DIRECT ESTIMATION PROCEDURES FOR ELICITING JUDGMENTS ABOUT UNCERTAIN EVENTS

Barbara C. Goodman

Michigan University

Prepared for:

Office of Naval Research

2 November 1973

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)* AD-770686

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 011313-5-T	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: DIRECT ESTIMATION PROCEDURES FOR ELICITING JUDGMENTS ABOUT UNCERTAIN EVENTS	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER: None
7. AUTHOR*(s)*: Barbara C. Goodman	8. CONTRACT OR GRANT NUMBER*(s)*: N00014-67-A-0181-0049
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Engineering Psychology Laboratory, Institute of Science and Technolgoy, University of Michigan, Ann Arbor, Michigan 48105	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 197-021; ARPA Order No. 2105
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Boulevard, Washington, D. C.]]]09	12. REPORT DATE: 2 November 1973
	13. NUMBER OF PAGES: ~~40~~ 44
14. MONITORING AGENCY NAME AND ADDRESS *(if different from Controlling Office)*: Engineering Psychology Programs, Office of Naval Research, Department of the Navy, Arlington, Virginia	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a DECLASSIFICATION/ DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Probabilistic Information Processing
Judgments
Elicitation
Bayes's Theorem
Estimation procedures

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This report re-analyses data from five studies concerned with methods for eliciting judgments about uncertain events. It focusses on response modes, such as odds, likelihood ratios, etc.: whether or not the response required the subject to aggregate items of data in his head; whether the scale on which the response was made logarithmic or linear; and whether the subject received additional feedback about the implications of his estimates in the course of making them. While no single experiment studies all these issues simultaneously, combination of the data from the

five experiments permits some strong conclusions:

1. Presence of additional feedback about the implications of estimates is probably the most powerful variable controlling the extremeness of these estimates; feedback makes estimates less extreme. Whether the less extreme estimates are closer to or further from correct Bayesian values depends on stimulus conditions.

2. Aggregated responses are consistently less extreme than nonaggregated responses.

3. Linear scales produce less extreme responses than logarithmic scales.

4. Likelihood ratio estimates are sometimes less extreme than odds estimates.

Other conclusions are also reviewed. Implications of these conclusions for the design of probabilistic information processing systems and for further research on response modes for information processing are discussed.

011313-5-T

DIRECT ESTIMATION PROCEDURES FOR ELICITING JUDGMENTS ABOUT UNCERTAIN EVENTS

Technical Report

2 November 1973

Barbara C. Goodman

Engineering Psychology Laboratory

The University of Michigan

Ann Arbor, Michigan

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Engineering Psychology Programs, Office of Naval Research under Contract No. N00014-67-A-0181-0049, Work Unit Number NR. 197-021.

INTRODUCTION

Within the domain of Bayesian research on Probabilistic Information Processing, only a few published studies have had the purpose of comparing different direct estimation response procedures (Kaplan & Newman, 1966; Phillips & Edwards, 1966; Fujii, 1967). Knowledge in this area becomes more important as the Bayesian techniques become more widely used in real world applications. Some of those using this technology have had access to the results of unpublished experiments conducted in University laboratories or have done their own research on this topic; others not.

This study analyzes the data from five separate research projects, only one previously published, in order to examine what we know and some of what we don't know about different direct estimation procedures for eliciting judgments about uncertainty.

Since all of these studies were done in the Engineering Psychology Laboratory, I have in each case been able to reanalyze the raw data. I am grateful to my colleagues for help and access.

THE FRAMEWORK

Four orthogonal independent variables have been manipulated in direct estimation response studies. These are the particular kinds of response mode in which S is asked to express his uncertainty judgment, the cumulative or noncumulative nature of the response, the particular nature of the scale being used to record the response, and the particular nature of the additional feedback given to S while S is in the process of deciding upon his assessment. The different response modes that have been examined include likelihood ratios (LR), odds (ODDS), and probabilities (PROB). Some experiments have systematically varied whether or not Ss aggregate their uncertainty assessments over more than one datum. The type of scale on which Ss record their estimates has also been investigated. Two types of scale need to be distinguished, predrawn logarithmically spaced scales and an everything else category. The variable Additional Feedback can itself be classified on the basis of three orthogonal considerations. These are whether the additional feedback is in graphic form or not, whether it is current system opinion based on just the one datum being evaluated or based on all the relevant data to date, and whether this additional feedback is in the form of ODDS or PROB.

These four variables form a four dimensional framework or taxonomy to classify all the response conditions in all of the experiments that will be discussed. This structure is presented in Table 1. The last dimension, Additional Feedback, in order to be orthogonal with the other three dimensions, specifically excludes those features of the response already contained in a description of the other three dimensions. Consider two examples. In the

TABLE 1

TAXONOMY

	Category	Definition
DIMENSION 1:		RESPONSE MODE
	ODDS	Ss express their uncertainty in odds judgments
	LR	Ss express their uncertainty in likelihood ratio judgments
	PROB	Ss express their uncertainty in probability judgments
DIMENSION 2:		AGGREGATION
	CUM	Ss' uncertainty judgments are aggregated over a set of data
	NONCUM	Ss' uncertainty judgments are for a single datum only
DIMENSION 3:		SCALE
	LOG	Ss record their uncertainty judgments on a predawn logarithmically spaced scale
	VERBAL	Ss verbally state their uncertainty judgments, or write them down in a blank on a form, or use a nonlogarithmic scale
DIMENSION 4:		ADDITIONAL FEEDBACK
	NONE	No additional feedback
	G-C-P (Graphic-Cumulative-Probability)	The additional feedback is a bar graph display of the probabilities of the hypotheses under consideration implied by the uncertainty judgments both for the current datum and for all previous data
	G-N-P (Graphic-Noncumulative-Probability)	The additional feedback is a bar graph display of the probabilities of the hypotheses under consideration implied by the uncertainty judgments for the current datum only
	V-C-P (Verbal-Cumulative-Probability)	The additional feedback is a nongraphic representation of the probabilities of the hypotheses under consideration implied by the uncertainty judgments both for the current datum and for all previous data
	V-C-O (Verbal-Cumulative-Odds)	The additional feedback is a nongraphic representation of the odds implied by the uncertainty judgments both for the current datum and for all previous data

first, S marks his ODDS assessments on a logarithmically spaced scale and he receives no additional feedback. In the second, S tells his LR assessments to E who first records them and then gives back to S the current odds based on S's assessments for all data. These two situations would be classified in the following way:

DIMENSION	ODDS ASSESSMENT	LR ASSESSMENT
1: Response Mode	ODDS	LR
2: Aggregation	CUM	NONCUM
3: Scale	LOG	VERBAL
4: Additional Feedback	NONE	V-C-O

The dimensions in this taxonomy, in addition to providing a framework for the classification of different response situations, capture the essence of a comparison of the Bayesian information processing approach with other approaches. Thus the distinctive features of the Bayesian approach provide the pattern which has guided the experiments on different direct estimation procedures, as well as the pattern that has guided other experimentation in this area.

