ARPA ORDER NO.: 189-1
3G10 Tactical Technology

R-1185-ARPA
July 1973

Quantifying Uncertainty Into Numerical Probabilities for The Reporting of Intelligence

Thomas A. Brown and Emir H. Shuford

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

The research described in this Report was sponsored by the Defense Advanced Research Projects Agency under contract No. DAHC15-73-C-0181. Reports of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

ARPA ORDER NO.: 189-1
3G10 Tactical Technology

R-1185-ARPA
July 1973

Quantifying Uncertainty Into Numerical Probabilities for The Reporting of Intelligence

Thomas A. Brown and Emir H. Shuford

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Published by The Rand Corporation

PREFACE

Within intelligence systems, it is important to assess the probability that certain events will occur and to determine the reliability of information on these events. Then, what is the best way to communicate any degree of uncertainty concerning these events, or information on them, to a decisionmaker?

Uncertainty can be quantified into numerical probabilities, which can then be readily and accurately communicated within the system. This report discusses the unique advantages, as well as the problems, involved in implementing such a method.

This report was prepared under The Rand Corporation's project on defense issues raised by technological and economic change, which was funded by the Defense Advanced Research Projects Agency.

SUMMARY

Interest has recently intensified in the area of better management and improved cost effectiveness of intelligence systems. Through the expansion of mathematics and the development of computers within the past two decades, many techniques for the analysis and synthesis of information have become available. An intelligence facility could communicate shades of uncertainty to decisionmakers more effectively if it would attach specific numerical probabilities to potentially confirmable statements.

In Sec. I of this report, we discuss the concept of quantifying the confirmability and certainty of information, using numerical probabilities. Next, we describe a number of scoring techniques that have been found to be useful in eliciting and assessing numerical probabilities (Secs. II and III). The fourth section examines the impact that these techniques will probably have on the motivation of individuals and organizations to make accurate reports of estimates and information. How the scoring techniques might be specifically applied to the reporting of intelligence and how they have been used elsewhere are discussed in Secs. V and VI, respectively. Finally, Sec. VII outlines an initial program for using numerical probabilities in transmitting information.

Four unique benefits stem from the elicitation, scoring, and calibration of numerical probabilities as described in this report. First, since numerical probabilities are quantitative, they can be given to the decisionmaker in a variety of forms (charts, graphs, tables, etc.). These methods have been used for centuries in business and science for transmission of quantitative data and it may now be appropriate for the intelligence community as well to benefit from these techniques.

Second, numerical probabilities are concise and relevant; they express uncertainty with greater precision and fewer syllables than do the verbal equivalents of the Kent Chart (see p. 33). Furthermore, many estimators today seem to try to communicate degrees of uncertainty by giving the reasons why they are uncertain. This may lead to long-winded documents that tell the decisionmaker more than he wants to know about

the estimator's internal processes of ratiocination without adding to his understanding of what the estimator thinks the chances are in the case at hand.

Third, the fact that numerical probabilities are scorable is very important. By keeping averages on the estimators' predicting accuracy, a person should, over time, be able to distinguish effective procedures and individuals from ineffective ones, increase the morale of the better estimators, improve the performance of those who are inexperienced or unskilled, and substantiate the credibility of the system as a whole.

Fourth, in order to score a numerical probability, it must be a forecast of a confirmable event. Thus, the introduction of such techniques will automatically tend to focus the attention of the intelligence system more on objective, confirmable events and less on metaphysical interpretations that may not really increase the decisionmaker's knowledge of what is happening or likely to happen in the real world.

Implementing the techniques described in this report will, of course, require that people be motivated and trained to express degrees of certainty and uncertainty in terms of numerical probabilities.

CONTENTS

I. ASSESSING THE CONFIRMABILITY AND CERTAINTY OF INFORMATION

The essential business of the intelligence establishment is to procure, evaluate, and transmit to the decisionmaker the information he requires to make wise decisions.

The statements which are passed from the intelligence system to the decisionmaker may be divided into two categories: confirmable and non-confirmable. A confirmable statement is one that can be judged as true or false by any reasonable person, given that all the facts regarding the statement are known.

Confirmable statements may concern future events. For example, the statement "Russia will have 65 ballistic missile submarines on January 1, 1974" is a confirmable statement, because when January 1, 1974 comes, it will either be true or false that Russia has 65 ballistic missile submarines, and anyone who has sufficient information will either affirm or deny the truth of the statement. Another example of a confirmable statement is "USC will win the Rose Bowl game in 1975." A non-confirmable statement expresses a value judgment: "USC will play better than their opponent in the Rose Bowl game in 1975." People often argue about whether or not the winner in a football game played better than the loser. It is, of course, possible to replace the vague concept of "playing better" with precise, objective concepts such as yards gained by passing and rushing, passes intercepted, fumbles, length of kick returns, and so on. Many superficially unconfirmable statements are simply short-hand expressions for a bundle of confirmable statements.

Ordinarily the decisionmaker should be more interested in confirmable statements, which are explicit, than in non-confirmable statements, which are ambiguous. The former deal with events in the real world; the latter generally deal with abstractions.

But sometimes it is impossible to make confirmable statements with a high degree of certainty. A natural way for the intelligence system to express this uncertainty is to attach a probability of truth to each confirmable statement it presents. Such a probability is a numerical measure of the uncertainty which arises from the lack of complete

information. Thus, a numerical probability is a complex function of the information available and taken into account at the time the probability is elicited. For this reason, a numerical probability for assessing the truth of a statement can be expected to vary over time with increases in the amount and types of information available. Figure 1 illustrates this.

Because each person or organization may have different sets of available information, their numerical probabilities for the same confirmable statement may be quite different. In this case, we are dealing with a personal probability, not an objective probability where everyone can be expected to agree on the precise numerical value. It should be noted, however, that these personal probabilities may converge to become objective probabilities when all people under consideration possess precisely the same information and have no bias in their evaluation of this information (Fig. 2).

Is a statement like "I believe there is a 70 percent chance that USC will win the Rose Bowl game in 1975," confirmable or not? On the one hand, it is unconfirmable. We know of no way to take apart a football game as could be done to a slot machine to determine the true probability of a given contingency taking place. Even after the game has been played, it cannot be determined whether the team *had* a 70 percent chance of winning that particular contest. But on the other hand, a "reproducing scoring system" may be applied (see Sec. II) where the accuracy of the probabilistic prediction is scored after the event. By accumulating these scores over time, it will undoubtedly be seen that some forecasters usually get higher scores than others. Thus the degree of accuracy of the forecasters' probability predictions will be discovered even if it is impossible to establish the "truth" or "falsehood" of a single, given statement.

The thesis of this report is that the intelligence community could more effectively communicate shades of uncertainty to decisionmakers if it would attach specific, numerical probabilities to potentially confirmable statements. The quality of this system's performance (or the performance of various sub-systems) would be partly evaluated on the basis of how well the system scored in terms of some suitable scoring

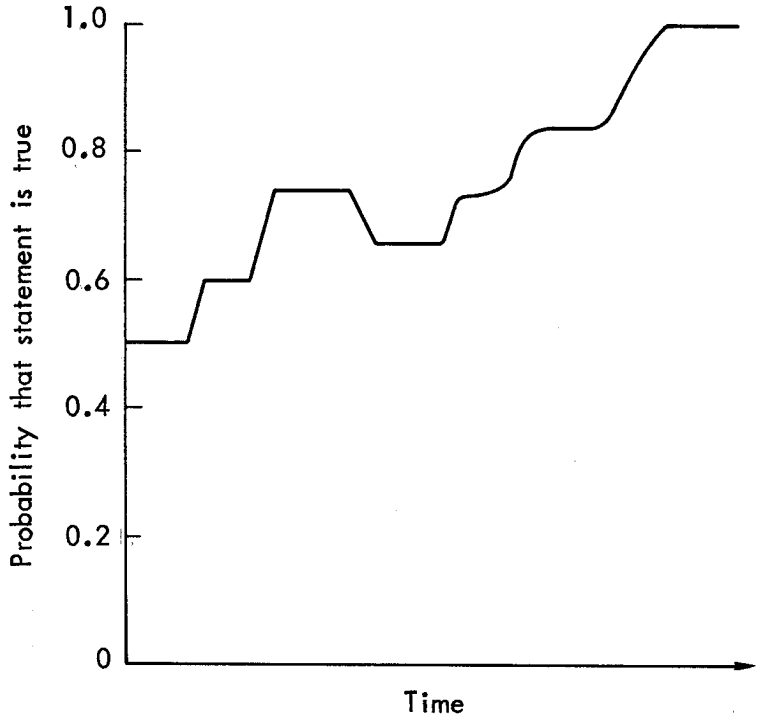

Fig.1 — The assessed probability that a statement is true as a function of time during which additional and finally completely sufficient information becomes available. (Note that probability could have progressed down to zero in the case of a statement that was disconfirmed)

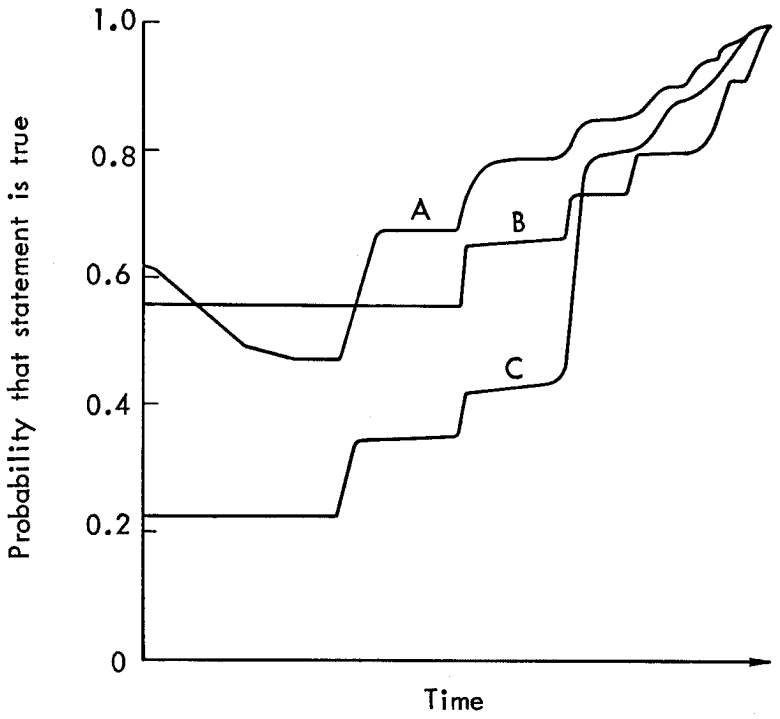

Fig.2 — Probabilities assessed by persons A, B, and C as a function of time. Probabilities differ because of different information available to A, B, and C and finally converge due to the impact of essential information, which becomes available to all.

scheme. If this policy were put into effect, a mechanism for feedback and control would be created, which could have two major functions:

1. Serve as a performance measure of the intelligence system and individuals within it. This policy would thus help to induce superior performance in gathering as much relevant information as possible to reduce uncertainty to a minimum, as well as in providing accurate and unbiased numerical probabilities to reflect the remaining uncertainty.
2. Serve as a track record of the performance of the intelligence system, a record which can enhance and support the credibility of the system. The record has the additional advantage of being a numerical summary that measures the value and accuracy of the communications provided by the intelligence system, without ever dealing in case histories or revealing sources of information.

