgment.

REFERENCES

AGNEW, N. M., & PYKE, S. W. *The science game: An introduction to research in the behavioral sciences.* Englewood Cliffs, N. J.: Prentice-Hall, 1969.

ANDERSON, N. H. Test of a model for opinion change. *Journal of Abnormal and Social Psychology,* 1959, **59,** 371–381.

ANDERSON, N. H. Application of an additive model to impression formation. *Science,* 1962, **138,** 817–818.

ANDERSON, N. H. Test of a model for number-averaging behavior. *Psychonomic Science,* 1964, **1,** 191–192.

ANDERSON, N. H. Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology,* 1965, **70,** 394–400. (a)

ANDERSON, N. H. Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology,* 1965, **2,** 1–9. (b)

ANDERSON, N. H. Component ratings in impression formation. *Psychonomic Science,* 1966, **6,** 279–280.

ANDERSON, N. H. Application of a weighted average model to a psychophysical averaging task. *Psychonomic Science,* 1967, **8,** 227–228. (a)

ANDERSON, N. H. Averaging model analysis of set-size effect in impression formation. *Journal of Experimental Psychology,* 1967, **75,** 158–165. (b)

ANDERSON, N. H. Application of a linear-serial model to a personality-impression task using serial presentation. *Journal of Personality and Social Psychology,* 1968, **10,** 354–362. (a)

ANDERSON, N. H. A simple model for information integration. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook.* Chicago: Rand McNally, 1968. (b)

ANDERSON, N. H. Comment on "An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment." *Psychological Bulletin,* 1969, **72,** 63–65.

ANDERSON, N. H. Functional measurement and psychophysical judgment. *Psychological Review,* 1970, **77,** 153–170.

ANDERSON, N. H. Integration theory and attitude change. *Psychological Review,* in press.

ANDERSON, N. H. Two more tests against change of meaning in adjective combinations. *Journal of Verbal Learning and Verbal Behavior,* 1971, **10,** 75–85.

ANDERSON, N. H., & BARRIOS, A. A. Primacy effects in personality impression formation. *Journal of Abnormal and Social Psychology,* 1961, **63,** 346–350.

ANDERSON, N. H., & HUBERT, S. Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 379–391.

ANDERSON, N. H., & JACOBSON, A. Effect of stimulus inconsistency and discounting instructions in personality impression formation. *Journal of Personality and Social Psychology,* 1965, **2,** 531–539.

ANDERSON, N. H., & JACOBSON, A. Further data on a weighted average model for judgment in a lifted weight task. *Perception and Psychophysics,* 1968, **4,** 81–84.

ANDERSON, N. H., & LAMPEL, A. K. Effect of context on ratings of personality traits. *Psychonomic Science,* 1965, **3,** 433–434.

ANDERSON, N. H., & NORMAN, A. Order effects in impression formation in four classes of stimuli. *Journal of Abnormal and Social Psychology,* 1964, **69,** 467–471.

ANDERSON, N. H., & SHANTEAU, J. C. Information integration in risky decision making. *Journal of Experimental Psychology,* 1970, **84,** 441–451.

ARMELIUS, B., & LENNTOFT, K. Effect of cue intercorrelation in a multiple cue probability learning task with different cue validities. Umeå Psychological Report No. 20, Department of Psychology, University of Umeå, 1970.

ASCH, S. E. Forming impressions of personality. *Journal of Abnormal and Social Psychology,* 1946, **41,** 258–290.

AZUMA, H., & CRONBACH, L. J. Cue-response correlations in the attainment of a scalar concept. *The American Journal of Psychology,* 1966, **79,** 38–49.

BEACH, L. R. Accuracy and consistency in the revision of subjective probabilities. *IEEE Transactions on Human Factors in Electronics,* 1966, **7,** 29–37.

BEACH, L. R. Probability magnitudes and conservative revision of subjective probabilities. *Journal of Experimental Psychology,* 1968, **77,** 57–63.

BEACH, L. R., & OLSON, J. B. Data sequences and subjective sampling distributions. *Psychonomic Science,* 1967, **9,** 309–310.

BEACH, L. R., & PHILLIPS, L. D. Subjective probabilities inferred from estimates and bets. *Journal of Experimental Psychology,* 1967, **75,** 354–359.

BEACH, L. R., & WISE, J. A. Subjective probability and decision strategy. *Journal of Experimental Psychology,* 1969, **79,** 133–138. (a)

BEACH, L. R., & WISE, J. A. Subjective probability revision and subsequent decisions. *Journal of Experimental Psychology,* 1969, **81,** 561–565. (b)

BEACH, L. R., WISE, J. A., & BARCLAY, S. Sample proportion and subjective probability revisions. *Organizational Behavior and Human Performance,* 1970, **5,** 183–190.

BIERI, J., ATKINS, A. L., BRIAR, S., LEAMAN, R. L., MILLER, H., & TRIPODI, T. *Clinical and social judgment: The discrimination of behavioral information.* New York: Wiley, 1966.

BJÖRKMAN, M. Learning of linear functions: Comparison between a positive and a negative slope. Report No. 183 from the Psychological Laboratories of the University of Stockholm, 1965.

BJÖRKMAN, M. The effect of training and number of stimuli on the response variance in correlation learning. Umeå Psychological Report No. 2, Department of Psychology, University of Umeå, 1968.

BOGARTZ, R. S., & WACKWITZ, J. H. Transforming response measures to remove interactions or other sources of variance. *Psychonomic Science,* 1970, **19,** 87–89.

BOWMAN, E. H. Consistency and optimality in managerial decision making. *Management Science,* 1963, **9,** 310–321.

