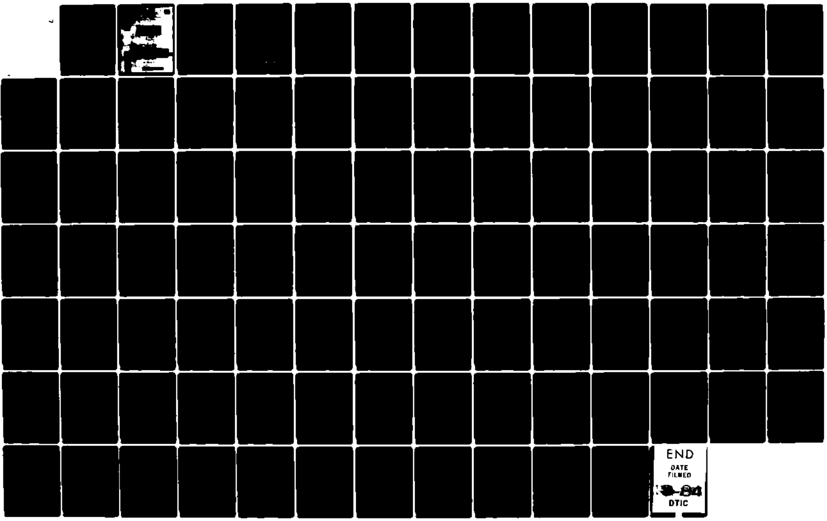
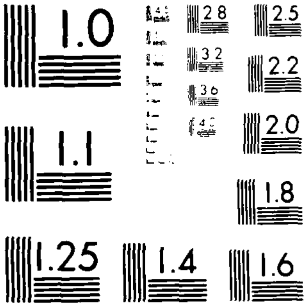
AD-A142 380 PROPHYLAXIS AND TREATMENT OF CYANIDE INTOXICATION 1/1
CYANIDE - MECHANISM OF PROPHYLAXIS(U) WASHINGTON STATE
UNIV PULLMAN J L WAY 15 APR 83 DAMD17-80-C-0134

UNCLASSIFIED F/G 6/20 NL



END
DATE
FILMED
8-84
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

THE EFFECT OF TYPE OF SELF-GENERATED EVIDENCE AND TYPE OF FEEDBACK ON OVER/UNDER CONFIDENCE

1982

CPT JOHN R. TIFFANY

1

THE EFFECT OF TYPE OF SELF-GENERATED EVIDENCE AND TYPE OF FEEDBACK ON OVER/UNDER CONFIDENCE

A Thesis

Presented to

the Faculty of the Graduate School

Indiana University of Pennsylvania

In Partial Fulfillment

of the Requirement for the Degree

Master of Arts

by

CPT John R. Tiffany

April 1982

Indiana University of Pennsylvania

The Graduate School

Psychology Department

We hereby approve the thesis of

CPT John R. Tiffany

Candidate for the Master of Arts

April 29, 1982 David E. Groner
Professor of Psychology, Chairperson
Psychology Department Chairperson

April 29, 1982
Dean of the School of Natural Sciences
and Mathematics

[illegible]
Professor of Psychology

April 29, 1982 Donald U. Robertson
Professor of Psychology

For my family

Acknowledgements

There are several persons to whom I am indebted for their assistance in completing this thesis. I thank Dr. David Grover, committee chairman, for his adivice and encouragement from topic selection through defense; Dr. Donald Robertson and Dr. Rio Sussmann, committee members, for their insight in interpreting data and for help in statistical analyses. Each with his own particular skills has made invaluable contributions to completion of this work. My thanks to fellow students Linda Baker, for helping write a computer program which provided feedback to subjects; to Maria Russo for helping conduct portions of the experiment; and to Colene Byrne for helping move test equipment. I thank Kimberlee Tiffany for reviewing and correcting errors in several drafts of this thesis.

I acknowledge the support of Perceptronics, Inc., for providing the general knowledge test questions used in Phase I of this experiment, and of Dr. Stan Halpin at Army Research Institute for providing reference materials and advice.

Finally, I thank Colonel Robert G. Krause and Captain Bobby Jones for writing recommendations that contributed to my selection for this educational program.

The funding for pursuit and completion of this phase of my education was provided under the Army Top 5% Fellowship Program.

Title: The Effect of Type of Self-Generated Evidence and Type of Feedback on Over/Under Confidence

Author: CPT John R. Tiffany

Thesis Chairperson: Dr. David E. Grover

Thesis Committee Members: Dr. Mario Sussmann,

Dr. Donald U. Robertson

A common error of decision makers is the failure to seek disconfirming evidence for hypotheses. Seeking only confirming evidence often leads to acceptance of incorrect hypotheses, and causes the decision maker to be overconfident in estimating correctness of the decisions. The primary objective of the present reserach was to determine the effects of types of self-generated evidence and presence or absence of feedback on reducing overconfidence. The secondary aim was to determine effects of feedback on generalization of decision making strategies. The following hypotheses were tested: (a) If only confirming evidence of a decision is generated by a subject, then the subject will be overconfident in estimating validity of the decision. (b) If only disconfirming evidence of a decision is generated by a subject, then subject overconfidence will be reduced. (c) If a subject is given feedback on accuracy of decisions, then overconfidence is reduced and generalization of decision making strategies across tasks will be enhanced.

A 5 x 2 x 3 factorial design utilizing 72 male and female undergraduate students was used to test the hypotheses.

Results were inconclusive in confirming or disconfirming the hypothesis that generating only confirming evidence leads to overconfidence. The hypothesis that disconfirming evidence would reduce overconfidence was not supported. The hypothesis that feedback would reduce overconfidence was disconfirmed. Unlike a previous study which used three blocks of questions in each treatment condition and found significant reduction in overconfidence scores (Koriat, Lichtenstein, & Fischhoff, 1980), this experiment used five blocks of questions. Consistent with the previous research, all treatment groups showed high levels of overconfidence in the first treatment block, and overconfidence generally declined through the third treatment block. The effect was only transitory. Overconfidence scores returned to previous high levels in all groups by the fourth or fifth treatment block. No generalization of the decision making strategies was demonstrated.

Table of Contents

List of Tables and Figures

Table Page

Figures Page

Chapter I

Introduction

Prior to 1970, decision theory was dominated by two types of behavioral models, the prescriptive models and the descriptive models. The prescriptive or normative models were designed to serve as instructions for decision makers, to set down rules which would lead to the making of ideal, rational decisions. The descriptive models were attempts at accurately representing how human beings really behave when making a decision. Both prescriptive and descriptive models were primarily composed of utility components and probability components. The utility component is a measure of whatever the decision maker (DM) attempts to maximize. The probability component is a measure of the decision maker's expectation that an event will occur.

Prescriptive Utility Models

The parameters of prescriptive utility models are discussed at length by Ellsburg (1961). According to Ellsburg, prescriptive models do not reflect actual behavior, but serve to help the decision maker behave the way he would like to behave in order to maximize gain or utility. Ellsburg suggests that in choosing a personal decision making model, the decision maker should make a decision according to a model and then decide whether or not the resulting decision is the best possible. If so, the model is suitable. If there is ever disagreement, the model should be discarded.

Early prescriptive utility models were designed in compliance with five axioms. The axioms are transitivity, comparability, dominance, irrelevance, and independence. A description of the axioms was provided by Allais (1953). Utility models followed or at least attempted to adhere to the axioms for almost twenty years. The article by Allais served as the definitive reference on the axioms.

Should the five axioms be accepted and applied in a prescriptive decision making mode, then for any decision, an outcome can be measured by its utility and each situation can be assigned a probabiltiy. A rational decision can then be made. When the utility and the probability of two outcomes are equal, either will be *chosen randomly.*

Prescriptive utility models developed prior to 1965 encouraged the DM to make a judgment of the possible gain or loss, compare the probability associated with each choice, and make a logical decision. The majority of these prescriptive models were models of Subjectively Expected Utility (SEU). The prescriptive SEU models can be applied to static situations, where only one decision is made or to dynamic situations, where a series of interrelated decisions are made. The most common method of testing SEU was developed by Allais (1953). Subjects were given a series of choices with amounts of money and odds for each possible gain. An example of the format developed by Allais is demonstrated in these two sample questions used by Kahneman and Tversky (1979, p. 265). The percentages who chose each option are shown in parentheses:

Choose between:

A. 2500 with probability .33	or	B. 2400 with certainty
2400 with probability .66		
0 with probability .01		
(18)		(82)

Choose between:

C. 2500 with probability .33	or	D. 2400 with probability .34
0 with probability .67		0 with probability .66
(83)		(17)

Descriptive Utility Models

Where prescriptive or normative models tried to advise a DM on how to behave, the descriptive models tried to reflect actual behavior. Many descriptive models were actually derived from prescriptive models; other descriptive models closely resembled prescriptive models but were developed by fitting mathematical explanations to observations of behavior. The descriptive utility models fall into three major categories--Algebraic Utility models, Constant Utility models, and Random Utility models.

The Algebraic decision models were derived from prescriptive SEU models. Like their normative counterparts, the algebraic models allow only that the DM will choose the alternative with the highest Subjectively Expected Utility. More than one choice is acceptable only if two or more alternatives have exactly the same expected utility. Algebraic models in such a restricted format were not extremely popular. According to Becker and McClintock (1967), most of the experimenters who utilized algebraic utility models "... introduced probabilistic choice modifications by employing statistical procedures to estimate the parameters of the model" (p. 260). The statistical modifications were not justified under a strict algebraic model, and led to the development of different kinds of models.

Modifications were made to algebraic models so that the models would be applicable to more situations. In doing so, the modified descriptive SEU models became extremely broad and practically useless as they retained little or no predictive power. An attempt at correcting the deficiencies of algebraic models were the Constant Utility models.

Three types of Constant Utility models were developed--the Weak Constant Utility model, the Strong Constant Utility model, and the Strict Constant Utility model. According to the Weak Constant Utility model (WCU), the probability that alternative 'a' will be chosen over 'b' *from pair 'm'* is (Becker & McClintock, 1967, p. 262):

$Pm^*(a)$ >or= 1/2 if and only if

$W(\Sigma Pa_x W_x)$ >or= $W(\Sigma Pb_x W_x)$

where '$W(\Sigma Pa_x W_x)$' and '$W(\Sigma Pa_x W_x)$' are strictly increasing functions of the terms in parentheses, 'Pa_x' is the WCU probability of receiving outcome 'x' when response 'a' is chosen; and 'W_x' is the WCU associated with outcome 'x'. '$Pm^*(a)$' is the WCU probability that 'a' will be chosen from pair '*m*'. The other element of pair '*m*' is '*b*'. *The* WCU probability that 'b' would be chosen is expressed as '$Pm^*(b)$'.

The equations for the Strong Constant Utility (SCU) model or Fechner model are not so simple. The Fechner model gave an exact value of the probability that 'a' would be chosen over 'b'.

Finally, the Strict Constant Utility (STCU) model or Luce model gives an exact value for predicting which alternative among several will be chosen.

The third category of descriptive models is the random utility collection. Random utility models differ from other descriptive models in that they assume that utility itself is a fluctuating variable. Random utility models assume that every decision is complex and that every decision involves a large set of considerations. Decision makers are capable of considering only a subset of the considerations at any one time. The subset considered at any moment can be determined by the DM's mood or other factors. The decision made depends on the subset under consideration when the DM is made to decide. Although some tests of the model have yielded positive results, no definitive test of the model has been designed.