The Bayesian approach to information processing has two distinctive features. First, this approach assumes that information processing is a 'divide and conquer' process. In other words, it assumes that the total assessment job should be broken down into smaller subtasks whereby the impact of each datum is assessed separately and the individual assessments are aggregated together mechanically. The proponents of the Bayesian viewpoint believe that this division of the whole inference task into smaller units not only makes the inference task easier, but also that it makes the final inference more accurate because S can make better assessments of the subunits. Thus the

'divide and conquer' label. The second feature of a Bayesian information processing approach is that change of opinion is additive on a logarithmic scale. This feature is in a sense the heart of Bayes's Theorem. The implication of this feature is that revision of opinion can be graphically represented by recording the series of likelihood assessments of a set of data on a logarithmically spaced scale of odds or probability. In fact by giving S a running record of the impact of his assessments on a logarithmically spaced odds or probability scale, one displays Bayes's Theorem to him.

If no differences result when Ss respond in LRs rather than odds, or when Ss respond by using nonaggregated assessments rather than aggregated assessments, or when Ss record their assessments on a log scale rather than a nonlogarithmic device, or when Ss do not receive Bayes's Theorem transformations of their assessments into some measure of the current likelihood of the hypotheses rather than receiving this feedback, then the formal introduction of Bayes's Theorem would serve no purpose for these Ss. These Ss would already by responding as Bayes's Theorem would predict. If, however, any of these experimental manipulations do result in assessment differences, then the formal introduction of Bayes's Theorem into the system would make a difference.

THE EXPERIMENTS

All five experiments that will be considered in this investigation had the common purpose of studying the effects of different direct estimation procedures for eliciting uncertainty judgments. Two of the studies were part of the large scale simulated strategic war setting experiments conducted by Ward Edwards and his colleagues at The University of Michigan to test the Bayesian information processing ideas. The simulation environment for these two studies was a simplified world ten years into the future. In this world only six nations played significant political and military roles. These were China, Japan, North America, Russia, United Arab Republic (a territory reaching from the Atlantic to India dominated by a prophet who sparked a Moslem revival), and the United Confederation of European States (a loose economic and military confederation). A 27-page summary of the history of the world gave Ss the background information they needed in order to become information processors in this future world. The history was designed to make different strategic war hypotheses, e.g., 'Russia and China are about to attack North America,' as well as a 'Peace will continue to prevail' hypothesis plausible. The list of hypotheses included four specific possible wars, "some other major conflict is about to break out," and the peace hypothesis. The Ss assumed the role of duty operators for the April 5th, 5PM to 6PM shift. These duty operators were part of the information processing system that served the Joint Chiefs of Staff. These processors were to assume that they were located in the basement of the Pentagon.

Three sensors delivered data to this information processing system. These

were the Ballistic Missile Early Warning System (BMEWS), the intelligence system, and a photo-reconnaissance satellite system. BMEWS, a large computerized radar system with three sites, was a degraded version of the present operational BMEWS system. The intelligence system was assumed to consist of spies, military attaches in U.S. embassies abroad, readers of foreign newspapers, and experts on foreign affairs. Each intelligence datum was a report of an event usually accompanied by a brief qualitative statement about the degree to which the event was surprising and what it meant. The photo-reconnaissance satellite system was assumed to consist of 20 satellites. The information processors received satellite system data that consisted of reports of particular events along with background information of the kind that might be obtained by comparing recent photographs with previous ones. A typical satellite system report was the following:

> "At 0630 this morning, two squadrons of conventional submarines sailed from Vladivostok. They steamed in a southerly direction until they were clear of the harbor, and then submerged. Evaluation: probably a routine exercise although this is an unusually large force."

At each session Ss assessed a set of five independent LRs for each of the 60 data comprising a scenario. Altogether there were nine scenarios.

The results of one of these experiments was published in 1968 (Edwards, Phillips, Hays & Goodman). This experiment, called PIPID, compared the effects of two kinds of additional feedback, G-N-P and G-C-P, on the performance of Ss making LR assessments on a log scale. Subjects recorded their assessments on a set of five scales, each having a lever mechanism that slid along

the scale. A set of scales consisted of six different ranges of logarithmically marked divisions. The S could turn a knob to select any one of the six different ranges. The first range extended from 1:1 to 10:1, the second extended from 10:1 to 100:1, and so on to 1,000,000:1. In front of Ss in both groups was a cathode ray tube (CRT) placed above the levers. The Ss in the group receiving the G-N-P additional feedback saw displayed on the CRT a bar graph representation of the posterior probabilities of the hypotheses based upon their current assessments only and equal prior probabilities. This transformation of the current set of LR judgments changed dynamically as an S either moved the levers, or reset a switch that indicated under which hypothesis the datum was more likely, or turned the knob that changed the scale range. The Ss in the group receiving the G-C-P additional feedback saw on the CRT a display of the posterior probabilities of the hypothesis based on their current LR assessments, all assessments for the previous data in that scenario, and a prior probability distribution of .10 for each of the war hypotheses and .50 for the peace hypothesis. This transformation of the current set of judgments, of all previous judgments, and of the unequal prior distribution also changed dynamically as S either moved the levers, or reset the switch, or turned the knob changing the scale range.

This experiment used a between-S design. Each of the 11 Ss, five in the G-N-P group and six in the G-C-P group, was trained, was run individually, completed one scenario per session, and had no more than one session per day.

According to the framework of this investigation, the two groups of Ss in this particular experiment are classified in the following way:

DIMENSION	Ss RECEIVING G-N-P	Ss RECEIVING G-C-P
1: Response Mode	LR	LR
2: Aggregation	NONCUM	NONCUM
3: Scale	LOG	LOG
4: Additional Feedback	G-N-P	G-C-P

In the other strategic war simulation experiment, called CORNOC, three different response situations were compared in a between-S design. One group of Ss, the DIS (display) group, had exactly the same response task as the group of Ss in the PIPID experiment that received the G-N-P additional feedback. Each S in this group assessed five LRs per datum for each of 60 data in the same nine scenarios. An S recorded his assessments on the same five sets of logarithmically marked scales. The display on the CRT fed back to S the posterior probabilities of the hypotheses under consideration based upon the LR assessments for the current datum and upon an equal prior distribution. Moreover, this display also changed dynamically as an S moved any one of the levers, or reset any of the switches, or turned any of the knobs to change a scale range. A second group of Ss, called the LEV (lever) group, had the same task as the DIS group except that there was no display on the CRT. Consequently, this group of Ss had no additional feedback. Each S recorded his 5 LRs per datum on the logarithmically spaced scales and then went on to assess the next datum in the scenario. The third group of Ss, called the NOC (no computer) group, did not have any aids. Each S in this group recorded the 5 LR assessments per datum on a single sheet of paper, in appropriately labeled blank spaces, and then began a new sheet for the next datum. The same nine scenarios used for the other groups of Ss were used also for this group.

