DOCUMENT RESUME

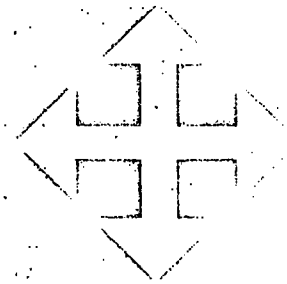
ED 045 723 TM 000 303

AUTHOR Slovic, Paul; Lichtenstein, Sarah
TITLE Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment.
INSTITUTION Oregon Research Inst., Eugene.
SPONS AGENCY Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
REPORT NO CRI-R-Monogr-Vol-10-No-1
PUB DATE Jul 70
NOTE 154p.

EDRS PRICE EDRS Price MF-$0.75 HC-$7.80
DESCRIPTORS Analysis of Variance, *Cognitive Processes, Comparative Analysis, *Cues, *Decision Making, Decision Making Skills, *Information Processing, Information Theory, Information Utilization, Models, Probability Theory, *Research Methodology, Task Analysis, Thought Processes

ABSTRACT

Most research on information utilization in judgment and decision making has followed two basic approaches: "regression" and "Bayesian." Each has characteristic tasks and characteristic information that must be processed to accomplish these tasks. There has been a tendency to work within a single approach with minimal communication between the resultant subgroups of workers. This analysis of the approaches examines (a) the models developed for prescribing and describing the use of information; (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; (c) the key independent variables that have been manipulated; and (d) the major empirical results and conclusions. Topics discussed include the configural use of information, task or environmental determinants of information utilization, learning to use information, sequential effects upon information processing, strategies for combining information, and techniques for aiding the decision-maker. Of particular interest is the degree to which the specific models and methods characteristic of different paradigms have directed attention to certain problem areas to the neglect of other equally important problems. Also of interest is whether a researcher studying a particular substantive problem could increase his understanding by employing other models and experimental methods. By laying bare the similarities and differences between each approach cross-method research may be facilitated. A comprehensive bibliography is provided. (Author/GS)



OREGON RESEARCH INSTITUTE

Basic Research in the Behavioral Sciences

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

Paul Slovic and Sarah Lichtenstein

ORI RESEARCH MONOGRAPH
Vol. 10 No. 1

This work was sponsored by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N0001-4-68-C-0131, Contract Authority No. NR 153-311, and by Grants MH-15414 and MH-12972 from the United States Public Health Service.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Oregon Research Institute 488 E. 11th Avenue Eugene, Oregon 97405	Unclassified 2b. GROUP NA

3. REPORT TITLE

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

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Oregon Research Institute Research Monograph, 1970, Vol. 10, No. 1.

5. AUTHOR(S) (First name, middle initial, last name)

Paul Slovic & Sarah Lichtenstein

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
July, 1970	148	268

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-68-C-0431 b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale.
Its distribution is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Personnel and Training Research Branch Washington, D. C. 20360

13. ABSTRACT Since 1960 there have been more than 600 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 in this chapter is to present a comparative analysis of these two broad methods of approach. Within each, we examine (a) the models that have been developed for prescribing and describing 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.

Some of the topics discussed include the configural use of information, task or environmental determinants of information utilization, learning to use information, sequential effects upon information processing, strategies for combining information, and techniques for aiding the decision-maker.

Of particular interest to us is the degree to which the specific models and methods characteristic of different paradigms have directed the researcher's attention to certain problem areas and caused him to neglect other problems that are equally important. Another question of interest is whether a researcher studying a particular substantive problem, could increase his understanding by employing other models and experimental methods. We hope that by laying bare the similarities and differences between each approach we can facilitate such cross-method research.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Decision processes						
Judgment						
Information processing						
Cue utilization						

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

by

Paul Slovic and Sarah Lichtenstein

Oregon Research Institute

Research Monograph
Vol. 10 No. 1
July, 1970

This work was sponsored by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, Under Contract No. N00014-68-C-0431, Contract Authority No. NR 153-311, and by Grants MH-15414 and MH-12972 from the United States Public Health Service.

Organizational Outline

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

Paul Slovic

and

Sarah Lichtenstein

Oregon Research Institute

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 synonymously.

Regardless of terminology, one thing is certain. Judgment is a fundamental cognitive activity that vitally effects the well being -- or more accurately, the survival -- of every human being. 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. No less important are the multitude of personal decisions made by individuals and affecting only themselves and a few others.

The difficulties attendant to decision making are usually blamed on the inadequacy of the available information, and, therefore, our technological sophistication has been mobilized to remedy this problem. Devices proliferate to supply the professional decision maker with an abundance of elegant data. Consider, for example, the sophisticated electronic sensors that provide information to the physician or the satellites relaying

masses of strategic data for military intelligence. However, the problem of interpreting and integrating this information has received surprisingly little attention. The decision maker is typically left to his own devices to utilize information to its best advantage. 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 -- intuition. And when you ask him what distinguishes a good judge from a poor one he might reply,

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

But 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 and relating this knowledge to the mainstream of cognitive psychology. The second problem is a more practical one and involves the attempt to make decision making more effective and efficient.

The most significant changes have been brought about by the advent and widespread availability of the digital computer. Anyone who thinks hard about the problems of integrating information into decisions wonders about the degree to which computerized systems can alleviate them. It seems obvious that effective automation of decision making requires knowledge concerning the operations best performed by man and those best done mechanically.

The Focus of This Paper

Information processing occurs at several levels. Our concern here is

not with microscopic 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 unitary decision. These are the deliberative processes commonly referred to by the terms "weighing," "balancing," or "trading off" information, and they include the activity known as inductive inference.

Prior to 1960 there was relatively little research on human information processing at this molar, judgmental level. Some notable exceptions include Brunswik's pioneering studies of inference in uncertain environments (Brunswik, 1956), the work on "probability learning" (Estes, 1959), Edwards' (1953, 1954a, b, c) investigations into gambling decisions, 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, the intellectual heritage of this early work has been supplemented by more than 600 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 importance 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 about five 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 comparative analysis of these two broad methods of approach. Within each, we shall examine (a) the models that have been developed for prescribing and describing 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.

Of particular interest to us is the degree to which the specific models and methods characteristic of different paradigms have directed the researcher's attention to certain problem areas and caused him to neglect other problems that are equally important. Another question of interest is whether a researcher studying a particular substantive problem, such as the use of inconsistent or conflicting information, could increase his understanding by employing other models and experimental methods. We hope that by laying bare the similarities and differences between each approach we can facilitate such cross-methods research.

Areas of Neglect

Limitations of space and of our own information-processing capabilities have forced us to neglect 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 detailed analysis of 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 a detailed coverage of this work the reader is referred to the chapter by Nancy Wiggins in this book.

There have been several attempts to apply information theory to the study of human judgment. One of the most notable 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 neglect here is that of probability learning, because it has been reviewed before and because it provides the decision maker with minimal opportunity to integrate information.

Lastly, we have not attempted to review signal detection theory, a Bayesian 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 by us 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 "functional measurement" paradigm.

The Correlational Paradigm

The correlational paradigm is characterized by its use of correlational statistics to describe a judge's integration of inherently probabilistic information. The basic approach requires the judge to make quantitative evaluations of a number of stimulus objects, 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 hopes 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 (1966) 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 (taken from Dudycha & Naylor) is shown in Figure 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

---

Insert Figure 1 about here

---

be quantifiable, if only to the extent of a 0-1 (e.g., high-low, yes-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

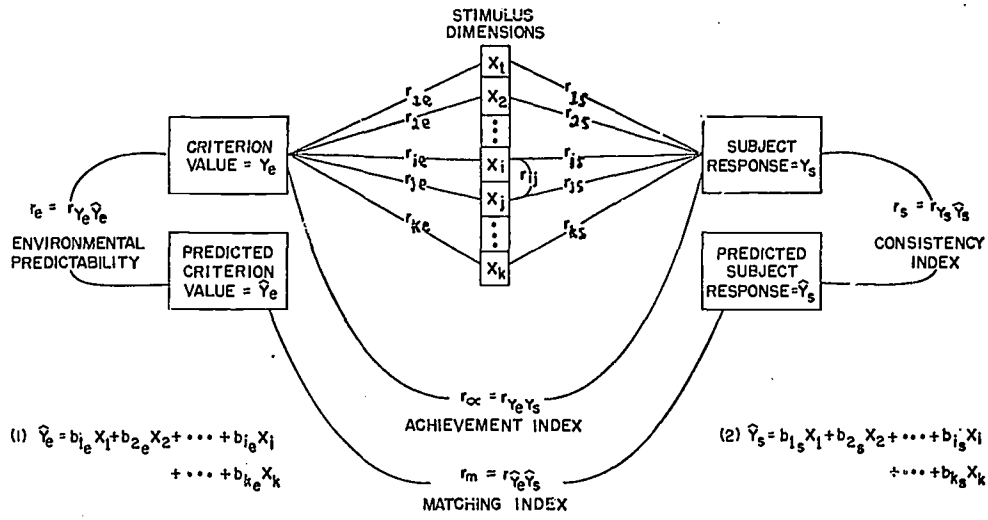


FIG. 1. Diagram of Lens Model showing the relationship among the cues, criteria, and subjects' responses.

(Taken from Dudycha and Naylor, 1966).

designated Y_e (for example, the student's actual grade point average). The relevance of the i^{th} information source 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 i^{th} 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 i^{th} cue is $r_{i,s}$, also known as his utiliztion coefficient for the i^{th} cue.

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

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

$$\hat{Y}_s = \sum_{i=1}^{k} b_{i,s} X_i \qquad (2)$$

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) represents the subject's decision-making strategy or policy. The multiple correlation coefficient, $r_s = r_{Y_s \hat{Y}_s}$, is a measure of how well his judgments can be predicted by a linear combination of cue values. It is also known as the subject's response consistency. The values of $b_{i,e}$ and $b_{i,s}$ provide measures of the importance of each cue in the ecology and for the judge.

The two most important summary measures of the judge's performance are:

$r_\alpha = r_{Y_e Y_s}$, the achievement index, and

$r_m = r_{\hat{Y}_e \hat{Y}_s}$, 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 non-linear 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:

$$r_\alpha = r_e r_s r_m + C\left[(1-r_e^2)(1-r_s^2)\right]^{\frac{1}{2}}. \qquad (3)$$

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 (r_m), and the extent to which nonlinear variance of one system is correlated with nonlinear variance of the other (C). As Hammond (1966) notes, the lens model permits a precise 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 component of the model is r_α, the achievement index. 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. Hoffman (1960) rationalizes the need for experimental control as follows:

> ". . . restricting the situation [by controlling the stimuli] assures that each person is evaluated with respect to the same information. Ambiguous and equivocal cues are removed, and all judges are thereby certain to have at their disposal the same information and no more. The inferences made beyond this point are certain to have their origins in the data provided" (p. 118).

A wide variety of mathematical models have been developed to capture the judge's policy. The first and most prominent of these is the linear model (Hoffman, 1960) as exemplified by Equation 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 simple linear combination of each of the available cues. When judgment is represented by the linear model, the $b_{i,s}$ values of Equation 2 and the utilization coefficients, $r_{i,s}$, are used to represent the relative importance given each cue. Hoffman (1960) proposed another 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 can be used to describe the relative contribution of each of the predictors as a proportion of the predictable linear variance.