II. THE SCORING OF PROBABILITIES

RATIONALE

If a forecaster says, "There is a 60 percent chance of rain tomorrow" and it does not, in fact, turn out to rain, how can his accuracy be assessed? It cannot flatly be said that the forecaster is wrong, but he certainly is not exactly right either. He deserves more credit than someone who had forecast an 80 percent chance of rain; he deserves less credit than someone who had predicted only a 40 percent chance of rain. But how much credit should he be given?

If the forecasters were gamblers, selecting from among various wagers at various odds on the chance of rain tomorrow, then the more accurate forecasters would, over a period of time, earn more money than the less accurate ones. Imagine a gambling house in which there were $[\phi(u)\, du]$ wagers available at odds of $1 - u{:}u$ (the "correct odds" for an event of probability u). If the forecaster believed that the probability of an event taking place was p, he would rationally accept all bets on the event taking place at odds better than those appropriate for probability p. Similarly, he would accept all bets on the event *not* taking place which were offered at odds better than those appropriate for probability $1 - p$. The net amount which the forecaster will win or lose may then be calculated:

$$\text{payoff (if event occurs)} = \underbrace{\int_0^p \phi(u)\, \frac{1-u}{u}\, du}_{\substack{\text{amount won on bets that the event } \textit{will} \text{ take place}}} - \underbrace{\int_0^{1-p} \phi(u)\, du}_{\substack{\text{amount lost on bets that the event } \textit{won't} \text{ take place}}}$$

$$\text{payoff (if event does not occur)} = \underbrace{\int_0^{1-p} \phi(u)\,\frac{1-u}{u}\,du}_{\text{amount won on bets that the event } won't \text{ take place}} - \underbrace{\int_0^{p} \phi(u)\,du}_{\text{amount lost on bets that the event } will \text{ take place}}$$

These formulas are easily extended to cover a forecaster choosing among n alternatives rather than two (for example: rain, snow, hail, or fair, instead of rain or no rain). If p_i is the probability of the i*th* alternative, then

$$\text{payoff (if } i\text{th event occurs)} = \int_0^{p_i} \phi(u)\,\frac{1-u}{u}\,du - \sum_{j\neq i}^{n} \int_0^{p_j} \phi(u)\,du$$

$$= \int_0^{p_i} \frac{\phi(u)\,du}{u} - \sum_{j=1}^{n} \int_0^{p_j} \phi(u)\,du$$

There is one serious intuitive difficulty with the above payoff scheme. For any positive ϕ (that is, in any gambling house where there are some wagers available at any probability), the forecaster will be able to make a profit (secure a positive payoff) by simply assuming equal probability for all alternatives. In other words, he will make money even if he is absolutely ignorant about the substantive nature of the events he is forecasting! We can easily adjust the system so that total ignorance corresponds to zero payoff by providing that the forecaster must take the odds on bets placed at probabilities greater than 1/n, but must offer the odds on bets he places at probabilities less than 1/n. Calculations similar to those above lead to the formula:

$$\text{payoff (if } i\text{th event occurs)} = \int_{\frac{1}{n}}^{p_i} \frac{\phi(u)\,du}{u} - \sum_{j=1}^{n} \int_{\frac{1}{n}}^{p_j} \phi(u)\,du$$

In applying these formulas it is important to note that there is always a time limit involved. That is, if a forecaster predicts rain on Monday, and it does not rain until Tuesday, the forecaster is counted wrong on his forecast; he gets no extra credit because of the rain which falls on Tuesday. Similarly, a war between A and B predicted by an intelligence analyst is a non-confirmable forecast. He must predict war with, say, probability 0.9 before the end of 1975 for his forecast to be confirmable or disconfirmable. If A and B are at peace until 2400 on December 31, 1975, then he gets the payoff appropriate to probability 0.1 regardless of what happens on January 1, 1976.

REPRODUCING SCORING SYSTEMS

It is easy to see that if a forecaster knows that he is to be rewarded according to the above scheme he should report the probabilities he really believes in rather than shading them one way or the other to exploit the scoring system. Any probability he reports which varies from his true belief will cause him to place some bets he considers unremunerative, or to fail to place some bets he considers remunerative. Such reward structures are called "reproducing scoring systems"; sometimes they are also called "proper scoring systems," "admissible scoring systems," or "scoring systems which encourage honesty."

The formulas we have derived so far are not completely explicit. It is necessary to specify the specific bets available (i.e., the function $\phi(u)$) before we can calculate the actual numerical payoffs. There are two possible choices for $\phi(u)$ which have been of particular importance in the development of the subject:

$\phi(u) = 1$ (logarithmic scoring system)

$\phi(u) = u$ (quadratic scoring system or "Brier score")

The Logarithmic Scoring System

Plugging $\phi(u) = 1$ into our previously derived expression, we obtain the following:

$$\text{payoff (if } i\textit{th} \text{ event occurs)} = \int_{\frac{1}{n}}^{p_i} \frac{du}{u} - \sum_{j=1}^{n} \int_{\frac{1}{n}}^{p_j} du$$

$$= \log (p_i) - \log (\tfrac{1}{n}) - \sum_{j=1}^{n} (p_j - \tfrac{1}{n})$$

$$= \log (np_i) \text{ (assuming } \sum_{j=1}^{n} p_j = 1).$$

This very simple payoff function has a number of desirable and unique properties. First of all, it is essentially the only reproducing system on more than two alternatives which depends only on the probability ascribed to the event which actually takes place. All reproducing scoring systems on more than two events, with the exception of the logarithmic scoring system and linear transformations of it, depend on both the probability ascribed to the event which actually takes place, and on the way the probability is divided among the events which do not take place [1]. A second desirable property of the logarithmic scoring system is the direct connection it establishes between the evaluation scheme and information theory.

Consider the profit which a forecaster expects to make from his forecast:

$$\text{Expected profit} = \sum_{i=1}^{n} p_i \log(np_i) = \log n - (- \sum_{i=1}^{n} p_i \log p_i).$$

In information theory, the quantity in brackets is called the "entropy" of the partition p_i. It represents the expected amount of information which will be conveyed by the event itself; or, in other words, the surprise content of the event. Thus the forecaster's average reward will be, over the long run, equal to the average amount by which he is able to reduce the surprise content of the events he is forecasting. This beautiful correspondence between the language of information theory

and the intuitive desiderata of a scoring system for forecasters is very satisfying, and a help in thinking about the meaning of the quantities involved.

A third important feature of the logarithmic scoring system is its close association with the maximum-likelihood method of statistical estimation. Suppose F forecasters have made forecasts on n different occasions. Let $p_{i,j}$ represent the probability ascribed by the i*th* forecaster to the event which actually took place on the j*th* occasion, and let k_j denote the number of available alternatives on the j*th* occasion. Then the total logarithmic score of the i*th* forecaster will be

$$\sum_{j=1}^{n} \log(k_j p_{ij}) = \sum_{j=1}^{n} \log k_j + \log\left(\prod_{j=1}^{n} p_{ij}\right).$$

Suppose a person wished to choose among the F hypotheses: "Forecaster i gives correct forecasts." To apply the maximum-likelihood method, which is probably the most important and general method known, one would find the value of i which makes $\Pi_{j=1}^{n} p_{ij}$ a maximum. This is precisely the same as choosing the forecaster who has scored the highest overall on the logarithmic scoring system! So we see that applying the logarithmic scoring system is quite consistent with the most efficient methods for statistical selection of accurate forecasters [2].

One disadvantage of the logarithmic scoring system is the fact that if a forecaster is ever unlucky enough to ascribe probability zero to an event which in fact takes place, the logarithmic scoring system prescribes that he pay an infinite penalty. This raises practical difficulties, because people will often say "probability zero" when in fact they mean "probability 0.01" or something similar. Therefore, when actually implementing a reproducing scoring system it is probably wise to truncate the logarithmic payoff function; that is, simply interpret "probability zero" as actually meaning "probability 0.01" or "probability 0.001" and reward your forecasters accordingly.