Testing the Axiomatic Models

The models of decision making developed through 1965 concentrated on mathematically representing ideals of rational behavior. Equations even for single-stage decisions were sometimes excessively complex, and led to representations of human behavior that were inaccurate due to extremely precise expectations of human behavior and failure to account for the normally large variance within and between groups. Research from 1965-1970 pointed out deficiencies in earlier models, found exceptions to the prescriptive axioms, and continued the trend in descriptive utility model development of processing volumes of data and fitting equations to the data. The period was not marked by development of original models, but by modification of earlier models. Models which had previously been applied to both static and multiple stage decisions were modified to apply specifically to one or the other. The bulk of research was done on single-stage tasks. Equations were somewhat simplified and notation was evidently standardized by consensus.

A refined Subjectively Expected Utility model dominated decision theory, the DM was assumed to maximize SEU, that is, to make decisions that would give him the largest possible gain. The DM would not consciously calculate SEU, but the fact that he attempted to maximize SEU was evidenced by his behavior. From the behavior, mathematical representations were constructed and functions defined to account for decision making. The mathematical representations could not only account for past behavior, but often had high predictive value.

Experiments designed to test SEU were concerned with testing one or more of the five normative postulates. According to Rapoport and Wallsten (1972), MacCrimmon (1968) tested all five postulates one at a time and found them generally valid. In addition, MacCrimmon did follow-up interviews with his subjects to find out why they made occasional decisions that did not conform to SEU axioms, and to give the DMs a chance to reconsider. MacCrimmon found that the DMs would indeed reconsider and decide in conformity with SEU, in most cases. The majority of other researchers did not find experimental support for all the axioms.

A proposed alternative to SEU was the Additive Difference Model (ADM) of Rapoport and Wallsten (1972). The ADM was designed to be used to account for intransitivity. This multidimensional model weighted alternatives, utility, and an undefined difference function. The formula for the ADM was somewhat complicated, which may the the reason ADM was never tested.

A more viable alternative to ADM was risk theory (Rapoport and Wallsten, 1972). Instead of concentrating on maximizing gain, risk theory concentrated on minimization of risk. According to Rapoport and Wallsten (1972, p. 143), there are three assumptions in risk theory:

a. risk is a property of options which affect choices among them,

b. options can be ordered with respect to their riskiness, and

c. the risk of an option is related to the variance of its outcomes.

Several experimenters have attempted to isolate the variables that affect riskiness. The major problem is that any number of DMs appear to preceive riskiness in any number of ways. A sample question used by Kahneman and Tversky (1979, p. 273) to test perception of riskiness is below:

Problem 12: In addition to whatever you own, you have been given 2000. You are now asked to choose between:

C. (-1000, .50) or D. (-500, 1.00)

(69) (31)

This particular example shows that most decision makers will risk twice the amount rather than lose an amount with certainty.

Non-Axiomatic Models

Two non-axiomatic approaches termed Functional Measurement and the Linear Model have been proposed (Rapoport and Wallsten, 1972). Functional measurement is a logical, statistical approach to decision making theory. Researchers using this method scale stimuli, measure large numbers of responses, and attempt to determine the rule relating stimuli to responses. Additive models provide the simplest use of functional measurement. One way to understand the additive model is to perceive it as a matrix (Anderson, 1970). The rows and columns of the matrix correspond to stimuli. The row stimuli are designated S1 through Si. The column stimuli are designated T1 through Tj. Each cell of the matrix corresponds to a pair of stimuli. Each of the stimuli Si and Tj have corresponding subjective values s_i and t_j. An equation for the additive model is (Anderson, 1970):

$$Rij = w_1 s_i + w_2 t_j$$

"... Rij is the theoretical response to the stimulus pair (Si, Tj), and w_1 and w_2 are the weight of the row and column dimensions, respectively" (p. 155).

The second non-axiomatic approach to decision theory is the linear model. The model gives a numerical value to the attractiveness of a goal based upon totaled regression weights of each dimension involved in the decision. The linear model has obvious advantages, including the allowance for determining the effects of any particular stimulus dimension. The difficulty is in scaling response data numerically.

Perhaps the simplest linear model, and one of the older descriptions in print was penned by Benjamin Franklin in 1772 (cited in Dawes and Corrigan, 1974, p.95):

> My way is to divide half a sheet of paper by a line into two columns; writing over the one Pro, and over the other Con. Then, doing three or four days consideration, I put down under the different heads short hints of the different motives, that at different times occur to me for or against the measure. When I have thus got them all together in one view, I endeavor to estimate the respective weights....

This is a popular normative decision making model that practically any DM can use. It holds appeal by its simplicity, not requiring any difficult mathematical calculations or the remembering of complex formulas.

<u>Cognitive Approaches</u>

Behavioral theories of decision making, whether normative or descriptive, tended to leave psychological limitations of the decision maker out of the decision process. The decision maker was typically characterized as a "black box" which was exposed to stimuli and emitted decisions. In the 1970s, psychologists and other decision theorists began to consider heuristics, cognitive limitations, the effects of context, and affective states. For the first time, decision theory began to attempt to take "subjective" psychological factors into account. Complicated mathematical models lost popularity; axiomatic approaches to model formulation were abandoned. A summarization of the changes which took place during that time was done by Tversky and Kahneman (1974). The question asked was not "How well do you perform?" but "How do you perform?" (Einhorn, 1980, p. 1).

To answer this question, many psychologists began to study not simply outcomes of decision making tasks, but processes which were used to make decisions. The approach utilized was to study the cognitive limitations imposed by memory and other information processing systems.

When determining probabilities that human decision makers would exhibit certain behaviors, decision theorists of the prior decades would typically measure the stimuli present prior to a response (decision), compute statistical relationships between the stimuli and responses, and then attempt to use the results in predicting decisions when given the antecedent stimuli. Human subjects in decision making experiments, however, showed large variation between subjects, and each decision maker displayed variation when repeatedly solving the same or similar problems. Under the behaviorist paradigm and prior to the development of cognitive psychology, most decision making theorists would have accounted for such variance by suggesting that the DM had not mastered adequately a normative model or that the variance allowed in a descriptive model was in need of a simple adjustment. Currently, under the cognitive paradigm, the failure of a DM to respond with consistency is, according to Pitz (1980), attributable to two sources. DM's possess an information processing system with a limited memory and a perceptual sensitivity that precludes certain strategies which may or may not be appropriate and may or may not change.

Consider a man who is Christmas shopping, looking at electronic toys to buy for his children. He picks up a game, reads the price, and recalls that he has seen another one in another store. He cannot remember if the price was lower or exactly where he saw the other game. He puts down the game and picks up another. He reads the price. He has never seen this electronic game before, but he decides that it is somewhat overpriced. He does not know how he reached that decision, only that the price is "too high". The man puts down the game and walks over to a desk top computer that challenges him to guess the rule it is using in forming a string of numbers. The computer displays the sequence 2-4-6, and asks the man to enter three strings of numbers. For each string the man enters, the computer will report whether or not the sequence fits the computer's rule. The man enters 4-6-8; the computer answers "CORRECT". The man enters 6-8-10, then 8-10-12. Each time, the computer answers "CORRECT". The computer then gives the man a display of four rules that might have been used:

A. Three ascending numbers.

B. Three even numbers.

C. Three prime numbers.

D. Three odd numbers.

The man chooses "B". The computer tells him that he is wrong. The man tries the problem again, repeating his earlier responses and choice. The computer again tells him that he is wrong. The man leaves the store, convinced that something is wrong with the computer.

The man in the toy store has demonstrated some of his cognitive limitations. He first displayed his memory limitations; then utilized a price judgment strategy that may or may not have been valid. Finally, he, like manydecision theorists prior to 1970, tested his hypothesis by looking only for supporting evidence. The computer was using rule A.

Decision makers demonstrate a variety of other shortcomings, most of which are resistant to thorough investigation either in or out of the laboratory. Ebbeson and Konecni (1980) demonstrated how laboratory simulations may provide data not applicable to real world tasks. In studying how judges determine amounts to be set as bail, the experimenters determined that judges were provided little more than a four part information brief on each person accused which included (1) prior record; (2) the extent to which the accused was tied to the local area; (3) a dollar amount for bail recommended by the district attorney; and (4) charges against the accused. Judges who had experience setting bail were given simulated cases with the four items of information and were asked to set bail exactly as they would in real cases. An analysis of the simulation data showed that all factors except the recommendation of the district attorney had significant effects on the judge's decision.

The experimenters then had trained observers unobtrusively observe the same judges, given the same information, in actual bail hearings. In the real world, the recommendation of the district attorney proved to be the most important factor in each judge's decision. Inconsistency has also been shown between laboratory and real world decisions involving sentencing of adult felons, deciding whether or not to turn an automobile in front of an oncoming car (Ebbeson & Konecni, 1980), and in judging swine (Phelps & Shanteau, 1978). Laboratory simulations, even though they may appear to contain all determining elements of their counterparts out of the laboratory, may simply be inaccurate.

Ebbesen and Konecni (1980) attempted to analyze the factors which are theoretically basic in decision making. In the case of bail setting, supportive evidence was found in the laboratory and then real world information was gathered as an afterthought. Doing further experimentation, the experimenters concluded that identical strategies are not employed by decision makers in and out of the laboratory, nor by different DMs in the same situation; nor by the same DM in similar situations. Any change in the environment can cause a different decision to be made. Such situation dependency outside the laboratory is probably attributable to numerous cues that are difficult to account for in laboratory experiments and are not naturally redundant. Any single cue or combination of cues can cause an unexpected decision.

Perhaps the most effective of these cues in determining often illogical decisions are those that cause affective reactions in the decision maker. This is one of the more controversial approaches used to account for unexpected DM behavior. Extensive research by Zajonc (1980) has shown that affective states play a powerful role in decision making. Zajonc argues that affect precedes cognition, that there is no evidence to indicate that cognitive processes occur first, or can occur without some affective component. Decision makers may prefer to believe their decisions rational, but making a choice between two alternatives is probably due to liking one more than the other, and " ... information collected about alternatives serves us less for making a decision than for justifying it afterward."(Zajonc, 1980, p. 151).

Since decision makers show such great variance, since laboratory derived rules are often not generalizable to tasks in the real world due to situation dependency, and since affective reactions may cause a DM to do something unpredictable anyway, it would appear that the problems in studying decision making are insurmountable. The problems of studying decision making are difficult to overcome, but knowledge of how decisions are made is possible. It may be true that people have as many unique methods as there are problems to be solved, and that a new decision rule called a heuristic is generated by the DM for each situation. If that is so, then one method of studying decisions, which would overcome the laboratory - real world inconsistencies due to situation dependency would be for decision theorists not to study the context dependent rules, but the rules that govern the generating of new rules (Einhorn, 1980). According to Einhorn, in addition to the context dependent heuristics there are generalizable heuristics, the metaheuristics, applicable to decisions of similar content or structure, and alterable only by mass accumulation of disconfirming evidence. A method suggested by Einhorn to determine what metaheuristic is being used, is to give negative feedback for problems of a specific strategy type and then determine which other specific decision strategies have changed.

Einhorn (1980) approaches decision theory by studying how decision makers utilize outcome feedback to modify decision strategies. A different method is suggested by Corbin (1980). She proposes studying decisions by examining prechoice behavior and the decisions that are never made. Corbin asserts that decision makers pass through several stages during which they theorize and reduce ambiguity before making a decision or deciding not to decide. Still another unique approach is offered by Fischhoff, Slovic, and Lichtenstein (1980). According to these experimenters, decisions made by an individual cannot be understood except by going beyond heuristics and discovering the values of the decision maker. Values are defined as "evaluative judgments regarding the relative or absolute worth or desirability of possible events." (Fischhoff et. al., 1980, p. 117).

These relatively unique approaches cited may indicate that there are potentially as many approaches to the study of decision making as there are theorists. The common ground now is that the DM is more than a "black box", that decisions are heavily context dependent, that heuristics and perhaps metaheuristics are employed in some way, and that humans perform in accordance with their limited memory and information processing capabilities.

