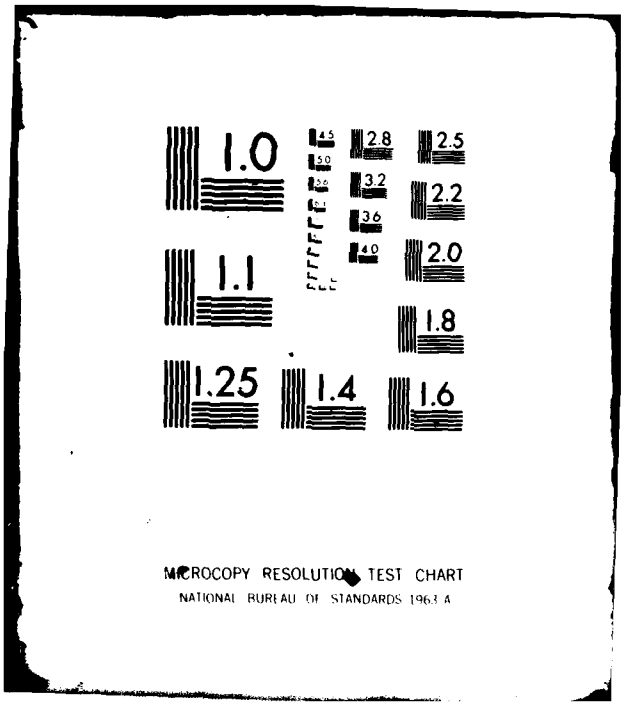
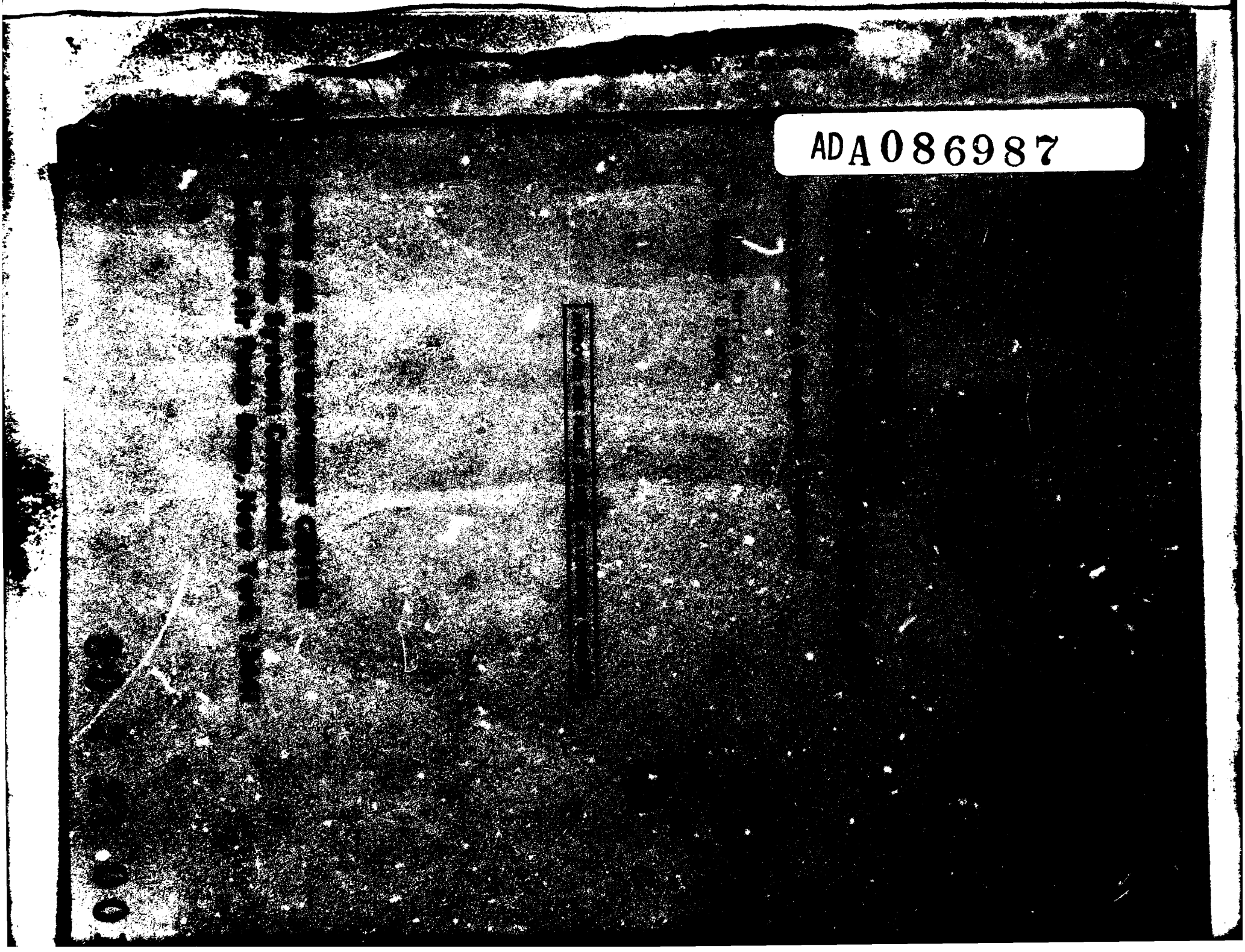
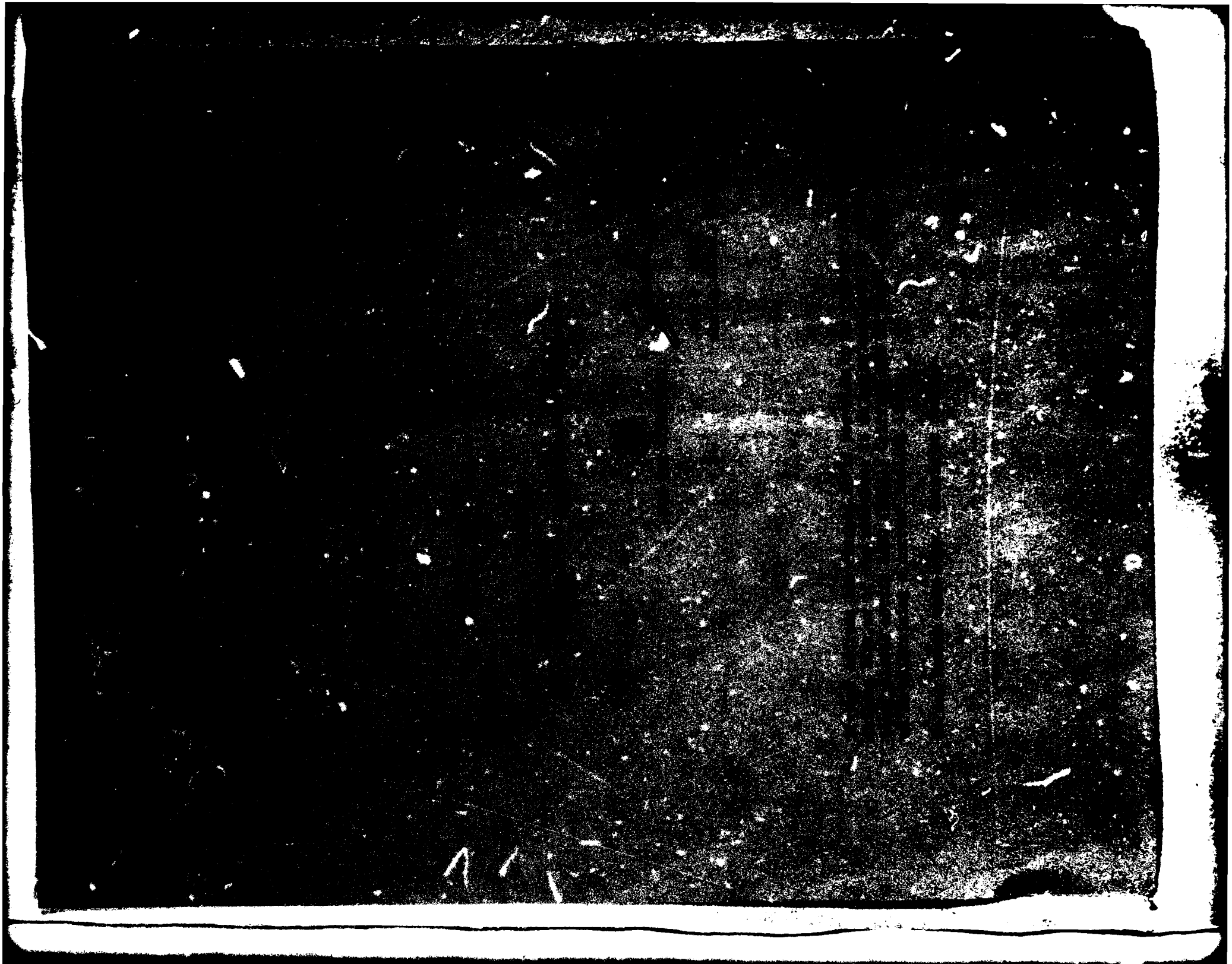
AD-A086 987 PATTERN ANALYSIS AND RECOGNITION CORP ROME NY F/G 15/4
AGGREGATING AND COMMUNICATING UNCERTAINTY.(U)
APR 80 J M MORRIS, R J D'AMORE F30602-78-C-0291
UNCLASSIFIED PAR-79-1 RADC-TR-80-113 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A