Twenty one male University of Michigan students served as Ss. There were 7 Ss in the DIS group, 6 Ss in the LEV group, and 8 Ss in the NOC group. Each

S was trained for approximately eight hours on the 'future world's' history and political environment and eight hours on how to do LR assessment. Each S was run individually, completed one scenario per session and had one session per day. All 21 Ss repeated at least one scenario and 19 of these Ss repeated four scenarios.

According to the taxonomy introduced, the three groups of Ss in this experiment are classified as follows:

DIMENSION	DIS Ss	LEV Ss	NOC Ss
1: Response Mode	LR	LR	LR
2: Aggregation	NONCUM	NONCUM	NONCUM
3: Scale	LOG	LOG	VERBAL
4: Additional Feedback	G-N-P	NONE	NONE

The third experiment in this investigation, labeled W&E, was done by Gloria Wheeler and Ward Edwards (in preparation). This experiment used a within-S design to investigate the effects of four different response situations. These situations were NONCUM LR, CUM LR, NONCUM ODDS, and CUM ODDS. The Ss in this experiment were assigned to one of four groups. Each group made the four different types of assessment in a different sequential order.

The stimuli in this experiment were 7" sticks painted with a blue part and a yellow part. The blue and yellow occurred in various proportions in different sticks. Two populations were defined. The populations were piles of sticks with the amount of blue and yellow paint as the random variable. Each of the populations had Gaussian (normal) distributions. One population had a mean length of blue of 4.5", the other had a mean length of blue of 2.5". Each population had a standard deviation of 1.25" of blue.

Let d' be defined as $m_1 - m_2/s$, where m_1 is the mean of one population,

m_2 is the mean of the second population, and s is the standard deviation of both populations. For this experiment d' = 1.6, which yields a veridical LR at the mean of either population of 5.00.

Eight sequences of ten sticks each were determined by drawings from a table of random normal deviates. Some of these sequences were then slightly modified so that there was a fairly large range in veridical final posterior odds and so that they looked random. From these eight 'random normal deviate' sequences, 16 physically different sequences were constructed, eight from the predominantly blue population and eight from the predominantly yellow population. Every S saw every sequence twice, for a total of 32 sequences. Each S for each different response situation saw eight sequences. The 32 sequences were ordered randomly and that same order was used throughout the whole experiment for every S, regardless of the order in which he made the different response assessments.

The population characteristics were displayed to the Ss by two charts. To prepare the charts, the cumulative normal distribution for each population was divided into 100 equally likely parts. The mean lengths of blue at 97 of the boundary points, randomly arranged, comprised the charts.

Responses were made in 10-page booklets, one response per page. On each page was printed, "_____:1 in favor of hypothesis _____." The Ss were briefly instructed in each new response situation prior to beginning that type of response. Thirty-six male University students, run individually or in pairs, served as Ss for this experiment.

The four different response situations in this experiment are classified in the following way:

DIMENSION					
1: Response Mode	LR	LR	ODDS	ODDS	
2: Aggregation	NONCUM	CUM	NONCUM	CUM	
3: Scale	VERBAL	VERBAL	VERBAL	VERBAL	
4: Additional Feedback	NONE	NONE	NONE	NONE	

The W&E experiment is the only experiment that I know about that studied a CUM LR response. Part of the instruction to each S was a written specification of the task on a "Cumulative LR Instruction Sheet." To illustrate the nature of the inference being asked for, I quote the following from this sheet:

> "...We will proceed as follows. First I will show you a single stick, and you will estimate a likelihood ratio for that stick. Then I will show you another stick, and I want you to evaluate the likelihood ratio of both sticks. Formally, I am asking what is the likelihood ratio for two sticks. Forget that you saw the first stick by itself, and forget the estimate you made. Pretend that I drew the two sticks simultaneously from somewhere...Which pile of sticks is more likely to produce those two sticks?..."

The fourth experiment in this investigation (Domas, Goodman & Peterson, 1972), examined the effects of six different response situations. This experiment, called D,G&P, used a between-S design with each S making only one type of response for all data. The six response conditions, classified according to the introductory framework, are as follows:

DIMENSION						
1: Response Mode	LR	LR	LR	ODDS	ODDS	ODDS
2: Aggregation	NONCUM	NONCUM	NONCUM	CUM	CUM	CUM
3: Scale	VERBAL	VERBAL	LOG	VERBAL	LOG	LOG
4: Additional Feedback	NONE	V-C-O	NONE	NONE	NONE	V-C-P

The stimulus environment simulated a commercial shipping problem. Each S was asked to assume the role of an analyst employed by a U.S. commercial shipping firm. The S's task was to predict whether competitor ships were destined for Port A or Port B. Port A was more distant than Port B and involved a long open-water voyage.

Date concerning the competitor ships fell into one of four categories: age of ship, gross tonnage, percent of capacity cargo load on board, and fuel taken on at the port of departure. Data samples were generated by drawing from pairs of Gaussian (normal) populations, each characterized by a specified d' level. For example, data about the age of competitor ships were drawn from two normal distributions with a d' of 0.40. The other d' levels, 0.80, 1.01, and 1.14, were associated with gross tonnage, percent of capacity cargo, and fuel taken on, respectively.

Twenty-six sequences, each consisting of four data, were generated randomly from the distributions associated with the four categories of data. Every sequence contained a datum from each of the four categories. Thirteen of the 26 sequences were drawn from distributions favoring Port A, the remaining 13 from distributions favoring Port B.

The population characteristics were displayed to the Ss by four sets of two charts. Each chart was a randomly arranged representative sample of the appropriate population. On a chart, each datum was a 7" vertical line 1/8" wide, partly black and partly white. The random variable had two interpretations. It was a specific scale value, e.g., 60% capacity cargo, or a length of black, e.g., 4.11".

The charts, eight in all, were arranged on a display board as follows:

all data about ships going to Port A were arranged from top to bottom on the left side. Similarly, all data about ships destined for Port B were arranged in a corresponding fashion, opposite Port A. There was a sliding scale between each of the four pairs of charts. The E could vary the length of a random sample from zero to 7" simply by moving the slider. To the S the datum could be interpreted as a length, relative to the other lengths of black on the charts, or as a number, representing the value of the data item in question.

There were three different sets of response sheets. For Ss making LR assessments on a verbal scale, each response sheet read as follows:

> "It is _____ times more likly that this datum would occur if the ship were going to Port _____ rather than to the other Port."

The Ss recorded their assessments for each datum on a separate page. Each response sheet for Ss making odds assessments read as follows:

> "Port _____ is favored. The odds favoring this Port are _____:1."

Here again Ss made written assessments, one per page. The log scale response sheet was a 17" by 22" sheet of paper with three logarithmically spaced scales from 1:1 to 1000:1 on the top half and symmetric values from 1:1 to 1:1000 on the bottom half. This sheet was further divided by 30 vertical lines with each line having the same logarithmic spacing. The Ss used one sheet for each sequence of four data, allowing one vertical line per datum.

Forty, University of Michigan, male students served as Ss. Ten Ss were run in each of two conditions, NONCUM LOG LR with no additional feedback and CUM LOG ODDS with V-C-P additional feedback. Five Ss took part in each of the

remaining four conditions. Each S was run individually.