The majority of current investigations in decision making theory conform to the research strategy proposed by Einhorn. Researchers have attempted to use outcome feedback to improve accuracy of decisions and modify decision making strategies. An illogical, uneconomical, but widely used decision making strategy first studied over twenty years ago has been a major topic of research.

An unusual phenomenon in decision making was evidenced in a rule guessing experiment by Wason (1960). A majority of subjects, after forming a tentative hypothesis, would seek only supporting evidence for the hypothesis and then make a final decision based only on that supporting evidence. The same failure by decision makers to seek disconfirming evidence for hypotheses was found by Einhorn (1980), Einhorn and Hogarth (1978), Estes (1976), and by Koriat, Lichtenstein, and Fischhoff (1980). Seeking only confirming evidence often leads to acceptance of incorrect hypotheses, and also causes the decision maker to be overconfident in estimating correctness of the decisions (Lichtenstein et. al., 1981). Measurements of calibration, the correspondence between the decision maker's estimated and actual accuracy of decisions, have shown overconfidence when decision makers sought only confirmation, no change when both confirming and disconfirming evidence was sought, and improvement approaching accuracy when only disconfirming evidence was assessed (Koriat, et. al., 1980). Slight improvement in calibration has been brought about by providing periodic feedback to the decision maker on the discrepancy between confidence judgments and the actual performance (Adams and Adams, 1958). A similar experiment involving feedback supported Adams and showed some generalization for other tasks of varying difficulty and content (Lichtenstein and Fischhoff, 1980).

A major endeavor in decision making research has been to improve calibration, that is, to minimize disparity between actual frequencies of occurrences and subjectively determined probabilities of the occurrences. For example, if a weather forecaster predicts a 70% chance of rain on each of ten consecutive days and there is rain on seven of those ten days, then the weather forecaster is perfectly calibrated. If he sould instead predict a 10% chance of rain for the same days with the same results, he would be poorly calibrated and possibly unemployed. The calibration attribute of decision makers is considered important to weather forecasters, stock brokers, intelligence analysts, and any other persons who are routinely required to make decisions and indicate the probabilites of the accuracy of those decisions. Improving DM calibration has been attempted by Oskamp (1962), Lichtenstein, Fischhoff, and Phillips (1981), Lichtenstein and Fischhoff (1980), and Koriat, Lichtenstein, and Fischhoff (1980). Two methods used have been to teach decision strategies involving seeking disconfirming evidence, and to provide feedback on appropriateness of confidence ratings.

Strategy modification has been used successfully in reducing overconfidence and improving calibration when DMs were required to seek disconfirming evidence for hypotheses (Koriat et. al., 1980). Feedback has been used with some success by Lichtenstein and Fischhoff (1980), but has led to little or no improvement in calibration in experiments by Adams and Adams (1958). Lichtenstein and Fischhoff (1980) point out that their study was unique in using intensive instructions, and in using sufficient responses -- two hundred questions per treatment block -- to ensure accurate feedback. The Lichtenstein and Fischhoff experiment demonstrates that rigorous laboratory conditions can be used to improve calibration. Additionally, continuous, accurate feedback provideed fairly rapidly after generation of probability estimates has led to real world improvement in calibration observed in weather forecasters by Murphy and Winkler (1977)(cited in Lichtenstein & Fischhoff, 1980).

The present research attempted to determine (a) the effects of type of self-generated evidence on over/under confidence and calibration, (b) the effects of presence or absence of feedback on over/under confidence and calibration, and the effects of feedback on over/under confidence, calibration, and generalization of decision making strategies. Three hypotheses were tested: (a) If only confirming evidence of a decision is generated by a subject, then the subject will be overconfident in estimating decision validity. (b) If only disconfirming evidence of a decision is generated by a subject, then overconfidence will be reduced. (c) If a subject is given feedback based on the accuracy of decisions, then overconfidence will be reduced and generalization of decision making strategies across tasks will be enhanced.

Chapter II

Method

Subjects

Subjects were 72 male and female undergraduate students participating as part of a requirement for an introductory course in psychology. Subjects were randomly assigned to one of six treatment groups In A 2 x 3 factorial design (type of feedback x type of evidence).

Test Materials

In Phase I of the experiment which tested effects of type of feedback and type of self-generated evidence on over/under confidence, a pamphlet was given to each subject consisting of five blocks of ten questions each selected from the general knowledge questions used in Experiment 3 of Lichtenstein and Fischhoff (1977). Blocks of questions were matched in difficulty. Each question was of a two-alternative format and questions covered a wide variety of topics. A random number gene^rator was then used to produce six different orders of question sets for the test booklets. For use in Phase II of the experiment, which tested for effects of type of feedback and type of self-generated evidence on generalization of decision making strategies across tasks, a blank sheet of paper for the "Concrete Reasoning" task (Wason & Johnson-Laird, 1972) and an answer sheet for the "Rule Guessing" task (Wason, 1960) were attached to each booklet in alternating order so that twelve unique test booklets resulted. Record sheets were blank forms with four columns headed "Numbers", "Reasons for Choice", "Conforms", and "Does Not Conform". Record sheets were comparable to those used by Wason (1960). One of these unique test booklets was used for each of the twelve subjects in each treatment condition. A printed sheet of five warm-up questions was inserted in each pamphlet prior to the first block of questions.

Apparatus

A 16 K-byte microcomputer (Radio Shack TRS-80 Model 26-1062) was used to provide visual feedback on a 20.4 x 25.9 cm black and white cathode ray tube (CRT). A Basic language computer program developed for this experiment (Appendix A) scored subject responses and determined the feedback display.

Dependent Measures

Over/Under Confidence scores in Phase I of the experiment were determined by subtracting the mean percentage of correct responses from the mean percentage of probability assessments across all scores for each subject, a method used by Koriat, Lichtenstein, and Fischhoff (1980). Calibration scores in Phase I were calculated using Oskamp's formula (Oskamp, 1962, p. 9):

$$A = \frac{\Sigma n \left| d_i \right|}{N}$$

where i is any point on the confidence scale from .5 to 1.0, d_i is the absolute difference between the point value on the scale and the percentage correct when given that point value; n_i is the number of judgements made at point i; and N is the total number of judgments made.

In Phase II, a "Rule Guessing" task and a "Concrete Reasoning" task were used to test for generalization of a learned decision making strategy to seek or not seek disconfirming evidence for hypotheses. In the "Rule Guessing" task, subjects were required to generate series of numbers and reasons or hypotheses accompanying each series. In the "Rule Guessing task of Phase II, seeking of disconfirming evidence was determined by comparing mathematical series given by a subject to the subject's current and previously stated hypotheses for the selection of the series. A series which was incompatible with either hypothesis was scored as evidence of seeking disconfirming evidence. This criterion was used by Wason (1960). In the "Concrete Reasoning" task, each subject was shown one side of four envelopes, each of which provided some information written or affixed to the visible side and had the potential, if turned over, to provide further information which could be used in making a decision. A subject was required to select envelopes one and four in order to be considered as seeking disconfirming evidence (Wason & Johnson-Laird, 1972).

Procedure

Prior to administration of the first test questions, five practice questions were given to each subject. The answers to these questions were entered into the microcumputer which computed calibration scores at each level of the confidence scale, an overall calibration score for each block of questions, and proportions of answers correct at each level of the confidence scale for which a judgment was made. Answers were analyzed and feedback was or was not given to the subject according to the instructions for his/her treatment condition. Data from the five practice questions was discarded.

In the first phase of the experiment, five blocks of ten questions were asked of each subject. A separate calibration score was computed for each individual's responses for each block of questions. Two independent variables (a) type of feedback, and (b) type of self-generated evidence were varied in a 2 x 3 factorial design to determine their effects on the dependent variables, over/under confidence and calibration. In the second phase of the experiment, two independent variables (a) type of feedback, and (b) type of self-generated evidence were varied in a 2 x 3 factorial design to determine their effects on the proportion of subjects seeking disconfirming evidence.

In Phase I, the six treatment groups were No Evidence (Control), No Evidence with Feedback, Confirming Evidence, Confirming Evidence with Feedback, Disconfirming Evidence, and Disconfirming Evidence with Feedback. No Evidence Groups answered the five blocks of questions with a choice and gave a confidence estimate ranging from .5 to 1.0. The Confirming Evidence Group answered the five blocks of questions by giving an answer, an estimate of confidence, and at least one reason for making the selection. The Disconfirming Evidence Group gave an answer, an estimate of confidence, and at least one reason why the choice might have been incorrect. The Confirming, Disconfirming, and No Evidence Groups with Feedback were given measures of individual calibration and proportions correct immediately after completing each block of ten questions. Instructions for the Feedback groups differed from those of the No Feedback groups in that the meaning and calculation of confidence scores was briefly explained and that subjects were told they would be given feedback on confidence scores at each point of the scale for which a judgment was made. Instructions for all treatment conditions were modeled after those used by Koriat et al. (1980)(see Appendix B).

The feedback consisted of each Feedback group subject viewing on the microcomputer CRT a table which presented the subject's overall calibration score, the calibration score and proportion correct at each percentage category, and the number of answers given in each category (see Figure 1). In addition, the experimenter briefly discussed the data with the subject.

The tabular presentation of calibration scores and proportions correct was immediately followed on the CRT by a graph depicting percentage of correct responses at each level of confidence estimates from .5 to 1.0 (see Figure 2). A diagonal line representing perfect calibration was imposed on the graph. The experimenter, refering to the diagonal line, explained to the subject how the subject percentage of correct scores at each confidence level showed overconfidence, underconfidence, or appropriate confidence.

In Phase II, each subject was given two additional tasks, balanced in order of presentation. In the "Rule Guessing" tasks, each subject was asked to guess a simple rule used to generate a series of three numbers. The experimenter gave a three number series that was generated by the rule that the experimenter had in mind. Each subject was required to then write down a set of three numbers and a reason for choosing the three numbers. For each set of three numbers generated by the subject, immediately after the subject wrote a reason, the experimenter told the subject whether or not the subject's series fit the rule. The subject recorded this feedback on the record sheet. The subject then continued generating number series until he was highly confident that he had discovered the rule. He then wrote down the rule across his record sheet ignoring column headings (see Appendix C for instructions).

In the second test of generalization, each subject was asked to perform a "reasoning problem" which was a version of the "concrete" problem used by Wason and Johnson-Laird (1972). In this task, four envelopes were shown to the subject: (a) a sealed envelope, (b) an open envelope, (c) an envelope with an affixed airmail stamp, and (d) an envelope with parcel post stamp. The subject was asked to decide if the following rule applied: "If a letter is sealed, then it has an airmail stamp on it." Subjects were instructed to indicate which envelope or envelopes that they would need to turn over in order to determine whether the rule was true or false.

Upon completion of the last task, subjects were instructed not to discuss the experiment or any portion of the experiment with any other persons for a period of one year.

Chapter III

Results

Manipulation Check

Test booklets were re-examined after all data had been gathered to ensure that instructions were followed. All subjects were found to have complied with instructions. The major concern was to verify that subjects in Confirming Evidence Groups generated confirming evidence and that those in Disconfirming Evidence Groups generated disconfirming evidence. Some subjects evidenced difficutly generating disconfirming evidence for .9 and 1.0 probability answers and would occasionally enter such reasons as "I can't think of any reason." or "I may be wrong because " and leave the sentence uncompleted. These errors were infrequent, and all data was retained for analysis.

Tallying of seeking of disconfirming evidence in the "Rule Guessing Task" was scored by the experimenter for evidence of seeking disconfirmation. These scores were validated by another experimenter who was blind to the treatment group of the subjects. The experimenters used the same criterion of determination prescribed by Wason (1960). Inter-rater agreement was then verified by computing a phi coefficient, $\phi = .83$.