AD A086987



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

18 | 1. REPORT NUMBER: RADC-TR-80-113 | 2. GOVT ACCESSION NO.: AD-A086987 | 9 | 3. RECIPIENT'S CATALOG NUMBER:

6 | 4. TITLE (and Subtitle): AGGREGATING AND COMMUNICATING UNCERTAINTY. | 5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report.

6. PERFORMING ORG. REPORT NUMBER: PAR Report 79-71

10 | 7. AUTHOR(s): John M. Morris, Raymond J. D'Amore | 15 | 8. CONTRACT OR GRANT NUMBER(s): F30602-78-C-0291

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Pattern Analysis and Recognition Corporation, 228 Liberty Plaza, Rome NY 13440

16 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 31021L | 17 | 1L

11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRDA), Griffiss AFB NY 13441

11 | 12. REPORT DATE: Apr[illegible]80

13. NUMBER OF PAGES: 388

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): Same

12 | 385

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

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): Same

14 | PAR-79-1

18. SUPPLEMENTARY NOTES: RADC Project Engineer: Patricia M. Langendorf (IRDA)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number):
Uncertainty
Decision Analysis
Estimative Intelligence
Strategic Intelligence
Subjective Probabilities

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

Strategic intelligence products contain a substantial element of uncertainty which must be communicated to the consumer if they are to be used effectively. As a social science, estimative intelligence uses methods which are largely global or intuitive in nature. Subjective probability assessments may be produced for these methods, combined, tested for consistency, and communicated to the consumer using techniques described here. Two computer-based demonstration systems, the Calibra-

(Cont'd)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

390101

Item 20 (Cont'd)

tion Assessment Package (CAP) and Subjective Probability Assessments (SPA), are described. General descriptions of several additional systems for data base management and probability assessment are also included.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

EVALUATION

This effort is a generally successful attempt to strenghten the theoretical foundation of estimative intelligence by identifying where quantitative uncertainties in data can be validly developed with currently available procedures.

In general, the fundamental assumptions underlying classical statistics do not hold in intelligence data analysis, and classical estimates of uncertainty are generally inapplicable to finished intelligence. This is a serious well-documented deficiency since intelligence consumers need information concerning uncertainty to effectively use estimates in crisis management, strategic planning, tactical planning, gaming & simulation and intelligence quality control.

The effort provides effective media for identifying and communicating the total uncertainty for specific intelligence estimates given the uncertainties of the data making up the estimate can be assessed. It also addresses communicating the significance of that uncertainty to decision makers. The results of this effort are being included in orientation classed for defense intelligence estimators.

Valid approaches to expressing uncertainty in terms of possibility theory and as non-parametric statistics are briefly discussed. Further development of these approaches requires basic theoretical research beyond the scope of the contract. Support to pursue fundamental research identified by the contract is being requested from AFOSR. The research is fundamental to improving intelligence data analytical methodology and credibility.

Patricia M. Langendorf

PATRICIA M. LANGENDORF
Project Engineer

SECTION 1

INTRODUCTION

"Estimating: An effort to appraise and analyze the future possibilities or courses of action in a situation under study and the various results or consequences of foreign or United States actions relating to that situation. This analysis of such a foreign situation would consider its development and trends to identify its major elements, interpret the significance of the situation, and evaluate the future possibilities and prospective results of various actions which might be taken ..." (Lyman Kirkpatrick, The Intelligence Community.)

Research to be reported here is intended to identify the disparate sources of uncertainty affecting defense estimates, define quantified measures of the uncertainties whenever possible, aggregate the uncertainties into total measures of uncertainty for projections, and provide effective media for communicating the total uncertainty and its significance to national decision makers.

Work was performed under contract #F30602-78-C-0291 and monitored by RADC/IRDA. It draws on research performed in the development of the Trend and Error Analysis Methodology System (TEAMS) under contract #F30602-76-C-0206.

The effort was intended to identify, quantify, aggregate and communicate uncertainties in defense estimative intelligence projections in a form which

is understandable to estimators with minimum technical and statistical background. The aggregated output uncertainty measures and the media used to present them are designed to be meaningful at the level of the national decision maker. Media are provided for the training of non-technically oriented estimators to the level required to use and understand the developed methodologies. The scope of this effort includes uncertainties associated with current methods of making defense intelligence estimates; it does not include the development of improved estimative intelligence methodologies.