BREHMER, B. Cognitive dependence on additive and configural cue-criterion relations. *The American Journal of Psychology,* 1969, **82,** 490–503. (a)

BREHMER, B. The roles of policy differences and inconsistency in policy conflict. Umeå Psychological Report No. 18, Department of Psychology, University of Umeå, 1969. (b) Also published as Program on Cognitive Processes Report No. 118, Institute of Behavioral Science, University of Colorado, 1969.

BREHMER, B. Inference behavior in a situation where the cues are not reliably perceived. *Organizational Behavior and Human Performance,* 1970, **5,** 330–347.

BREHMER, B., & LINDBERG, L. A. The relation between cue dependency and cue

validity in single-cue probability learning with scaled cue and criterion variables. *Organizational Behavior and Human Performance,* 1970, **5,** 542–554.

BRIGGS, G. E., & SCHUM, D. A. Automated Bayesian hypothesis-selection in a simulated threat-diagnosis system. In J. Spiegel & D. E. Walker (Eds.) *Information systems sciences: Proceedings of the second congress.* Washington, D. C.: Spartan Books, 1965, Pp. 169–176.

BRODY, N. The effect of commitment to correct and incorrect decisions on confidence in a sequential decision-task. *American Journal of Psychology,* 1965, **78,** 251–256.

BROWN, T. R. The judgment of suicide lethality: A comparison of judgmental models obtained under contrived versus natural conditions. Unpublished doctoral dissertation, University of Oregon, 1970.

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking.* New York: Wiley, 1956.

BRUNSWIK, E. *The conceptual framework of psychology.* Chicago: University of Chicago Press, 1952.

BRUNSWIK, E. Representative design and probabilistic theory in a functional psychology. *Psychological Review,* 1955, **62,** 193–217.

BRUNSWIK, E. *Perception and the representative design of experiments.* Berkeley: University of California Press, 1956.

CARROLL, J. D. Functional learning: The learning of continuous functional mappings relating stimulus and response continua. Research Bulletin (RB-63-26), Princeton, N. J.: Educational Testing Service, 1963.

CHRISTAL, R. E. JAN: A technique for analyzing group judgment. Technical Documentary Report PRL-TDR-63-3, Personnel Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Lackland AFB, Texas, 1963.

CLARKSON, G. P. E. *Portfolio selection: A simulation of trust investment.* Englewood Cliffs, N. J.: Prentice-Hall, 1962.

COCHRAN, W. G., & COX, G. M. *Experimental designs.* (2nd ed.) New York: Wiley, 1957.

COHEN, J. Multiple regression as a general data-analytic system. *Psychological Bulletin,* 1968, **70,** 426–443.

COOMBS, C. H. *A theory of data.* New York: Wiley, 1964.

COOMBS, C. H., & HUANG, L. C. Polynomial psychophysics of risk. *Journal of Mathematical Psychology,* 1970, **7,** 317–338.

DALE, H. C. A. Weighing evidence: An attempt to assess the efficiency of the human operator. *Ergonomics,* 1968, **11,** 215–230.

DARLINGTON, R. B. Multiple regression in psychological research and practice. *Psychological Bulletin,* 1968, **69,** 161–182.

DAWES, R. M. Social selection based on multidimensional criteria. *Journal of Abnormal and Social Psychology,* 1964, **68,** 104–109.

DAWES, R. M. A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist,* 1971, **26,** 180–188.

DAWES, R. M., & DILLER, R. D. The prediction of bootstrapping and a method of amalgamation. *Oregon Research Institute Research Bulletin,* 1970, Vol. 10, No. 6.

DUCHARME, W. M. A response bias explanation of conservative human inference. *Journal of Experimental Psychology,* 1970, **85,** 66–74.

DUCHARME, W. M., & PETERSON, C. R. Intuitive inference about normally distributed populations. *Journal of Experimental Psychology,* 1968, **78,** 269–275.

DUDYCHA, A. L. A Monte Carlo evaluation of JAN: A technique for capturing and

clustering rater's policies. *Organizational Behavior and Human Performance,* 1970, **5,** 501–516.

DUDYCHA, A. L., & NAYLOR, J. C. The effect of variations in the cue R matrix upon the obtained policy equation of judges. *Educational and Psychological Measurement,* 1966, **26,** 583–603. (a)

DUDYCHA, L. W., & NAYLOR, J. C. Characteristics of the human inference process in complex choice behavior situations. *Organizational Behavior and Human Performance,* 1966, **1,** 110–128. (b)

DUSTIN, D. S., & BALDWIN, P. M. Redundancy in impression formation. *Journal of Personality and Social Psychology,* 1966, **3,** 500–506.

EARLE, T. C. Task learning, interpersonal learning, and cognitive complexity. *Oregon Research Institute Research Bulletin,* 1970, Vol. 10, No. 2.

ECKENRODE, R. T. Weighting multiple criteria. *Management Science,* 1965, **12,** 180–192.

EDWARDS, W. The theory of decision making. *Psychological Bulletin,* 1954, **51,** 380–418.

EDWARDS, W. Dynamic decision theory and probabilistic information processing. *Human Factors,* 1962, **4,** 59–73.

EDWARDS, W. Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology,* 1965, **2,** 312–329.

EDWARDS, W. Nonconservative probabilistic information processing systems. Report from Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, USAF, ESD-TR-66-404, 1966.

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

EDWARDS, W., LINDMAN, H., & PHILLIPS, L. D. Emerging technologies for making decisions. *New directions in psychology II.* New York: Holt, Rinehart & Winston, 1965.

EDWARDS, W., LINDMAN, H., & SAVAGE, L. J. Bayesian statistical inference for psychological research. *Psychological Review,* 1963, **70,** 193–242.