The Quadratic Scoring System

Now let us turn to the second important reproducing scoring system,

the quadratic system. Plugging $\phi(u) = u$ into our general expression yields

$$\text{payoff (if } i\text{th event occurs)} = \int_{\frac{1}{n}}^{p_i} du - \sum_{j=1}^{n} \int_{\frac{1}{n}}^{p_i} udu$$

$$= p_i - \frac{1}{n} - \sum_{j=1}^{n} \frac{p_j^2 - (\frac{1}{n})^2}{2}$$

$$= p_i - \frac{1}{2} \sum_{j=1}^{n} p_j^2 - \frac{1}{2n} .$$

This may be rewritten as

$$\text{payoff (if } i\text{th event occurs)} = \frac{n-1}{2n} - \frac{1}{2} \sum_{j=1}^{n} (e_j - p_j)^2$$

$$e_j = 1 \quad \text{if} \quad j = i$$
$$= 0 \quad \text{if} \quad j \neq i.$$

This second way of writing the quadratic scoring system payoff makes it clear that the forecaster who minimizes the squared difference between his a priori forecast and the a posteriori distribution (i.e., the actual outcome of the event) gets the best score. Thus the quadratic scoring system is closely related to the traditional concept of least-squares optimization.

Some practical advantages of the quadratic scoring system are that it is very easy to calculate (although this is perhaps a negligible advantage in the age of computers), and never calls for infinite payoffs or penalties (unlike the logarithmic scoring system). The fact that the quadratic scoring system is a reproducing scoring system was discovered about 1950 by a meteorologist named Brier [3]. Since that time this

score has often been used as a routine tool in evaluating alternative weather forecasting techniques.

Scoring Continuous Probability Distributions

In forecasting political and economic events, it is often interesting to estimate the size of a quantity (votes for a particular candidate, GNP, tank production in given quarter, etc.) for sometime in the future rather than choosing among a finite number of alternatives. To cope with such cases, any reproducing scoring system on a finite number of alternatives may be converted into a scoring system applicable to probability density functions on a continuum by a simple limiting process. If $p(t)\,dt$ represents the probability density function presented by a forecaster, either of the following two scoring systems will have the reproducing property:

$$\log K\, p(t) \qquad \text{(logarithmic)}$$

$$2p(t) - \int_{-\infty}^{+\infty} p(u)^2\, du \qquad \text{(quadratic)}$$

In the above, K is any constant, and t represents the true value achieved by the quantity in question. It has been found in independent experiments,* that when students are asked to give probability density functions for uncertain quantities, the distributions they give tend to be too "tight." Instead of the two percent that would be expected, about 40 percent of the time the true answer falls outside the 0.01 - 0.99 percentile band of any given respondent's distribution. This poor performance can undoubtedly be corrected by suitable training, but it does influence one to prefer the quadratic to the logarithmic scoring system because the former does not depend so strongly on what hypotheses

*Results given in two unpublished papers, one by M. Alpert and H. Raiffa, and the other by T. A. Brown.

are made about the distribution of probability mass in the tails of the distribution. Another alternative is to convert the probability estimation task to a discrete one by eliciting the probability that the true quantity will be found to lie in each of several intervals. By assuming a functional form for the continuous distribution of uncertainty, the assigned probabilities can be used to estimate the parameters, and, thus, the complete distribution. This method might serve to reduce the bias reported above.

III. CALIBRATING PROBABILITIES AND THE DETECTION OF BIAS

CALIBRATING THE PROCESS OF ASSIGNING PROBABILITIES

Up to now we have been considering the basic requirement that must be met by any method for eliciting probabilities, which is how to provide an incentive for the reduction of uncertainty and for accurate estimation of probabilities. The scoring systems described above must, of course, be the basis of any incentive system provided to an individual or organization. It must be taken as the direct and basic measure of the value of the information contained in the probabilities. There is, however, another way of viewing the process of estimating probabilities--a way which provides not only an operational definition of probabilities, but also a way of providing knowledge of results to an individual or organization in order to allow them to improve their estimation of probabilities and in return to achieve a higher score as measured by the scoring systems discussed above. There are several ways of doing this, each way representing a somewhat different, but not contradictory, view of calibrating the accuracy of numerical probabilities.

The External Validity Graph

Suppose an individual defines probabilities both as to the occurrence and nonoccurrence of each of a large number of confirmable statements. Suppose further that enough time has passed so that each statement can be unequivocably confirmed or disconfirmed. There is a way of taking data of this type and calibrating the external validity of the numerical probabilities. This can be done by using the probability assignments to define subclasses of events and by examining the relative frequency of occurrence and nonoccurrence of each of these subclasses. To be more specific, consider each time a numerical probability of 0.80 was assigned. Say this happened 1,000 times. Now, the 1,000 events in this subclass are all characterized by the fact that the individual has made a probabilistic forecast or prediction that there was an 80 percent chance that the event would be confirmed. So we can proceed by counting how many of these 1,000 events did in fact

occur. Taking the probability of 0.80 at face value, we would expect to see that 800 of the events occurred out of the total number of 1,000, yielding a relative frequency of 0.80. This process may be repeated for each probability level from 0 to 1 and, in the ideal case, we would hope to find that the empirically determined relative frequency is equal to the probability, except for sampling variability, over the whole range from 0 to 1.00. If such were the case, we would evaluate the probability assessor as an unbiased and realistic assessor of probabilities. Furthermore, these probabilistic predictions have a direct empirical interpretation. Whenever the realistic assessor says that the probability is 0.80, then 80 percent of the time, no matter what events he is predicting, the event will occur, or if his probabilistic prediction is 0.10, then 10 percent of the time the event will be confirmed, and so on. The graph of such data, as shown in Fig. 3, is called an *external validity graph* because the probabilistic predictions made by the individual are tied down and related to confirmation and disconfirmation of the events in the external world.

Now, to the extent that the graph relating relative frequency to numerical probability deviates from a straight line with unit slope, the probability assessor is biased in his assessment of the value of the information available to him, and, in the extreme case, where this graph has a slope of zero, the probabilistic predictions of the probability assessor would be absolutely worthless and have no relation to the external world. If, however, the slope is greater than zero, there does exist a relation, and there is evidence that the probability assessor has some useful information for discriminating degrees of uncertainty about the external world.

Estimating the Realism Function

Under suitable conditions it appears that the function relating relative frequency to numerical probability can be approximated by a straight line. The least-squares procedure for estimating the slope, a, and intercept, b, of this straight line is given below.

Let

$r_1 < r_2 < \ldots < r_L$ be the levels of probability assigned,

u_i = number of times r_i used and event was confirmed, and

v_i = number of times r_i used and event was disconfirmed.

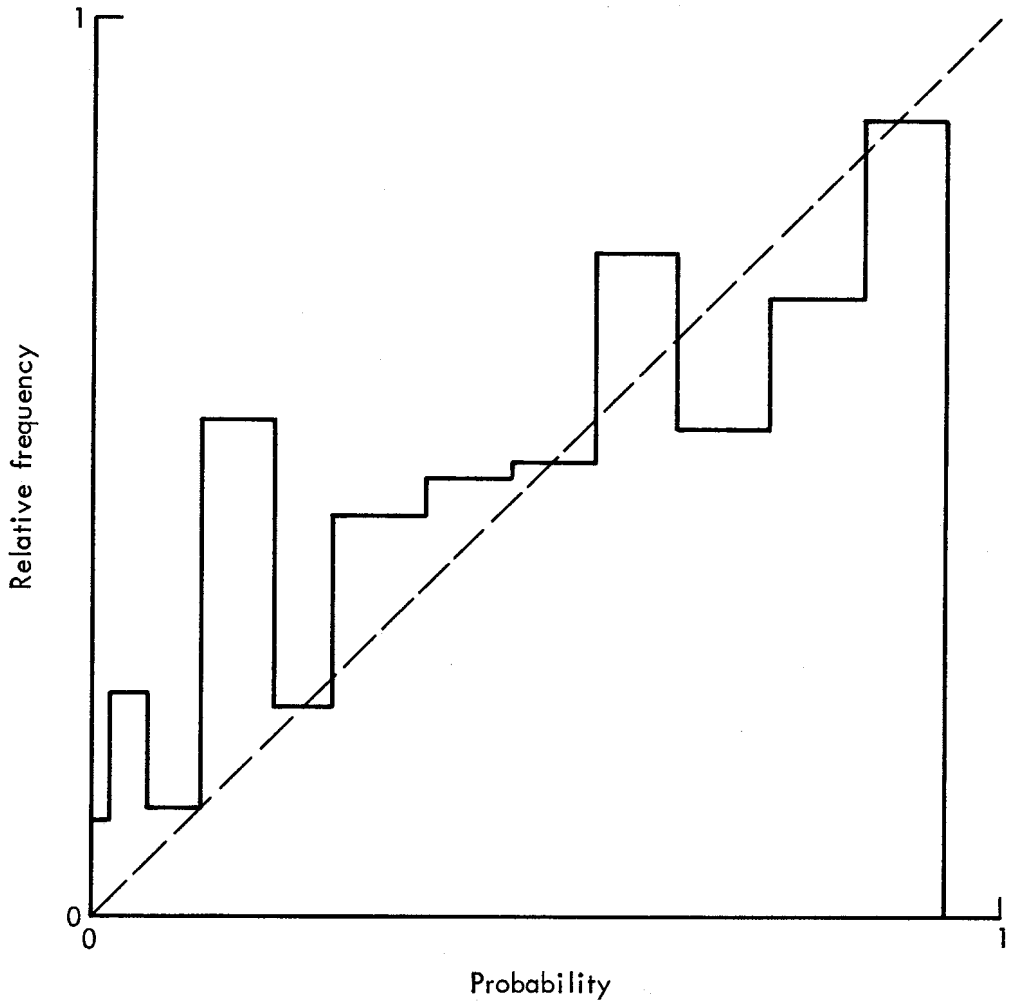

Fig. 3--Illustration of an external validity graph showing relative frequency of correctness for different groups of probability assignment. Dashed line shows ideal match between relative frequencies and probabilities

We want to find a and b so as to minimize

$$\sum_{i=1}^{L} (u_i + v_i)\left(\frac{u_i}{u_i + v_i} - ar_i - b\right)^2$$

The least-squares estimators are

$$\hat{a} = \frac{\sum(u_i + v_i)\sum u_i r_i - \sum(u_i + v_i)r_i \sum u_i}{\sum(u_i + v_i)r_i^2 \sum(u_i + v_i) - [\sum(u_i + v_i)r_i]^2}$$

$$\hat{b} = \frac{-\sum(u_i + v_i)r_i \sum u_i r_i + \sum(u_i + v_i)r_i^2 \sum u_i}{\sum(u_i + v_i)r_i^2 \sum(u_i + v_i) - [\sum(u_i + v_i)r_i]^2}$$

The advantage of this linear estimation procedure is that now we can calibrate a person's numerical probabilities using much less data than before. For example, experience has shown that with the external validity graph, several hundred if not thousands of observations are required. With the least-squares estimation procedure, stable estimates may be obtained with as few as twenty or more probability estimates.