The last experiment in this investigation, called S,P&M, was performed by Kurt Snapper, Cameron Peterson & Allan Murphy. The two Ss in this experiment were University of Michigan faculty members who had been weather forecasters. Their task was to assess the precipitation probability forecast for each of 25 sequences of historical data. Each sequence contained one sample of each of nine different sources of information, presented in chart form. The sources selected were the most important items of information that would be received in a forecast period. Moreover, these data appeared in the order that they would be received at a U.S. Weather Bureau Station. These sources included such items as isoprobability curves, upper and lower atmospheric charts, and barometric charts. Because these sources of information are conditionally dependent sources given the precipitation, no precipitation hypotheses each S's task was to estimate conditionally dependent probabilities. Their task, the task of assessing the precipitation probability forecast more strictly defined, was the task of assessing the probability of at least .01" of precipitation within the forecast period.

This experiment used a within-S design. The two Ss assessed the precipitation probability forecast in two different ways. However, each time they were given the same 25 sequences of nine data.

In one situation S was asked for a cumulative conditional probability judgment. In other words, he was asked for his present assessment of the probability of precipitation given his previous probability assessment and the new information just presented in the current datum. At the beginning of a new sequence, S was to assume that the precipitation probability was .50.

"The precipitation probability forecast, based on Chart #102 is _____.

The revised probability forecast, based on Chart #'s 8 and 9 is _____.

The revised probability forecast, based on Chart #12 is _____.

⋮

(and finally)

The revised probability forecast, based on Chart #41 is _____."

In the other response situation S was asked for a noncumulative conditional probability assessment. In this situation, for each assessment S was to assume that all the previous data had led him to a probability assessment of .50. His task was to revise this probability on the basis of the new information presented to him in the current datum. In this condition S filled out nine sheets of paper for each sequence. The first sheet read as follows:

"The forecast, based only on the information in Chart #102 is _____."

The second sheet read in the following way:

"Assume that the previous precipitation probability forecast, based on all previous information was .50.

The revised forecast, based on only the new information in Chart #'s 8 and 9 is _____."

The ninth sheet read as follows:

"Assume that the previous precipitation probability forecast, based on all previous information was .50.

The revised forecast, based on only the new information in Chart #41 is _____."

The S received no additional feedback in either response condition. Each S was run individually.

According to the framework of this investigation, the two response situations in this experiment are classified in the following way:

DIMENSION			
1:	Response Mode	PROB	PROB
2:	Aggregation	CUM	NONCUM
3:	Scale	VERBAL	VERBAL
4:	Additional Feedback	NONE	NONE

RESULTS

The following discussion uses two conventions; both are typically also embodied in the experiments being reanalyzed and in the original reports of them. One, only a convention, is that odds are expressed as numbers of the form X:1, where $X \geq 1$. This can always be accomplished by appropriate choice of which hypothesis enters into the numerator and which into the denominator of the odds ratio. A similar convention applies to likelihood ratios, except in contexts in which consistency with the preceding convention for odds requires expressing the likelihood ratio as a number less than 1. The second convention is that the conclusions are expressed as though the prior odds before the first datum in each sequence were 1:1. In most of the experiments here reanalyzed, that was true. Where it was not, appropriate attention was paid to the question during the data analysis.

In general, the following data analyses were done on final log odds, regardless of the response mode actually used by Ss. (An exception is noted in the text.) "Final" means that the response is the last odds estimate or other estimate of a cumulative quantity associated with a sequence of data that led to successive revisions of that quantity, if a cumulative response mode was used, or was the appropriate corresponding number calculated from a sequence of noncumulative estimates, if a noncumulative response mode was used. So a statement such as "noncumulative LR responses are larger than cumulative odds responses" translates into "final log odds calculated from noncumulative LR responses are more extreme, in the appropriate direction, than the logarithms of final cumulative odds responses."

For this investigation one group of responses is considered significantly larger than another group of responses when the following two conditions are met: (1) the correlation coefficient between the two sets of assessments is high enough so that it can be assumed that the two groups of responses are linearly related (high enough is arbitrarily defined as greater than .900); and (2) and 95% confidence interval for $\hat{\beta}$, defined as the estimated slope of the linear structural relation between these two random variables, does not contain the point 1.000.

There is a basic difference between a linear structural relation analysis and a linear regression analysis (see Isaac, 1970, and Kendall & Stuart, 1961). The purpose of a linear regression analysis is to find the parameters, α and β, of the line that best predicts the values of the dependent variable given the values of the independent variable. It is assumed that the dependent variable is a random variable and that the independent variable is fixed and measured without error. The purpose of a linear structural relation analysis, however, is to determine the interrelationship between two variables when both are random and when either or both of them are measured with error. The interrelationship is measured by the slope, β, of the straight line relating one dependent variable to the other.

The model for a linear structural relation is given by the following formula:

$$y = \alpha + \beta(x-\epsilon_x) + \epsilon_y , \qquad (1)$$

where ϵ_x is the random error in measuring x and ϵ_y is the random error in measuring y.

When λ = variance of ϵ_y/variance of ϵ_x is known, the formula for measuring $\hat{\beta}$, as given by Isaac (1970) is as follows:

$$\hat{\beta} = \frac{\text{var } y - \lambda \text{ var } x + [(\text{var } y - \lambda \text{ var } x)^2 + 4\lambda \text{ Cov}^2(x,y)]^{1/2}}{2 \text{ Cov}(x,y)} \quad (2)$$

The confidence limits for $\hat{\beta}$ are given by the following formula taken from Kendall & Stuart (1961):

$$\tan\left\{\hat{\theta} \pm \frac{1}{2} \arcsin\left[2t\left\{\frac{\text{var } x \cdot \text{var } y - \text{Cov}^2(x,y)}{(n-2)[(\text{var } x - \text{var } y)^2 + 4 \text{ Cov}^2(x,y)]}\right\}^{1/2}\right]\right\} \quad (3)$$

where $\hat{\theta}$ = arc tan $\hat{\beta}$ and t is the appropriate "Student's" deviate for (n-2) degrees of freedom and the confidence level chosen.

In order to calculate $\hat{\beta}$ and the confidence interval for $\hat{\beta}$, λ, the ratio of the error variances, must be determined. The error variances were estimated differently in the different experiments. In the first experiment, PIPID, no scenarios were repated. Therefore, the slope $\hat{\beta}$ for the two groups in this study was estimated for two values of λ, $\lambda = 1$ and $\lambda = 0$. These values of λ are extreme values given the λ values estimated in the other experiments where reliability information was available. (Here and in the Snapper, Peterson & Murphy experiment discussed below, the choice of which dimension to call x and which to call y was made to keep $\lambda < 1$.)

In the second experiment, CORNOC, the error variances were estimated from repeated data values. All of the 21 Ss in CORNOC repeated at least one of the nine scenarios and 19 of these Ss repeated four scenarios. Since there were five final odds per scenario and four scenarios with repeat data, there were 20 points for which I calculated both the first round average group final log

odds assessments and the second round average group final log odds assessments. The error variances were then estimated to be the variances of the difference between the average first round and second round final log odds assessments. The necessary λ values were then calculated by taking the appropriate ratios of these error variance estimates.