EDWARDS, W., & PHILLIPS, L. D. Man as transducer for probabilities in Bayesian command and control systems. In G. L. Bryan and M. W. Shelley (Eds.), *Human judgments and optimality.* New York: Wiley, 1964.

EDWARDS, W., PHILLIPS, L. D., HAYS, W. L., & GOODMAN, B. C. Probabilistic information processing systems: Design and evaluation. *IEEE Transactions on Systems Science and Cybernetics,* 1968, Vol. SSC-4, 248–265.

EINHORN, H. J. The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin,* 1970, **73,** 221–230.

EINHORN, H. J. Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance,* 1971, **6,** 1–27.

ESTES, W. K. The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: A study of a science.* New York: McGraw-Hill, 1959, II.

FISHBEIN, M., & HUNTER, R. Summation versus balance in attitude organization and change. *Journal of Abnormal and Social Psychology,* 1964, **69,** 505–510.

FITTS, P. M., & DEININGER, R. L. S-R compatibility: Correspondence among paired elements within stimulus and response codes. *Journal of Experimental Psychology,* 1954, **48,** 483–492.

FRIED, L. S., & PETERSON, C. R. Information seeking: Optional vs. fixed stopping. *Journal of Experimental Psychology,* 1969, **80,** 525–529.

GALANTER, E., & HOLMAN, G. L. Some invariances of the iso-sensitivity function and their implications for the utility function of money. *Journal of Experimental Psychology,* 1967, **73,** 333–339.

GELLER, E. S., & PITZ, G. F. Confidence and decision speed in the revision of opinion. *Organizational Behavior and Human Performance,* 1968, **3,** 190–201.

GETTYS, C. F., & MANLEY, C. W. The probability of an event and estimates of posterior probability based upon its occurrence. *Psychonomic Science,* 1968, **11,** 47–48.

GIBSON, R. S., & NICHOL, E. H. The modifiability of decisions made in a changing environment. Report from Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, USAF, ESD-TR-64-657, 1964.

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

GOLDBERG, L. R. Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin,* 1970, **73,** 422–432.

GOLLOB, H. F. Impression formation and word combination in sentences. *Journal of Personality and Social Psychology,* 1968, **10,** 341–353.

GORDON, J. R. M. A multi-model analysis of an aggregate scheduling decision. Unpublished Ph.D. Dissertation, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, 1966.

GRAY, C. W. Predicting with intuitive correlations. *Psychonomic Science,* 1968, **11,** 41–43.

GRAY, C. W., BARNES, C. B., & WILKINSON, E. F. The process of prediction as a function of the correlation between two scaled variables. *Psychonomic Science,* 1965, **3,** 231–232.

GREEN, B. F., JR. Descriptions and explanations: A comment on papers by Hoffman and Edwards. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

GREEN, D. M., & SWETS, J. A. *Signal detection theory and psychophysics.* New York: Wiley, 1966.

GRINNELL, M., KEELEY, S., & DOHERTY, M. E. Bayesian predictions of faculty judgments of graduate school success. Paper presented at the meeting of the Midwestern Psychological Association, Cincinnati, 1970.

GUSTAFSON, D. H. Evaluation of probabilistic information processing in medical decision making. *Organizational Behavior and Human Performance,* 1969, **4,** 20–34.

GUSTAFSON, D. H., EDWARDS, W., PHILLIPS, L. D., & SLACK, W. V. Subjective probabilities in medical diagnosis. *IEEE Transactions on Man-Machine Systems,* 1969, MMS-**10**(3), 61–65.

HALPERN, J., & ULEHLA, Z. J. The effect of multiple responses and certainty estimates on the integration of visual information. *Perception and Psychophysics,* 1970, **7,** 129–132.

HAMMOND, K. R. Probabilistic functioning and the clinical method. *Psychological Review,* 1955, **62,** 255–262.

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

HAMMOND, K. R. Probabilistic functionalism: Egon Brunswik's integration of the history, theory and method of psychology. In K. R. Hammond (Ed.), *The Psychology of Egon Brunswik,* New York: Holt, Rinehart and Winston, 1966.

HAMMOND, K. R., & BOYLE, P. J. R. Quasi-rationality, quarrels, and new conceptions of feedback. Program on Cognitive Processes Report No. 130, Institute of Behavioral Science, University of Colorado, 1970.

HAMMOND, K. R., & BREHMER, B. The quasi-rational nature of quarrels about policy. In J. Hellmuth (Ed.), *Cognitive Studies,* Vol. 2, *Deficits in Cognition.* New York: Brunner/Mazel, Inc., in press.

HAMMOND, K. R., HURSCH, C. J., & TODD, F. J. Analyzing the components of clinical inference. *Psychological Review,* 1964, **71,** 438–456.

HAMMOND, K. R., KELLY, K. J., SCHNEIDER, R. J., & VANCINI, M. Clinical inference in nursing: Revising judgments. *Nursing Research,* 1967, **16,** 36–45.

HAMMOND, K. R., & SUMMERS, D. A. Cognitive dependence on linear and nonlinear cues. *Psychological Review,* 1965, **72,** 215–234.

HAMMOND, K. R., TODD, F. J., WILKINS, M. M., & MITCHELL, T. O. Cognitive conflict between persons: Application of the "Lens Model" paradigm. *Journal of Experimental Social Psychology,* 1966, **2,** 343–360.

HAMMOND, K. R., WILKINS, M. M., & TODD, F. J. A research paradigm for the study of interpersonal learning. *Psychological Bulletin,* 1966, **65,** 221–232.

HAYES, J. R. Human data processing limits in decision making. In E. Bennett (Ed.), *Information system science and engineering. Proceedings of the First Congress on the Information Systems Sciences.* New York: McGraw-Hill, 1964.