Darlington (1968) has recently pointed out the unfortunate fact that indices of relative weight become suspect when the factors are intercorrelated. This problem has led many researchers to work with sets of stimuli across 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, being the antithesis of representative design. Brunswik observed (1955; pp. 204-205) that factorial designs may produce certain combinations of values that are incompatible in nature or otherwise unrealistic and disruptive of the very process they were meant to disclose. This criticism cannot be taken lightly and, as we shall see, some evidence does exist that judgment processes differ as a function of cue interrelationships. But, for the researcher who is primarily interested in relative weights, rather than in r_a, orthogonal designs often seem preferable to those in which the cues are representatively correlated. 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 their belief that 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, this quote 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).

Curvlinear functions such as this quote suggests 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 excellent example of configural reasoning involving price changes, volume of trading, and market cycle is 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 professional decision makers state that their judgments are associated with complex, sequential, and interrelated rules, chances are 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 rather unsatisfactory from a descriptive standpoint because of its lack of specificity. For example, a random, a linear, or an invalid configural strategy will each result in a C value of zero. Even a non-zero C coefficient does not indicate the form of

the nonlinear cue utilization.

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 $b_{12}X_1X_2$ will have to 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 becomes 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 judgment 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). However, the ANOVA model typically imposes two important restrictions on the factors that describe the cases being judged: (a) the levels of the factors must be categorical (e.g., good vs. average vs. poor; up vs. down, etc.) rather than continuous variables; and (b) the factors must be orthogonal. The usual way to produce orthogonality is to construct all possible combinations of the cue levels in a completely crossed factorial design. In return for these restrictions, the ANOVA model efficiently sorts the information about linear and nonlinear judgment processes into non-overlapping and meaningful portions.

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. Provided sufficient levels of the factor were included in the design, 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; that is, 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 judgment 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. The index ω^2, described by Hays (1963, pp. 324, 382, 407), provides an estimate of the proportion of the total variation in a person's judgments that can be predicted from a knowledge of the particular levels of a given cue or a pattern of cues. It includes linear and nonlinear variance and, therefore, it is analogous to, but more general than, Hoffman's index of relative weight.

One difficulty with the ANOVA technique is that a complete crossing of all possible combinations of cues becomes unmanageable when the number of cues increases above a relatively small number, 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 zero, 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, 1969; and Slovic, 1969).

The Functional Measurement Paradigm

The technique of functional measurement can be considered an extension and generalization of the correlational paradigm. As such, it has formed the basis of an intensive program of research on information processing over the past decade. The essential ideas and representative results stem from the work

of Norman Anderson, and are summarized in Anderson (1968b; 1969; 1970). A parallel approach, conjoint measurement, is outlined in papers by Luce and Tukey (1964) and Tversky (1967).

Functional measurement attempts to perform several jobs simultaneously; the scaling of stimulus attributes and response measures and the determination of the psychological law or function relating the two. Its basic premise is that measurement scales and substantive theory are integrally related. In functional measurement studies of information processing, the subject receives several items of information that are to be integrated into a single judgment. The theoretical problem is to relate this judgment to the psychological scale values and the weights of each item of information. Special attention has been given to judgmental tasks in which a simple algebraic model, involving adding, averaging, subtracting, or multiplying the informational input, serves as the substantive theory.

The main technical features of functional measurement are use of factorial designs, quantitative responses, and a procedure for monotonically rescaling these responses. The use of factorial designs arises from the fact that the theoretical models studied thus far have almost always been reducible to an ANOVA model. Therefore, ANOVA has been the principle analytical tool, serving to represent the theoretical postulates and providing a goodness of fit test of the models. In addition, ANOVA provides estimates of such theoretical parameters as the psychological values of the information items.

Some examples may serve to illustrate the method. Consider the simplest application of functional measurement -- to additive models. It is convenient to describe the model as applied to a two-way factorial design. The rows of the design would correspond to stimulus aspects (items of information) from Set $S = [S_1, S_2 \ldots S_i]$ and the columns to aspects from

Set $T = [T_1, T_2 \ldots T_j]$. Items within each set could represent the same sorts of information, as in studies of impression formation where Sets S and T each contain different adjectives descriptive of a person (Anderson, 1962). Or they could represent different stimulus dimensions, as in the correlational paradigm. An example of this is the study by Sidowski and Anderson (1967) where 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 (two adjectives or a city-occupation combination) that the judge is to integrate.

The subjective values of S_i and T_j are denoted by the corresponding lower-case letters s_i and t_j, respectively. The equation for the basic model is then:

$$R_{ij} = w_1 s_i + w_2 t_j \quad , \tag{4}$$

where R_{ij} is the theoretical response to the stimulus described by the pair of items (S_i, T_j) and w_1 and w_2 are the weights of the row and column dimensions. It is usually assumed that the subjective value of an item is independent of the other item with which it is paired and that w_1 and w_2 are constant over row and column stimuli, respectively.

Equation (4) implies that the row by column interaction is 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_i and T_j. For example, if Equation (4) is averaged over columns, the mean for Row i is

$$R_{i\cdot} = w_1 s_i + \text{constant}, \tag{5}$$

where the dot subscript on R denotes the average over the column index. The constant expression is $w_2 t_\cdot$, the same for all rows. Equation (5) says that the row means form a linear function of the subjective values of the row stimuli.

In terms of the raw data, then, the row means constitute an interval scale of the row stimuli. Similarly, the column means constitute an interval scale of the column stimuli.

All of the above results hold for an additive model. An averaging model constrains the weights to sum to unity and this constraint provides a basis for estimating both scale values and weights (Anderson, 1970).

Anderson (1969) notes that caution is required in interpreting the meaning of significant interactions when these occur. Interactions may occur as a result of cognitive configurality that violates the model or from defects in the measurement scale of the response such as floor and ceiling effects, response preferences, and anchor effects. In some cases, a monotonic rescaling of the response can be used to eliminate the interaction and save the model.

Anderson and colleagues have applied these techniques in a number of ingenious ways to study various substantive problems of information processing. For example, they have studied tasks in which stimuli were presented in serial order and the serial positions corresponded to factors in the design. The weights indicated by the main effects thus produce a serial position curve that can be used to assess primacy or recency effects in information combination (Anderson, 1965b; 1968a; Shanteau, 1969). When information is presented serially, Anderson (1965b) noted that 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 the k^{th} item of information is received, is given by

$$R_k = R_{k-1} + C_k (S_k - R_{k-1}) \quad , \qquad (6)$$

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

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 principal 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 the world of business and economics by Schlaifer (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 these two papers.

The Bayesian approach is thoroughly embedded within the framework of decision theory. Its basic tenets are that probability is orderly opinion and that the optimal revision of such opinion, in the light of relevant new information, is 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. 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, D, Bayes' theorem states that:

$$P(H_i/D) = \frac{P(D/H_i)\ P(H_i)}{P(D)} \quad . \tag{7}$$

In Equation (7), $P(H_i/D)$ 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. P(D), the probability of the datum, serves as a normalizing constant, and is equal to $\Sigma_i\ P(D/H_i)\ P(H_i)$. Although Equation (7) is appropriate for discrete hypotheses, it can be rewritten, using integrals, to handle a continuous set of hypotheses and continuously varying data (Edwards, 1966).

It is often convenient to form the ratio of Equation (7) taken with respect to two hypotheses, H_i and H_j:

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

For this ratio form, new symbols are introduced:

$$\Omega_1 = LR_D \cdot \Omega_o \quad ;$$

the posterior odds,

$$\Omega_1 = \frac{P(H_i/D)}{P(H_j/D)} \quad ,$$

are equal to the product of the likelihood ratio of the datum,

$$LR_D = \frac{P(D/H_i)}{P(D/H_j)} \quad ,$$

times the prior odds,

$$\Omega_o = \frac{P(H_i)}{P(H_j)} \quad .$$

Bayes' theorem can be used sequentially to measure the impact of several data. The posterior probability computed for just 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. For n data, D_k, the final posterior odds, given all the data, are

$$\Omega_n = \prod_{k=1}^{n} LR_{D_k} \cdot \Omega_o \quad . \qquad (8)$$

Equation (8) 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.

Although the ratio form of Bayes' theorem summarizes all the information only in the case of two hypotheses, Bayes' theorem can be used with any number of hypotheses, in which case one ends with a set of posterior odds that can be translated into a distribution of posterior probabilities across all the hypotheses.

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

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

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

$$P(H_i/D_1,D_2) = \frac{P(D_2/H_i,D_1)\ P(D_1/H_i)\ P(H_i)}{\text{normalizing constant}} \quad . \qquad (9)$$

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

The meaning of the conditional indpendence assumption might be clarified by some examples. 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. The evidence from these cues could not be combined via Bayes' theorem without altering it as shown in Equation (9). 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 model. 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. The other bag holds 30 red chips and 70 blue chips -- but the subject does not know which bag is which. The experimenter flips a coin to choose one of the bags. He 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 has in front of him a device for recording his responses: two upright rods and on them, 100 washers. Behind the rods is a board calibrated so that the subject can easily tell how many washers are on each rod. One rod is labeled "70% Red," the other, "70% Blue."

The subject is asked to use the washers to express his opinion as to the probability that the predominantly red or predominantly blue bag is the one being sampled from. When the subject puts 75 washers on the "70% Red" rod and 25 on the "70% Blue" rod he is indicating his opinion that the chances are 75 in 100 that the predominantly red bag was the one chosen.

At the start, before the first chip is drawn, the subject is required to place 50 washers on each rod, 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 moving from one rod to the other as many washers as he wishes. The subject sees 10 successive chips drawn; the basic information for the data analysis is the 10 responses 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 1 (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 urn:

$$LR_{\text{Red Chip}} = \frac{P(\text{Red Chip}/H_{70\% \text{ Red}})}{P(\text{Red Chip}/H_{70\% \text{ Blue}})} = \frac{.7}{.3} \quad ;$$

$$LR_{\text{Blue Chip}} = \frac{P(\text{Blue Chip}/H_{70\% \text{ Red}})}{P(\text{Blue Chip}/H_{70\% \text{ Blue}})} = \frac{.3}{.7} \quad .$$

Thus the posterior odds of the predominently red urn having been chosen, given a sample of, say, 6 red chips and 4 blue chips are:

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

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

The primary data analysis compares subjects' probability revision 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:

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

where SLLR is the log likelihood ratio inferred from the subjects' probability estimates and BLLR is the optimal (Bayesian) log likelihood ratio. When log likelihood ratios are used, the optimal responses become linear with the amount of evidence favoring one hypothesis over the other. The AR can be viewed as the slope of the best fitting line on a plot of the log of subjects' responses against the log of optimal responses. The AR will be equal to 1

when the subject revises optimally, will be greater than 1 when the subject's revisions imply greater certainty about the truth of one hypothesis than is justified by the data, and less than 1 when the reverse is true. In the rare case when a subject treats a datum as if it pointed to one hypothesis while the datum in fact pointed to the other, the AR would be negative.

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.

Other Bayesian experiments have employed multinomial distributions to generate samples of data. Table 1 provides a hypothetical illustration.