Using this procedure, we have found that some people (about one in ten) can, when tested out using one of the scoring systems, produce probabilities which are best fitted by a line with an intercept of zero and a slope of one. In other words, they can produce unbiased and realistic probabilities. We find, however, that the *realism function*, the least-squares fitted function, deviates from this ideal line for most people, as shown in Fig. 4. In some cases, the slope is less than one, while in other cases it is greater than one.

If the slope is less than one, the person appears to be over-valuing his information. A slope of less than one implies that when a probability of one is assigned to an event, the relative frequency of occurrence is not one, but some value less than one. The event is not as likely to occur as the person thinks it is. On the other hand, when

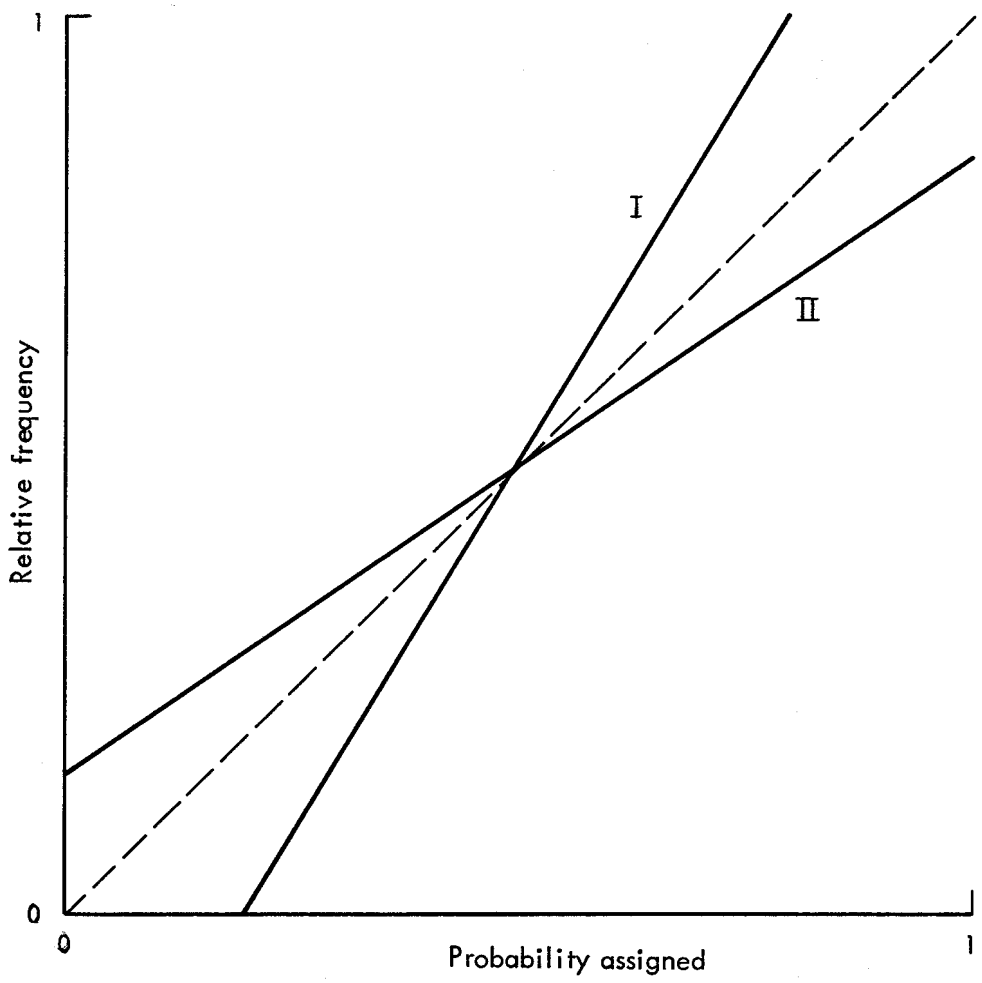

Fig. 4--Two realism functions based on probability assignments for two-outcome questions. Person I undervalues his information while person II overvalues his information.

he assigns a probability of zero to an event, the relative frequency of occurrence, instead of being zero, is some positive value indicating that his information on the nonoccurrence of an event is not worth as much as he thinks it is.

By similar reasoning one can see that if the slope is greater than one, the person is undervaluing or underutilizing the information available to him. He is, for example, assigning a probability of only 0.9 to events when he really has information which would justify an assignment of one to the event since the relative frequency is one even when

he assigns the probability of 0.9. At the other extreme he may be assigning a probability of occurrence of 0.05 to events that have a zero relative frequency of occurrence, in which case he would be justified in assigning a zero probability to them, and so on.

Uncertainty in Assigning Probabilities

Most people, when called upon to assess a probability, feel varying degrees of uncertainty or unsureness as to what the exact value of the probability should be. For this reason, even the most outspoken advocates of using personal probability in business decisionmaking, military decisionmaking, etc., have hesitated from seriously eliciting and using personal probabilities of this sort. However, if the techniques for calibrating and operationally defining probability that have just been described work in practice, then this hesitation will be an error in judgment.

More explicitly, no matter how uncomfortable one may feel at estimating probabilities, if one has been properly trained and turns out to be capable of giving unbiased and realistic probability estimates which are in contact with external reality in the sense described above, then this should be sufficient and complete justification for the use of probabilities. In other words, the probabilities given by this person are useful and accurate. The fact that personal feelings of unsureness or uncomfortableness are involved may be a psychological characteristic of giving probability estimates but does not contradict the validity of these probability estimates. Realizing this, one might be well advised to discount his feelings of uncertainty and unsureness and to accept the probability estimates at face value on the basis of their demonstrated validity and usefulness.

MEASURES OF BIAS

Perceived Versus Actual Amount of Information

In the calibration described above, the probability assessor assigns a probability distribution over each of, say, n possible outcomes. We can use these probabilities to measure the amount of information that the

probability assessor perceives that he possesses with respect to the question at hand. In the language of information theory, the amount of this information is measured by

$$\log n + \sum_{j=1}^{n} p_i \log p_i$$

The probability assessor is maximally uncertain or has, in his view, zero information when his probabilities are all equal. He has maximal or complete information and minimal uncertainty when he assigns a probability of one to one of the outcomes.

If the questions used for the evaluation are independent, then the total amount of the information that the probability assessor perceives he possesses with respect to the subject matter under evaluation is measured by

$$k \log n + \sum_{j=1}^{k} \sum_{i=1}^{n} p_{ij} \log p_{ij}$$

where k is the number of questions and p_{ij} is the assessor's perception of the probability that the *ith* alternative on the *jth* question is correct.

It should be clear that this perceived amount of information can be computed from the person's probability values. It should be noted that without the use of a scoring system such as those described above, there is nothing to prevent the probability assessor from claiming to have no uncertainty whatsoever about the subject matter, or, equivalently, complete information. Remember, the second expression given above is maximized by assigning probabilities of one and zero across all the events.

Now another possibility is to use the realism function as estimated earlier (p. 16) to correct on an *ex post facto* basis the probability assignments given by the person and to use these corrected probabilities to estimate the *actual* amount of information possessed by the person.

This quantity may be either higher, lower, or the same as the *perceived* amount of information.

If the realism function has a slope less than one, indicating that the person overvalues his information, then the new information measure will be computed using probabilities which are less extreme in the direction of one and zero than those of the perceived amount of information measure. Thus, the actual amount of information possessed by the person will be less than he perceives it to be.

If, on the other hand, the realism function has a slope greater than one because the person undervalues his information, the actual amount of information measure will be based on probabilities which are more extreme than those used in the perceived amount of information measure. The person's information is better than he perceives it to be.

Expected Total Score

Another way of defining realism is to evaluate the actual score earned by the probability assessor with respect to his distribution of expected total score. On each trial, the probability assessor allocates a probability distribution over the possible outcomes. This implies an expected score for that trial, and if the trials are independent, the sequence of distributions implies a distribution of total score for the test as a whole. The mean of this distribution, based on the scoring function $f(p_{ij})$, is

$$m = \sum_{j=1}^{k} \sum_{i=1}^{n} p_{ij} \, f(p_{ij})$$

and the variance v of the distribution is

$$v = \sum_{j=1}^{k} \sum_{i=1}^{n} p_{ij} \, f(p_{ij})^2 - \sum_{j=1}^{k} \left[\sum_{i=1}^{n} p_{ij} \, f(p_{ij}) \right]^2$$

Now, if the actual score obtained by the probability assessor falls in the lower tail of this distribution, one can argue that the probability assessor is overvaluing his information and that this overvaluation might well be measured by the probability (based on the expected total score distribution) of getting an observed total score as small or smaller than that one actually obtained. If, on the other hand, the observed total score for the evaluation falls in the upper tail of the distribution, then the probability assessor is undervaluing his knowledge and the degree of undervaluation may be assessed by the upper tail probability based on the expected total score distribution.

Another approach along these lines might be to derive the expected total score distribution based on the corrected probabilities by using the realism function described above. In this case, one can compare the likelihood of the actual total score using the probabilities actually provided by the probability assessor with the likelihood of the *different* correct total score based on the corrected probabilities. A likelihood ratio measure would test the hypothesis that the corrected probabilities are a perfect description of the data versus the hypothesis that the uncorrected probabilities are a perfect description of the data. When the logarithmic scoring function is used, the likelihood ratio can be computed on the basis of simple transformations of the actual and the adjusted total test scores.