All Ss in the Wheeler & Edwards experiment repeated eight of the 16 scenarios. Each S repeated two scenarios in each of the four different response conditions. The experimental design is such that for each scenario under each response condition, there are from 7 to 10 persons with repeat data. The estimated error variance for each response condition is the variance of the difference between the first and second round average final log odds measurements for the eight repeated scenarios.

In the Domas, Goodman & Peterson experiment, no scenarios were repeated. However, of the 104 stimuli, 13 pairs (26 stimuli) were such that the true LR of one member of the pair was within .01 of the true value of the other member. These 26 data or 13 pairs were used to estimate error variances for each group of Ss. The first occurrence of one of the pair was considered the first round estimate, while the occurrence of the second member of this pair was considered the second round estimate. Each of the first and second round assessments was then averaged across the Ss within a group. The error variance for each of the six groups of Ss was estimated to be the variance of the difference between what is considered the first and second round average log LR assessments for the 13 pairs of data.

There were no repeat data in the Shapper, Peterson & Murphy experiment. Consequently, $\hat{\beta}$ was estimated twice, using the two λ values of one and zero.

The summary of the linear structural relation analysis statistics for the different groups of Ss in the different experiments, when only a single dimension was varied in the response conditions between the groups being compared, is contained in Table 2. The dependent variable for all comparisons is the average final log odds for each group. The final odds in each case was either estimated directly or calculated by means of Bayes's Theorem.

On the basis of the evaluation criteria initially established, i.e., $r > .900$ and the 95% confidence interval for $\hat{\beta}$ not containing 1.000, several conclusions can be drawn from this analysis. First, odds assessments are sometimes significantly larger than LR assessments. Verbal, cumulative odds estimates were significantly larger than verbal, cumulative LR estimates, when there was no additional feedback to either group. However, odds were not significantly larger than LRs when both were verbal, noncumulative estimates with no additional feedback.

A second conclusions is that aggregation makes a difference for verbal responses with no additional feedback. Nonaggregated judgments, whether LRs or odds, resulted in significantly larger final posterior odds than did aggregated judgments. There is evidence that this same conclusion may hold for conditional probabilities. In the Shapper, Peterson & Murphy experiment, the data for each forecaster were analyzed separately. The $\hat{\beta}$ values were estimated for cumulative probability plotted on the y-axis and noncumulative probability plotted on the x-axis. For one forecaster $\hat{\beta}$ is .292 and $\lambda = 1$, and .639 when $\lambda = 0$. The 95% confidence interval for the larger value is (.460, .854). The slope for the other forecaster is .218 when $\lambda = 1$, and .396 when $\lambda = 0$. The 95% confidence interval around .396 is (.302, .501). These values were not

TABLE 2

LINEAR STRUCTURAL RELATION STATISTICS FOR GROUP COMPARISONS WHEN SINGLE DIMENSIONS VARIED

Dimension Varied	Dimension Information					N: The Number of FINAL LOG ODDS	Number of Data Aggregated Per Point	Correlation Coefficient	$\hat{\beta}$	95% Confidence Interval for $\hat{\beta}$	y-axis	x-axis
	Experiment	Response Mode	Aggregation	Scale	Additional Feedback							
RESPONSE MODE.....	W&E	LR ODDS	NONCUM CUM	VERBAL	NONE	16	10	.970	1.068	(.926, 1.255)	NONCUM ODDS	NONCUM LR
	W&E	LR ODDS	CUM	VERBAL	NONE	16	10	.987	1.247	(1.138, 1.371)	CUM ODDS	CUM LR
AGGREGATION.....	W&E	LR	NONCUM CUM	VERBAL	NONE	16	10	.961	.293	(.239, .323)	CUM LR	NONCUM LR
	W&E	ODDS	NONCUM CUM	VERBAL	NONE	16	10	.994	.348	(.317, .380)	CUM ODDS	NONCUM ODDS
SCALE.....	CORNOC	LR	NONCUM	LOG VERBAL	NONE	45	60	.943	1.488	(1.337, 1.664)	LOG LR	VERBAL LR
	D,G&P	LR	NONCUM	LOG VERBAL	NONE	26	4	.984	1.176	(1.091, 1.266)	LOG LR	VERBAL LR
	D,G&P	ODDS	CUM	LOG VERBAL	NONE	26	4	.951	1.531	(1.337, 1.767)	LOG ODDS	VERBAL ODDS
ADDITIONAL FEEDBACK...	CORNOC	LR	NONCUM	LOG	G-N-P NONE	45	60	.921	.500	(.439, .566)	G-N-P	NONE
	D,G&P	LR	NONCUM	VERBAL	V-C-O NONE	26	4	.996	.944	(.886, 1.006)*	V-C-O	NONE
	D,G&P	ODDS	CUM	LOG	V-C-P NONE	26	4	.984	.669	(.620, .721)	V-C-P	NONE
	FIPID	LR	NONCUM	LOG	G-C-P G-N-P	45	60	.904	$\lambda = 1$:.693 $\lambda = 0$:.796	(.595, .799) (.689, .915)	G-C-P	G-N-P

*Analysis performed on log LR data yielded significant result in same direction.

put into Table 2 because the correlation coefficients, .629 and .727, for these two sets of assessments were considerably lower than the correlations for the other functions. However, this correlational analysis was done in probabilities and not log odds, thereby constraining the range of the scales. Moreover, each of these two analyses was based on individual data, not averaged data. Since the upper bound of the 95% confidence interval for $\hat{\beta}$ is less than one for both forecasters, the data strongly suggest that aggregation reduces the size of probability assessments. This issue needs to be tested in controlled experiments with many Ss, using data that would lead to final cumulative probabilities that span the range from .5 to at least .999. Covering this range densely enough to permit separate analyses of more and less extreme regions is important because many researchers believe that Ss don't know how to use the extreme regions of the probability scale correctly. Therefore, experimental tests that are based solely on responses appropriate to the ends of the scale may be susceptible to scale errors in addition to whatever assessment errors may exist. To guard against very nonhomogeneous 'random response error' in different parts of the probability scale, stimuli should be selected from all parts, and data analyses performed on subsets of the data.

A third conclusion is that the scale makes a significant difference when there is no additional feedback. Log scale responses were significantly larger than verbal scale responses for both NONCUM LRs and CUM ODDS, when there was no additional feedback.

A fourth conclusion from Table 2 is that additional feedback of the posterior probabilities of the hypotheses under consideration makes a significant difference regardless of whether that feedback is just for the current datum

or for all previous data. There is also some evidence that cumulative odds feedback may make a significant difference for NONCUM LR responses on a verbal scale, even though the 95% confidence interval for $\hat{\beta}$ contains 1.000. In the Domas, Goodman & Peterson experiment, in addition to doing the linear structural relation analyses on the 26 average final log odds for each group, I also did these same analyses using the 104 average log LRs as points. The conclusions to be drawn in Table 2 from the average log LR analyses were the same as from the average final log odds analyses except for the comparison of NONCUM LR assessment on a verbal scale with and without V-C-O additional feedback. In this case, the log LR analysis resulted in a significant difference between the two groups. The group without the additional feedback made larger estimates than the group with the V-C-O additional feedback. Here is another hypothesis that can be put to experimental test. But this hypothesis warrants further testing not only because the different dependent variables resulted in different levels of significance, but also because the slopes were so very close to one, .944 for the final log odds analysis and .907 for the log LR analysis.