Confidence

Over/Under Confidence scores were determined by subtracting the mean percentage of correct responses from the mean percentage of probability assessments across all scores for each subject (Koriat et. al., 1981). Mean Over/Under Confidence scores (see Figure 3) show that subjects were over-confident, that the Disconfirming Evidence Group Without Feedback showed the highest overconfidence, and that the greatest effects of feedback were in reducing overconfidence in the Disconfirming Evidence condition and in increasing overconfidence in the No Evidence condition (see Table 1).

Differences between conditions were tested for significance by means of a 5 x 2 x 3 (treatment block x type of feedback x type of self-generated evidence) factorial analysis of variance (see Table 3). The change in confidence scores across treatment blocks was significant, $\underline{F}(4, 264) = 3.02$, $p<.02$. All other interactions and main effects were non-significant although the interaction of Type of Evidence x Type of Feedback did approach significance, $\underline{F}(2, 66) = 2.94$, $p<.058$.

Mean Over/Under Confidence scores for each treatment condition were plotted for each treatment block. The plotted data (see Figure 5) revealed a similar trend in each treatment condition. All treatment groups showed high levels of overconfidence in the first treatment block;

overconfidence declined through the third treatment block, then returned to high levels by the fourth or fifth block. The degree to which the interaction of the confidence scores across treatment blocks were related in linear, quadratic, and cubic components was obtained and tested for significance by means of a 6 x 5 (treatment condition x treatment block) factorial trend analysis. The quadratic trend across treatment blocks was significant, F(1,284) = 10.78, p<.005. Results for interaction and main effects were non-significant.

Calibration

Calibration scores were calculated using Oskamp's formula (Oskamp, 1962) and measured appropriateness of confidence. Analysis of calibration scores (see Figure 4) yielded results corresponding to the mean Over/Under Confidence scores. Subjects in all treatment conditions were poorly calibrated. The highest group mean calibration score, indicating poorest calibration, was in the Disconfirming Evidence Group Without Feedback. Table 2 provides a summary of the means and standard deviations.

Differences between conditions were tested for significance by means of a 5 x 2 x 3 (treatment block x type of feedback x type of self-generated evidence) factorial analysis of variance (see Table 3). The Feedback x Evidence interaction was significant, F(2,66) = 5.36, p<.007. All other interactions and main effects were non-significant. An analysis of simple main effects for each type of Self-Generated Evidence showed that the mean of the Feedback Group was significantly lower than the mean of the No Feedback Group in the Disconfirming Evidence condition, $F(1,66) = 8.24$, $p<.01$. The mean of the Feedback Group was significantly higher than the mean of the No Feedback Groups in the No Evidence condition, $F(1,66) = 4.91$, $p<.05$.

Disconfirming Evidence

The data analyzed were the proportion of subjects in each treatment condition who showed evidence of seeking disconfirming evidence. In the "Rule Guessing Task", the criterion for selection were provided by Wason (1960). For the "Concrete Reasoning Task", the criterion were described by Wason & Johnson-Laird (1972).

Responses were coded so that each individual in each condition and task received a numeric score, "1" for seeking disconfirming evidence and "0" for not seeking disconfirming evidence. Differences between conditions with regard to these scores were then tested for significance by means of a 2 x 3 (type of feedback x type of self-generated evidence) factorial analysis of of variance (see Table 4). Main effects and interaction effects for both tasks were non-significant.

Chapter IV

Discussion

The hypothesis that if only confirming evidence of a decision is generated by a subject, then the subject will be overconfident in estimating decision validity was neither confirmed nor disconfirmed by the data. The hypothesis that if a subject is given feedback based on accuracy of decisions, then overconfidence will be reduced and generalization will be enhanced was not supported by the data. The hypothesis that if only disconfirming evidence of a decision is generated by a subject, then subject overconfidence will be reduced was not supported by the data.

The consensus of current research in decision theory is that decision makers tend to be overconfident in estimating the probabilities that their decisions are correct. Koriat et al. (1980, p. 4) proposed an information processing mechanism which would account for overconfidence. The mechanism suggests a predisposition during memory search and retrieval to "... rely more heavily on considerations consistent with a chosen answer than on considerations contradicting it." The predisposition is made evident when the decision maker (DM) is required to support/refute a decision or assess confidence in the decision. Accordingly, the DM can readily produce reasons for a choice, experiences difficulty generating reasons against the choice, and is overconfident in the decisions made. These proposals compliment those of Wason (1968) who asserts that a DM makes a decision based on available information. If the available information supports one alternative, then that alternative is chosen. Once the alternative is chosen, the DM is likely to seek further evidence supporting the decision. Having once amassed support for a decision, the DM is then reluctant to admit plausibility of disconfirming evidence.

Wason (1960) also noted that most subjects in simple mathematical reasoning tasks seek confirming evidence exclusively. Such a common decision making strategy would account for the finding by Koriat et al. (1980) that subjects who are required to generate no evidence for or against their decisions are overconfident, to approximately the same degree as subjects who are required to generate confirming evidence for their decisions.

The hypothesis that feedback would reduce overconfidence and enhance generalization was disconfirmed. Mean over/under confidence scores showed a reduction in overconfidence approaching but not reaching significance for the Disconfirming Evidence Group With Feedback in comparison to the Disconfirming Evidence Group Without Feedback. Upon first inspection, feedback appears to have had a reverse effect on reducing overconfidence for the No Evidence Groups and to have had little effect on reducing overconfidence on the Confirming Evidence Groups. However, after plotting the mean confidence scores of treatment groups across the five blocks of questions, the predicted reduction in confidence is evidenced through the first three blocks. The confidence scores then return to previous high levels by the fourth or fifth block of questions. The result of not achieving overall significant reduction in overconfidence conflicts with that of Lichtenstein & Fischhoff (1980) who found that intensive use of feedback based on large numbers of responses would significantly reduce overconfidence scores of subjects responding to two-alternative items. In contrast to this experiment, Lichtenstein & Fischhoff found that overconfidence scores measured across 12 blocks were significantly reduced with feedback training and remained relatively constant. Inexplicably, most or all of the reduction occured after the first trial.

Transitory effects in this experiment may be due to the relatively small number of test questions, ten, per block, and the brief instructions and feedback discussion.

The Lichtenstein and Fischhoff experiment used 200 questions per block, five typed single-spaced pages of instructions, eleven training sessions with feedback, and extensive experimenter/subject feedback discussions. As in the present investigation, computer-generated feedback was immediately presented to each subject upon completion of the last question in each treatment block. Comparing the intensity of training in the Lichtenstein and Fischhoff experiment to that used in this experiment, it is not surprising that training effects reported by those authors were more lasting. Building on the Koriat er al. (1980) information processing mechanism, the different results can be explained without contradiction. The DM in the No Evidence Group With Feedback or Confirming Evidence Group With Feedback, while responding to the first block of questions, performs a memory search and retrieval and is predisposed to place greater reliance on confirming evidence. Using the somewhat biased results, the DM then employs an already existing heuristic to determine a numeric equivalent for a feeling of confidence. When feedback is provided to the DM as the means of reducing overconfidence, the DM uses it, not to reduce the bias in his method of memory search and retrieval, but to fit numbers more appropriately to his feeling of confidence.

The decision maker employs a heuristic when matching numeric values to feelings of confidence. That heuristic is generally resistant to change. Since decision makers are quite often encouraged to display confidence or give overconfident estimates (Fischhoff, 1981), it is not surprising that heuristics employed for matching numeric values reflect the overconfidence.

An overconfidence reduction training program that uses brief instructions, few questions, feedback, and brief discussions of the feedback cannot overcome a heuristic developed over a several-year time span. The heuristic, though it may be temporarily overshadowed, quickly returns to use. Overconfidence reduction is better accomplished with more intense training which may cause immediate suppression and eventual modification of an unrelaistic heuristic.

The continuation of the second hypothesis, that feedback would enhance generalization, was not supported by the data. Generalization of learned calibration practices has been evidenced by slight improvements in calibration when training and test items were similar artificial game tasks (Pickhardt and Wallace, 1974). Subject scores in more realistic game settings, however, showed no improvement in calibration. Lichtenstein & Fischhoff (1980) did report generalization of calibration and overconfidence reduction training in tasks similar to training items in content and form, but differing in level of difficulty. Generalization tasks failed totally when response modes were not similar, for example, a two-alternative response mode and a four-alternative response mode. When training and test modes both use similar responses which consist of a choice and a confidence measure, the element of training which generalizes may be the ability to fit a numerical value to a feeling of confidence, or may be the generalization of a training strategy to seek either confirming, disconfirming, or both types of evidence. In contrast to the successful Lichtenstein and Fischhoff (1980) generalzation experiments, this experiment used dissimilar training response modes and test response modes.

The purpose of one portion of this experiment was to determine whether or not training strategies specifically would generalize. Solution of test items required subjects to actively seek or generate confirming and/or disconfirming evidence. No evidence of generalization was found. This result may be due to insufficient intensity and duration of training as in the overconfidence reduction training or may indicate that the metaheuristic which controls memory search and retrieval is more resistant to change than the heuristics which regulate fitting of numeric values to situational feelings of confidence. Evidence for change in a metaheuristic would be shown by changes in responses across tasks which utilized dependent heuristics (Einhorn, 1980).

The hypothesis that subject generation of disconfirming evidence would lead to reduced overconfidence was disconfirmed. Subject data from the Disconfirming Evidence Group Without Feedback showed no significant reduction in mean overconfidence scores across the five training blocks. This result appears to conflict with the results of Koriat et al. (1980). The Koriat et al. experiment referred to as Experiment 2 used three treatment conditions and a control condition. Treatment conditions called for generating one reason supporting each choice, one reason contradicting each choice, or both one supporting and one contradicting reason. The control subjects were not required to generate reasons. A within subjects design was used; each subject answered three sets of ten questions each under the control condition, then three blocks of ten questions each in one of the treatment conditions.

Results of Experiment 2 showed that subjects in the control and supporting conditions were equally overconfident and poorly calibrated. Subjects in the both condition showed no significant difference from the control condition. The contradicting evidence subjects showed a significant improvement in calibration and approached significance in reduction of overconfidence.

This experiment and the Koriat et al. experiment used brief sets of instructions, ten questions per training block, and required brief responses by the subject. The conflict in findings is then readily explained by attributing the reduction in overconfidence to a transitory effect of training. The Koriat et al. experiment found a reduction in overconfidence scores in the three sets of ten questions each used in the contradicting evidence condition. Similarly, this experiment found a reduction in overconfidence scores in the first three treatment blocks of ten question each for the disconfirming evidence without feedback condition. In the fourth and fifth blocks, however, overconfidence scores again increased to initial high levels. Unexpectedly, confidence scores in the disconfirming evidence groups were higher than confidence scores in the no evidence or confirming evidence groups. This overconfidence probably develops from DM memory search and retrieval activities. The decision maker searches for disconfirming evidence, finds little or none, and is even more confident in the decision. This explanation holds in that students who served as subjects reported a lack of familiarity with most items, some difficulty making decisions, and greater difficulty in generating disconfirming evidence. Typically, DMs who perform tasks perceived as difficult or impossible are the persons with the most extreme overconfidence (Nickerson & McGoldrich, 1965).

In this experiment, reduction of overconfidence was attempted by varying type of feedback and type of self-generated evidence. Feedback was associated with a transitory reduction in overconfidence; training to seek disconfirming evidence had a similar transitory reduction effect. Combined effects of type of feedback and type of self-generated evidence caused no significant reduction in overconfidence although the reduction in overconfidence approached significance. The reduction in overconfidence across the five treatment blocks was somewhat transitory, but was significant, indicating that the effects of feedback and seeking disconfiming evidence on reduction of overconfidence are additive. The transitory effect of training was most interesting in that if only three blocks of question had been used in each condition, then the effects of feedback could easily have been overestimated.