HAYES, J. R. Strategies in judgmental research. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

HAYS, W. L. *Statistics for psychologists.* New York: Holt, Reinhart and Winston, 1963.

HENDRICK, C. Averaging *vs.* summation in impression formation. *Perceptual and Motor Skills,* 1968, **27,** 1295–1302.

HENDRICK, C., & COSTANTINI, A. F. Effects of varying trait inconsistency and response requirements on the primacy effect in impression formation. *Journal of Personality and Social Psychology,* 1970, **15,** 158–164. (a)

HENDRICK, C., & COSTANTINI, A. F. Number averaging behavior: A primacy effect. *Psychonomic Science,* 1970, **19,** 121–122. (b)

HIMMELFARB, S. The impact of neutral information about a person. Unpublished manuscript, University of California at LaJolla, 1970.

HIMMELFARB, S., & SENN, D. J. Forming impressions of social class: Two tests of an averaging model. *Journal of Personality and Social Psychology,* 1969, **12,** 38–51.

HOEPFL, R. T., & HUBER, G. P. A study of self-explicated utility models. *Behavioral Science,* 1970, **15,** 408–414.

HOFFMAN, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin,* 1960, **47,** 116–131.

HOFFMAN, P. J. Cue-consistency and configurality in human judgment. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

HOFFMAN, P. J., & BLANCHARD, W. A. A study of the effects of varying amounts of predictor information on judgment. *Oregon Research Institute Research Bulletin,* 1961.

HOFFMAN, P. J., SLOVIC, P., & RORER, L. G. An analysis-of-variance model for the

assessment of configural cue utilization in clinical judgment. *Psychological Bulletin,* 1968, **69,** 338–349.

HOWELL, W. C. Some principles for the design of decision systems: A review of six years of research on a command-control system simulation. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio. AMRL-TR-67-136, 1967.

HUBER, G. P., SAHNEY, V. K., & FORD, D. L. A study of subjective evaluation models. *Behavioral Science,* 1969, **14,** 483–489.

HURSCH, C., HAMMOND, K. R., & HURSCH, J. L. Some methodological considerations in multiple cue probability studies. *Psychological Review,* 1964, **71,** 42–60.

HURST, E. G., JR., & MCNAMARA, A. B. Heuristic scheduling in a woolen mill. *Management Science,* 1967, **14,** B-182-B-203.

JONES, C. Parametric production planning. *Management Science,* 1967, **13,** 843–866.

JONES, M. R. From probability learning to sequential processing: A critical review. *Psychological Bulletin,* in press.

KAHNEMAN, D., & TVERSKY, A. Subjective probability: A judgment of representativeness. *Oregon Research Institute Research Bulletin,* 1970, Vol. 10, No. 5.

KAPLAN, R. J., & NEWMAN, J. R. Studies in probabilistic information processing. *IEEE Transactions on Human Factors in Electronics,* 1966, **7,** 49–63.

KATES, R. W. Hazard and choice perception in flood plain management. Department of Geography Research Paper No. 78, University of Chicago, 1962.

KATONA, G. *Psychological Analysis of Economic Behavior.* New York: McGraw-Hill, 1951.

KERRICK, J. S. The effect of relevant and non-relevant sources on attitude change. *Journal of Social Psychology,* 1958, **47,** 15–20.

KLAHR, D. Decision making in a complex environment: The use of similarity judgments to predict preferences. *Management Science,* 1969, **15,** 595–618.

KLEINMUNTZ, B. The processing of clinical information by man and machine. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

KNOX, R. E., & HOFFMAN, P. J. Effects of variation of profile format on intelligence and sociability judgments. *Journal of Applied Psychology,* 1962, **46,** 14–20.

KOFORD, J. S., & GRONER, G. F. The use of an adaptive threshold element to design a linear optimal pattern classifier. *IEEE Transactions on Information Theory,* 1966, Vol. IT-12, 42–50.

KORT, F. A nonlinear model for the analysis of judicial decisions. *The American Political Science Review,* 1968, **62,** 546–555.

KRANTZ, D. H., & TVERSKY, A. Conjoint measurement analysis of composition rules in psychology. *Psychological Review,* 1971, **78,** 151–169.

KRIZ, J. Der Likelihood Quotient zur erfassung des subjektiven Signifikanzniveaus. Forschungsbericht No. 9, Institute for Advanced Studies, Vienna, 1967.

KUNREUTHER, H. Extensions of Bowman's theory on managerial decision-making. *Management Science,* 1969, **15,** 415–439.

LAMPEL, A. K., & ANDERSON, N. H. Combining visual and verbal information in an impression-formation task. *Journal of Personality and Social Psychology,* 1968, **9,** 1–6.

LEDLEY, R. S., & LUSTED, L. B. Reasoning foundations of medical diagnosis. *Science,* 1959, **130,** 9–21.

LEE, J. C., & TUCKER, R. B. An investigation of clinical judgment: A study in method. *Journal of Abnormal and Social Psychology,* 1962, **64,** 272–280.

LEVIN, I. P., & SCHMIDT, C. F. Sequential effects in impression formation with binary intermittent responding. *Journal of Experimental Psychology,* 1969, **79,** 283–287.

LICHTENSTEIN, S., & FEENEY, G. J. The importance of the data-generating model in probability estimation. *Organizational Behavior and Human Performance,* 1968, **3,** 62–67.

LINDBLOM, C. E. The science of "muddling through." In W. J. Gore and J. W. Dyson (Eds.), *The making of decisions.* London: Collier-MacMillan, Ltd., 1964.

LOEB, G. *The battle for investment survival.* New York: Simon & Schuster, 1965.

LUCE, R. D., & TUKEY, J. Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology,* 1964, **1,** 1–27.