The previous conclusions were based on comparisons between group estimates without considering optimality. For all the data in both the W&E and D,G&P experiments, the true value of each LR was known. By applying Bayes's Theorem, the final odds was calculated for each sequence of data in both experiments. Linear regression analyses were performed whereby the average group final log odds for all sequences of data were compared with the Bayesian final log odds. Given high correlation coefficients and intercepts close to zero, the closer a particular regression slope is to one, the more optimal is the group's

estimates. The summary of these linear regression analyses is found in Table 3. However, since no statistical tests were performed on the differences between these slopes, we cannot conclude that any one slope is significantly more accurate than any other slope.

In the W&E experiment all four groups used verbal scales and had no additional feedback. Under these conditions, the cumulative odds assessments were significantly larger than the cumulative LR assessments, and the slope for the cumulative odds versus Bayesian odds regression was .558 as contrasted with a slope of .283 for the cumulative LR group. Therefore, we don't know whether there is a tendency for cumulative odds to be larger or more accurate than cumulative LRs. The hypothesis that needs to be tested is that cumulative odds assessments are more accurate than cumulative LR assessments. A test of this kind requires an experimental situation where truth is known, where the veridical cumulative odds judgments run the range of values from 1:1 to very large numbers of at least 1000:1. Furthermore, separate data analyses should be done over different intervals of the range.

In this same experiment the noncumulative odds judgments were larger, but not significantly larger, than the noncumulative LR judgments. However, the slope for the noncumulative odds group in the Bayesian regression was farther from one, 1.166, than the slope, 1.041, for the noncumulative LR group. Consequently, for the nonaggregated condition, the data suggest that while odds assessments may be larger than LR assessments, they may also be less accurate. This hypothesis needs to be tested using situations where truth is known, where the veridical noncumulative likelihood judgments run over a large range of values and where separate data analyses are done over different portions of the range.

TABLE 3

LINEAR REGRESSION STATISTICS FOR GROUP FINAL LOG ODDS,
AS DEPENDENT VARIABLE,
COMPARED WITH BAYESIAN FINAL LOG ODDS

Experiment	Dimension Information				Correlation Coefficient	Regression Slope
	Response Mode	Aggregation	Scale	Additional Feedback		
W&E	LR	NONCUM	VERBAL	NONE	.980	1.041
	LR	CUM	VERBAL	NONE	.910	.283
	ODDS	NONCUM	VERBAL	NONE	.991	1.166
	ODDS	CUM	VERBAL	NONE	.957	.358
D,G&P	LR	NONCUM	LOG	NONE	.980	1.463
	LR	NONCUM	VERBAL	NONE	.983	1.176
	ODDS	CUM	LOG	NONE	.980	1.304
	ODDS	CUM	VERBAL	NONE	.960	.845
	LR	NONCUM	VERBAL	V-C-O	.986	1.116
	ODDS	CUM	LOG	V-C-P	.985	.876

In the D,G&P experiment accuracy and size comparisons can be made for the scale dimension. Assessments made on a log scale were significantly larger than assessments made on a verbal scale for both the noncumulative LR and cumulative odds response groups that received no additional feedback. Moreover, these log scale assessments resulted in slopes further away from 1.000 in the Bayesian regression. These results suggest that assessments made on a log scale may be larger than assessments made on verbal scale, irrespective of accuracy, when there is no additional feedback. This hypothesis needs further testing in situations where truth is known, where the true values extend over a wide range, and where separate data analyses are done over different parts of this range.

Data from the D,G&P experiment also suggest that while cumulative odds assessments on a log scale with no additional feedback are significantly larger than this same kind of assessment with V-C-P additional feedback, they may also be less accurate. Thus cumulative posterior probability additional feedback may increase accuracy, at least for an aggregated odds response made on a log scale. This hypothesis needs to be tested.

While linear structural relation analyses comparing just those groups where a single dimension in the response condition is varied at any one time identifies those dimensions that affect the size of assessments, such analyses do not give any information about the relative strengths of the effects of response conditions on response magnitudes. To determine an exact ordering of the different response conditions based either on the size or accuracy of the assessments would require a very large factorial experiment—which has not been done. However, it is possible to get some information about an ordering

of the different response conditions based on size of assessments by doing linear structural relation analyses on groups in which two or more response dimensions vary simultaneously. The statistics summarizing certain of these linear structural relation analyses are presented in Table 4.

An order relation, >, is defined as follows over two sets of assessments, A and B, where the A values are plotted on the y-axis and the B values on the x-axis. First, A > B, if $\hat{\beta}$ in the structural relation analysis is greater than one and if the 95% confidence interval for $\hat{\beta}$ does not span 1.000. Second, A > B, if several different structural analyses compare A and B, A > B by assumption 1 for at least one comparison, and not B > A for any of the others. If not A > B and not B > A, then A ? B.

The simultaneous varying of response mode and aggregation in the W&E experiment provides much information. From the single dimensional analyses of these data we know that ODDS > LRs and that NONCUM responses > CUM responses. These inequalities suggest one of the two following orderings:

NONCUM ODDS > NONCUM LR > CUM ODDS > CUM LR

or

NONCUM ODDS > CUM ODDS > NONCUM LR > CUM LR .

To determine the exact ordering of these four groups, the following six comparisons are necessary: NONCUM ODDS & NONCUM LR, NONCUM ODDS & CUM ODDS, NONCUM ODDS & CUM LR, NONCUM LR & CUM ODDS, NONCUM LR & CUM LR, and CUM ODDS & CUM LR. The single dimensional analyses show that NONCUM ODDS ? NONCUM LR, NONCUM ODDS > CUM ODDS, NONCUM LR > CUM LR and CUM ODDS > CUM LR. Varying these two dimensions simultaneously results in NONCUM LR > CUM ODDS and