Table 1

Some Multinomial Data Generating Hypotheses

Data Class	Subclasses	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. 15 - 18	.30	.30	.30
	4. Above 18	.30	.15	.30

Note.--Cell entries are $P(D_{jk}/H_i)$ values.

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 the resulting evidence-hypothesis matrix are conditional probabilities of the form $P(D_{jk}/H_i)$, i.e., the probability that the k^{th} 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.

Table 1 may be used to generate data sequences (hypothetical students) whose GPA classification is to be predicted by subjects.

Schum (1966b,p. 35) describes the experimental paradigm, based on a multinomial task, for studying information processing:

> "Samples of evidence, . . . , form the basic input to subjects. In experimental inference situations subjects are asked to assume some prior probability distribution across the hypothesis set. Then subjects aggregate and process the samples of evidence in order to revise their prior opinions about the likelihood that the various hypotheses had generated the evidence. In order to do this they need indications of the impact of each item of evidence [i.e., $P(D_{jk}/H_i)$]. The experimental paradigm described

above can rest upon an objective probability base in which an objective $P(D_{jk}/H_i)$ matrix has been prscribed by the experimenter. Such a matrix can be given directly to subjects in which case their only task is to aggregate items of evidence with prescribed impact. In another case, subjects might be required to estimate the multinomial $P(D_{jk}/H_i)$ distributions under each hypothesis from relative frequencies of occurrence of the various subclasses. Subsequent knowledge of which hypothesis in fact generated each sample of evidence is necessary in this case so that the evidence-hypothesis relationships or diagnostic impact of each item of evidence can be learned by the subject. In either case, subjective probability revisions (on the basis of evidence) are in the form of posterior probabilities $[P(H_i/D)]$ or some analog such as posterior odds.

The direction and size of these revisions can be compared with the theoretical revisions prescribed by 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 for making a good decision. In seeking additional information, the decision maker must weight the relative advantage of the information to be purchased against its cost. When the probabilistic characteristics of the task are well specified, 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, specify a stopping point that will be 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 inasmuch as 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,c), Rapoport (1969), and Wallsten (1968) for examples of this and related 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.

Points of Similarity and Difference

First and foremost, each paradigm is based on a theoretical model of the process whereby information is integrated into a judgment or decision. Furthermore, these various models are closely related. The simple linear model plays a key role in both correlational and functional measurement studies. Bayes' theorem is also a linear model under a logarithmic transformation (i.e., Equation (8) translates into $\log \Omega_n = \sum_{k=1}^{n} \log LR_{D_K} + \log \Omega_o$). In addition, the proportional change model of impression formation

(Equation 6) is Bayesian in spirit since it conceptualizes the step-by-step buildup of judgment in terms of a weighted combination of the present datum and prior response. Each of these models contains analogous descriptive parameters for the purpose of assessing the relevance of data dimensions or data items for the judge. Thus correlational studies describe correlations ($r_{i,s}$), regression weights ($b_{i,s}$), or relative weights ($RW_{i,s}$); ANOVA studies estimate $\omega^2_{i,s}$ values; functional measurement models produce estimates of w_k; and Bayesians infer subjective likelihood ratios. Despite the different terminology, the similarity of purpose is marked.

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. However, the meaning of optimality differs in the two models. The optimality of multiple regression rests on the acceptance of a built-in payoff function: the least squares criterion of goodness-of-fit. The Bayesian model is optimal in a different sense: an idealized rational decision maker can be satisfied that he is logically consistent. The resulting posterior probabilities can be combined with <u>any</u> 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). In contrast to the lens and Bayesian models, the functional measurement approach and the stream of correlational research exemplified by Hoffman's work, differ by being exclusively descriptive in intent.

The data that serve as input to the decision maker vary somewhat both within and across each approach. The correlational paradigm typically involves <u>dimensions</u> of data, usually monotonically related to the criterion.

Data processed within the functional measurement and Bayesian studies, by contrast, are typically discrete, categorical particles, although these approaches can also process dimensional data. As to the relationsips among data sources, functional measurement requires factorially combined data elements and workers within the descriptive stream of correlational research also prefer orthogonal structure. Lens model research often uses data that are correlated in a fashion representative of the real world. Bayes' theorem, however, requires conditionally independent data. A rough translation of this requirement in correlational terms would be that the correlations between cues, with the criterion dimension partialled out, must be zero.

The response required of the subject also differs across paradigms. The correlational and functional measurement 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 (for example, 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 principal, 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).

The Bayesian paradigm looks at fixed hypotheses and examines the manner in which their subjective likelihood is revised in the light of new information about the world. For this reason it has 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 is unchanged over this period. However, the static vs. dynamic distinction is not inherent in the models. A good example of this point is illustrated within the functional measurement paradigm where the weighted average model takes both a static form (Equation 4) and a dynamic form -- the "proportional change model" of Equation 6. In like manner, a regression equation can handle information sequentially and the item-by-item revision of judgment can be compared to optimal revisions specified by the equation.

Testing the Models

Although the models are similar, the attitudes of researchers towards testing them differ somewhat. 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:

> ". . . Ss 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 Ss' revision rule is essentially Bayes' theorem. That is to say, Ss' 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 functional measurement paradigm, has chided other 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 of this 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 meaningful interaction.

The Dilemma of Paramorphic Representation

The mathematical models we have been discussing serve, at the very least, to provide a quantified, overall, descriptive summary of the manner in which information is weighted and combined. To what extent do they represent the actual cognitive operations performed by the judge? Hoffman (1960, 1968) raised a problem particularly germane to this question. 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 minerology, Hoffman introduced the term "paramorphic representation" to remind us 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 problem in connection with models based upon correlational techniques, Bayesian models also face this dilemma.

Empirical Research

As the preceeding discussion indicated, judgment researchers are studying similar phenomena but with somewhat different methods. In the remainder of this chapter 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 functional measurement paradigms. We have further categorized the work according to five broad problem areas relating to the use of information by the decision maker.

Insert Table 2 about here

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 problem of whether to include non-linear terms in the policy equation. The focal topic of functional measurement is the distinction between two variants of the linear model, the summation model and the averaging model, in impression formation. In Bayesian research, the focal topic is a particular form of sub-optimal performance called conservatism. These topics are not closely inter-related. They are emphasized here simply because they have received so much attention in the three research areas.

The second research 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 task variable of number of

Table 2. Overview of Topics in Bayesian and Regression Studies of Judgment

	Regression Studies		Bayesian Studies
	Correlational Paradigm	Functional Measurement Paradigm	Probability Estimation Paradigm
Typical Dependent Variables	r_s, $r_{i,s}$, $b_{i,s}$, r_α	weights (w_i) and scale values (s_i and t_j); 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	response mode payoffs data diagnosticity prior probabilities sequence length
III. Sequential Determinants of Information Use	little or no research	primacy vs. recency	primacy vs. recency the inertia effect
IV. Learning to Use Information	single-cue functional learning positive vs. negative cues multiple-cue learning type of feedback interpersonal learning use of non-metric cues	little or no research	effects of payoffs % and type of feedback
V. Descriptive Strategies: What is the Judge Really Doing?	strategies in correlational research	little or no research	strategies for estimating P(H/D)
VI. Aiding the Decision Maker	bootstrapping	little or no research	probabilistic information processing systems

items of information. In the correlational paradigm, Einhorn (in press) has shown a decrease in r_s, subjects' consistency, as the number of cues increased. Anderson (1965a) used varying set sizes in a functional measurement paradigm to test predictions of his 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 -- e.g., 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 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 sequential determinants of information use, learning to use information, strategies for combining information, and techniques for aiding the decision maker.

Focal Topic of Correlational and ANOVA Research: Modeling a Judge's Policy

The Linear Model

A large number of studies have attempted to represent the judge's idiosyncratic weighting policy by means of the linear model (Equation 2). Examination of more than thirty 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, in press; 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). In some cases, the stimuli are artificial and the judges are unfamiliar with the task. Typical of these is a study by Knox & Hoffman (1962), who had college students 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 had students 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 binary 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. Another example is the work of Dawes (1970) who built 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. Most of these models were not cross-validated. However, in the few studies that have applied the linear model derived from one sample of judgments to predict a second sample, there has been remarkably little shrinkage -- usually only a few points (Einhorn, in press; Slovic & Lichtenstein, 1968; Summers & Stewart, 1968; and Wiggins & Hoffman, 1968).

Capturing a Judge's Policy

The various dimensions of a stimulus object are certainly not equally important and judges do not weight them equally. One of the purposes of using the linear model to represent the judgment 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 (1967) examined the policies whereby hospital personnel granted weekend passes to patients at a mental hospital. They noted that: "For five of the six items (cues) there was at least one judge for whom it was the most important and at least one for whom it was nonsignificant" (p. 196). 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 categories 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(in press)describes statistical evidence obtained by other researchers indicating just the opposite -- that large craters are more likely than small ones to be malignant. In a similar fashion, Hammond, Hursch, and Todd (1964) reanalyzed data obtained by Grebstein (1963) in which clinical

psychologists with varying levels of experience judged IQ on the basis of Rorschach signs. By examining policy equations they were able to trace the performance decrement shown by inexperienced judges to their misuse of one particular sign.

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; 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 repeated 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. Thus, it would appear to have tremendous potential for providing insight into expert judgment in many areas.

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 versus 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 good 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, 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 configurality using the C coefficient, although the ambiguous nature of low C values does not preclude the possibility of configural judgment processes (lack of nonlinearity in the environment or a difference between the nonlinearity in the environment and judgmental systems are sufficient to insure low C values).

In light of the simple but compelling arithmetic underlying the "main effect approximation" one would expect that the results of this early research should not have been too surprising. Yet the search continued, buoyed by (a) the repeated assertions of human 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. Their data, 861 MMPI profiles from 7 hospitals and clinics, was selected because MMPI lore considered it 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 ($X_i \cdot X_j$). 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 thirteen subjects were best described by the sign model, 3 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).

Summers and Stewart (1968) applied a similar tactic to search for nonlinearity in a new domain. They had undergraduate subjects predict the long-range effects of various foreign policies on the basis of four cues. Application of a linear and quadratic model in derivation and cross-validation samples produced results quite congruent with those of Wiggins and Hoffman. Studies in which judges 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 configural and curvilinear 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 human judgment 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 highlights their importance with regard to theoretical and general explanatory considerations. Their philosophy is typified by Green's argument to the effect that :

> "Nonlinear relationships and interactions are to a first approximation linear and additive . . . . If the goal is prediction . . . an adequate description will serve. But if the goal is to understand the process, then we must beware of analyses that mask complexities" (Green, 1968, p. 98).

To illustrate the complexity inherent in judgments that are quite predictable with a linear model, consider the data from the previously-described study of ulcer diagnosis conducted by Hoffman, Slovic, and Rorer (1968). An ANOVA technique showed that each of the nine radiologists who served as subjects exhibited at least two statistically significant interactions. One

showed thirteen. 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 occurs 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 stock-brokers, and Kort (1968), in his study of workmen's compensation decisions, did and both uncovered information that provided worthwhile insights into 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) had subjects 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. The judgments of each city-job combination were found to be a weighted sum of the values for the two components except that the attractiveness of being a teacher was more dependent upon the attractiveness of the city than were the other occupations. This may be because teachers are in more direct contact with the cities' socio-economic milieu. Other interesting examples of interactive cue utilization have been found by Lampel and Anderson (1968) and Gollob (1968).