Making the training effect more permanent while maintaining a degree of economy is a logical next step in this research. At some point between the five training blocks of ten items each which lead to transitory effects and the two or thirteen blocks of 200 questions each used by Koriat et al. (1980) which lead to more lasting training effects there may be an economical treatment which will have relatively permanent effects. On the other hand, the return to previous high levels of overconfidence may itself be transitory. An experiment of similar design, but extending the number of treatment blocks would confirm or disconfirm this possibility.

Summary

In conclusion, the purpose of this experiment was to determine effects of presence or absence of feedback and type of self-generated evidence on DM over/under confidence and on generalization. Each type of feedback and type of self-generated evidence separately led to at best transitory effects on reducing overconfidence. The combined application of feedback and self-generated disconfirming evidence led to a reduction in overconfidence that approached but did not reach significance. These findings, with exception of the transitory effect of training, are in agreement with the consensus of recent decision making research. The discrepancy is explained in terms of a currently accepted information processing mechanism. Treatment groups showed no differences in performance on tasks used to test generalization of decision making strategies across tasks. The failure to generalize is explained in terms of the same information processing mechanism. Suggestions for future research are provided.

References

Adams, J.K. & Adams, P.A. Realism in confidence judgments. Psyychological Review, 1961, 68, 33-45.

Allais, M. The comportment of rational man. Econometrica, 1953, 21, 503-546

Anderson, N.H. Information integration theory. A brief survey. In D.H. Krantz, R.C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psyhology (Vol. 2). San Francisco: Freeman, 1974.

Anderson, N.H. Functional measurement in psychophysical judgment. Psychological Review, 1970, 77, 153-170.

Aumann, R.J. Utility theory without the completeness axiom. Econometrica, 1962, 30, 445-462.

Becker, G.M. & McClintock, C.G. Value: behavioral decision theory. Annual Review of Psychology, 1967, 12, 239-286.

Corbin. R.M. Decisions that might not get made. In T.S. Wallsten (Ed.), Cognitive Processes in Choice and Decision Behavior. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Dawes, R.M. & Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Ebbesen, E.B. & Konecni, V.J. On the external validity of decision making research. In T.S. Wallsten (Ed.), Cognitive Processes in Choice and Decision Behavior. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Einhorn, H.J. Learning from experience and suboptimal rules in decision making. In T.S. Wallsten (Ed.), Cognitive Processes in Choice and Decision Behavior. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Einhorn, H.J. & Hogarth, R. Confidence in judgment: Persistence of the illusion of validity. Psychological Review, 1978, 85, 395-416.

Ellsburg, D. Risk, ambiguity, and the Savage axioms. Quarterly Journal of Economics, 1961, 75, 643-669.

Estes, W.K. The cognitive side of probability learning. Psychological Review, 1976, 83, 37-64.

Fischhoff, B. Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1981.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Knowing what you want: measuring labile values. In T.S. Wallsten (Ed.), Cognitive Processes in Choice and Decision Behavior. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. Econometrica, 1979, 47, 263-291.

Koriat, A , Lichtenstein, S., & Fischhoff, B. Reasons for confidence. Journal of Experimental Psychology: Human learning and Memory, 1980, 6, 107-118.

Lichtenstein, S. & Fischhoff, B. Do those who know more also know more about how much they know? The calibration of probability judgments. Organizational Behavior and Human Performance, 1977, 20, 159-183.

Lichtenstein, S. & Fischhoff, B. Training for calibration. Organizational Behavior and Human Performance, 1980, 26, 149-171.

Lichtenstein, S., Fischhoff, B., & Phillips, L.D. Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press, 1981.

Nickerson, R.S., & McGoldrick, C.C., Jr. Confidence ratings and level of performance on a judgmental task. Perceptual and Motor Skills, 1965, 20, 311-316.

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

Phelps, R.H., & Shanteau, J. Livestock judges: how much information can an expert use? Organizational Behavior and Human Performance, 1978, 21, 209-219.

Pickhardt, R.C., & Wallace, J.B. A study of the performance of subjective probability assessors. Decision Sciences, 1974, 5, 347-363.

Pitz, G.F. The very guide of life: the use of probabilistic information for making decisions. In T.S. Wallsten (Ed.), Cognitive Processes in Choice and Decision Making. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Rapoport, A., & Wallsten, T.S. Individual decision behavior. Annual Review of Psychology, 1972, 23, 131-176.

Tversky, A., & Kahneman, D. Judgment under uncertainty: heuristics and biases. Science, 1974, 185, 1124-1131.

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

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

Wason, P.C., & Johnson-Laird, P.N. Psychology of Reasoning. Cambridge, Mass.: Harvard University Press, 1972.

Zajonc, R.B. Feeling and thinking: preferences need no inferences. American Psychologist, 1980, 35, 151-173.

Appendix A

Basic Language Computer Program for Scoring Subject Responses and Providing Visual Feedback on Over/Under Confidence Scores