LUCHINS, A. S. Experimental attempts to minimize the impact of first impressions. In C. I. Hovland (Ed.), *The order of presentation in persuasion.* New Haven: Yale University Press, 1957.

LUCHINS, A. S. Definitiveness of impression and primacy-recency in communications. *The Journal of Social Psychology,* 1958, **48,** 275–290.

LUSTED, L. B. *Introduction to medical decision making.* Springfield, Ill.: Charles C. Thomas, 1968.

MADDEN, J. M. An application to job evaluation of a policy-capturing model for analyzing individual and group judgment. 6570th Personnel Research Laboratory Aerospace Medical Division Air Force Systems Command, PRL-TDR-63-15, May, 1963.

MAGNUSSON, D., & NYSTEDT, L. Cue relevance and feedback in a clinical prediction task. Report No. 272 from the Psychological Laboratories, University of Stockholm, 1969.

MAGUIRE, T. O., & GLASS, G. V. Component profile analysis (COPAN)—An alternative to PROF. *Educational and Psychological Measurement,* 1968, **28,** 1021–1033.

MANIS, M., GLEASON, T. C., & DAWES, R. M. The evaluation of complex social stimuli. *Journal of Personality and Social Psychology,* 1966, **4,** 404–419.

MARTIN, D. W. Data conflict in a multinomial decision task. *Journal of Experimental Psychology,* 1969, **82,** 4–8.

MARTIN, D. W., & GETTYS, C. F. Feedback and response mode in performing a Bayesian decision task. *Journal of Applied Psychology,* 1969, **53,** 413–418.

MARTIN, H. T., JR. The nature of clinical judgment. Unpublished doctoral dissertation, Washington State College, 1957.

MCEACHERN, A. W., & NEWMAN, J. R. A system for computer-aided probation decision-making. *Journal of Research on Crime and Delinquency,* 1969, **6,** 184–198.

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

MILLER, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review,* 1956, **63,** 81–97.

MILLER, M. J., & SARAFINO, E. The effects of intercorrelated cues on multiple probability learning. *Organizational Behavior and Human Performance,* in press.

MORRISON, H. W., & SLOVIC, P. Effects of context on relative judgments of area. Paper presented at the meeting of the Eastern Psychological Association,

Atlantic City, 1962. (Also in IBM Research Note NC-104, Watson Research Center, 1962.)

NAYLOR, J. C., & CLARK, R. D. Intuitive inference strategies in interval learning tasks as a function of validity magnitude and sign. *Organizational Behavior and Human Performance,* 1968, **3,** 378–399.

NAYLOR, J. C., & SCHENCK, E. A. The influence of cue redundancy upon the human inference process for tasks of varying degrees of predictability. *Organizational Behavior and Human Performance,* 1968, **3,** 47–61.

NAYLOR, J. C., & WHERRY, R. J., SR. The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. *Educational and Psychological Measurement,* 1965, **25,** 969–986.

NEWELL, A., SHAW, J. C., & SIMON, H. A. Elements of a theory of human problem solving. *Psychological Review,* 1958, **65,** 151–166.

NEWTON, J. R. Judgment and feedback in a quasi-clinical situation. *Journal of Personality and Social Psychology,* 1965, **1,** 336–342.

ODEN, G. C., & ANDERSON, N. H. Differential weighting in integration theory. *Journal of Experimental Psychology,* in press.

OSGOOD, C. E., & TANNENBAUM, P. H. The principle of congruity in the prediction of attitude change. *Psychological Review,* 1955, **62,** 42–55.

OSKAMP, S. How clinicians make decisions from the MMPI: An empirical study. Paper presented at the American Psychological Association, St. Louis, 1962.

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

PANKOFF, L. D., & ROBERTS, H. V. Bayesian synthesis of clinical and statistical prediction. *Psychological Bulletin,* 1968, **70,** 762–773.

PARDUCCI, A., THALER, H., & ANDERSON, N. H. Stimulus averaging and the context for judgment. *Perception and Psychophysics,* 1968, **3,** 145–150.

PETERSON, C. R. Aggregating information about signals and noise. *Proceedings, 76th Annual Convention, APA,* 1968, 123–124.

PETERSON, C. R., & BEACH, L. R. Man as an intuitive statistician. *Psychological Bulletin,* 1967, **68,** 29–46.

PETERSON, C. R., & DUCHARME, W. M. A primacy effect in subjective probability revision. *Journal of Experimental Psychology,* 1967, **73,** 61–65.

PETERSON, C. R., DUCHARME, W. M., & EDWARDS, W. Sampling distributions and probability revisions. *Journal of Experimental Psychology,* 1968, **76,** 236–243.

PETERSON, C. R., HAMMOND, K. R., & SUMMERS, D. A. Multiple probability learning with shifting cue weights. *American Journal of Psychology,* 1965, **78,** 660–663. (a)

PETERSON, C. R., HAMMOND, K. R., & SUMMERS, D. A. Optimal responding in multiple-cue probability learning. *Journal of Experimental Psychology,* 1965, **70,** 270–276. (b)

PETERSON, C. R., & MILLER, A. J. Sensitivity of subjective probability revision. *Journal of Experimental Psychology,* 1965, **70,** 117–121.

PETERSON, C. R., & PHILLIPS, L. D. Revision of continuous subjective probability distributions. *IEEE Transactions on Human Factors in Electronics,* 1966, HFE-**7,** 19–22.

PETERSON, C. R., SCHNEIDER, R. J., & MILLER, A. J. Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology,* 1965, **69,** 522–527.

PETERSON, C. R., & SWENSSON, R. G. Intuitive statistical inferences about diffuse hypotheses. *Organizational Behavior and Human Performance,* 1968, **3,** 1–11.