Hoffman (1968) 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 looked to other cues for resolution of the conflict. This study and related studies reported in Hoffman (1968) and Anderson and Jacobson (1965) indicate that the linear model needs to be ammended to include a term sensitive to the level of incompatibility among important cues.

Tversky (1969) and Einhorn (1970, in press) also hypothesized, and found, specific nonlinear uses of information. Tversky showed that 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 ξ. If the difference was less than or equal to ξ, 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 ξ on the probability dimension.

Einhorn made a valuable contribution to our understanding of nonlinear judgment by developing mathematical functions that could be incorporated into a prediction equation to approximate conjunctive and disjunctive 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 is a function of the maximum value of the stimulus

or one of its attributes. For example, a scout for a pro 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 weighs 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. Using a cross-validation sample of judgments, he found that many subjects were fit better by the nonlinear, noncompensatory models than by the linear model. The conjunctive model proved superior to the disjunctive model, especially as the number of cues increased. 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 a judge's responses. We think of these as computed or objective policies. Judges in a number of studies, after a policy was computed on the basis of their responses, were asked to describe the relative weights they were using 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 the 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. 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 certain individual judges.

One type of error in self insight has emerged in all of these studies. Judges universally and 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. This question is an intriguing one and needs to be investigated with more precision than was done in this study. It may be that the most experienced judges produce verbal rationales for their evaluations that are less trustworthy than those of their inexperienced colleagues!

Task Determinants in Correlational and ANOVA 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, subjects used both and weighted them equally. When they disagreed, subjects focused on one of the cues or looked to other cues for resolution of the conflict. Also, in situations of higher cue conflict, r_s was considerably lower. These effects were found both when cue-conflict and cue-intercorrelations varied together and when conflict was varied holding intercorrelations constant. Dudycha and Naylor (1966) have also studied the effect of varying the intercorrelations among cues upon policy equations. They found that profiles of $r_{i,s}$ values showed less scatter and r_s decreased as the correlation between cues decreased.

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, will lead to greater weighting of this cue by a judge. This increased weight will persist, they proposed, even among subsets of stimuli for which this cue is not unusually variable. Underlying this hypothesis was the assumption that the judge is motivated to differentiate among stimuli along the criterion dimension and that cues which increase his ability to differentiate will reinforce and increase 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 one 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. For one group of subjects a highly valid cue was manipulated in this way. For a second group, the variability of a minor cue was altered. The hypothesized effect was found in seven out of ten subjects when a strong cue was manipulated. Increasing the variability of the minor cue had no effect upon its subsequent use. The authors came to the tentative conclusion that the judge may alter his system of judgment because of the characteristics of samples he judges.

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 scores were presented as T scores with a mean of 50 and standard deviation of 10. A second condition presented the information in percentile scores. The latter profiles showed considerably more scatter and had more extreme scores than the former 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. Beta weights did not differ significantly between formats.

Number of Cues

There has been surprisingly little correlational research done on the

effects of varying the number of cues upon a judge's performance. A pilot study by Hoffman and Blanchard (1961) suggested some interesting effects but was limited by a small number of subjects. They had subjects predict a person's weight on the basis of physical characteristics. They judged the same stimulus at different times, seeing either 2, 5, or 7 cues. 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 (in press) also found that response 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 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. It is obvious, however, that more work is needed in this area.

Cue-Response Compatibility

Fitts and Deininger (1954) introduced the concept of stimulus-response compatibility to explain the results of several paired-associates learning and reaction time experiments. 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 a more recent 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 -- namely equating the gamble with an amount of money such that they would be indifferent between playing the gamble and receiving the stated amount for certain. 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 a general hypothesis to the effect 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 Functional Measurement: 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. Most 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 and statistical techniques such as ANOVA within the conceptual framework of

functional measurement.

One of the first studies to test the additive model empirically was done 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.

Anderson's (1962) study 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. Most of the 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 of modeling.

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 affect 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 bi-polar adjective scales to evaluate the stimulus persons. The judgments became more favorable as the amount of information increased thus supporting the summation model. The Fishbein and Hunter study has been criticized by Rosenberg (1968) who argued that presenting the most favorable adjectives first permits possible sequential effects to influence the results.

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. Another result of this 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. At first glance this seems to support a summation model but Anderson showed how it could be accommodated using an averaging model that incorporates an initial impression with non-zero weight and scale value s_o. The final impression is assumed to be an average of this initial impression and the scale values of the traits. Algebraically,

$$J_n = \frac{nws + (1-w)\, s_o}{nw + (1-w)}$$

where J_n is the judgment based on n adjectives each of scale value s and weight w. The term s_o represents the initial or neutral impression. Its

relative importance is (1-w)/[nw + (1-w)], a value that decreases as more information accumulates. 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 Kerrick (1958), Manis, Gleason, and Dawes (1966), 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.

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) for each adjective (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 only 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 not be as valid as the others. Anderson and Jacobson argued that the averaging model might have to include differential weights to accommodate the reduced impact of inconsistent information. However, they also pointed out that

their results might be caused by an order effect of the sort that gives more weight to information earlier in the sequence.

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 $P_{AB} - P_A P_B$, 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 here 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 subjective personality traits.

Other Contextual Effects

Anderson and Lampel (1965) and Anderson (1966) had subjects form an impression based on three adjectives and then rate the likability of 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. Anderson (1966) noted that this type of deviation from an additive model does not show up as an interaction effect. Thus, lack of interaction

does not unequivocably support the additive model. Anderson suggests that the most 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) found evidence to support this change of meaning interpretation over several competing hypotheses and Chalmers (1969) argued that change of meaning could readily be accommodated by Anderson's weighted average model.

Concluding Comments

The additive model dominates the area of impression formation much as the linear model dominates correlational research. Like linearity, additivity is not a completely satisfactory concept, however, and there are many subtle factors competing with one another to determine the deviations in the data. The contradictory findings in this area are difficult to evaluate due to the considerable variation in types of stimuli, response modes, and instructions across studies. Manis, Gleason, and Dawes (1966; p. 418) seem to summarize the present state of affairs most aptly with their comment:

> "It seems clear that our main need at the present time is for more research concerning those variables (e.g., topic, situation, subject characteristics) which determine the combinatorial model that is applied in the evaluation of complex social stimuli; available evidence suggests that there is no single model which can be universally applied."

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 change after one observation.

Much of the Bayesian research has been motivated by a desire to better understand 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 principle competing explanations as to the "locus of conservatism" are the misperception, misaggregation, and artifact hypotheses (Edwards, 1968).

Misperception

Perhaps subjects don't understand the data generator underlying, or producing, the probabilistic data. 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 Bientema (1967) and Vlek and van der Heijden (1967) showed that it does. Vlek and Bientema presented subjects with samples (e.g., 5 black and 4 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. Beach (1966) showed that subjects do exhibit this consistency even when poorly trained in the characteristics of the task. He concluded that ". . . even though subjects' subjective probabilities were inaccurate, they were still the bases for decisions . . ." (p. 35). Peterson, Ulehla, Miller, Bourne, and Stilson (1965) asked subjects in a binomial dice task "what is the probability that Die W is generating the data?" and "what is the probability that the next roll will be White?" Their answers were conservative but consistent. Peterson, DuCharme and Edwards (1968) had subjects estimate P(H/D), then P(D/H). Then they were instructed in P(D/H) by being shown several theoretical sampling distributions from an urn 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 the use of subjects' estimates of P(D/H) to predict estimates of P(H/D) accounted for all the conservatism of the latter responses. 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 distributions available at the time they made post-instruction 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 lies in the perception of 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 plaintive 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 occurrance: the likelihood ratio. The violation of this likelihood principle has been shown by Vlek (1965) and Vlek and van der Heijden (1967), who show a systematic decrease in the Accuracy Ratio as a function of the rarity of the data, and it can serve as an explanation to Lichtenstein and Feeney's (1968) results. Beach (1968) directly tested Vlek's idea. 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, 6 red and 3 blue chips, and asked to state the probability that such a sample might occur, the subject must, in a sense, aggregate the separate impacts 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 events 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. This same result was shown by Peterson and Swensson (1968), using the usual binomial task and an unusual diffuse-hypothesis task. For the latter, the subjects were told to imagine an urn containing many thousand red and blue chips. The proportion of red chips was decided by a random procedure such that all percentages were equally likely. A defining sample (varying from 1 to 19 chips) was then shown; the subjects were asked to make a point estimate of the proportion of red chips in the urn. Following each such estimation process, the subjects were asked to imagine a second urn with proportions of chips just the reverse of the present (unknown) urn. They were then shown an informational sample (one chip or 5 sequential chips) and asked to give posterior odds regarding which of the two urns had been sampled. It was these estimates, as well as the estimates made to the more usual binomial task, that were very nearly optimal for one datum and the first datum out of five, but more conservative for data two through five.

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 approach, since in that study the one datum had to be aggregated with a varying value for the prior probability of the hypothesis. In addition, Peterson and Miller used a probability response mode. This mode, as will be discussed later, is highly susceptible to a non-optimal but simple strategy which produces artifactual results. Peterson and Phillips (1966) also found first-trial conservatism using a probability response mode. 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.

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).

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.

Artifact

The third explanation of conservatism, that conservatism is artifactual, was originally suggested by Peterson (see Edwards, 1968), and has been recently supported and renamed response bias by DuCharme (in press). 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, across all trials, was marked 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. Phillips and Edwards (1966) compared four different direct estimation modes in a bookbag and poker chip task. The "probability" response was made by distributing 100 white discs in two vertical troughs, the height of the discs in each trough indicating the probabilities of the two hypotheses. The "verbal odds" response was a verbal statement of the posterior odds after each datum. The "log odds" group estimated posterior odds by setting a sliding pointer on a scale of odds spaced logarithmically; the scale ran from 1:1 to 10,000:1. The "log probability" subjects used a similar sliding scale, labeled in probabilities rather than odds, where the spacing of the probabilities was determined by converting the probabilities to odds and scaling the odds logarithmically.

The motivation for using odds and log odds as response alternatives rested upon an uneasiness with the properties of a probability scale. The amount of change in posterior probabilities induced by a single datum decreases as the probabilities prior to the receipt of that datum become more extreme.

In addition to this problem of nonlinearity between data and response, there is a potential problem with ceiling effects because of the boundedness of the probability scale at zero and one. The subject may be reluctant, in a long task, to give an extreme response early in the sequence, for fear of "using up" the scale before the last data arrive. When odds and log odds are used, however, both these difficulties are avoided: odds bear a constant multiplicative relationship to binomial data, while log odds are linearly related to such data. In addition, neither scale has a ceiling.

The Phillips and Edwards results were consistent with the above reasoning: for three different bag-compositions and for five different sequences of 20 chips from each, the "verbal odds" and "log odds" groups showed the least conservatism.