Loss in Score Due to Lack of Realism

With the logarithmic scoring system, the score depends solely upon the probability assigned to the correct event. When it has been asserted that the probability assessor loses score because of his bias, this can be shown to him by using the realism function to adjust his probabilities on an *ex post facto* basis in order to compute a new score, the one he could have made if he had been unbiased in his assessment of probabilities. To the extent that his probabilities are biased, this new score will be higher than the score actually achieved by the probability assessor and the difference represents his loss due to his inability to correctly assess the value of his information. This loss is sometimes called the labeling error. The difference between the new corrected

score and a perfect score represents the probability assessor's loss due to the lack of perfect and complete knowledge of the events under assessment. Some analysts call this loss the sorting error. Figure 5 shows a representative graph provided for feedback to a probability assessor using the Rand Videographic System for eliciting probabilities.

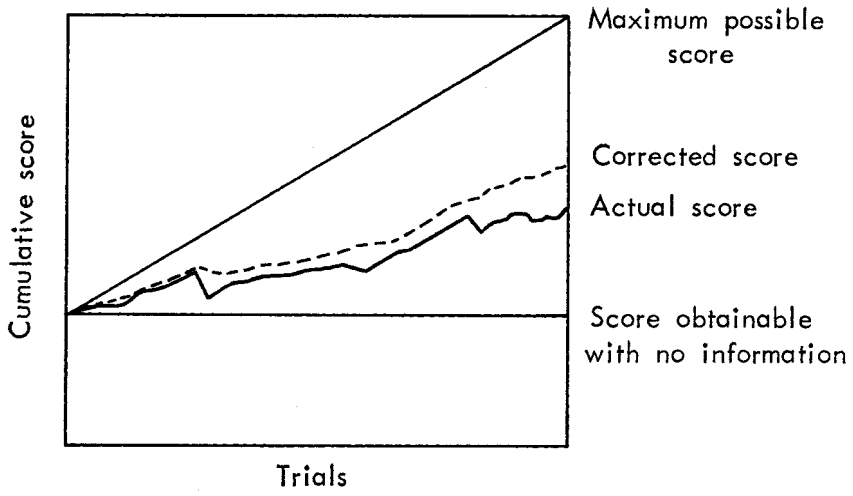

Fig. 5— One way of illustrating loss in score due to lack of realism

When to Correct Probabilities

Although we have been talking about using the realism function to correct and to adjust on an *ex post facto* basis the probabilities yielded by a probability assessor, we recommend that this be used only as a basis for providing feedback so that the assessor might learn to become more realistic in his evaluation of his information.

It may not be a good idea to correct the probabilities yielded by the probability assessor on the basis of a realism function obtained in the past. The danger here is that with incentive systems based on one of the scoring systems we have described, the probability assessor is constantly striving to improve the accuracy and realism of his assessments. Thus if we used data based on previous behavior to adjust new probability estimates from the probability assessor, we may

be making an incorrect adjustment. The bias may no longer be there or it may not be of such a magnitude or it may not fall in such a direction.

For these reasons, we suspect that it is better to take current numerical probabilities at face value. However, it must be recognized that the wisdom of this depends on the rapidity with which probability assessors are able to overcome their biases. If actual experience indicates that some individuals stubbornly continue to overvalue or undervalue their information (perhaps because of some deep-seated propensity for risk taking or risk aversion), then it may be wise to use their realism function to "correct" probability estimates which they provide.

IV. INCENTIVES TO ENCOURAGE ACCURACY

IMPORTANCE OF EMPHASIZING SCORES

In our experience it is absolutely essential that the person assigning probabilities focus upon the conditional scores as provided by a reproducing scoring system, rather than on the probabilities themselves. If for some reason the person ignores the scores and focuses upon the probabilities, the resulting probability assignments will almost certainly be biased and there will be a loss of potential information.

There appear to be several reasons for this. First, use of the conditional scores provides structure to the task and serves to define the probabilities in a relevant fashion. Second, without the conditional scores according to a reproducing scoring system, the person is likely to take the probability assignments themselves as the measure of his performance. When he does so, the optimal strategy is, of course, to assign a probability of one to the most likely outcome and a probability of zero to all other outcomes so the person is driven to appear to overvalue his information and fails to discriminate degrees of uncertainty. Third, to the extent that the conditional scores are really important to the person, he can concentrate on maximizing his expected score and is better able to ignore other utilities or values that might enter in to influence his behavior. To take just one example, some people are reluctant to admit that they possess less than complete and perfect information about the subject at hand, and thus are reluctant to assign any probability less than one. This reluctance might be overcome by associating money, bonuses, or career development incentives to the scores yielded by a reproducing scoring system. If such hard incentives were attached to these scores, great care would be necessary to insure that comparisons between individuals who had made their scores answering different lists of questions were carried out on a fair basis.

INCENTIVES FOR GATHERING ADDITIONAL INFORMATION

Given that the person cares about his score as yielded by one of the reproducing scoring systems, what incentive do they provide for

doing a careful job of information gathering and analysis? To understand this, we must take a closer look at the reproducing scoring systems. Figure 6 shows how two reproducing scoring systems vary as a function of the probability assigned to the event which in fact occurred.

In a working situation, a person does not have to be content with his current probabilities. He can, for example, attempt to gather additional information and do further research in order to drive his probabilities toward either one or zero and this, of course, is precisely the behavior we would like to encourage. What gain accrues to the person who puts out this extra effort?

Consider first the currently dominant incentive system whereby the person, in effect, makes an absolute prediction that a statement will or will not be confirmed and is rewarded for his correct predictions. Suppose that we scale these rewards so that when the person is maximally uncertain about which of two events will occur, the expected reward is zero, while if the person has completely sufficient information corresponding to a probability of either zero or one, his expected gain is equal to one. In this case, the person's expected gain from making his prediction is proportional to the distance between his probability and one half as shown in Fig. 7. By exerting additional effort to drive his probability toward either zero or one, the person increases his expected gain.

The first reaction of many people to the notion of assigning probabilities and using reproducing scoring systems is that this would represent a policy of being lenient on individuals and would encourage sloppy work. The true state of affairs may be determined by examining the expected gain functions produced by the reproducing scoring systems as shown in Fig. 7. Notice that these are U-shaped functions such that the person gains relatively little by making a slight effort which pushes his probability moderately away from one half; in order to make large gains he must gather enough solid information to move his probability much closer to zero or to one. So, rather than being a more lenient reward structure, the new ones provide much greater incentives

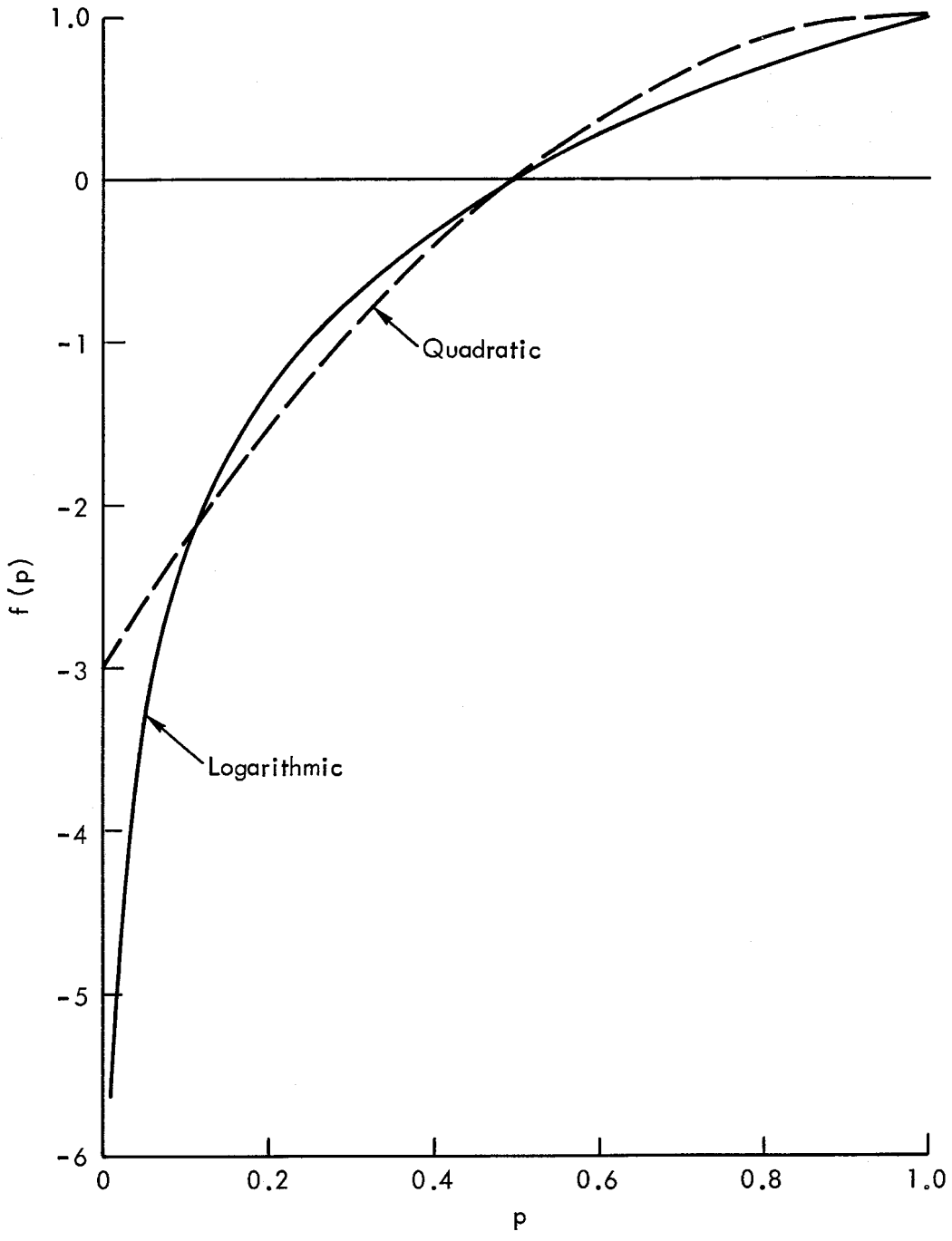

Fig.6--Score, f(p), as a function of the probability, p, assigned to the correct one of two events shown for the two reproducing scoring systems discussed in this report. Scoring systems have been normalized so that completely sufficient information yields a score of one while no information yields a score of zero