TABLE 4

LINEAR STRUCTURAL RELATION STATISTICS FOR GROUP COMPARISONS WHEN MULTIPLE DIMENSIONS VARIED

Dimensions Varied	Dimension Information: Experiment	Response Mode	Aggregation	Scale	Additional Feedback	N: The Number of FINAL LOG ODDS	Number of Data Aggregated Per Point	Correlation Coefficient	$\hat{\beta}$	95% Confidence Interval for β	y-axis	x-axis
RESPONSE MODE + AGGREGATION	D,G&P	LR / ODDS	NONCUM / CUM	VERBAL	NONE	26	4	.962	.727	(.642, .819)	CUM ODDS	NONCUM LR
	W&E	LR / ODDS	NONCUM / CUM	VERBAL	NONE	16	10	.976	.369	(.323, .416)	CUM ODDS	NONCUM LR
	W&E	LR / ODDS	CUM / NONCUM	VERBAL	NONE	16	10	.989	.278	(.256, .301)	CUM LR	NONCUM ODDS
SCALE + ADDITIONAL FEEDBACK	CORNOC	LR	NONCUM	VERBAL / LOG	NONE / G-N-P	45	60	.955	.733	(.663, .808)	LOG G-N-P	VERBAL NONE
	D,G&P	LR	NONCUM	VERBAL / LOG	V-C-O / NONE	26	4	.983	1.238	(1.142, 1.347)	LOG NONE	VERBAL V-C-O
	D,G&P	ODDS	CUM	VERBAL / LOG	NONE / V-C-P	26	4	.970	1.020	(.916, 1.138)*	LOG V-C-P	VERBAL NONE
ALL FOUR DIMENSIONS	D,G&P	LR / ODDS	NONCUM / CUM	VERBAL / LOG	V-C-O / NONE	26	4	.981	1.171	(1.076, 1.275)	LOG ODDS	VERBAL LR
	D,G&P	LR / ODDS	NONCUM / CUM	VERBAL / LOG	V-C-O / V-C-P	26	4	.994	.785	(.748, .821)	LOG ODDS	VERBAL LR

*Analysis performed on log LR data yielded significant result in opposite direction.

NONCUM ODDS > CUM LR. Therefore, the ordering that emerges from the W&E experiment is NONCUM ODDS > NONCUM LR; NONCUM ODDS,NONCUM LR > CUM ODDS > CUM LR. This ordering implies that the aggregation dimension is more important than the response mode dimension in determining the size of assessments.

In the D,G&P experiment these same two dimensions were simultaneously varied. When the scale and additional feedback conditions were the same as in the W&E experiment, i.e., verbal scale and no additional feedback, the same interactive ordering emerged, NONCUM LR > CUM ODDS. When the log scale was used as a recording mechanism instead of the verbal scale, the ordering was NONCUM LR ? CUM ODDS with final log odds as the dependent variable in the structural analysis and NONCUM LR > CUM ODDS with log LRs as the dependent variable. Thus both experiments confirm the finding that aggregation is more important than response mode in determining the size of assessments.

The scale and additional feedback dimensions were simultaneously varied in both the CORNOC and D,G&P experiments. The single dimensional structural analyses of the CORNOC experiment yielded these inequalities: LOG > VERBAL and NONE > G-N-P for NONCUM LR responses. These two inequalities imply one of the two following orderings:

LOG,NONE > LOG,G-N-P > VERBAL,NONE > VERBAL,G-N-P

or

LOG,NONE > VERBAL,NONE > LOG,G-N-P > VERBAL,G-N-P .

To completely determine the correct ordering six comparisons are necessary. The single dimensional analyses yielded LOG,NONE > VERBAL,NONE and LOG,NONE > LOG,G-N-P. The analysis in Table 4 showed that VERBAL,NONE > LOG,G-N-P.

Without all six comparisons the ordering cannot be completely determined. However, the ordering suggested by the data is as follows:

LOG,NONE > VERBAL,NONE > LOG,G-N-P > VERBAL,G-N-P .

This ordering implies that the additional feedback of the posterior probabilities over the hypotheses, even for a single datum assessment, may be more important than the log scale in determining the size of noncumulative LR estimates. This hypothesis needs testing.

The single dimensional analyses of the D,G&P experiment resulted in the following inequalities: LOG,NONE > VERBAL,NONE for both CUM ODDS and NONCUM LR; LOG,NONE > LOG,V-C-P for CUM ODDS; and VERBAL,NONE > VERBAL,V-C-O for NONCUM LR. The analyses in which two response dimensions were varied simultaneously yielded LOG,NONE > VERBAL,V-C-O for NONCUM LRs and ambiguous results for the LOG,V-C-P and VERBAL,NONE comparison for CUM ODDS. The structural relation analysis of average final log odds yielded a $\hat{\beta}$ of 1.020 with a 95% confidence interval of (.916, 1.138) when LOG,V-C-P is plotted on the y-axis and VERBAL,NONE on the x-axis. The structural relation analysis of average log LR yielded a $\hat{\beta}$ of .882 with a 95% confidence interval of (.790, .983) for this same comparison. These findings for the D,G&P experiment show that the LOG, NONE condition yielded the largest estimates of all and that the log scale is more important than V-C-O additional feedback in determining the size of NONCUM LR responses.

The CORNOC and D,G&P experiments together imply that when considering just the log scale and additional feedback dimensions, one is not more important than the other. The order of these two dimensions depends upon the type

of feedback given. The results were that odds additional feedback is not more important than the log scale in determining the size of LR estimates, but that probability feedback may be more important than the log scale in determining the size of CUM ODDS estimates. This latter assertion is a hypothesis in need of a test.

Simultaneously varying all four dimensions in the D,G&P experiment yielded some interesting information. The ODDS,CUM,LOG,NONE response condition resulted in larger estimates than the LR,NONCUM,VERBAL,V-C-O response condition. The LR,NONCUM,VERBAL,V-C-O response condition resulted in larger estimates than the ODDS,CUM,LOG,V-C-P response condition. If we are willing to assume transitivity, then these two results together, ODDS,CUM,LOG,NONE > LR,NONCUM,VERBAL,V-C-O > ODDS,CUM,LOG,V-C-P, imply that probability feedback is more important than the combined effects of the response mode, aggregation, and scale dimensions.

Reliability information was used in the linear structural relation analyses to estimate error variances. However, the reliability of LR and ODDS estimates is a topic worthy of independent consideration. Some researchers have collected repeat data on Ss who performed in Bayesian information processing experiments. No one, to my knowledge, has reported any of it, however. The data in Table 5 fill this void somewhat.

I analyzed the repeat data for Ss in the CORNOC, W&E, and D,G&P experiments. From each experiment the particular data points that are used for these analyses are the same data points that were used to determine the error variances for the linear structural analyses. The different statistics incorporated into Table 5 are the following: (1) the difference between the first