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 across different diagnosticity levels, between the direct, verbal odds mode and the choice among bets response. 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 in a situation requiring subjects to learn the probabilities associated with the flashing of seven lights. They found that estimated and inferred probabilities correlated .93 (slope = 1.06), averaged across 20 subjects. Strong agreement between probability estimates and probabilities inferred from bids, has also been found in two studies by Beach and Wise (1969a, c). 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, in a bookbag and poker chip task. Prior to each sampled chip, subjects predicted the color of the chip; after the chip was shown, subjects guessed which hypothesis was true (it was this decision that stopped the clock); then subjects indicated their certainty in the decision by assigning a confidence judgment to the chosen hypothesis. These confidence judgments have been shown to be consistently related to probability responses (Beach & Wise, 1969b). 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 confidence.

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 (1969c), 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, using sequences of ten data, did find differences attributable to repeated responses. Four of his groups gave confidence ratings after each datum; the groups differed in the degree to which the responses made on previous trials were available to them. The fifth group gave confidence ratings only after seeing all ten data. The group that made repeated responses in a way that was most difficult to remember performed similarly to the group making only a single, final response, but the results from other groups with repeated responses showed a non-optimal sequence effect. Halpern and Ulehla (1970), using a signal detection task, also found differences between repeated responses and a single, final response; the latter more closely matched an internal-consistency prediction derived from signal detection theory than did the former.

Nominal vs. probability responses. Another question of interest is whether there is 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). Unfortunately, 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:

Quadratic: $v(p) = 10{,}000 - 10{,}000\,(1-p)^2$

Logarithmic: $v(p) = 10{,}000 + 5{,}000\,\log_{10} p$

Linear: $v(p) = 10{,}000p$

The quadratic and log payoffs share the characteristic that the subject can expect to win the most points by reporting his true subjective probability (Toda, 1963). For the linear payoff, the subject ought always to estimate 1.0 for the more likely hypothesis.

The results indicated that payoffs help to decrease conservatism, but do not eliminate it. The log group was better than the quadratic group, which differed little from the control group. The linear group made many extreme estimates (reported probabilities larger than the Bayesian probabilities), reflecting a tendency in the direction of the optimal strategy of reporting all estimates as 1.0 or 0. 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. Three payoffs were used, based on the subject's estimated posterior probability at the end of each sequence.

Logarithmic: $v(p) = 10 + 12.851 \log_{10} p$,

Linear: $v(p) = 12p - 2$, and

All-or-none: S received 10 points if the hypothesis to which he assigned the largest posterior probability was correct; otherwise he received 0 points.

In the all-or-none payoff scheme the size of the posterior probabilities was irrelevant, and the payoff provided no instructional feedback to the subject regarding the size of his response. Nevertheless, the all-or-none group was only slightly inferior to the log group; both groups were conservative except with the short (four-item) sequences. The linear group was not, on the average, conservative, but the responses were highly variable: the posterior odds inferred from subjects' responses 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.

Whereas the studies just described varied payoffs as a function of slight differences in probability estimates, Pitz and Downing (1967) studied the effects of payoffs on a much grosser level of response -- namely binary predictions. 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, relative to the payoffs associated with the other prediction. The introduction of biased payoffs is a second way in which this study differs from those described above. The subjects were highly optimal when using the symmetric matrix. But although they altered their predictions as a function of varying payoffs, 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, 20 green or 20 red, 80 green than when the possible contents of the bag are more similar, say 60 red, 40 green versus 40 red, 60 green. Several experiments (Peterson, DuCharme, & Edwards, 1968; Peterson & Miller, 1965; Phillips & Edwards, 1966; Pitz, Downing, & Reinhold, 1967; and Vlek, 1965) have manipulated diagnosticity in this way and all have found 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 or scenarios 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 scenarios of higher overall diagnosticity lead to greater conservatism. Martin and Gettys found that their least diagnostic scenarios 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. 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 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 4 reds and 2 blues in a sample of 6 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) show that when this difference is held constant, the larger sample sizes yield 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 differs in the binomial and multinomial task. In the binomial task this is effected by including several data with equal but opposite diagnostic values (red and blue chips) which cancel each other out. But in Schum's task with six hypotheses and data of varied diagnosticity, total diagnosticity can be held constant only by using individual data in the 18-data sample that are each, on the average, of low diagnosticity compared to the data used in the 6-data 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 4 data each, with 4 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 simultaneous than with successive presentation of data.

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 scenario. They used two different data-generating models, Model A and Model B. Both models were simpler than the multinomial models used in other research in that every possible datum favored just one hypothesis, with the five other hypotheses being equally less likely. The impact of a single datum upon the hypothesis favored by

that datum was always the same within a model, but differed between models. Each datum from Model A was more diagnostic than each datum from Model B. Any scenario of six data could be characterized by the number of data favoring the most-favored hypothesis. At one extreme, each datum could favor a different hypothesis; then the posterior odds between any pair of hypotheses was 1.0. At the other extreme, all six data could favor the same hypothesis; this would be the most diagnostic case. There were nine other cases of intermediate diagnosticity.

Each subject gave posterior probability responses to 264 scenarios (six data presented simultaneously); there were 12 different scenarios for each of the 11 diagnosticity cases for each of the 2 models. The appropriate conditional probability matrix (either Model A or Model B) was always displayed to subjects. Subjective log likelihood ratios were computed from the probability estimates and compared with Bayesian log likelihood ratios. The results from Model A were typical of the diagnosticity studies previously mentioned -- subjects were sensitive to changes in the diagnosticity, but as diagnosticity increased, subjects became increasingly conservative.

The results from Model B scenarios represented a unique finding -- as diagnosticity increased in Model B scenarios, extremity of response increased. Seven of the eight subjects showed this effect; the other subject was slightly more variable but otherwise nearly optimal. 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 strategies look conservative in one case and extreme in the other case.

The Effects of Manipulating Prior Probabilities

The results of several studies in which the prior probabilities were systematically varied are mixed. Three studies found no systematic or significant effect of this variable upon subjects' responses: Phillips and Edwards (1966) reported no effect on Accuracy Ratio attributable to five different prior probability levels in a bookbag and poker chip experiment. Phillips, Hays, and Edwards (1966), and Schum (1966b) reported no effects due to change in priors in multinomial tasks.

Strub (1969), using a binomial task, observed that subjects' terminal posterior probability estimates after 100 data were higher for priors of .90 - .10 than for priors of .50 - .50, but he did not report his data in sufficient detail to determine whether one condition produced more optimal behavior than the other.

Peterson and Miller (1965), recognizing that the place to look for the effect of priors is right at the beginning of data accumulation, rolled one of two dice just once for every "sequence." They used nine levels of prior probabilities, from .1 to .9, and found that subjects' Accuracy Ratios increased (became less conservative) as the priors became more extreme (departed from .5). This clear-cut 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. The above findings

regarding priors are too inconclusive to fit in this mold, but exactly this result of varying priors has been found using signal detection models by Ulehla (1966) and Galanter and Holman (1967). Wendt (1969) also found partial sensitivity to priors. 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 priors 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. In one, 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(D/H) to be greater for shorter sequences than for longer ones. Similarly, Shanteau (1969) 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.

Sequential Determinants of Information Use

More often than not information is presented to a judge or decision maker in sequential order. In some cases, the decision is evaluated after each new item of information. In others the decision maker must wait until all the information has been presented before responding. In both cases it is of interest to determine whether the way that information is used depends upon the order in which it appears in the sequence.

Primacy and Recency Effects

Without a doubt the most thoroughly investigated type of sequential effect aims 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. The issue seems to have been studied first by Lund (1925) who presented evidence supporting a "Law of Primacy in Persuasion" but the modern impetus can be attributed to the work of Asch (1946). Asch had subjects judge the favorableness of a person described by six adjectives. When these adjectives were read in decreasing order of favorableness (i.e., intelligent, industrious, forceful, critical, stubborn, envious) the final impression was more favorable than when the reverse order was used, indicating a primacy effect. Asch hypothesized that the initial adjectives set up a "directed impression" that caused the later adjectives to shift their meanings to conform to the existing set. This research stimulated a body of research on order effects in persuasion, summarized in a book edited by Hovland (1957). Even by this early date it was evident that there was no completely general principle of primacy and recent work has borne this out, focusing instead on delineating some of the situational

factors influencing order effects.

Table 3 presents an overview of more than two dozen studies of primacy and recency. We have grouped these studies into three categories according to the type of stimulus information with which the judge had to deal. The first group of studies involves verbal items, such as adjectives to be integrated into an overall impression of a person, foods descriptive of a meal, headlines descriptive of a paper, etc. In the second class of studies, subjects were presented with simple quantitative or perceptual inputs, either numbers, weights, or lengths of lines, which had to be averaged. Group III consists of studies where the subject had to make probability estimates or predictions about the true state of the world on the basis of sample data.

Insert Table 3 about here

Within each major class, the studies have been further subdivided into two categories, depending on whether the judgment was made only after the final item of the information sequence (coded F) as opposed to being made after each item of information or after several but not all of the data were received. These latter two conditions have been coded I, for intermittent.

Category I; Verbal information. Studies involving verbal items of information have typically employed some version of the following design to assess order effects. First the items are scaled individually with respect to the criterion. These sets are then presented in ascending scale order and vice versa, as in the Asch study. A related procedure first sorts items 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 a HHLL sequence to be higher than that for a

Table 3

Studies of Primacy and Recency

Type of Information	Stimulus Items	Total Set Size	Response	Time of Response[a] I or F	Effect Found
I. Verbal Items					
Asch (1946)	adjectives	6	written evaluative rating	F	Primacy
Anderson & Barrios (1961)	adjectives	2,6	written -4 to +4 rating	F	Primacy - with 6 items only
Anderson & Hubert (1963)	adjectives	6,8	spoken 1-8 rating	F	Recency[c]
Anderson & Norman (1964)	adjectives	6	written 1-8 rating	F	Primacy
Anderson (1965b)	adjectives	6,9	spoken 1-8 rating	F	Primacy
Anderson (1968a)	adjectives	3,6	slash mark on continuous line	F	Recency
Hendrick & Constantini (1970a)	adjectives	6	spoken 1-8 rating	F	Primacy and Recency[d]
Stewart (1965)	adjectives	4,6,8	spoken 1-8 rating	F	Primacy
Anderson & Norman (1964)	foods	6	written 1-8 rating	F	Primacy
Anderson & Norman (1964)	newspaper headlines	6	written 1-8 rating	F	Some Primacy
Anderson & Norman (1964)	life events during one week	6	written 1-8 rating	F	No Effect
Luchins (1957)	descriptive paragraph	-	miscellaneous questionnaire[b]	F	Primacy
Stewart (1965)	adjectives	4,6,8	spoken 1-8 rating	I	Recency
Anderson (1968a)	adjectives	3,6	slash mark on continuous line	I	Recency
Levin & Schmidt (1969)	adjectives	6	spoken binary response	I	Recency
Rhine (1968)	adjectives	2-5	slash mark on continuous line	I	Recency
Rosenkrantz & Crockett (1965)	adjectives	16	miscellaneous questionnaire[b]	I	Recency
Luchins (1958)	descriptive paragraph	-	miscellaneous questionnaire[b]	I	Recency
Anderson (1959)	arguments in a trial	16	11-category written scale of guilt	I	Recency early, Primacy later
II. Numbers, Weights, Lines					
Anderson (1964)	2-digit numbers	7,8	spoken estimate of cumulative average	I	Recency
Hendrick & Constantini (1970b)	2- and 3-digit numbers	6	spoken estimate of cumulative average	F	Primacy
Weiss & Anderson (1969)	line lengths	6	average length - on metered bar	I	Recency
Weiss & Anderson (1969)	line lengths	6	average length - by method of reproduction	F	Recency
Anderson (1967a)	lifted weights	6	spoken average of heaviness	F	Recency
Anderson & Jacobson (1968)	lifted weights	3	spoken average of heaviness	F	Recency
III. Probabilistic Samples					
Peterson & DuCharme (1967)	rolls of a die	100	P(H/D) estimates on a bar	I	Primacy
Peterson & DuCharme (1967)	samples from a multinomial population	100	P(H/D) estimates on a bar	I	Primacy
Roby (1967)	binary samples from an urn	20	written P(H/D) estimates	I	Primacy
Dale (1968)	multinomial samples of intelligence info.	20	P(H/D) estimates - washers on pegs	I	Primacy
Pitz & Reinhold (1968)	binary samples from an urn	5	spoken binary prediction	I	Recency
Shanteau (1969)	binary samples from an urn	5,9,15	inferences about sample proportions and P(H/D)	I	Recency

[a] I stands for intermittent and includes studies where subjects made responses after each new item of information. F stands for response only after the final item.