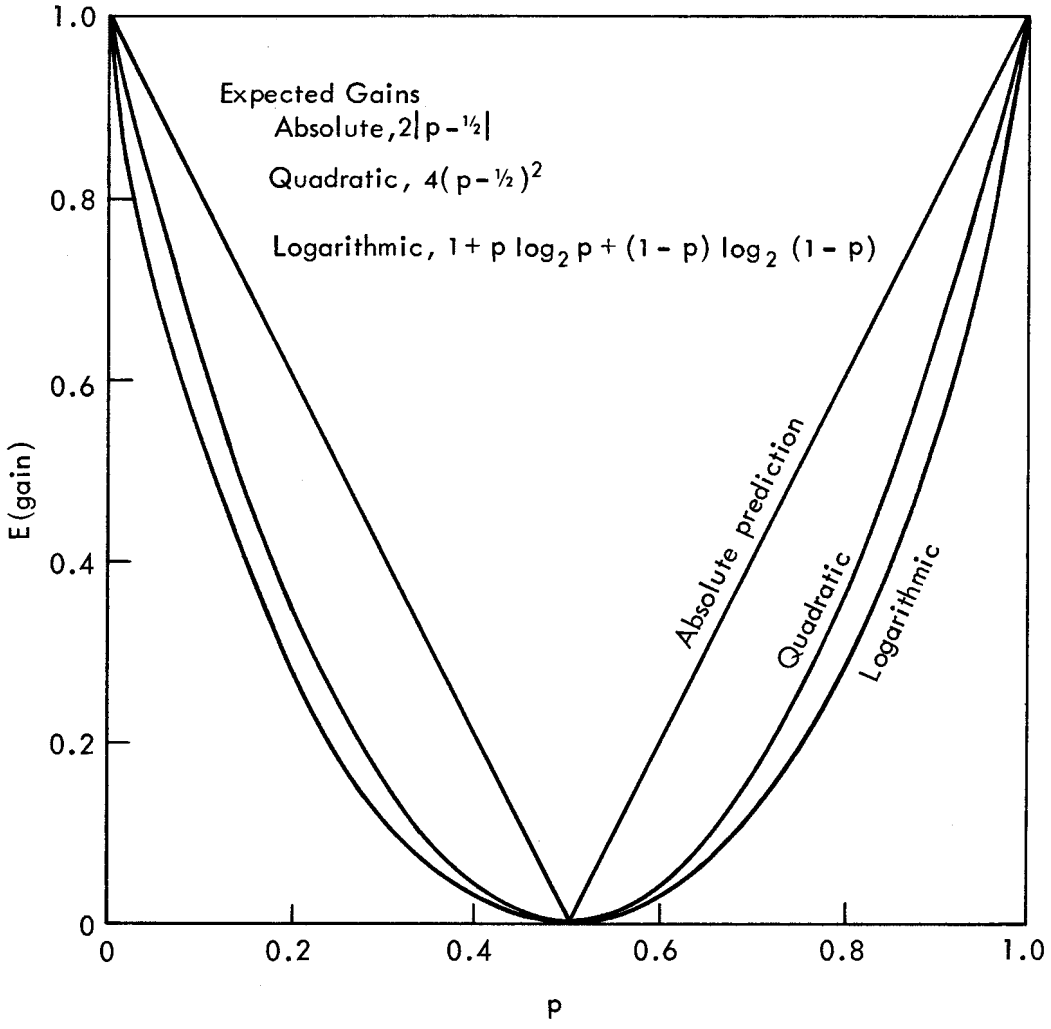

Fig. 7--Expected gain, E(gain), as a function of probability, p, for two events shown for the two reproducing scoring systems and for absolute prediction

toward gathering solid information. The logarithmic score is strongest in this respect and the slope of its expected gain function approaches infinity as the probabilities approach either one or zero.

V. APPLICATIONS TO INTELLIGENCE PROBLEMS

Intelligence work is customarily divided into three categories: collection, production, and dissemination. We will now discuss the application of explicit probabilistic forecasting and reproducing scoring systems to each of these activities:

THE COLLECTION PROCESS

Collection Activities

At first glance it might appear that probabilistic forecasts or estimates are rather inappropriate for the collection process. Collectors fall roughly into three categories: technical systems, overt collectors such as attachés, ambassadors, consuls, and travelers; and clandestine collectors.

The technical system ordinarily do not produce probabilistic statement; they produce a variety of records, tapes, and similar documents, but it is hard to see how reproducing scoring systems could be applied directly to score any of these documents. When an expert looks at these documents, and draws conclusions (often uncertain) from them, he is ordinarily said to be producing rather than collecting intelligence.

Overt collectors on the other hand are often called upon to integrate the numerous bits and snippets of information with which they come in contact (including informal opinions from, for example, responsible officials or communications media) into an overall assessment of who will win a given election, whether or not a given bill will pass, or whether some specific action will be taken by the government with which they are most concerned. It could be argued that when one is asked to make such an assessment, he is acting as an intelligence producer rather than as a collector. On the other hand, the assessment produced will be used as an input to analysts working with [redacted] information from a variety of sources to produce "finished intelligence."

A Reinterpretation of Rating Systems

A method of rating raw intelligence reports from the field has been widely applied in the United Kingdom, United States, and other countries. Each report has a letter attached to it (A, B, C, D, E, or F) and a number (1, 2, 3, 4, 5, or 6). The letter indicates the quality of the source, and the number indicates an initial appraisal of the content. The code definitions are:

Grading of Source		Appraisal of Content	
A	Completely reliable	1	Confirmed by other sources
B	Usually reliable	2	Probably true
C	Fairly reliable	3	Possibly true
D	Not usually reliable	4	Truth doubtful
E	Not reliable	5	Improbable
F	Reliability cannot be judged	6	Truth cannot be judged

The *letter* is assigned on the basis of an agent's past record, and is transcribed from his dossier. If agents were routinely reporting explicit probabilistic estimates, this letter rating could be replaced by his realism function. The *number* is not actually a measure of the truth of a report, but rather a measure of how well the report fits in with previously available information. In a probabilistic system, this number could be replaced by the agent's expected score on the assumption that the previously estimated probability of the event in question was correct, that is, $H(p,q) = \sum p_i f(q_i)$, where p = previously supposed probability and q = agent's report. If this score was large and positive, it would indicate a definite report highly consistent with previous estimates; if it was small, it would indicate a vague report; and if it was large and negative, it would indicate a definite report at variance with previous estimates. Reports from reliable sources with large negative appraisal scores should be singled out for special attention, for they may indicate a considerable step forward in the knowledge concerning some situation.

Under certain circumstances, reproducing scoring systems might be useful mechanical tools in attempting to analyze enemy deception operations. For example, if a person had some information that he *knew* had been fed to him by the enemy, then letting p in H(p,q) be the probability vector of what had been inferred from just the information provided by the enemy gives a measure of the coherence of any new report with information the enemy has been providing. If report X is more coherent with what the enemy evidently wishes him to believe than it is with what can be inferred from all channels, then there are grounds for suspecting that X is somehow controlled by the enemy.

THE PRODUCTION PROCESS

Evaluation and Its Impact

The biggest potential payoff for the systematic application of reproducing scoring systems in intelligence is in the production side. Imagine a system in which several independent analysts, working either from the same information or from slightly different bodies of information, make explicit probability assessments of a set of future confirmable events. By keeping score on these forecasts using a reproducing scoring system, it could be determined which of these analysts was more skillful. By examining the extent to which they individually under- or overvalue their information, they could be counselled as to how to improve their forecasts.

Simply knowing that their work is being evaluated by such an objective yardstick as a reproducing scoring system may have an immediate favorable effect on the morale of intelligence producers, and on the quality of their work. Many producers today are, we understand, haunted by the fear that no use is made of their output, and that they are cogs in a machine grinding out meaningless results. Some intelligence producers may fear that the only way to get ahead is to generate estimates that correspond with the prevailing attitudes of higher-ups. The institution of a systematic grading and feedback routine based on appropriate scores would hopefully relieve these concerns, and stimulate the

producers to more careful and precise estimates (see Sec. IV). In numerical form, these estimates could be readily summarized, in contrast to more informal written estimates that can hardly be combined and summarized without great effort and probable distortions of some sort.

Possibilities for External Production

The existence of an objective measure for uncertain forecasts suggests that it might be profitable to put the intelligence community in competition, in certain areas, with the body of informed citizens at large. For example, if a person is interested in forecasting events in Ruritania over a period of time, he could write to professors, businessmen, politicians, and others in the United States who are known to be especially knowledgeable about Ruritanian affairs, soliciting their assistance. Those who respond favorably would be mailed questionnaires asking them to assign probabilities to various alternative future events in Ruritania, and could be paid on the basis of what score they make. The cost of this endeavor would depend on how efficiently it was administered, but it might be fairly low, since the level of pay required would only cover the time spent filling out the questionnaire, not the time spent accumulating the background and knowledge required to do so. The mean response of such a broad panel of outside experts might well be a more accurate view of future events than the response of a smaller group of regular government employees. Since such an operation has never, to our knowledge, been tried, we are simply not in a position to judge.

THE DISSEMINATION PROCESS

Verbal Descriptions of Uncertainty

Another major potential payoff is in the field of dissemination. For a long time intelligence experts have wrestled with the problem of finding verbal equivalents for the various states of uncertainty, which are expressed so clearly and unambiguously by numerical probabilities.