TABLE 5

RELIABILITY INFORMATION

Experiment	Dimension Information: Response Mode	Aggregation	Scale	Additional Feedback	Dependent Variable	N: The Number of FINAL LOG ODDS or LOG LR	Number of Data Aggregated Per Point	Number of Ss Per Group	$\overline{x_1-x_2}$	$100\% \times \frac{\lvert\overline{x_1-x_2}\rvert}{\text{Range (1)}}$	$100\% \times \frac{\lvert\overline{x_1-x_2}\rvert}{\text{Range (2)}}$	$\overline{s^2_{x_1-x_2}}$	r
CORNOC	LR	NONCUM	VERBAL	NONE	FINAL LOG ODDS	20	60	8	-.027	1.1%	1.1%	.027	.971
	LR	NONCUM	LOG	NONE	FINAL LOG ODDS	20	60	6	-.139	4.6%	4.5%	.041	.923
	LR	NONCUM	LOG	G-N-P	FINAL LOG ODDS	20	60	7	.015	0.7%	0.9%	.018	.997
D,G&F	LR	NONCUM	VERBAL	NONE	LOG LR	13	1	5	.056	4.1%	4.1%	.020	.959
	LR	NONCUM	VERBAL	V-C-O	LOG LR	13	1	5	-.002	0.2%	0.2%	.013	.964
	LR	NONCUM	LOG	NONE	LOG LR	13	1	10	.027	1.9%	1.6%	.023	.973
	ODDS	CUM	VERBAL	NONE	LOG LR	13	1	5	-.098	8.3%	8.5%	.016	.662
	ODDS	CUM	LOG	NONE	LOG LR	13	1	5	-.038	2.3%	3.2%	.021	.885
	ODDS	CUM	LOG	V-C-P	LOG LR	13	1	10	.010	1.2%	1.2%	.007	.962
W&E	LR	NONCUM	VERBAL	NONE	FINAL LOG ODDS	8	10	7-10*	.170	4.9%	5.5%	.251	.687
	ODDS	NONCUM	VERBAL	NONE	FINAL LOG ODDS	8	10	8-10*	.113	2.9%	2.5%	.422	.890
	LR	CUM	VERBAL	NONE	FINAL LOG ODDS	8	10	8-10*	-.067	4.6%	5.3%	.042	.977
	ODDS	CUM	VERBAL	NONE	FINAL LOG ODDS	8	10	8-10*	-.119	15.7%	13.6%	.019	.657

*The exact number of Ss in the average depends upon the particular point.

round and second round assessments averaged over the number of points in the function; (2) the percentage of the absolute value of this average difference divided by the average range of assessments in the first round; (3) the percentage of the absolute value of the average difference divided by the average range of assessments in the second round; (4) the variance of the difference between the first and second round assessments averaged over the number of points in each function; and (5) the correlation coefficient of the function relating the first round assessments to the second round assessments.

By any criterion, most of the groups of Ss in all three experiments were very reliable.

DISCUSSION AND CONCLUSIONS

This study has attempted to organize the information available on different direct estimation procedures investigated in experiments on Bayesian information processing. A taxonomy classified different response situations on the basis of four independent dimensions—response mode, aggregation, scale, and additional feedback.

Linear structural analyses were performed in which the different response situations were compared. Two criteria were established, a correlation coefficient of greater than .900 and a 95% confidence interval about the slope not containing the point 1.000. When these criteria were met, the set of estimates made under one response condition was considered significantly larger than the set of estimates made in the other response condition.

On the basis of these criteria and the regression analyses performed when truth was known, the following results have been demonstrated:

1. The response mode sometimes makes a difference. CUM ODDS assessments were significantly larger than CUM LR assessments when both sets of estimates were recorded on verbal scales and there was no additional feedback.

2. Aggregation makes a difference. Nonaggregated LRs or ODDS were significantly larger than the judgments made in the corresponding aggregated conditions when the responses were made on verbal scales and there was no additional feedback. There is data to suggest the hypothesis that this finding may hold true for PROBs.

3. The scale used makes a difference. Log scale recording devices resulted in significantly larger assessments than verbal scale recording devices for both the NONCUM LR and CUM ODDS groups when there was no additional feedback.

4. Additional feedback of the posterior probabilities of the hypotheses does make a significant difference in the size of NONCUM LR and CUM ODDS judgments. There is some evidence to suggest the hypothesis that V-C-O additional feedback may make a difference for NONCUM LRs on a verbal scale.

5. The use of log recording scales results in significantly larger assessments than the use of other methods of recording estimates. It may be that this result holds regardless of accuracy. This hypothesis needs to be tested.

6. Cumulative posterior probability additional feedback may increase the accuracy for CUM ODDS assessments made on a log scale. This is one more hypothesis to be investigated.

7. Aggregation is more important than response mode in determining the size of assessments.

8. The log scale is more important than V-C-O additional feedback in determining the size of NONCUM LR responses.

9. The hypothesis that cumulative probability additional feedback may be more important than the combined effects of the response mode, aggregation and scale dimensions is the most exciting hypothesis that these data suggest for me. But this issue needs very careful examination.

The findings of this investigation are supported by previous research. Kaplan and Newman (1966) found that the computer-aggregated posterior probabilities for Ss making nonaggregated P(D|H) judgments were significantly larger than posterior probabilities directly assessed by Ss making aggregated P(H|D) judgments. The data from one of the three experiments reported in Phillips & Edwards (1966) suggest that log scales increase the size of probability judgments. Tsuneko Fujii, who ran parts of this experiments, reanalyzed the data and found that groups that estimated on log scales, odds or probabilities, made larger and more accurate estimates than groups using non-log scales. These results are reported in Fujii (1967).

Since the experimental manipulations in the studies reported in this investigation resulted in significant differences in the size of assessments, all Ss who served as information processors did not perform as Bayes's Theorem would predict. Thus the particular dimensions suggested by the Bayesian approach and incorporated into the taxonomy do make a difference.

The finding that both LR and ODDS assessments are extremely reliable was most encouraging. However, these analyses were performed on averaged data, not individual data.

The results of this investigation suggest some advice for persons applying the Bayesian approach in real world applications. Don't have Ss give their uncertainty judgments on LOG scales if no feedback is given to them about the implications of their assessments, at least until there is further testing of the hypothesis that LOG scales result in larger assessments regardless of accuracy. Be cautious about feeding back the posterior probabilities of the hypotheses under consideration. This feedback may be the most potent variable

studied. Use either the LR or ODDS response mode, whichever is more convenient, when you ask for noncumulative responses. Verbally aggregated responses with no additional feedback result in very conservative assessments, and should consequently be avoided.

BIBLIOGRAPHY

Domas, P. A., Goodman, B. C. & Peterson, C. R. Bayes's Theorem: Response scales and feedback. Technical Report, The University of Michigan, Engineering Psychology Laboratory, No. 037230-5-T, September 1972.

Edwards, W., Phillips, L. D., Hays, W. & Goodman, B. C. Probabilistic information processing systems: Design and evaluation. IEEE Transactions on Systems Science and Cybernetics, 1968, Vol. SSC-4, 248-265.

Fujii, T. Conservatism and discriminability in probability estimation as a function of response model. Jap. Psychol. Res., 1967, 9, 42-47.

Isaac, P. D. Linear regression, structural relations, and measurement error. Psych. Bull., 1970, Vol. 74, No. 3, 213-218.

Kaplan, R. J. & Newman, J. R. Studies in probabilistic information processing. IEEE Transactions on Human Factors in Electronics, 1966, Vol. HFE-7, No. 1, 49-63.

Kendall, M. G. & Stuart, A. The Advanced Theory of Statistics. Vol. II - Inference and Relationship. London: C. Griffin and Co., 1961, 375-391.

Phillips, L. D. & Edwards, W. Conservatism in a simple probability inference task. J. exp. Psychol., 1966, 346-354.

Wheeler, G. E. & Edwards, W. Misaggregation explains conservative inference about normally distributed populations (in preparation).