[b] This miscellaneous questionnaire required subjects to write a descriptive paragraph, check other adjectives, make predictions about the target stimulus, etc.

[c] Recency was found in one of two experiments where subjects had to recall the adjectives just after making their ratings. The other experiment found no sequential effect.

[d] Recency was obtained when subjects had to pronounce each adjective after it was read. Otherwise, primacy occurred.

LLHH sequence. Recency would produce the opposite effect. Another related design was employed by Anderson (1965b) who interpolated a block of LLL or HHH items into a sequence containing 3 or 6 other items, all with opposite scale values. For example, in the sequences:

1. LLLHHH
2. HLLLHH
3. HHLLLH
4. HHHLLL

primacy would result in increasing judgment as one proceeded from sequence 1 to 4 and the LLL block moved towards the latter part of the sequence.

The results of studies in Category I can be summarized as follows. When only one judgment is made, at the end of the sequence, ten studies reported primacy effects, three observed recency, and two found no effect. One study showing recency, Anderson and Hubert (1963), required subjects to recall the adjectives just after making their rating. When recall was not required, primacy occurred. These results were interpreted as indicating that primacy was caused by decreased attention to the later adjectives. Recall presumably eliminated primacy by forcing attention to the later adjectives. Anderson (1968a) also found recency in a study that departed slightly from the typical design. Instead of using high and low items in the same sequence, Anderson used moderately high (M+) and high (H) or moderately low (M-) and low (L) items, thus reducing the glaring inconsistencies that usually occur when H and L items are included together. The fact that primacy was not obtained here led Anderson to propose that its existence in the other studies was due to subjects discounting the later, inconsistent evidence, much as occurred in the Anderson and Jacobson (1965) study of stimulus inconsistency. Hendrick and Constantini (1970a) examined both the attention decrement and inconsistency explanations of primacy. They varied the degree of perceived

inconsistency among sets of three high- and three low-valued adjectives. They also required half of their subjects to repeat each adjective after it was presented. Variation of inconsistency was found to have no effect. A strong primacy effect was observed except in the situation where subjects had to pronounce the adjectives. There, recency occurred. The authors concluded that these results supported the attention decrement hypothesis and they argued that the recency found in the Anderson (1968a) study occurred not because of the low degree of inconsistency in that study but because Anderson's subjects also pronounced the adjectives.

A study by Anderson (1965b) indicated linear primacy effects. The earlier the information was presented in the sequence, the greater its effect, by a constant amount, in the sequence. Anderson proposed a weighted average model to account for the data, where the weights declined as a linear function of ordinal position in the set.

Although Anderson and Norman (1964) argued that primacy seems unlikely to stem from a change in meaning effect, later studies (Anderson, 1966; Anderson & Lampel, 1965; Wyer & Watson, 1969) have shown that such contextual effects do occur and Chalmers (1969) proposed that change of meaning be incorporated formally into Anderson's weighted average model.

Turning to the studies in Category I, where subjects responded intermittently as stimulus information was acquired, a radically different picture emerges -- recency strongly predominates. It is not clear why making judgments during the sequence should lead to recency whereas making only one judgment at the end of the sequence generally produces primacy. Luchins (1958) presented subjects with two blocks of highly inconsistent information about an individual, in paragraph form. They filled out two detailed questionnaires about the subject of the paragraph, one after each block. Luchins argued that the inconsistencies were accentuated by the first

questionnaire, making it difficult for subjects to assimilate the later information and causing them to respond to the second questionnaire in terms of the second block of information, hence a recency effect. Why this clearly inconsistent information was not discounted, however, as seemed to occur in an earlier study (Luchins, 1957) where only final responses were given, is an unanswered question. Luchins (1958) observed that his subjects did not regard themselves as committed to the opinions they had expressed on the first questionnaire, often giving diametrically opposite answers on the second administration.

Stewart (1965) argued that responding after each new adjective forces subjects to pay equal attention to each one and to weight the new information and the old impression equally. Although equal attention and equal weighting might seem to predict neither primacy nor recency, Anderson (1959) showed that equal weighting of new information and an old impression, in a situation where there is an initial impression prior to seeing the new information, will necessarily produce recency.

The study by Anderson (1959) deserves special mention because it explicitly tested for order effects across a long sequence (16 items) of relatively complex arguments in a trial setting. A recency effect was observed early but decayed and was replaced by a primacy effect later in the sequence. Anderson hypothesized that opinion is made up of two parts, a superficial component that is quite labile and produces recency, and a basal component which forms slowly and is then relatively little influenced by new communications. This resistance to change results in a primacy effect.

Category II; Numbers, weights, and lines. Stimuli used by studies in this category are described by information that is unlikely to produce strong feelings of incongruity in subjects and in this way they differ from

those in Categories I and III. The six studies in which subjects averaged numbers, weights, or lines have typically employed factorial designs to assess the effects of serial position.

With only one exception, the studies in this category found recency effects. Whether or not judgments were made intermittently seemed to make little difference. The exception is a study by Hendrick and Constantini (1970b) that obtained primacy in number averaging when subjects responded only after the final stimulus. In this respect, number averaging is similar to the integration of verbal items (Category I).

Anderson (1967a) noted that contrast effects might be one cause of the pervasive recency phenomenon for lifted weights. For example, if weight L is felt lighter in sequence HHL than in LHH, the data would show a recency effect.

Weiss and Anderson (1969) hypothesized that memory and storage requirements might determine the recency they found in intermittent judgments of average length of lines. To carry this idea further, it may be that subjects do not preserve the individual memories of previous lines, numbers, or weights when responding intermittently. These may tend to lose their identity when integrated into an earlier impression. When the time comes to integrate another item into the impression, subjects give the new item more nearly equal weight rather than weighting it by the reciprocal of n, the number of items in the sequence. In other words, if the subject does not keep in mind the number of previous items, his subjective perception of n may undergo temporal decay. Although Weiss and Anderson conducted one test of this hypothesis without obtaining substantiating results, it would appear to merit further consideration. Weiss and Anderson did find less recency when judgments were made only at the end of a sequence.

Category III; Probabilistic information. Studies investigating sequential use of probabilistic information have generally required subjects to make judgments after each new datum is presented. Five such studies have reported primacy effects and two have obtained recency. The dominance of primacy here contrasts with the recency effects generally found when subjects make intermittent judgments upon receipt of verbal or numerical information. Why? One possible explanation is the fact that the studies by Peterson and DuCharme (1967), Roby (1967), and Dale (1968), 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 strongly inconsistent data sequences.

Summary of primacy-recency studies. It appears that order effects are highly pervasive phenomenon, appearing in studies employing quite diverse stimuli and response modes. Whether recency or primacy occurs seems very much dependent upon the task characteristics. Primacy is usually found when the subject responds only at the end of the sequence and the later information is highly incongruent with the earlier data. When recall or pronunciation of the stimuli is required or when judgments are made during the sequence itself, recency predominates. However, when strong committments have been developed on the basis of early information and the recent information is extremely implausible (Category III) even intermittent responding produces primacy. When the information is homogeneous in nature and not likely to create feelings of incongruency (Category II), recency is observed. Although many hypotheses have been proposed to account for these data, their

causes remain to be precisely determined.

An Inertia Effect in Bayesian Research

The property of inertia was attributed to opinions by Anderson (1959) in the course of discussing the concept of a "basal component" -- that part of an opinion which becomes increasingly resistant to change as information accumulates. More recently Pitz and his associates have conducted a series of studies demonstrating the existence of "inertia" in studies where opinions 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 was labeled an "inertia effect" and was also found by Geller and Pitz (1968).

Geller and Pitz investigated two possible explanations of the inertia effect. The first says that inertia stems from strong committment 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 committment 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 committment 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 9 draws produce 6 red and 3 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 non-predicted disconfirming events and this was taken as support for the second hypothesis.

Further evidence for the committment 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 attributed 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.

The inertia effect can be thought of 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 some of the primacy-recency studies discussed above. It is perhaps relevant that most of the studies in Category I (see Table 3) 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, each of the studies in Category III that obtained primacy (Dale, 1968; Peterson & DuCharme, 1967; Roby, 1967) had subjects make estimates on some mechanical device that preserved the previous response and required it to be physically manipulated when changes were made. All this is obviously "post hoc" analysis but it seems to indicate that future research on primacy and recency should take a close look at the effect of the way in which the previous response in the sequence is made and stored.

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 (see, for example, the chapters by Naylor and Bjorkman in this book). 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 their learning task.

Single cue learning. Research with single cues has focussed upon what Carroll (1963) 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 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 best. Later work by Bjorkman (1965) and Naylor and Clark (1968) centered around the relative ease of learning positive vs. negative linear functions both in deterministic and probabilistic settings. In the latter studies the degree of predictability was manipulated and described in terms of the absolute magnitude of r_e (note that, in single-cue studies, $r_{i,e}$ is equivalent to r_e; similarly, $r_{i,s}$ equals r_s and $b_{i,e}$ and $b_{i,s}$ equal b_e and b_s respectively). The results of these studies indicated that positive relationships between cue and criterion are learned much more readily than negative ones.

Bjorkman (1968) was interested in what he called "correlation learning,"

defined as functional learning where error (r_e < 1.00) was involved. He observed that correlational learning requires a subject to learn both a function and the probability distributions around the regression curve. In one experiment he found that the variance of a subject's responses about their 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, Bjorkman concluded that correlational tasks are learned through a two-stage process involving both functional learning and probability learning with the former being temporally prior to the latter.

Conservatism in single-cue learning. A particularly interesting issue in single-cue learning is concerned with determining whether or not subjects in these studies exhibit conservatism such as is evidenced in Bayesian studies of performance. Several results have been brought to bear upon this matter but they must be viewed cautiously because of the problems of 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 sub-ranges. The assumptions underlying this measure are (a) that the criterion distribution reflects the true probabilities of the various hypothesis states within each sub-range 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 -- particularly in the upper

and lower thirds of the cue distribution and for high values of r_e.