PHILLIPS, L. D. Some components of probabilistic inference. Technical Report No. 1, Human Performance Center, University of Michigan, 1966.

PHILLIPS, L. D., & EDWARDS, W. Conservatism in a simple probability inference task. *Journal of Experimental Psychology,* 1966, **72,** 346–357.

PHILLIPS, L. D., HAYS, W. L., & EDWARDS, W. Conservatism in complex probabilistic inference. *IEEE Transactions on Human Factors in Electronics,* 1966, HFE-**7,** 7–18.

PITZ, G. F. The sequential judgment of proportion. *Psychonomic Science,* 1966, **4,** 397–398.

PITZ, G. F. Sample size, likelihood, and confidence in a decision. *Psychonomic Science,* 1967, **8,** 257–258.

PITZ, G. F. An inertia effect (resistance to change) in the revision of opinion. *Canadian Journal of Psychology,* 1969, **23,** 24–33. (a)

PITZ, G. F. The influence of prior probabilities on information seeking and decision making. *Organizational Behavior and Human Performance,* 1969, **4,** 213–226. (b)

PITZ, G. F. Use of response times to evaluate strategies of information seeking. *Journal of Experimental Psychology,* 1969, **80,** 553–557. (c)

PITZ, G. F., & DOWNING, L. Optimal behavior in a decision-making task as a function of instructions and payoffs. *Journal of Experimental Psychology,* 1967, **73,** 549–555.

PITZ, G. F., DOWNING, L., & REINHOLD, H. Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology,* 1967, **21,** 381–393.

PITZ, G. F., & REINHOLD, H. Payoff effects in sequential decision-making. *Journal of Experimental Psychology,* 1968, **77,** 249–257.

PLATT, J. R. Strong inference. *Science,* 1964, **146,** 347–353.

PODELL, J. E. The impression as a quantitative concept. *American Psychologist,* 1962, **17,** 308. (Abstract)

PODELL, H. A., & PODELL, J. E. Quantitative connotation of a concept. *Journal of Abnormal and Social Psychology,* 1963, **67,** 509–513.

POLLACK, I. Action selection and the Yntema-Torgerson worth function. In E. Bennett (Ed.), *Information system science and engineering: Proceedings of the First Congress on the Information Systems Sciences.* New York: McGraw-Hill, 1964.

PRUITT, D. G. Informational requirements in making decisions. *American Journal of Psychology,* 1961, **74,** 433–439.

RAIFFA, H., & SCHLAIFER, R. *Applied statistical decision theory.* Boston: Harvard University, Graduate School of Business Administration, Division of Research, 1961.

RAPOPORT, A., & BURKHEIMER, G. J. Sequential decision making: Descriptive models, sensitivity analysis, and numerical results. University of North Carolina Psychonomic laboratory, No. 83, 1970.

RAPPOPORT, L. Interpersonal conflict in cooperative and uncertain situations. *Journal of Experimental Social Psychology,* 1965, **1,** 323–333.

RHINE, R. J. Test of models and impression formation. Paper presented at the meeting of the Western Psychological Association, San Diego, March, 1968.

ROBY, T. B. Belief states and sequential evidence. *Journal of Experimental Psychology,* 1967, **75,** 236–245.

RODWAN, A. S., & HAKE, H. W. The discriminant-function as a model for perception. *American Journal of Psychology,* 1964, **26,** 380–392.

ROREB, L. G., HOFFMAN, P. J., DICKMAN, H. D., & SLOVIC, P. Configural judgments revealed. *Proceedings of the 75th Annual Convention of the American Psychological Association,* 1967, **2,** 195–196.

ROSENBERG, S. Mathematical models of social behavior. In G. Lindzey and E. Aronson (Eds.), *The Handbook of Social Psychology,* 1968, **1,** 186–203.

ROSENKRANTZ, P. S., & CROCKETT, W. H. Some factors influencing the assimilation of disparate information in impression formation. *Journal of Personality and Social Psychology,* 1965, **2,** 397–402.

RUSSELL, C. S. Losses from natural hazards. Working Paper No. 10, Natural Hazard Research Program, Department of Geography, University of Toronto, 1969.

SARBIN, T. R. A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology,* 1942, **48,** 593–602.

SARBIN, T. R., & BAILEY, D. E. The immediacy postulate in the light of modern cognitive psychology. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik.* New York: Holt, Rinehart and Winston, 1966.

SAVAGE, L. J. *The foundations of statistics.* New York: Wiley, 1954.

SCHENCK, E. A., & NAYLOR, J. C. A cautionary note concerning the use of regression analysis for capturing the strategies of people. *Educational and Psychological Measurement,* 1968, **28,** 3–7.

SCHLAIFER, R. *Probability and statistics for business decisions.* New York: McGraw-Hill, 1959.

SCHMIDT, C. F. Personality impression formation as a function of relatedness of information and length of set. *Journal of Personality and Social Psychology,* 1969, **12,** 6–11.

SCHUM, D. A. Inferences on the basis of conditionally nonindependent data. *Journal of Experimental Psychology,* 1966, **72,** 401–409. (a)

SCHUM, D. A. Prior uncertainty and amount of diagnostic evidence as variables in a probabilistic inference task. *Organizational Behavior and Human Performance,* 1966, **1,** 31–54. (b)

SCHUM, D. A. Concerning the evaluation and aggregation of probabilistic evidence by man-machine systems. In D. E. Walker (Ed.), *Information System Science and Technology.* Washington, D. C.: Thompson Book Co., 1967.

SCHUM, D. A. Behavioral decision theory and man-machine systems. Report No. 46-4, Interdisciplinary Program in Applied Mathematics and Systems Theory. Houston: Rice University, 1968.