Basic Language Computer Program

```
7 'LPRINT CHR$(27)"E":LPRINT CHR$(27)"G"
9 LPRINT"":LPRINT""
10 LPRINT TAB(67)"48"
11 LPRINT CHR$(154)
12 LPRINT "":LPRINT""
15 LPRINT TAB(20)"Basic Language Computer Program"
16 LPRINT TAB(15)"for Scoring Responses and Providing Feedback"
17 LPRINT"":LPRINT"":LPRINT""
18 LPRINT CHR$(15)
19 LLIST 20-
20 'A BASIC LANGUAGE COMPUTER PROGRAM TO SCORE RESPONSES
21 'AND PROVIDE FEEDBACK FOR THE "SIX SETS OF TEN QUESTIONS
22 'EACH" COMPILED BY KORIAT, LICHTENSTEIN, AND FISCHHOFF(1980)
25 'PROGRAM DEVELOPED BY
26 'JOHN R. TIFFANY AND LINDA S. BAKER
27 '
30 L1$="A":L2$="A":L3$="B":L4$="B":L5$="B"
40 A1$="B":A2$="A":A3$="B":A4$="B":A5$="B":A6$="A":A7$="A":A8$="A":A9$="B":A0$="A"
50 B1$="B":B2$="B":B3$="A":B4$="B":B5$="B":B6$="A":B7$="A":B8$="B":B9$="A":B0$="B"
60 C1$="B":C2$="B":C3$="A":C4$="A":C5$="B":C6$="B":C7$="A":C8$="B":C9$="A":C0$="B"
70 D1$="B":D2$="A":D3$="A":D4$="A":D5$="B":D6$="B":D7$="A":D8$="B":D9$="A":D0$="B"
80 E1$="A":E2$="A":E3$="B":E4$="A":E5$="A":E6$="A":E7$="B":E8$="B":E9$="B":E0$="B"
90 F1$="B":F2$="A":F3$="A":F4$="A":F5$="A":F6$="B":F7$="B":F8$="A":F9$="B":F0$="A"
100 INPUT "WHICH BLOCK OF QUESTIONS IS TO BE ANSWERED";I$
101 I5=0:I6=0:I7=0:I8=0:I9=0:I0=0
102 C5=0:C6=0:C7=0:C8=0:C9=0:C0=0
103 P5=0:P6=0:P7=0:P8=0:P9=0:P0=0
109 IF I$="NONE"THEN5000
110 IF I$="A" GOTO 300
120 IF I$="B" GOTO 400
130 IF I$="C" GOTO 500
140 IF I$="D" GOTO 600
150 IF I$="E" GOTO 700
155 IF I$="F" GOTO900
160 IF I$="L" GOTO 800
161 IF I$="T" GOSUB3000
162 GOSUB4000
163 GOTO165
165 PRINT CHR$(254)
170 PRINT "ANSWER USING LETTERS A--F OR L " :PRINT CHR$(254):GOTO100
300 INPUT "WHAT IS THE ANSWER TO A1";A$
```

```
310 INPUT "WHAT IS THE PROBABILITY OF A1";N
311 IF A$=A1$THENX=1ELSEX=0
312 GOSUB 2000
313 INPUT "ANS A2";A$
314 INPUT "PROB A2";N
314 INPUT "PROB A2";N
315 IF A$=A2$THEN X=1ELSEX=0
316 GOSUB2000
317 INPUT "ANS A3";A$
318 INPUT "PROB A3";N
319 IF A$=A3$THEN X=1ELSEX=0
320 GOSUB2000
321 INPUT "ANS A4";A$
322 INPUT "PROB A4";N
323 IF A$=A4$THENX=1ELSEX=0
324 GOSUB2000
325 INPUT "ANS A5";A$
326 INPUT "PROB A5";N
327 IF A$=A5$THENX=1ELSEX=0
328 GOSUB2000
329 INPUT "ANS A6";A$
330 INPUT "PROB A6";N
340 IF A$=A6$THENX=1ELSEX=0
341 GOSUB 2000
342 INPUT "ANS A7";A$
343 INPUT "PROB A7";N
344 IF A$=A7$THENX=1ELSEX=0
345 GOSUB 2000
346 INPUT "ANS A8";A$
347 INPUT "PROB A8";N
348 IF A$=A8$THENX=1ELSEX=0
349 GOSUB 2000
350 INPUT "ANS A9";A$
351 INPUT "PROB A9";N
352 IF A$=A9$THENX=1ELSEX=0
353 GOSUB 2000
354 INPUT "ANS A0";A$
355 INPUT "PROB A0";N
356 IF A$=A0$THENX=1ELSEX=0
357 GOSUB2000
358 GOSUB3000
359 GOSUB4000
360 GOTO100
400 INPUT "ANS B1";B$
410 INPUT "PROB B1";N
411 IF B$=B1$THENX=1ELSEX=0
412 GOSUB2000
413 INPUT "ANS B2";B$
414 INPUT "PROB B2";N
415 IF B$=B2$THENX=1ELSEX=0
416 GOSUB2000
417 INPUT "ANS B3";B$
```

```
418 INPUT "PROB B3";N
419 IF B$=B3$THENX=1ELSEX=0
420 GOSUB2000
421 INPUT "ANS B4";B$
422 INPUT "PROB B4";N
423 IF B$=B4$THENX=1ELSEX=0
424 GOSUB 2000
425 INPUT "ANS B5";B$
426 INPUT "PROB B5";N
427 IF B$=B5$THENX=1ELSEX=0
428 GOSUB 2000
429 INPUT "ANS B6";B$
430 INPUT "PROB B6";N
431 IF B$=B6$THENX=1ELSEX=0
432 GOSUB 2000
433 INPUT "ANS B7";B$
434 INPUT "PROB B7";N
435 IFB$=B7$THENX=1ELSEX=0
436 GOSUB2000
437 INPUT "ANS B8";B$
438 INPUT "PROB B8";N
439 IF B$=B8$THENX=1ELSEX=0
440 GOSUB 2000
441 INPUT "ANS B9";B$
442 INPUT "PROB B9";N
443 IF B$=B9$THENX=1ELSEX=0
444 GOSUB 2000
445 INPUT "ANS B0";B$
446 INPUT "PROB B0":N
447 IF B$=B0$THENX=1ELSEX=0
448 GOSUB2000
449 GOSUB3000
450 GOSUB4000
451 GOTO100
500 INPUT "ANS C1";C$
510 INPUT "PROB C1";N
520 IFC$=C1$THENX=1ELSEX=0
530 GOSUB2000
531 INPUT "ANS C2";C$
532 INPUT "PROB C2";N
533 IFC$=C2$THENX=1ELSEX=0
534 GOSUB2000
535 INPUT "ANS C3";C$
536 INPUT "PROB C3";N
537 IFC$=C3$THENX=1ELSEX=0
538 GOSUB2000
539 INPUT "ANS C4";C$
540 INPUT "PROB C4";N
```

```
541 IFC$=C4$THENX=1ELSEX=0
542 GOSUB2000
543 INPUT "ANS C5";C$
544 INPUT "PROB C5";N
545 IF C$=C5$THENX=1ELSEX=0
546 GOSUB2000
547 INPUT "ANS C6";C$
548 -637
549 IFC$=C6$THENX=1ELSEX=0
550 GOSUB2000
551 INPUT "ANS C7";C$
552 INPUT "PROB C7";N
553 IFC$=C7$THENX=1ELSEX=0
554 GOSUB2000
555 INPUT "ANS C8";C$
556 INPUT "PROB C8";N
557 IFC$=C8$THENX=1ELSEX=0
558 GOSUB2000
559 INPUT "ANS C9";C$
560 INPUT "PROB C9";N
561 IFC$=C9$THENX=1ELSEX=0
562 GOSUB2000
563 INPUT "ANS C0";C$
564 INPUT "PROB C0";N
565 IFC$=C0$THENX=1ELSEX=0
566 GOSUB2000
567 GOSUB3000
568 GOSUB4000
569 GOTO100
600 INPUT "ANS D1";D$
610 INPUT "PROB D1";N
611 IFD$=D1$THENX=1ELSEX=0
612 GOSUB2000
613 INPUT "ANS D2";D$
614 INPUT "PROB D2";N
615 IF D$=D2$THENX=1ELSEX=0
616 GOSUB2000
617 INPUT "ANS D3";D$
618 INPUT "PROB D3";N
619 IFD$=D3$THENX=1ELSEX=0
620 GOSUB2000
621 INPUT "ANS D4";D$
622 INPUT "PROB D4";N
623 IFD$=D4$THENX=1ELSEX=0
624 GOSUB2000
625 INPUT "ANS D5";D$
```

```
626 INPUT "PROB D5";N
627 IFD$=D5$THENX=1ELSEX=0
628 GOSUB2000
629 INPUT "ANS D6";D$
630 INPUT "PROB D6";N
631 IFD$=D6$THENX=1ELSEX=0
632 GOSUB2000
633 INPUT"ANS D7";D$
634 INPUT"PROB D7";N
635 IFD$=D7$THENX=1ELSEX=0
636 GOSUB2000
637 INPUT"ANS D8";D$
638 INPUT"PROB D8";N
639 IFD$=D8$THENX=1ELSEX=0
640 GOSUB2000
641 INPUT"ANS D9";D$
642 INPUT"PROB D9";N
643 IFD$=D9$THENX=1ELSEX=0
644 GOSUB2000
645 INPUT"ANS D0";D$
646 INPUT"PROB D0";N
647 IFD$=D0$THENX=1ELSEX=0
648 GOSUB2000
649 GOSUB3000
650 GOSUB4000
651 GOTC100
700 INPUT "ANS E1";E$
710 INPUT "PROB E1";N
711 IFE$=E1$THENX=1ELSEX=0
712 GOSUB2000
713 INPUT "ANS E2";E$
714 INPUT "PROB E2";N
715 IFE$=E2$THENX=1ELSEX=0
716 GOSUB2000
717 INPUT "ANS E3";E$
718 INPUT "PROB E3";N
719 IFE$=E3$THENX=1ELSEX=0
720 GOSUB2000
721 INPUT "ANS E4";E$
722 INPUT "PROB E4";N
723 IFE$=E4$THENX=1ELSEX=0
724 GOSUB2000
725 INPUT "ANS E5";E$
726 INPUT "PROB E5";N
727 IFE$=E5$THENX=1ELSEX=0
728 GOSUB2000
729 INPUT "ANS E6";E$
730 INPUT "PROB E6";N
```

```
731 IFE$=E6$THENX=1ELSEX=0
732 GOSUB2000
733 INPUT "ANS E7";E$
734 INPUT "PROB E7";N
735 IFE$=E7$THENX=1ELSEX=0
736 GOSUB2000
737 INPUT "ANS E8";E$
738 INPUT "PROB E8";N
739 IFE$=E8$THENX=1ELSEX=0
740 GOSUB2000
741 INPUT "ANS E9";E$
742 INPUT "PROB E9";N
743 IFE$=E9$THENX=1ELSEX=0
744 GOSUB2000
745 INPUT "ANS E0";E$
746 INPUT "PROB E0";N
747 IFE$=E0$THENX=1ELSEX=0
748 GOSUB2000
749 GOSUB3000
750 GOSUB4000
751 GOTO100
800 INPUT "WHAT IS THE ANSWER TO L1";L$
810 INPUT "WHAT IS THE PROBABILITY FOR L1";N
820 IF L$=L1$ THEN X=1 ELSE X=0
830 GOSUB 2000
831 INPUT "ANSWER L2";L$
832 INPUT "PROBABILITY L2";N
833 IF L$=L2$THENX=1ELSEX=0
834 GOSUB 2000
835 INPUT "ANSWER L3";L$
836 INPUT "PROBABILITY L3";N
837 IF L$=L3$THENX=1ELSEX=0
838 GOSUB 2000
840 INPUT "ANSWER L4";L$
841 INPUT "PROBABILITY L4";N
842 IF L$=L4$THENX=1ELSEX=0
843 GOSUB 2000
845 INPUT "ANSWER L5";L$
846 INPUT "PROBALILITY L5";N
847 IF L$=L5$THENX=1ELSEX=0
848 GOSUB 2000
849 GOSUB 3000
850 GOSUB4000
851 GOTO100
900 INPUT "WHAT IS THE ANSWER TO F1";F$
910 INPUT "WHAT IS THE PROBABILITY OF F1";N
911 IF F$=F1$THENX=1ELSEX=0
912 GOSUB2000
913 INPUT "ANS F2";F$
914 INPUT "PROB F2";N
```

```
915 IF F$=F2$THENX=1ELSEX=0
916 GOSUB2000
917 INPUT "ANS F3";F$
918 INPUT "PROB F3";N
919 IFF$=F3$THENX=1ELSEX=0
920 GOSUB2000
921 INPUT"ANS F4";F$
922 INPUT"PROB F4";N
923 IFF$=F4$THENX=1ELSEX=0
924 GOSUB2000
925 INPUT"ANS F5";F$
926 INPUT"PROB F5";N
927 IFF$=F5$THENX=1ELSEX=0
928 GOSUB2000
929 INPUT"ANS F6";F$
930 INPUT"PROB F6";N
931 IFF$=F6$THENX=1ELSEX=0
932 GOSUB2000
933 INPUT"ANS F7";F$
934 INPUT"PROB F7";N
935 IFF$=F7$THENX=1ELSEX=0
936 GOSUB2000
937 INPUT"ANS F8";F$
938 INPUT"PROB F8";N
939 IFF$=F8$THENX=1ELSEX=0
940 GOSUB2000
941 INPUT"ANS F9";F$
942 INPUT"PROB F9";N
943 IFF$=F9$THENX=1ELSEX=0
944 GOSUB2000
945 INPUT"ANS F0";F$
946 INPUT"PROB F0";N
947 IFF$=F0$THENX=1ELSEX=0
948 GOSUB2000
949 GOSUB3000
950 GOSUB4000
951 GOTO100
999 GOTO4000
2000 IF N>.5THEN2100ELSEGOTO2010
2010 P5=P5+1
2020 IF X=1THENC5=C5+1
2030 IF X=0THENI5=I5+1
2040 GOTO2600
2100 IFN>.6THEN2200
2110 P6=P6+1
2120 IFX=1THENC6=C6+1
2130 IFX=0THENI6=I6+1
2140 GOTO2600
2200 IFN>.7THEN2300
```

```
2210 P7=P7+1
2220 IFX=1THENC7=C7+1
2230 IFX=0THENI7=I7+1
2240 GOTO2600
2300 IFN>.8THEN2400
2310 P8=P8+1
2320 IFX=1THENC8=C8+1
2330 IFX=0THENI8=I8+1
2340 GOTO2600
2400 IFN>.9THEN2500
2410 P9=P9+1
2420 IFX=1THENC9=C9+1
2430 IFX=0THENI9=I9+1
2440 GOTO2600
2500 P0=P0+1
2510 IFX=1THENC0=C0+1
2520 IFX=0THENI0=I0+1
2600 RETURN
3000 PRINT
3010 PRINT "% CATEGORY";TAB(16);"# OF ANS IN CAT";TAB(40);"PROPORTION CORRECT"
3029 IFP5=0THEN3039
3030 PRINT ".50";TAB(16);P5;TAB(40);C5/(C5+I5)
3031 GR(0)=C5/(C5+I5)
3032 OS(5)=P5*ABS(.5-GR(0))
3039 IFP6=0THEN3049
3040 PRINT ".60";TAB(16);P6;TAB(40);C6/(C6+I6)
3041 GR(1)=C6/(C6+I6)
3042 OS(6)=P6*ABS(.6-GR(1))
3049 IFP7=0THEN3059
3050 PRINT ".70";TAB(16);P7;TAB(40);C7/(C7+I7)
3051 GR(2)=C7/(C7+I7)
3052 OS(7)=P7*ABS(.7-GR(2))
3059 IFP8=0THEN3069
3060 PRINT ".80";TAB(16);P8;TAB(40);C8/(C8+I8)
3061 GR(3)=C8/(C8+I8)
3062 OS(8)=P8*ABS(.8-GR(3))
3069 IFP9=0THEN3079
3070 PRINT ".90";TAB(16);P9;TAB(40);C9/(C9+I9)
3071 GR(4)=C9/(C9+I9)
3072 OS(9)=P9*ABS(.9-GR(4))
3079 IFP0=0THEN3110
3080 PRINT "1.0";TAB(16);P0;TAB(40);C0/(C0+I0)
3081 GR(5)=C0/(C0+I0)
3082 OS(10)=P0*ABS(1.-GR(5))
3110 PRINT "CALIBRATION SCORE AT .5 = ";OS(5)
3120 PRINT "CALIBRATION SCORE AT .6 = ";OS(6)
3130 PRINT "CALIBRATION SCORE AT .7 =";OS(7)
3140 PRINT "CALIBRATION SCORE AT .8 =";OS(8)
3150 PPINT "CALIBRATION SCORE AT .9 =";OS(9)
3160 PRINT "CALIBRATION SCORE AT 1. =";OS(10)
3200 OK=0
3210 FOR K=5TO10
3220 OK=OK+OS(K)
```

```
3240 FM=OK/10
3250 PRINT "OVERALL CALIBRATION SCORE = ";FM
3910 INPUT "READY";Z$
3990 RETURN
4000 CLS
4001 N=100
4002 FOR G=202 TO 842 STEP 64
4003 PRINT @G,N
4004 N=N-10
4005 NEXT G
4010 FORX=30TO120
4020 Y=41
4030 SET(X,Y)
4040 NEXT X
4050 FOR Y=9TO41
4060 X=30
4070 SET(X,Y)
4080 NEXTY
4090 P=911
4100 FORN=.5TO1.1STEP.1
4110 PRINT@P,N
4120 P=P+8
4130 NEXT N
4140 X=32:Y=24
4150 IFX>110THEN4300
4160 SET(X,Y)
4170 X=X+5:Y=Y-1
4180 GOTO4150
4200 PRINT
4300 FOR D=0 TO 5
4301 A=D
4308  B=INT(GR(D)*10):B=10-B
4309 X=34+(A*16)
4310 Y=INT(10+(B*3)):Y=Y-1
4311 SET (X,Y)
4314 NEXT D
4600 PRINT TAB(27)"CONFIDENCE ESTIMATES"
4620 PRINT
4621 PRINT@320,"% CORRECT";
4622 PRINT@900," ";
4625 GOTO30010
4630 INPUT "READY";V$
4640 PRINT CHR$(234)
4999 RETURN
5000 END
```

Appendix B

Instructions for Phase I

(Subject Instructions for Completing Five Blocks of Two-Alternative General Knowledge Questions)

Instructions for No Evidence Group Without Feedback

This study is concerned with human decision making. We are interested in how people make decisions and how accurately they can estimate their chances of being correct.