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 Bjorkman (1965), Gray (1968), Gray, Barnes, and Wilkinson (1965), and Schenck and Naylor (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 (in press) 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.

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 has taken a great variety of forms. Most of the studies rely upon the lens model for conceptual and analytical guidance. Several have varied 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, in press; Hammond & Summers, 1965; 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) when relationships are linear and r_e is held constant, subjects do better as cue intercorrelations (redundancy) increase (Naylor & Schenck, 1968); (d) subjects can learn to detect changes in relative cue weights over time although they do so slowly (Peterson, Hammond, & Summers, 1965a); (e) 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 especially difficult (Summers, 1967); (f) 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, 1966); and (g) subjects can learn to use valid cues even when they

are not reliably perceived (Brehmer, in press).

Conservatism has not been an explicit concern in many multiple-cue learning studies. However, one study, by 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.

A number of studies have investigated the effects of different modes of feedback upon correlational learning. Outcome feedback works but is relatively slow. Lens model feedback, indicating how a subject's cue utilization coefficients compare with the ecological validities, is far more effective (Newton, 1965; Todd & Hammond, 1965).

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 subject's 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. Brehmer argued that it is necessary to invent methods to display to the subjects the real sources of their disagreement. Another interesting finding from this area is that persons initially trained to have linear policies are less likely to change than are persons with more complex, nonlinear policies (Brehmer, 1969b; Earle, 1970).

Non-metric stimuli, events and responses. Bjorkman (1967) has applied the lens model to the learning of non-metric stimuli that are predictive of non-metric events. Bjorkman makes the distinction between "event learning" (also known as probability learning) where the subject must learn to predict by means of relative frequencies of different events and "stimulus-event learning" where stimuli function as cues for events. Bjorkman (1969a) studied performance in a 2 x 2 task (two cues and two events) and found a substantial degree of differential maximizing, a strategy whereby the subject always predicts the most likely event, given the particular cue. This strategy can be considered an optimal one in the sense that it maximizes the number of correct predictions. A similarly high level of optimality was observed by Summers (in press) in a 2 x 2 task and by Beach (1964) and Howell and Funaro (1965) in more complex prediction tasks. The latter study used scaled cue values. These results contrast with the phenomenon of probability matching whereby the subject matches his response probabilities to the relative frequencies of the events. Probability matching is commonly observed in simpler event-learning tasks. The optimal strategy of predicting the most probable event every time is a tedious chore in these simple event-learning tasks. In stimulus-event learning, subjects can maximize without being completely repititious and this may account for their increased optimality in these more complex tasks.

One other finding of interest in these kinds of tasks is that subjects' responses become much more consistent as soon as feedback is removed (Azuma & Cronbach, 1966; Bjorkman, 1969b; Bruner, Goodnow, & Austin, 1956). The presence of feedback apparently promotes hypothesis testing wherein subjects attempt to outguess the random sequence of events.

Bayesian Studies of Learning

Bayesian researchers have been notably uninterested in the topic of learning; they usually treat learning as a confounding to be avoided. 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 an article by Peterson (1968). Peterson's subjects responded to more than 8000 four-data sequences, but the study does not mention whether feedback was given (presumably it was not), and all analyses are based on all the data, without any attention paid to changes over time. Peterson, like most other Bayesian researchers, is interested in how subjects behave -- not how they learn.

A few studies do look at learning and 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 the 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.

Goldstein, Emanuel, and Howell (1968) varied diagnosticity, percent of feedback, and specificity of feedback. They found that learning, as evidenced by increased optimality of response over time, occurred only for the high diagnosticity condition, not for the more difficult conditions. The performance with more difficult data started at and stayed at the best level finally obtained in the easy, i.e., highly diagnostic, condition. They also found no differences in optimality between 100% feedback, 67% feedback, 33% feedback, and no feedback. These peculiar results were attributed to the unusual task employed: guessing whether a number drawn from one normal distribution was larger than or smaller than a number drawn (but not exposed) from another normal distribution (this is like asking "if this female is 5' 8", will a randomly selected male be taller or shorter?"). The distribution from which the exposed number was drawn alternated on each trial. A simple but non-optimal strategy of using the same cut-off point to determine one's response, regardless of which population was sampled first, gave excellent results when the means of the two populations were close together (low diagnosticity), but worked badly when the means were farther apart (high diagnosticity). Thus the latter group was forced to learn and adopt a more complicated strategy, using two cut-off points, one for each distribution. As to the ineffectiveness of feedback, the authors note that, since the subjects saw one number from each distribution on every other trial, they apparently learned enough about the situation even when they were not told whether their answer was correct.

Martin and Gettys (1969) gave subjects either nominal feedback (H_1 generated the data) or probabilistic feedback (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 the probability feedback produced better 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. 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 later 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. Staël von Holstein (1969) studied the ability to predict the price of 12 well-known stocks two weeks in the future. His 72 subjects, who included bankers, stock market experts, statisticians, business teachers and business administration students, made probability estimates for each stock across five hypotheses (decrease more than 3%, decrease 1 to 3%, change less than 1%, increase 1 to 3% and increase more than 3%), for 10 consecutive two-week periods. At the end of each period, the subjects received outcome feedback. The task proved to be exceptionally difficult. Only two of the 72 subjects performed as well as a hypothetical, totally ignorant subject who always assigned a probability of .2 to every hypothesis. The subjects would have done better by acknowledging their own inabilities in stock-market prediction, thus giving more diffuse estimates. The subjects did not learn to improve their performance over the ten periods, but they did apparently learn, to some extent, how poor they were: the spread of their probability estimates across the hypotheses

increased over time.

Schum (1966a)showed that subjects can learn and utilize existing conditional non-independence in multinomial data. The subjects were warned which data sources might be non-independent, 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 non-independence 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 non-independence than a model in which independence was falsely assumed.

Two additional learning studies were oriented to the previously-discussed misperception explanation of conservatism. In order to heighten the point that subjects' conservatism resulted from their misunderstanding of the data generator, Peterson, DuCharme, and Edwards (1968) showed that subjects were less conservative after they had seen 100 illustrative samples of data from the binomial data generator. Wheeler and Beach (1968) amplified this finding. They 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. The effects of such training were seen in increased accuracy of subjects' estimated sampling distributions and decreased conservatism.

Descriptive Strategies: What is the Judge Really Doing?

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. In this section we shall examine a few studies that point to the deficiencies of these models with regard to providing insight into cognition. These studies aim to uncover specific strategies or cognitive rules that subjects employ in order to produce the judgments demanded of them.

Forewarning of the descriptive problems of models was provided by Hoffman (1960), whose paper not only provided much of the impetus for correlational research but also presented a cogent discussion of the distinction between simulating behavior and actually capturing the ongoing psychological processes. Nevertheless, with but a few exceptions, the ensuing research has not been oriented towards uncovering strategies. Instead most research has implicitly assumed that the various regression and Bayesian models that summarize the data so well actually describe cognitive processes.

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 -- namely, 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 3/4 of the amount I would expect to win. If the reverse were true, I would ask the experimenter to pay me about . . . 1/2 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 conduct an extensive and still unfinished study to uncover strategies by which subjects average just 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 simple strategies. For example, Azuma and Cronbach (1966) studied the ability of subjects to learn 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 sub-universes and employed different rules within each of these. If correlations are to be used, they argued, they should be calculated separately for each sub-universe of stimuli.

Strategies for Estimating P(H/D)

The modern theory of probability was conceived during the 17th Century when an aristocratic Frenchman, the Chevalier de Méré, realized that reason could be substituted for painful experience in determining one's chances at the gambling tables. Since that time, students of the theory have been continually amazed at its subtlety and the extent to which answers 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 some researchers. Beach (1966) and others attributed conservatism to erroneous subjective probabilities rather than an inadequate 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 even in this). After that, the success he obtains may be purely a matter of coincidence -- a fortuitious 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 us 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 nonoptional 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 sample 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 the many of the effects observed there.

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 data with whatever striking aspect of the situation the experimenter has provided. This strategy was observed by Dale (1968) in a pseudo-military task involving four hypotheses. The values of $P(D_j/H_i)$ were displayed as histograms. As the subjects received the ten data reports, they often physically arranged the data reports to form a histogram which they then compared with the conditional probability display. 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 notes 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.

The binomial task provides the subject with the least amount of explicit information against which the subject can compare the sample. In such tasks, several independent studies have shown that the subjects make their responses match the sample. For example, Beach, Wise, and Barclay (1970), using a binomial task with a simultaneous sample of n items, found a remarkably close relationship between the sample proportion and the posterior probability estimates. Several of their subjects remarked that sample proportions are very compelling because they are available (and somehow relevant) numbers in a very difficult and foreign task. Studies by Kriz (1967) and Shanteau (1969) have reported similar use of sample proportions as the basis for P(H/D) estimates. This simplifying strategy does not take into account the likelihood of the data, as specified by the population proportions. Subjects thus would not change their responses across tasks that vary 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, the strategy of using the sample proportion to estimate 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 others. Examples of improper and 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, appeal to

authority, and appeal 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.

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

Probabilistic Information Processing Systems

A great deal of Bayesian research has centered about some new ideas for putting probability assessments to use in diagnostic systems. Edwards (1962) introduced the notion of a probabilistic information processing (PIP) system because of his concern about 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?). His proposal was simple. 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, 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 thus added considerable impetus to the development of

PIP systems.

Edwards and Phillips (1964) promoted the PIP system as a promising alternative to traditional command and control systems. They hypothesized that PIP 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. Edwards and Phillips observed, for example, that calculation is inadequate to assess the probability that Russia would have launched 25 reconnaissance satellites in the last three days if she planned a missle attack on the United States. Only human judgment can evaluate this type of information; PIP systems obtain and use such judgments systematically.

Given the basic idea of a PIP, much experimental research was needed before it could be implemented effectively. 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 and carefully done studies was by Edwards, Phillips, Hays, and Goodman (1968). They constructed an artificial future world (complete with "history" up to 1975) and wrote 18 scenarios, each with 60 data items. The subjects related the 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. All subjects then responded to the 18 scenarios. 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 60 data in each scenario. In PIP, this aggregation was done by machine to compute final odds.

No optimal model can be devised for this simulation. The "true" hypothesis for any scenario was not known. Results showed, however, that the PIP group arrived at larger final odds than other groups. When PIP 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 solved ingeniously 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 editor) or ending a word (like looked). 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 bigrans, 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 out of three 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 third 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). Unlike the PIP simulations of Edwards and Kaplan and Newman, the Ohio State research employed a frequentistic environment where the experimenters specified a P(D/H) matrix that governed the sampling from a limited set of data to form a number of scenarios. 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 six 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 producing posterior probabilities. Recent endeavors have attempted to step up the complexity of the simulations in an attempt to narrow the gap between them and real-world diagnostic systems. Schum (1969) discusses some of the problems that must be solved as more realistic complexity is introduced into the system. For example, in systems that periodically experience high rates of data accumulation, experts who assess P(D/H) may have to aggregate their judgments over a series of data (i.e., judge $P(D_1D_2D_3 \ldots D_n/H)$. When data items are nonindependent, these conditional probabilities can become quite complex. Three experiments reported by Schum, Southard, and Wombolt (1969) found that men could adequately aggregate diagnostic import across small samples of such nonindependent data. In addition, PIP was increasingly superior to POP when scenarios were either large or highly diagnostic or both. There is no longer any doubt that PIP is a viable concept for the design of decision systems. Future work will most likely see the extension of PIP to non-military settings along with greater attention to the practical details of implementing such systems in the real world. 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.