SCHUM, D. A. Concerning the simulation of diagnostic systems which process complex probabilistic evidence sets. Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, Technical Report 69-10, April, 1969.

SCHUM, D. A., GOLDSTEIN, I. L., HOWELL, W. C., & SOUTHARD, J. F. Subjective probability revisions under several cost-payoff arrangements. *Organizational Behavior and Human Performance,* 1967, **2,** 84–104.

SCHUM, D. A., GOLDSTEIN, I. L., & SOUTHARD, J. F. Research on a simulated Bayesian information-processing system. *IEEE Transactions on Human Factors in Electronics,* 1966, HFE-7, 37–48.

SCHUM, D. A., & MARTIN, D. W. Human processing of inconclusive evidence from multinomial probability distributions. *Organizational Behavior and Human Performance,* 1968, **3,** 353–365.

SCHUM, D. A., SOUTHARD, J. F., & WOMBOLT, L. F. Aided human processing of

inconclusive evidence in diagnostic systems: A summary of experimental evaluations. AMRL-Technical Report-69-11, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, May, 1969.

SHANTEAU, J. C. An additive model for sequential decision making. *Journal of Experimental Psychology*, 1970, **85**, 181–191.

SHANTEAU, J. C., & ANDERSON, N. H. Test of a conflict model for preference judgment. *Journal of Mathematical Psychology*, 1969, **6**, 312–325.

SHEETS, C., & MILLER, M. J. The effect of cue-criterion function form on multiple cue probability learning. *American Journal of Psychology*, in press.

SHEPARD, R. N. On subjectively optimum selection among multiattribute alternatives. In M. W. Shelly, II, and G. L. Bryan (Eds.), *Human judgments and optimality*. New York: Wiley, 1964.

SIDOWSKI, J. B., & ANDERSON, N. H. Judgments of city-occupation combinations. *Psychonomic Science*, 1967, **7**, 279–280.

SIMON, H. A. Rational choice and the structure of the environment. *Psychological Review*, 1956, **63**, 129–138.

SIMON, H. A. *The sciences of the artificial*. Cambridge, Mass.: MIT Press, 1969.

SLOVIC, P. Cue consistency and cue utilization in judgment. *American Journal of Psychology*, 1966, **79**, 427–434.

SLOVIC, P. Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. *Journal of Applied Psychology*, 1969, **53**, 255–263.

SLOVIC, P., FLEISSNER, D., & BAUMAN, W. S. Quantitative analysis of investment decisions. *Journal of Business*, in press.

SLOVIC, P., & LICHTENSTEIN, S. C. The relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology Monograph Supplement*, 1968, **78**, No. 3, Part 2.

SLOVIC, P., RORER, L. G., & HOFFMAN, P. J. Analyzing the use of diagnostic signs. *Investigative Radiology*, 1971, **6**, 18–27.

SMEDSLUND, J. *Multiple-probability learning*. Oslo: Akademisk Forlag, 1955.

SMITH, A. *The money game*. New York: Random House, 1968.

STEWART, R. H. Effect of continuous responding on the order effect in personality impression formation. *Journal of Personality and Social Psychology*, 1965, **1**, 161–165.

SUMMERS, D. A. Rule versus cue learning in multiple probability tasks. *Proceedings of the 75th Annual Convention of the American Psychological Association*, 1967, **2**, 43–44.

SUMMERS, D. A. Conflict, compromise, and belief change in a decision-making task. *Journal of Conflict Resolution*, 1968, **12**, 215–221.

SUMMERS, D. A. Adaptation to change in multiple probability tasks. *American Journal of Psychology*, 1969, **82**, 235–240.

SUMMERS, D. A., & HAMMOND, K. R. Inference behavior in multiple-cue tasks involving both linear and nonlinear relations. *Journal of Experimental Psychology*, 1966, **71**, 751–757.

SUMMERS, D. A., & STEWART, T. R. Regression models of foreign policy judgments. *Proceedings of the 76th Annual Convention of the American Psychological Association*, 1968, **3**, 195–196.

SUMMERS, S. A. The learning of responses to multiple weighted cues. *Journal of Experimental Psychology*, 1962, **64**, 29–34.

SUMMERS, S. A., SUMMERS, R. C., & KARKAU, V. T. Judgments based on different functional relationships between interacting cues and a criterion. *American Journal of Psychology*, 1969, **82**, 203–211.

SWETS, J. A. (Ed.) *Signal detection and recognition by human observers: Contemporary readings*. New York: Wiley, 1964.

SWETS, J. A., & BIRDSALL, T. G. Deferred decision in human signal detection: A preliminary experiment. *Perception and Psychophysics*, 1967, **2**, 15–28.

TODA, M. Measurement of subjective probability distribution. Report No. 3, Pennsylvania State College, Institute of Research, Division of Mathematical Psychology, 1963.

TODD, F. J., & HAMMOND, K. R. Differential feedback in two multiple-cue probability learning tasks. *Behavioral Science*, 1965, **10**, 429–435.

TUCKER, L. R. A suggested alternative formulation in the development of Hursch, Hammond, & Hursch, and by Hammond, Hursch, and Todd. *Psychological Review*, 1964, **71**, 528–530.

TVERSKY, A. Additivity, utility, and subjective probability. *Journal of Mathematical Psychology*, 1967, **4**, 175–202. (a)

TVERSKY, A. A general theory of polynomial conjoint measurement. *Journal of Mathematical Psychology*, 1967, **4**, 1–20. (b)

TVERSKY, A. Utility theory and additivity analysis of risky choices. *Journal of Experimental Psychology*, 1967, **75**, 27–36. (c)

TVERSKY, A. Intransitivity of preferences. *Psychological Review*, 1969, **76**, 31–48.