The well-known intelligence expert Sherman Kent has often advocated nailing down the phrases used to connote various levels of uncertainty to precise numerical ranges. One attempt to do this is shown in Table 1. This table is based on discussions with analysts as to how they use words and phrases such as "certain," "probable," "almost certain," and "impossible." Tables of this sort are called Kent Charts. There are fundamental problems with any such effort to find verbal equivalents for numerical probabilities.

Table 1

A "KENT CHART"

Probability (percent)	Verbal Equivalent
100	It is certain that ...
85-99	It is almost certain that ...
60-84	It is probable that ...
40-59	The chances are about even that ...
15-39	It is probable that ... not ...
1-14	It is almost certain that ... not ...
0	It is impossible that ...

First of all, if the analyst thinks there is a 61-percent chance of an event happening and translates this into "it is probable that ...," the consumer could well interpret this as an 80 percent probability. Thus, the translation from a numerical statement to a verbal one always loses a certain amount of information. But the situation is even more serious than this. Most consumers are unaware of what particular verbal equivalents the producers have elected to use, and thus may construe "probable" as meaning 50-70 percent instead of the intended 60-84 percent, or think that "it is almost certain that ... not ..." means 10-30 percent instead of the intended 1-14 percent. Indeed, surveys of intelligence consumers have shown that their interpretation of the precise meaning of the words and phrases included on the Kent Chart covers an extremely wide range. It is very hard to see what virtue there is in replacing a numerical probability with a verbal equivalent, if one is interested in precise communication.

Numerical Probabilities and the Track Record

By expressing estimates in numerical terms rather than verbal equivalents it will be possible to keep scores on disseminated material. The construction of a track record should greatly increase the credibility of finished intelligence; under the present system spectacular failures are remembered while steady accuracy tends to be ignored. By accumulating a batting average for the various components of the intelligence system we should insure that the community output is taken as seriously as it deserves to be.

VI. RELATED APPLICATIONS

In what other areas of application have the techniques of eliciting, scoring, and calibrating numerical probabilities been employed? Not very many.

Why is this so? The concept of probability has certainly been around for a long time. However, the success of actuarial science and statistics has focused attention upon probabilities obtained as relative frequencies calculated on large aggregates of data.

More recently, a great deal of attention has been focused on using logic and mathematics to guide decisionmaking. Out of this work has come the realization that decisions must often be made in the absence of historical data and with respect to unique events for which there is no real possibility of indefinite repetition in order to observe the outcome of identical trials. Even so, probability theory and decision analysis can be a useful guide for behavior if the relevant probabilities are interpreted more generally as numerical probabilities incorporating uncertainty in the sense described in this report.

We can distinguish three areas where formal use is being made of numerical probabilities as a measure of uncertainty. They are weather forecasting, drilling decisions, and educational testing.

WEATHER FORECASTING

In 1950, Brier [3] proposed that weather forecasts be expressed in terms of probability and that they be scored by the quadratic scoring system (see Sec. II). In 1960, the U.S. Weather Bureau began internal use of probabilities for precipitation forecasts. In 1965, the Weather Bureau began reporting to the public its probabilistic forecasts, e.g., "There is a 10 percent chance of rain tonight, rising to 30 percent tomorrow."

While the Weather Bureau apparently does not use the quadratic scoring system to evaluate and motivate individual forecasters, this scoring system is used to answer questions such as

1. What is the level of skill in present-day temperature and precipitation forecasts?
2. What skill is contributed to the public forecast by each of the forecast echelons?
3. Which of the forecast areas are performing with acceptable accuracy? [4]
4. What are the "recent trends in the accuracy and quality of the Weather Bureau forecasting service?" [5]

Although the Weather Bureau has had at least 23 years of experience with the notion of scoring probabilistic forecasts, a certain lack of understanding and conceptual clarity apparently remains, perhaps indicating a need for better training in this area [6-8].

Daily precipitation forecasts were, of course, prime candidates for the application of numerical probabilities and reproducing scoring systems. The prediction is made each day and the event is confirmed or disconfirmed within a very short period of time.

In intelligence applications, the delay between prediction and confirmation will typically be much longer. There are, however, a number of similarities between the functions of the Weather Bureau and those of an intelligence system. The most important is that the major purpose of both organizations is to furnish information to others for use in making decisions. The Weather Bureau serves the needs of its users by attaching a numerical (probability) measure of uncertainty to each of its forecasts.

DRILLING DECISIONS

In 1960, C. Jackson Grayson published his landmark study of drilling decisions by oil and gas operators [9]. The three objectives of his research were:

1. To describe decision problems in a business situation where uncertainties are great.
2. To learn how businessmen are making decisions in this setting.

3. To explore the possibilities of applying "decision theories" to aid the decisionmaker.

He found that verbal, rating, ranking, and other equivalents for probability were often used and misinterpreted by operators, geologists, engineers, and landmen [9, p. 225]. His formal methods of decision analysis required that uncertainty be expressed as numerical probabilities which he obtained by having the individual choose between gambles. Many such choices between gambles were needed to obtain each probability. It should be noted that the reproducing scoring system method (see Sec. II) is conceptually related to the method used by Grayson, but much less awkward in that the individual makes one response (his probability) which specifies how he would bet in all gambles.

Grayson was also concerned with the possibility of falsification of the probabilities [9, p. 261] and, in fact, gathered evidence that among the individuals and branches of the larger organizations there exist quite different perceptions of the company goals [9, Chap. 6], which may lead to inconsistent decisions and distortion of the information passed to higher echelons. He considered the possibility of incentives for accurate probability estimation but he was apparently unaware of the existence of reproducing scoring systems which could have satisfied most of his concerns.

A major theme to come out of Grayson's work, which has possibly great relevance to intelligence applications, is the need to educate the individuals involved in an implementation in order to gain acceptance for and effective use of formal methods. The next application suggests that we may be in a better position than ever before to carry out the necessary education and training.

EDUCATIONAL TESTING

Whenever a student is faced with a question on, say, a multiple choice test, he may encounter a varying amount of uncertainty depending upon his knowledge of the subject matter. The student must, in effect, base his choice of answer upon probabilistic predictions as to which answer is correct. If the student responded to the test question by

assigning these probabilities directly to each possible answer instead of by choosing the answer he perceived most likely to be correct, we would have a much better indication of his knowledge of the subject matter. By using the logarithmic scoring system (see Sec. II), we can encourage the student to reveal his true state of knowledge. By calibrating his probabilities we can give him the information he needs to become more realistic in assessing the value of his facts and reasoning. By tying incentives to the logarithmic test scores we can reward the student for studying more and achieving higher levels of mastery.* In brief, we would have a better output measure for education and training.

In order to determine the potential of this method of probability measurement for application to education and training in the military, The Rand Corporation, under the sponsorship of the Defense Advanced Research Projects Agency, is conducting theoretical studies* and developing on-line computer-based techniques† for training military personnel to understand probability measurement and for administering both written and performance tests to military personnel. At present, the techniques are under development on the Rand Videographic Computer System and, as a consequence, are limited to use within the Rand Santa Monica facility. Within a year, however, The Rand Corporation will transfer these techniques to the PLATO IV computer utility which has remote terminals available to the Air Force, the Army, and the Navy, and at other locations in the U.S. At that time, it will be possible to train personnel on a large scale in the procedures of probability measurement as used for education and training.

These computer-based techniques make it possible for an individual to gain, very rapidly and easily, a great deal of experience dealing with uncertainty and expressing it in numerical probabilities. By taking tests and immediately receiving knowledge of results in terms of correct answer, test score, and bias in assessing uncertainty, the individual can increase his understanding of and his skill with numerical probabilities.

---

*In a forthcoming Rand report by E. H. Shuford and T. A. Brown.

†In a forthcoming Rand report by W. L. Sibley.

Such an experience would appear to be a necessary component of any educational program intended to increase understanding and acceptance of the use of numerical probabilities by individuals within the intelligence system. The experience provided by the Rand system is entirely consistent with the principles developed in this report and suggested for intelligence applications.

VII. STEPS TOWARD IMPLEMENTATION

PROBLEMS WITH EXPLICIT PROBABILISTIC FORECASTS

In previous sections we have discussed the advantages of using explicit probabilistic forecasts evaluated by reproducing scoring systems as a routine intelligence tool. In this section we will discuss some of the objections and problems which may arise under such a system.

Irrational Decisionmakers

First of all, it is a manifest fact confirmed by numerous psychological experiments that many people are very irrational gamblers in the sense that they do not act in such a way as to maximize their expected gain. Experiments by Slovic and Lichtenstein at the Oregon Research Institute [10,11], for example, have shown that subjects will often seek to maximize their *possible* gain, or minimize the probability of *any* loss, rather than seeking to maximize their expected gain. An example of the former behavior, which is "irrational" is a person who enters lotteries; the latter behavior is exemplified by people who pay a dollar to get the "full coverage" rather than the "$100 deductible" on a rented car. Thus, under the system we are proposing, expert A might outshine expert B not because he knows more, but only because he is more skilled at expressing his knowledge in terms of "money-making" bets than expert B. Some people might regard this as an injustice. But if expert B truly has a tendency to overstate his position or to hedge his estimates too much, this will distort the quality of the information he conveys to the intelligence system *regardless* of whether he states his position in explicitly probabilistic terms or not. Only by insisting on such explicit statements, and keeping score, will the system be able to detect these tendencies early and counsel the expert to correct them. And there is some evidence that gambling skill is trainable, that people can learn fairly quickly to express their hunches about future confirmable events quite accurately in terms of numerical probabilities. Use of teaching machines like the Rand

Videographic Three-Alternative Test Program of Sibley* and Shuford is one way to do this.

Distortion Due to Unwanted Utilities

A second problem with keeping score is that its effectiveness will be limited by the fact that some individuals will always be playing some private game of their own rather than the one you have designed for them to play. For example, some experts may want to emerge with the top score at all costs and will run risks of getting a very bad score in order to increase their chance of coming out number one. Such an individual will look like someone who is overvaluing his information. This "Caesar or Nothing" syndrome may be counteracted by making the forecasting task one of indeterminate length, and by deemphasizing any kind of special recognition for the very highest scorers. The financial, or other, payoff to the individual must be simply proportional to his score.