You will be asked to answer fifty general knowledge questions, ten questions at a time. To answer a question, just put a circle around either a or b next to the question in this test booklet. The question will look like this. (Show example question on a sheet as instructions are read.) After you circle a or b, you will be asked to estimate the probability that your answer is correct. You are to limit your estimates to .5 through 1.0. Please respond in even tenths, that is .5, .6, .7, .8, .9, 1.0. Write your estimate on the line marked 'Probability' in your answer booklet. To give you an idea of what you are to do, if you are absolutely certain that your answer is correct you should write 1.0, if you are almost certain, write .9. If you think there is only a 50-50 chance that you are correct, write .5. If you might be correct, write .6, and so. Get the idea? Any questions?

After each block of ten questions, I will record your answers.

Let's begin with five practice questions. Turn to the first page of your booklet. Remember, read the question, choose your answer, circle a or b, and estimate the probability that you are correct.

Instructions for Confirming Evidence Group Without Feedback

This study is concerned with human decision making. We are interested in how people make decisions and how accurately they can estimate their chances of being correct.

You will be asked to answer fifty general knowledge questions, ten questions at a time. To answer a question, just put a circle around either a or b next to the question in this test booklet. The question will look like this. (Show example question on a sheet as instructions are read.) After you circle a or b, then write at least one reason why your answer could be right. For example, reasons may include facts that you know, things you vaguely remember, assumptions that make you believe that your answer is likely to be correct, feelings, or associations. After you write down your reason or reasons, you will be asked to estimate the probability that your answer is correct. You are to limit your estimates to .5 through 1.0. Please respond in even tenths, that is .5, .6, .7, .8, .9, 1.0. Write your estimate on the line marked 'Probability' in your answer booklet. To give you an idea of what you are to do, if you are absolutely certain that your answer is correct you should write 1.0, if you are almost certain, write .9. If you think there is only a 50-50 chance that you are correct, write .5. If you might be correct, write .6, and so. Get the idea? Any questions?

After each block of ten questions, I will record your answers.

Let's begin with five practice questions. Turn to the first page of your booklet. Remember, read the question, choose your answer, circle a or b, write at least one reason for the choice, and then estimate the probability that you are correct.

Instructions for Disconfirming Evidence Group Without Feedback

This study is concerned with human decision making. We are interested in how people make decisions and how accurately they can estimate their chances of being correct.

You will be asked to answer fifty general knowledge questions, ten questions at a time. To answer a question, just put a circle around either a or b next to the question in this test booklet. The question will look like this. (Show example question on a sheet as instructions are read.) After you circle a or b, then write at least one reason why your answer could be wrong. For example, reasons may include facts that you know, things you vaguely remember, assumptions that give you some doubt that your answer s correct, feelings, or associations. After you write down your reason or reasons, you will be asked to estimate the probability that your answer is correct. You are to limit your estimates to .5 through 1.0. Please respond in even tenths, that is .5, .6, .7, .8, .9, 1.0. Write your estimate on the line marked 'Probability' in your answer booklet. To give you an idea of what you are to do, if you are absolutely certain that your answer is

correct you should write 1.0, if you are almost certain, write .9. If you think there is only a 50-50 chance that you are correct, write .5. If you might be correct, write .6, and so. Get the idea? Any questions?

After each block of ten questions, I will record your answers. Let's begin with five practice questions. Turn to the first page of your booklet. Remember, read the question, choose your answer, circle a or b, write at least one reason why your answer might be wrong, and estimate the probability that your answer is correct.

Instructions for No Evidence Group
With Feedback

This study is concerned with human decision making. We are interested in how people make decisions and how accurately they can estimate their chances of being correct.

You will be asked to answer fifty general knowledge questions, ten questions at a time. To answer a question, just put a circle around either a or b next to the question in this test booklet. The question will look like this. (Show example question on a sheet as instructions are read.) After you circle a or b, you will be asked to estimate the probability that your answer is correct. You are to limit your estimates to .5 through 1.0. Please respond in even tenths, that is .5, .6, .7, .8, .9, 1.0. Write your estimate on the line marked 'Probability' in your answer booklet. To give you an idea of what you are to do, if you are absolutely certain that your answer is correct you should write 1.0, if you are almost certain, write .9. If you think there is only a 50-50 chance that you are correct, write .5. If you might be correct, write .6, and so. Get the idea? Any questions?

After each block of ten questions, I will record your answers and give you feedback on the accuracy of your estimates of how likely it was that you were correct, the probabilities you provided on your answer sheet. I will tell you if you were overconfident or underconfident at each of the probability levels from .5 to 1.0.

For example, if you said .5 for two answers and one of those answers was correct, your confidence estimate is accurate since you were right 50% of the time. If you gave a probability of .8 for three answers and missed one of them, you would be over-confident since you were only right 67% of the time. Any questions?

Let's begin with five practice questions. Turn to the first page of your booklet. Remember, read the question, choose your answer, circle a or b, and estimate the probability that your answer is correct.

Instructions for Confirming Evidence Group With Feedback

This study is concerned with human decision making. We are interested in how people make decisions and how accurately they can estimate their chances of being correct.

You will be asked to answer fifty general knowledge questions, ten questions at a time. To answer a question, just put a circle around either a or b next to the question in this test booklet. The question will look like this. (Show example question on a sheet as instructions are read.) After you circle a or b, then write at least one reason why your answer could be right. For example, reasons may include facts that you know, things you vaguely remember, assumptions that make you believe that your answer is correct, feelings, or associations. After you write down your reason or reasons, you will be asked to estimate the probability that your answer is correct. You are to limit your estimates to .5 through 1.0. Please respond in even tenths, that is .5, .6, .7, .8, .9, 1.0. Write your estimate on the line marked 'Probability' in your answer booklet. To give you an idea of what you are to do, if you are absolutely certain that your answer is

correct you should write 1.0, if you are almost certain, write .9. If you think there is only a 50-50 chance that you are correct, write .5. If you might be correct, write .6, and so. Get the idea? Any questions?

After each block of ten questions, I will record your answers and give you feedback on the accuracy of your estimates of how likely it was that you were correct, the probabilities you provided on your answer sheet. I will tell you if you were overconfident or underconfident at each of the probability levels from .5 to 1.0.

For example, if you said .5 for two answers and one of those answers was correct, your confidence estimate is accurate since you were right 50% of the time. If you gave a probability of .8 for three answers and missed one of them, you would be over-confident since you were only right 67% of the time. Any questions?

Let's begin with five practice questions. Turn to the first page of your booklet. Remember, read the question, choose your answer, circle a or b, write at least one reason why your answer might be right, and estimate the probability that your answer is correct.

Instructions for Disconfirming Evidence Group
With Feedback

This study is concerned with human decision making. We are interested in how people make decisions and how accurately they can estimate their chances of being correct.

You will be asked to answer fifty general knowledge questions, ten questions at a time. To answer a question, just put a circle around either a or b next to the question in this test booklet. The question will look like this. (Show example question on a sheet as instructions are read.) After you circle a or b, then write at least one reason why your answer could be wrong. For example, reasons may include facts that you know, things you vaguely remember, assumptions that give you some doubt that your answer is correct, feelings, or associations. After you write down your reason or reasons, you will be asked to estimate the probability that your answer is correct. You are to limit your estimates to .5 through 1.0. Please respond in even tenths, that is .5, .6, .7, .8, .9, 1.0. Write your estimate on the line marked 'Probability' in your answer booklet. To give you an idea of what you are to do, if you are absolutely certain that your answer is

correct you should write 1.0, if you are almost certain, write .9. If you think there is only a 50-50 chance that you are correct, write .5. If you might be correct, write .6, and so. Get the idea? Any questions?

After each block of ten questions, I will record your answers and give you feedback on the accuracy of your estimates of how likely it was that you were correct, the probabilities you provided on your answer sheet. I will tell you if you were overconfident or underconfident at each of the probability levels from .5 to 1.0.

For example, if you said .5 for two answers and one of those answers was correct, your confidence estimate is accurate since you were right 50% of the time. If you gave a probability of .8 for three answers and missed one of them, you would be over-confident since you were only right 67% of the time. Any questions?

Let's begin with five practice questions. Turn to the first page of your booklet. Remember, read the question, choose your answer, circle a or b, write at least one reason why your answer might be wrong, and estimate the probability that your answer is correct.

Appendix C

Instructions for Phase II

(Subject Instructions for Completing
The "Concrete Reasoning" Task
and "Rule Guessing" Task)

Instructions for "Rule Guessing" Task (Wason, 1960)

You will be given three numbers which conform to a simple rule that I have in mind. This rule is concerned with a relation between any three numbers and not with their absolute magnitude, i.e., it is not a rule like all numbers above (or below) 50, etc. Your aim is to discover this rule by writing down sets of three numbers, together with reasons for your choice of them. After you have written down each set, I shall tell you *whether your numbers conform* to the rule or not, and you can make a note of this outcome on the record sheet provided. There is not time limit but you should try to discover this rule by citing the minimum sets of numbers. Remember that your aim is not simply to find numbers which conform to the rule, but to discover the rule itself. When you feel highly confident that you have discovered it, and not before, you are to write it down and tell me what it is.

You will write the rule across the record sheet ignoring column headings. You will be allowed to make only one guess at the rule. When you have made your guess, the task is over. Do you have any questions?

Instructions for "Concrete Reasoning" Task

Before you are four envelopes. The first is obviously sealed; the second is obviously open. The third has an airmail stamp; the fourth has an ordinary postage stamp. A rule is printed above the envelopes. That rule is: "If a letter is sealed, then it has an airmail stamp on it." Your task is to list the envelope or envelopes, that envelope or envelopes only, that need to be turned over in order to determine whether the rule is true or false. When you have made your decision, write down the number or numbers of the envelope or envelopes that you would need to turn over. A blank page has been provided in your answer book for your response.

Do you have any questions?