TVERSKY, A., & KAHNEMAN, D. The belief in the law of small numbers. *Psychological Bulletin*, in press.

TVERSKY, A., & KAHNEMAN, D. The judgment of probability by retrieval and construction of instances. *Oregon Research Institute Research Bulletin*, 1971, Vol. 11, No. 2.

TVERSKY, A., & KRANTZ, D. H. Similarity of schematic faces: A test of interdimensional additivity. *Perception & Psychophysics*, 1969, **5**, 124–128.

UHL, C. Learning of interval concepts. I. Effects of differences in stimulus weights *Journal of Experimental Psychology*, 1963, **66**, 264–273.

UHL, C. N., & HOFFMAN, P. J. Contagion effects and the stability of judgment. Paper read at Western Psychological Association, Monterey, California, 1958.

ULEHLA, Z. J. Optimality of perceptual decision criteria. *Journal of Experimental Psychology*, 1966, **71**, 564–569.

ULEHLA, Z. J., CANGES, L., & WACKWITZ, F. Integration of conceptual information. *Psychonomic Science*, 1967, **8**, 223–224.

ULMER, S. S. The discriminant function and a theoretical context for its use in estimating the votes of judges. In J. B. Grossman and J. Tanenhaus (Eds.), *Frontiers of judicial research*. New York: Wiley, 1969.

VLEK, C. A. J. The use of probabilistic information in decision making. Psychological Institute Report No. 009-65, University of Leiden, The Netherlands, 1965.

VLEK, C. A. J., & BEINTEMA, K. A. Subjective likelihoods in posterior probability estimation. Psychological Institute Report No. E 014-67, University of Leiden, The Netherlands, 1967.

VLEK, C. A. J., & van der HEIJDEN, L. H. C. Subjective likelihood functions and variations in the accuracy of probabilistic information processing. Psychological Institute Report No. E 017-67, University of Leiden, The Netherlands, 1967.

VON NEUMANN, J., & MORGENSTERN, O. *Theory of games and economic behavior.* (3rd ed., 1953) Princeton: Princeton University Press, 1947.

WALD, A. *Sequential analysis.* New York: Wiley, 1947.

WALLSTEN, T. S. Failure of predictions from subjectively expected utility theory in a Bayesian decision task. *Organizational Behavior and Human Performance,* 1968, **3,** 239–252.

WALLSTEN, T. S. The likelihood-ratio principle and conjoint measurement. Paper read at the tenth annual meeting of the Psychonomic Society, San Antonio, November, 1970.

WARD, J. H., JR., & DAVIS, K. Teaching a digital computer to assist in making decisions. 6570th Personnel Research Laboratory Aerospace Medical Division Air Force Systems Command, Technical Documentary Report, PRL-TDR-63-16, Lackland AFB, Texas, June, 1963.

WEISS, D. J., & ANDERSON, N. H. Subjective averaging of length with serial presentation. *Journal of Experimental Psychology,* 1969, **82,** 52–63.

WEISS, W. Scale judgments of triplets of opinion statements. *Journal of Abnormal and Social Psychology,* 1963, **66,** 471–479.

WENDT, D. Value of information for decisions. *Journal of Mathematical Psychology,* 1969, **6,** 430–443.

WHEELER, G., & BEACH, L. R. Subjective sampling distributions and conservatism. *Organizational Behavior and Human Performance,* 1968, **3,** 36–46.

WHERRY, R. J., SR., & NAYLOR, J. C. Comparison of two approaches–JAN and PROF–for capturing rater strategies. *Educational and Psychological Measurement,* 1966, **26,** 267–286.

WHITE, G. F. Optimal flood damage management: Retrospect and prospect. In A. V. Kneese and S. C. Smith (Eds.), *Water research.* Baltimore: Johns Hopkins Press, 1966.

WIGGINS, N. Multivariate models for cue-utilization: An individual differences approach. In L. Rappoport and D. Summers (Eds.), *Human judgment and social interaction.* New York: Holt, Rinehart & Winston, in press.

WIGGINS, N., & HOFFMAN, P. J. Three models of clinical judgment. *Journal of Abnormal Psychology,* 1968, **73,** 70–77.

WILLIAMS, J. D., HARLOW, S. D., LINDEM, A., & GAB, D. A judgment analysis program for clustering similar judgmental systems. *Educational and Psychological Measurement,* 1970, **30,** 171–173.

WILLIS, R. H. Stimulus pooling and social perception. *Journal of Abnormal and Social Psychology,* 1960, **60,** 365–373.

WINKLER, R. L., & MURPHY, A. H. "Good" probability assessors. *Journal of Applied Meteorology,* 1968, **7,** 751–758.

WOHLSTETTER, R. *Pearl Harbor: Warning and decision.* Stanford, California: Stanford University Press, 1962.

WYER, R. S., JR. The effects of information redundancy on evaluations of social stimuli. *Psychonomic Science,* 1968, **13,** 245–246.

WYER, R. S., JR. Information redundancy, inconsistency, and novelty and their role in impression formation. *Journal of Experimental Social Psychology,* 1970, **6,** 111–127.

WYER, R. S., JR., & WATSON, S. F. Context effects in impression formation. *Journal of Personality and Social Psychology,* 1969, **12,** 22–33.

YNTEMA, D. B., & KLEM, L. Telling a computer how to evaluate alternatives

as one would evaluate them himself. In E. Bennett (Ed.), *Information system science and engineering: Proceedings of the First Congress on the Information System Sciences.* New York: McGraw-Hill, 1964.

YNTEMA, D. B., & TORGERSON, W. S. Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics,* 1961, HFE **2**(1), 20–26.

RECEIVED: DECEMBER 10, 1970