Another game some experts may incorrigibly play is "influence the decisionmaker." That is, they will attempt to make their estimates in such a way that the decisionmaker will be influenced to pursue a policy which the expert favors. This appears to be a common vice in the present system, perhaps because it's the only game in town. By introducing a new game, that of scoring forecasts as to their accuracy, we will probably reduce the prevalence of this vice even if we do not eradicate it entirely.

Recasting Intelligence Questions into Confirmable Propositions

The most serious problem with making widespread use of explicit probability estimates and reproducing scoring systems in the intelligence establishment is the apparent fact that many very important intelligence questions may be very hard to cast into unambiguous confirmable propositions. For example, an expert might reasonably forecast "the rapid growth of chauvinist nationalism in country X" during the coming year. This could be reflected in a wide variety of events: withdrawal of country X from the United Nations, expropriation of foreign property,

*Described in a forthcoming Rand report by W. L. Sibley.

persecution or expulsion of ethnic minorities, higher military expenditures, antiforeign riots, changes in school curriculum, and so on. Each individual event is confirmable, but the question of how many such events are required to add up to "*rapid* growth of chauvinist nationalism" is a subjective matter about which reasonable people may differ. Is it reasonable to break down such a statement into confirmable propositions? On the one hand, one could argue that what the decisionmaker is really interested in is the general statement, not the probabilities of the myriad of possible incidents which could be viewed as validating the general statement. On the other hand, one could argue as we did in the first section of this paper that if a general statement is not equivalent to some collection of confirmable statements, then it is not really meaningful, and therefore the attempt to reduce all statements to confirmable propositions is in and of itself a very useful exercise. One way to clarify whether this is really a serious problem would be to analyze a set of past intelligence estimates (e.g., all NIE's from 1960-65) and see how many confirmable propositions as opposed to unconfirmable propositions they contain, and how many of the superficially unconfirmable propositions could be translated easily into a collection of confirmable statements. A sentence-by-sentence analysis of a sample body of intelligence output by the present system would probably identify five kinds of sentences:

1. Confirmable propositions which could be readily cast as explicit probabilistic forecasts.
2. Non-confirmable propositions which could be restated as a collection of confirmable statements.
3. Non-confirmable propositions which are meaningful, but cannot be restated as a collection of confirmable statements.
4. Statements of fact.
5. Connective statements, platitudes, and so on.

If the output of the current system consists entirely of sentences of types 3, 4, and 5, then the task of introducing explicit probabilistic forecasts of confirmable events may be very difficult. If, however,

a substantial portion of the output consists of sentences of types 1 and 2, then the new methods merely represent a change in language.

Emphasis on Trivial but Predictable Events

If individual forecasters are left free to select the questions on which their probabilistic forecasts will be judged, they will almost certainly have a tendency to drift toward forecasts of events which are relatively predictable but not necessarily of much interest to the decisionmaker. For example, a weather forecaster would tend to predict "the sun will rise tomorrow" rather than risking a prediction on precipitation. For this reason, it is probably necessary for generators of intelligence requirements to give the various forecasters the questions on which they will be formally scored. The requirements generators must, of course, select these questions carefully in order to avoid giving away information which they wish to withhold; but hopefully it is possible to do this.

PROGRAMS TO SUPPORT IMPLEMENTATION

Two types of activities appear to be essential for the successful implementation of the techniques described in this report. One is in the area of system development, while the other is concerned with analysis.

System Development

People untrained in the logic of probability and decision theory do not necessarily obey these rules and behave in a manner consistent with their own best interests. Training should be provided both to those who will be asked to assign probabilities and to those who will use probabilities to make decisions. The training must go beyond a formal course in probability and decision theory to offer extensive experience in assigning probabilities in uncertain situations with knowledge of results in terms of a reproducing scoring system and in terms of bias in the process of assigning probabilities. Such training is desirable in order to remove bias from the probability estimates and, by allowing people to understand the essential logic of the process, to gain acceptance for the use of probabilities.

A computer-based system for the elicitation and assessment of probabilities such as that currently realized on the Rand Videographic System appears to be a very useful component of the training program. By taking "tests" on this system, the student can very rapidly exercise and improve his skill at evaluating uncertainty and assigning appropriate probabilities. Rapid presentation and immediate computation and feedback of results are necessary for effective learning of this skill. It is difficult to see how these features could be duplicated with equivalent speed using manual operations.

We recommend, therefore, that this computer-based system be further developed along with other course materials required for effective training in the logic of probability and decisionmaking and that this training program be implemented on a pilot basis at a school for intelligence analysts. This seems to be an efficient way to test out and further improve the appropriateness and effectiveness of the training program and to observe the degree of acceptance of quantitative probabilities by the people involved. The details of how the computer-based system would be integrated with the curriculum of whatever intelligence school is selected for a trial would have to be worked out, of course, in close consultation and cooperation with the faculty of the school in question. It could probably be combined with material already being taught at the school in such a way that the time required to complete the course would not be lengthened, and the interest of the students in the material taught would be increased. If this training program proved successful, then we would have an answer to the problem of individuals being poor probability assessors as discussed in Sec. I.

We recommend at the next state of implementation that the training program and computer-based system be joined with the actual operations of an intelligence activity so that the working analysts may be trained and evaluated according to their track record over a period of time. This effort will attack the problem of distortions due to unwanted utilities mentioned above and will undoubtedly bring new problems and solutions with respect to implementation.

Analysis

While the system development just described is taking place, a certain amount of analysis should be underway, aimed at detecting particular deficiencies in current intelligence operations in order to guide implementation decisions. Part of this effort would be concerned with examining intelligence output to estimate the difficulty of recasting into confirmable propositions. If this appears to be a major problem, then the implementation strategy should focus on this dimension. If, on the other hand, this is a minor problem and much of the intelligence product is composed of confirmable statements, then a different implementation strategy is called for.

VIII. CONCLUSIONS AND RECOMMENDATIONS

The systematic use of explicit probabilistic forecasts and estimates, scored retrospectively by an appropriate reproducing scoring system, would represent a far-reaching reform in the intelligence service which could have multiple benefits, both direct and indirect. Four major benefits spring from the fact that such estimates are quantitative, concise, scorable, and confirmable.

Since such estimates are quantitative, they can be communicated to the decisionmaker in a variety of methods (charts, graphs, tables, etc.). Quantitative data have been transmitted via these methods in business and science for centuries, and any method which makes this tradition available to the intelligence community has a major point in its favor.

Explicit quantitative estimates are concise; they express uncertainty with greater precision and fewer syllables than do the verbal equivalents of the Kent Chart. Furthermore, many estimators today seem to try to communicate degrees of uncertainty by giving the reasons why they are uncertain. This may lead to long-winded documents which tell the decisionmaker more than he wants to know about the estimator's internal processes of ratiocination without adding to his understanding of what the estimator thinks the chances are in the case at hand.

The fact that quantitative probability estimates are objectively scorable is a tremendous factor in their favor. By keeping objective batting averages on estimators, over time it will be possible to distinguish effective procedures and individuals from ineffective ones, increase the morale of the better estimators, and increase the credibility of the system as a whole.

Finally, scoring a probabilistic forecast requires that it be a forecast of a confirmable event. Thus the introduction of such techniques will automatically tend to focus the attention of the intelligence system more on objective, confirmable events and less on metaphysical interpretations which, however attractive they may be on the surface, do not really increase the decisionmaker's knowledge of what is happening or likely to happen in the real world.

Two types of activities are recommended as essential for the successful implementation of the techniques described in this report. First, a training system incorporating computer-aided elicitation and assessment of probabilities should be developed and tried out in a school for intelligence analysts and then joined with an intelligence operation. Second, a limited amount of analysis of current intelligence operations and output should be conducted in a pilot program in order to help guide implementation strategy.

REFERENCES

1. Shuford, E. H., Jr., A. Albert, and H. E. Massingill, Jr., "Admissible Probability Measurement Procedures," *Psychometrika*, 31, 1966, pp. 125-145.

2. Winkler, R. L., "Scoring Rules and the Evaluation of Probability Assessors," *Journal of the American Statistical Association*, 64, 1969, pp. 1073-1078.

3. Brier, G. W., "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1950, pp. 1-3.

4. Roberts, C. F., J. M. Porter, and G. F. Cobb, "Report on the Forecast Performance of Selected Weather Bureau Offices for 1966-1967," Weather Bureau Technical Memorandum FCST-9, December 1967.

5. Roberts, C. F., and J. M. Porter, "Recent Trends in the Accuracy and Quality of Weather Bureau Forecasting Service," Weather Bureau Technical Memorandum FCST-8, November 1967.

6. Hughes, L. A., "On the Use and Misuse of the Brier Verification Score," ESSA Technical Memorandum WBTM CR-18, August 1967.

7. Murphy, A. H., "Scalar and Vector Partitions of the Probability Score: Part I. Two-State Situation," *Journal of Applied Meteorology*, Vol. 11, March 1972, pp. 273-282.

8. Murphy, A. H., "Scalar and Vector Partitions of the Probability Score: Part II. N-State Situation," *Journal of Applied Meteorology*, Vol. 11, December 1972, pp. 1183-1192.

9. Grayson, C. J., *Decisions Under Uncertainty: Drilling Decisions by Oil and Gas Operators*, Boston: Harvard Business School, Division of Research, 1960.

10. Slovic, P., and S. Lichtenstein, "Importance of Various Preferences in Gambling Decisions," *Journal of Experimental Psychology*, 78, No. 4, 1968, pp. 646-654.

11. Slovic, P., and S. Lichtenstein, "Relative Importance of Probabilities and Pay-Offs in Risk Taking," *Journal of Experimental Psychology Monograph*, 78, No. 3, Part 2, November 1968.