Appendix D

Summary Tables for Analyses of Variance, Means, and Standard Deviations for Dependent Measures

Table 1

Cell Means and Standard Deviations of

Over/Under Confidence Scores

On Type of Feedback by Type of Self-Generated Evidence

		Type of Self-Generated Evidence		
		None	Confirming	Disconfirming
Type of Feedback				
No Feedback	M =	7.57	10.15	15.23
	SD =	6.29	7.39	8.67
Feedback	M =	12.97	9.03	10.32
	SD =	6.31	6.54	7.32

Table 2

Cell Means and Standard Deviations of

Calibration Scores On Type of Feedback

by Type of Self-Generated Evidence

		Type of Self-Generated Evidence		
		None	Confirming	Disconfirming
Type of Feedback				
No Feedback	M =	244.50	269.00	302.00
	SD =	53.63	41.32	47.38
Feedback	M =	287.83	251.00	245.83
	SD =	57.49	32.42	50.95

Table 3

Summary Table of Analyses of Variance

for Dependent Measures in Phase I

Response Measure	DF	MS	F	p
Confidence Scores				
B (Type Evidence)	2	337.62	1.22	NS
C (Type Feedback)	1	4.01	.01	NS
BxC (Interaction)	2	816.70	2.94	NS
S/AB (Between Grps)	66	277.78		
A (Treatment Block)	4	822.20	3.02	p<.018
AxB (Interaction)	8	98.45	.36	NS
AxC (Interaction)	4	270.50	.99	NS
AxBxC (Interaction)	8	154.32	.57	NS
S/ABC (Within Grps)	264	272.09		
Calibration Scores				
B (Type Evidence)	2	10443.71	.87	NS
C (Type Feedback)	1	4840.76	.40	NS
BxC (Interaction)	2	64212.96	5.36	p<.007
S/AB (Between Grps)	66	11977.53		
A (Treatment block)	4	16656.72	1.30	NS
AxB (Interaction)	8	4868.93	.93	NS
AxC (Interaction)	4	13475.23	1.05	NS
AxBxC (Interaction)	8	7978.01	.62	NS
S/ABC (Within Grps)	264	12803.09		

Table 4

Summary Table of Analysis of Variance

for Dependent Measures in Phase II

Response Measure	SS	DF	MS	F	*p*
Rule Guessing Scores					
A (Type Feedback)	.528	2	.264	1.05	NS
B (Type Evidence)	.056	1	.056	0.22	NS
AxB (Interaction)	.694	2	.347	1.38	NS
S/AB (Within Grps)	16.667	66	.253		
Concrete Reasoning Scores					
A	.083	2	.042	.36	NS
B	.125	1	.125	1.09	NS
AxB	.083	2	.042	.36	NS
S/AB	7.583	66	.115		

Table 5

Summary Table of Trend Analysis

for Over/Under Confidence Scores

Across Treatment Blocks

Response Measure	DF	MS	F	p
A (Treatment Blocks)				
Linear	1	321.33	1.24	NS
Quadratic	1	2830.08	10.78	p <.005
Cubic	1	75.40	.28	NS
S/A (Within Groups)	284	262.43		

Figure 1

Representation of Tabular Feedback

Provided to Subjects on CRT

```
PROB BO? .5

% CATEGORY          # OF ANS IN CAT             PROPORTION CORRECT
.50                  3                           .666667
.60                  2                           .5
.70                  1                           1
.90                  1                           0
1.0                  3                           1
CALIBRATION SCORE AT .5 =  .5
CALIBRATION SCORE AT .6 =  .2
CALIBRATION SCORE AT .7 = .3
CALIBRATION SCORE AT .8 = 0
CALIBRATION SCORE AT .9 = .9
CALIBRATION SCORE AT 1. = 0
OVERALL CALIBRATION SCORE =  .19
```

Figure 2

Representation of Graph of Subject

Data Feedback on CRT

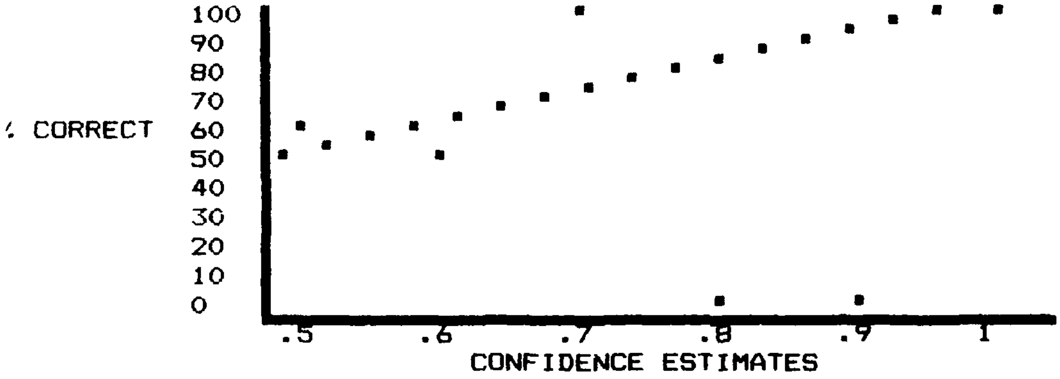
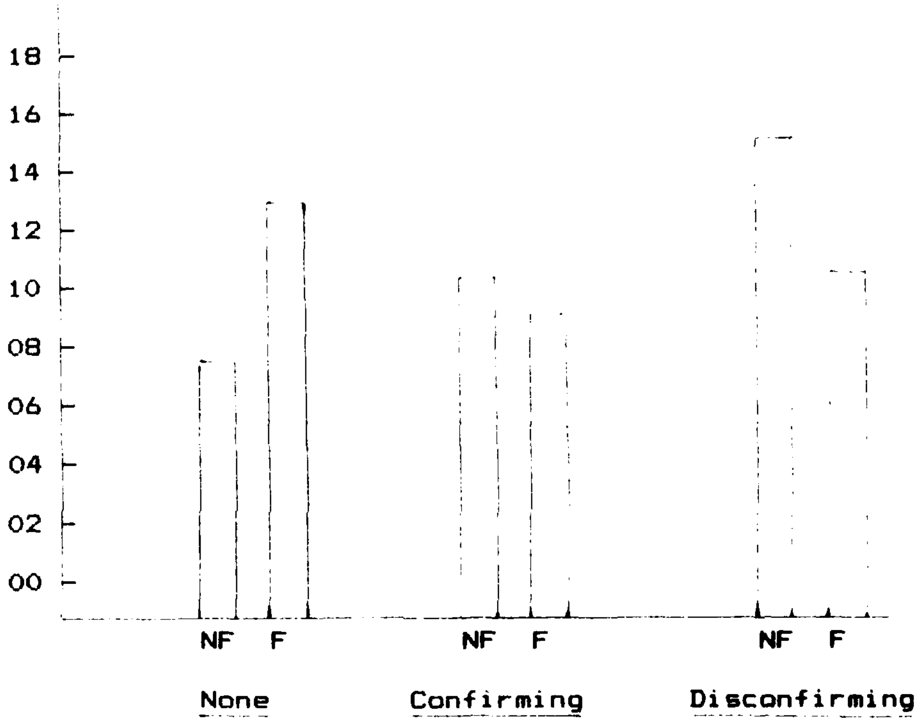


Figure 3

Mean Over/under Confidence

on All Treatment Conditions

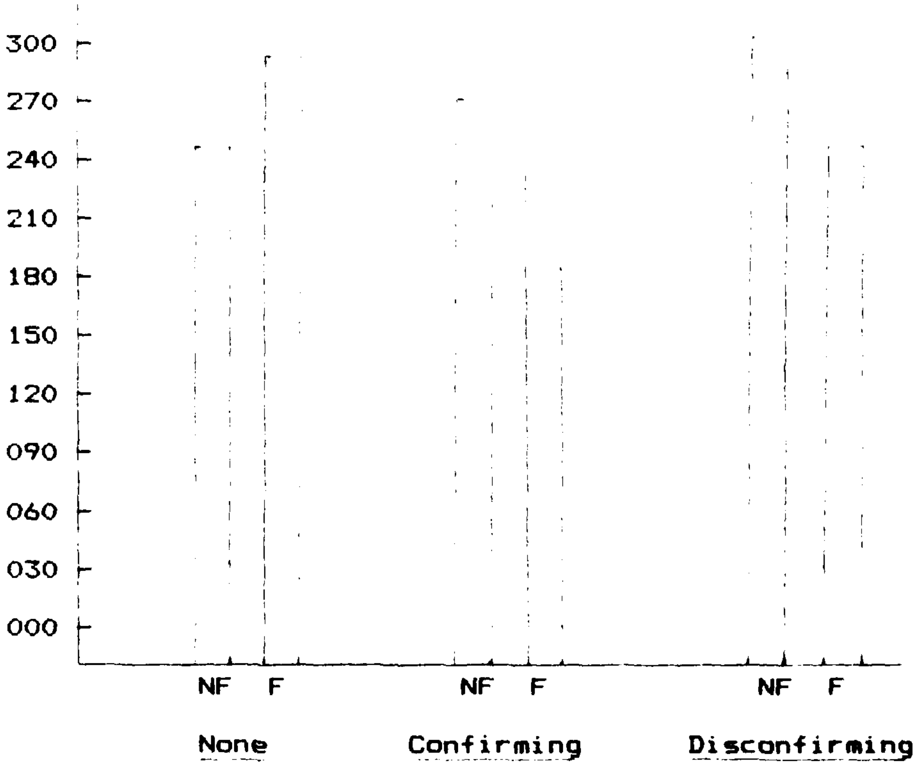


Type of Self-Generated Evidence

Figure 4

Mean Calibration Scores

on All Treatment Conditions

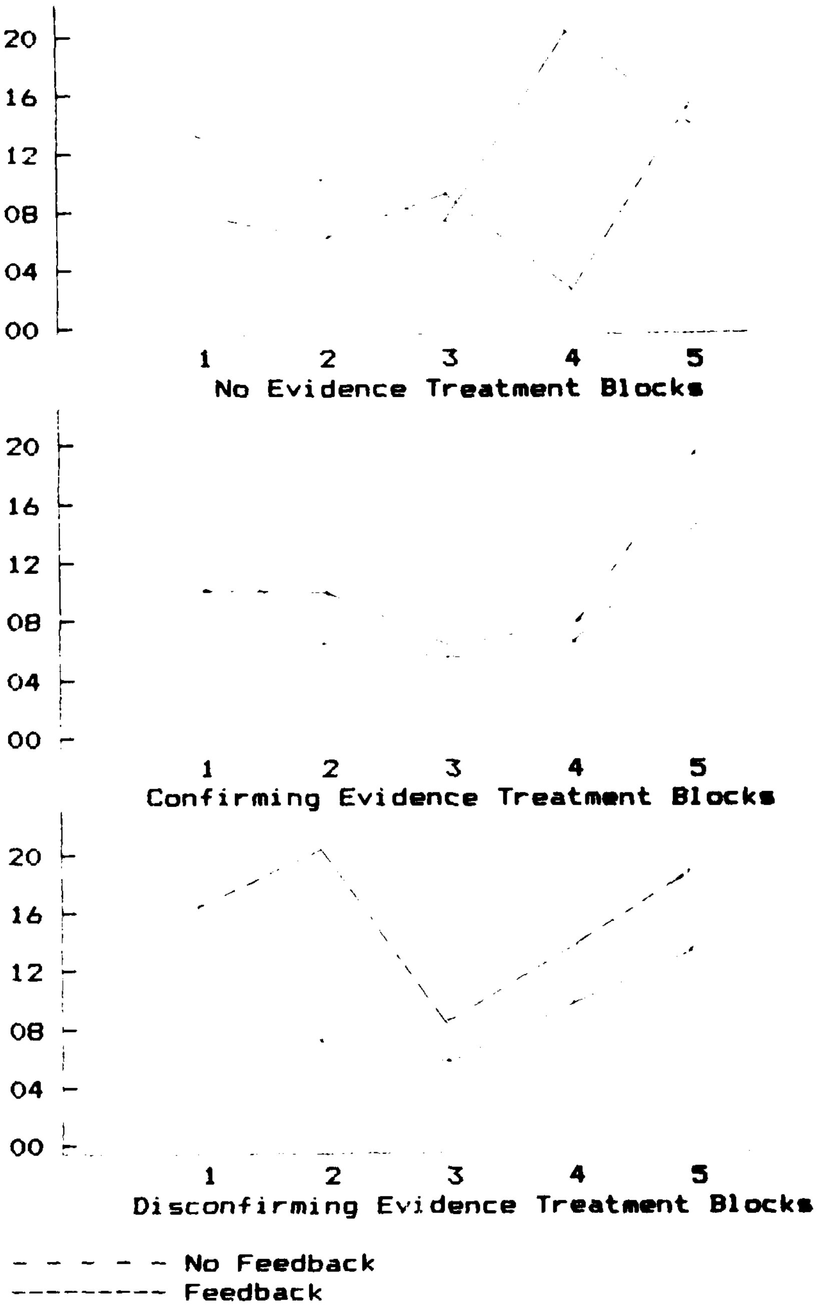


Type of Self-Generated Evidence

Figure 5

Mean Confidence Scores

Across Treatment Blocks


No Evidence Treatment Blocks
Confirming Evidence Treatment Blocks
Disconfirming Evidence Treatment Blocks
- - - - - No Feedback
--------- Feedback