Bootstrapping

Can a system be designed to aid the decision maker that takes as input 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) puts it,

> "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 rather nonlinear criterion. After 12 days of practice, they were given a set of test trials and 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 non-linearities 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. More recently, Dudycha and Naylor (1966) 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 such strategies.

Bowman (1963) outlined a bootstrapping approach within the context of managerial decision making that has stimulated considerable empirical research (see Gordon, 1966; Hurst and 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 out of the possible set of 250,000. 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 recent demonstration of bootstrapping comes from a study of a graduate-student admissions committee by Dawes (1970). 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 each of a new 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 cut 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 eliminated 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 could result in an annual savings of about 18 million dollars worth of professional time across the nation's graduate schools.

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 preceeding pages? Several generalizations seem appropriate. First, it is evident that the judge responds in a highly predictable manner to the information available to him. Furthermore, much of what we used to call intuition can be explicated in a precise and quantitative manner. With regard to this point, it appears that one's self insight into his own cognitive processes is deficient and there is much to be gained by appropriate feedback of the quantitative aspects of one's judgment behavior.

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. The specific form of sequential effects that occur is very much 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 interfacing this area with the mainstream of cognitive psychology -- work on concept formation,

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, 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 available their own optimal model in the multiple regression equation but have shown little interest in comparing subjects with it (except for a substantial number of learning studies). Instead, they have spent a great deal of effort using correlational methods to describe a judge's idiosyncratic weighting process -- an enterprise in which Bayesians and functional measurement researchers have been uninterested. Researchers using functional measurement 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 adding 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, we can see that this is true. A correlationalist would find it difficult to use, as his cues, intelligence reports: "General Tsing was seen last Monday lunching with Ambassador Ptui." Instead, he will feel more comfortable with conceptually continuous cues such as MMPI scores, or Grade Point Averages. Similarly, at the level of the criterion, 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 one hand, we see neglected problems for which a model is perfectly well suited. Why have the Bayesians neglected learning? 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 persistant, even stubborn, pursuit of topics for which the model is awkward. Correlationalists have been devoted 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.

Under the intellectual leadership of researchers such as Brunswik, Edwards, Anderson, Hammond, and Hoffman, several excellent 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 -- and it may be that they are missing some important research opportunities by limiting their approach to a single paradigm.

Towards an Integration of Research Efforts.

We suggest that researchers should employ a multiparadigm approach, searching for the most appropriate tasks, models, and analysis techniques 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 dangers of staying within a single model, and the potential value of diversity, are illustrated by research on primacy and recency effects. Hendricks and Constantini (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. However, they were apparently unaware of a number of Bayesian studies that did obtain primacy when early and late data were highly inconsistent (Dale, 1968; Peterson & DuCharme, 1967; Roby, 1967) and recency when later data were not so inconsistent (Pitz & Reinhold, 1968; Shanteau, 1969). The discrepancy between the Hendricks and Constantini data and the Bayesian results needs to be explained.

The study by Shanteau (1969) provides a nice example of the utility of

applying methods and tasks from different paradigms. Shanteau used a functional measurement technique to study sequential effects in 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.

Novelty. How do subjects handle data that is rare or novel? Wyer (1970) examined the effects of novelty, defined in terms of the unconditional 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 non-specific 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. 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 (non-learning) studies where diagnosticity was varied. Bayesian research on this topic has been extensive and has pointed up the difficulties subjects have in integrating probabilistic information. It is important to investigate the generality of these difficulties; the different data and response formats possible within the correlational paradigm provide an excellent opportunity to do this.

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 are Bayesian in spirit, inasmuch as they view human judgments as essential and attempt to blend them optimally (see Pankoff & Roberts, 1968, for an elaboration of this point). However, PIP assumes that the aggregation process is faulty and attempts to circumvent this by having subjects estimate P(D/H) values and letting a machine combine them. Bootstrapping assumes that subjects can aggregate information appropriately except for unreliability that must be filtered out. Actually, one could incorporate the bootstrapping notion into a PIP system by having subjects make a series of posterior probability judgments from which their implicit P(D/H) opinions could be inferred. These inferred values could then be processed by Bayes' theorem. Alternatively, one could apply the PIP assumption to bootstrapping in a correlational framework by asking subjects to estimate the regression coefficients directly. The success of bootstrapping and PIP systems suggests that the assumptions of both are probably correct -- 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 excell bootstrapping or vice-versa.

Conclusions

It is obvious that large gaps exist in our understanding of information processing in human judgment -- despite several energetic research programs over the last decade. We hope that, in the future, researchers will not be bound unnecessarily by the constraints of one particular experimental paradigm, and will, instead, approach substantive problems with an awareness of the diverse array of models, methods, and tasks that are available.

Footnote

1. 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.

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., & 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.

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. Cue probabilism and inference behavior. Psychological Monographs: General and Applied, 1964, 78.

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 estimates and confidnece ratings. Journal of Experimental Psychology, 1969, 79, 438-444. (b)

Beach, L. R., & Wise, J. A. Subjective probability revision and subsequent decisions. Journal of Experimental Psychology, 1969, 81, 561-565. (c)

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.

Bjorkman, 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.

Bjorkman, M. Stimulus-event learning and event learning as concurrent processes. Organizational Behavior and Human Performance, 1967, 2, 219-236.

Bjorkman, 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.

Bjorkman, M. Individual performances in a single-cue probability learning task. Scandinavian Journal of Psychology, 1969, 10, 113-123. (a)

Bjorkman, M. Policy formation in a non-metric task when training is followed by non-feedback trials. Umeå Psychological Reports, No. 6, Department of Psychology, University of Umeå, 1969. (b)

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, in press.

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, in press.

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

Chalmers, D. K. Meanings, impressions, and attitudes: A model of the evaluation process. Psychological Review, 1969, 76, 450-460.

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.

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. Toward a general framework for evaluation. Report No. MMPP-64-7, Michigan Mathematical Psychology Program, University of Michigan, 1964.

Dawes, R. M. Graduate admissions: A case study. Unpublished paper, Oregon Research Institute, Eugene, Oregon, (1970).

DuCharme, W. M. A response bias explanation of conservative human inference. Journal of Experimental Psychology, in press.

DuCharme, W. M., & Peterson, C. R. Intuitive inference about normally distributed populations. Journal of Experimental Psychology, 1968, 78, 269-275.

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.

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, No. 10.

Edwards, W. Probability-preferences in gambling. American Journal of Psychology, 1953, 66, 349-364.

Edwards, W. Probability preferences among bets with differing expected values. American Journal of Psychology, 1954, 67, 56-67. (a)

Edwards, W. The reliability of probability preferences. American Journal of Psychology, 1954, 67, 68-95. (b)

Edwards, W. Variance preferences in gambling. American Journal of Psychology, 1954, 67, 441-452. (c)

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, Pp. 17-52.

Edwards, W., Lindman, H., & Phillips, L. D. Emerging technologies for making decisions. New directions in psychology II. New York: Holt, Rinehart, & Winston, 1965, Pp. 261-325.

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 & M. W. Shelley (Eds.), Human judgments and optimality. New York: Wiley, 1964, Pp. 360-401.

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, in press.

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, Pp. 383-491.

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., & Nicol, 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.

Goldstein, I. L., Emanuel, J. T., & Howell, W. C. Effect of percentage and specificity of feedback on choice behavior in a probabilistic information-processing task. Journal of Applied Psychology, 1968, 52, 163-168.

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 an 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.

Grebstein, L. C. Relative accuracy of actuarial prediction, experienced clinicians, and graduate students in a clinical judgment task. Journal of Consulting Psychology, 1963, 27, 127-132.

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, Pp. 91-96.

Green, D. M., & Swets, J. A. Signal detection theory and psychophysics. New York: Wiley, 1966.

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, Pp. 15-80.

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., & Summers, D. A. Cognitive dependence on linear and nonlinar 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 system 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, Pp. 251-267.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart & Winston, 1963.

Hendrick, C., & Constantini, 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., & Constantini, A. F. Number averaging behavior: primacy effect. Psychonomic Science, 1970, 19, 121-122. (b)

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.

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, Pp. 53-90.

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.

Hovland, C. I., (Ed.), The order of presentation in persuasion. New Haven: Yale University Press, 1957.

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.

Howell, W. C., & Funaro, J. F. Prediction on the basis of conditional probabilities. Journal of Experimental Psychology, 1965, 69, 92-99.

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.

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.

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.

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.

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.

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.

Peterson, C. R., Ulehla, Z. J., Miller, A. J., Bourne, L. E., & Stilson, D. W. Internal consistency of subjective probabilities. Journal of Experimental Psychology, 1965, 70, 526-533.

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.

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 system 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. Effects of observation cost on sequential search behavior. Perception & Psychophysics, 1969, 6, 234-240.

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.

Rorer, 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, & 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 & Winston, 1966, Pp. 159-203.

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

Schenck, A., & Naylor, J. C. The effect of cue intercorrelation on performance in a multiple-cue choice situation. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May, 1965.

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 eivdence 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 decision-making model for sequential estimation and inference judgments. Unpublished paper, Center for Human Information Processing, 1969.

Shepard, R. N. On subjectively optimum selection among multiattribute alternatives. IN M. W. Shelly, II, & G. L. Bryan (Eds.), Human judgments and optimality. New York: Wiley, 1964, Pp. 257-281.

Sidowski, J. B., & Anderson, N. H. Judgments of city-occupation combinations. Psychonomic Science, 1967, 7, 279-280.

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, in press.

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

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

Stael von Holstein, C.-A. S. The assessment of discrete probability distributions -- An experimental study. Institute of Mathematical Statistics Research Report No. 41, University of Stockholm, 1969.

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.

Strub, M. H. Experience and prior probability in a complex decision task. Journal of Applied Psychology, 1969, 53, 112-117.

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., & 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. Alternative bases for choice in probabilistic discrimination. Journal of Experimental Psychology, in press.

Summers, S. A., Summers, R. C., & Karkau, V. T. Judgments based on different functional relationships between interacting cues and a criterion. The 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 by Hursch, Hammond, & Hursch, and by Hammond, Hursch & Todd. Psychological Review, 1964, 71, 528-530.

Tversky, A. A general theory of polynomial conjoint measurement. Journal of Mathematical Psychology, 1967, 4, 1-20.

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.

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 & J. Tanenhaus (Eds.), Frontiers of judicial research. New York: Wiley, 1969, Pp. 335-369.

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

Vlek, C. A. J., & Bientema, 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.

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, PRL-TDR-63-16, June, 1963.

Weiss, W. Scale judgments of triplets of opinion statements. Journal of Abnormal and Social Psychology, 1963, 66, 471-479.

Weiss, D. J., & Anderson, N. H. Subjective averaging of length with serial presentation. Journal of Experimental Psychology, 1969, 82, 52-63.

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

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., & 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).