Research is based upon three types of investigation:

1. Interviews conducted with Defense Intelligence Agency Directorate of Estimates (DIA/DE) personnel specifically for this project and for the preceding TEAMS project.

2. A review of research documents dealing with strategic intelligence methods, statistical decision analysis, and related areas, as listed in the bibliography provided as Appendix A to this report.

3. The application of techniques described in the literature to the estimative intelligence methods as defined by the interviews.

The purpose of this Final Report is to review and present the research which constitutes this effort. Two additional products are provided:

1. A Training Manual for the use of DIA/DE in teaching methods for quantifying, aggregating and communicating uncertainties in estimative intelligence products. A copy of this Training Manual is included under separate cover as Appendix B of the Final Report.

2. Computer programs for the demonstration of techniques for aggregating and communicating uncertainties. These are being provided in machine-readable form to DIA/DE, together with appropriate documentation, as a separate data item.

1.1. SCOPE OF THIS REPORT

The quotation with which this report begins suggests the scope of estimative intelligence: appraisal and analysis of future possibilities, and of the results of prospective actions. Although it emphasizes future developments, estimative intelligence shares its primary goal with other types of strategic intelligence: determination of the capabilities and intentions of a prospective adversary. Thus, an estimate is correct when it identifies the capabilities and intentions correctly, and incorrect when it does not.

This interpretation of the correctness of an estimate is not the one which is normally used. In most evaluations of intellegence estimates, they have been treated as predictions of future events and future force levels, rather than as appraisals of capabilities and intentions. When a <u>prediction</u> fails to come true, then it can be regarded as an error; but an <u>estimate</u> may have correctly identified the intentions and capabilities of an adversary,

only to find that political, economic, technical, or other changes have modified those intentions and capabilities before they could be realized. Ideally, then, evaluations of estimative intelligence should be based on the correctness with which capabilities and intentions have been identified.

Over the past five years, there has been increasing interest in evaluating the quality of DE's work, with particular emphasis on projections of future force levels for the USSR. Critiques of these projections have frequently centered around underestimates of the size and composition of Soviet ballistic missile forces. A series of underestimates was identified and widely reported in the popular press. While there is no doubt that underestimates and overestimates should be identified and eliminated, it is also important to develop better methods for indicating the uncertainty present in DE's projections, to prevent misinterpretations of the degree of confidence with which they are asserted. Estimates should not only identify the force levels as accurately as possible; they should indicate the degree of uncertainty remaining in the estimate. It is to this latter problem that the research reported here was directed.

Estimators have been asked to provide numerical measures of the degree of uncertainty present in projections of two types: (1) a range of values, generally including a high, low, and best estimate, such that roughly 75 percent of the actual values will be found to lie between the high and the low estimate; and (2) a numerical estimate of the probability that a projected event will occur.

Interviews with estimators indicated that they found it extremely difficult to describe the methods that they used in obtaining these numerical measures of uncertainty. They frequently spoke in terms of "intuition," which seemed to represent nothing more than guesswork. Figures were plucked out of the air, or they were inserted simply to fill the blanks in required reports. Estimators felt that there was little reason to develop better methods since, in their opinion, intelligence consumers made little or no use of the uncertainty figures that were included in their reports.

Measures of uncertainty are nevertheless of importance to the consumers. Decisions concerning the commitment of U.S. funds depend crucially on expected levels of development in the USSR and elsewhere. Composition and deployment of U.S. forces must be planned to meet an expected threat; the realism of such plans will depend on the probabilities attached to various potential threats. In short, in every area in which DIA estimates are used, some recognition of the uncertainties present in these estimates is required.

This report is therefore intended to provide methods which should prove more reliable and consistent than unguided intuition. In addition, methods for combining or aggregating uncertainties from a variety of sources are described. Techniques for communicating uncertainties to the consumers of estimative intelligence are also included.

This Final Report consists of twelve sections, as follows:

1. Introduction.

2. Background.

3. The uses of uncertainty. The role that uncertainty plays in the use of estimates.

4. Uncertainty and probability. A discussion of the meaning of uncertainty in terms of probabilities.

5. Estimative intelligence methods. A detailed discussion of methods in use at DE for the production of finished estimates.

6. Statistical methods. A review of decision analytic (Bayesian) and other methods which have been suggested for use in estimating uncertainties.

7. Probability assessments. Methods for obtaining probability assessments for use in reporting projections.

8. Calibrating uncertainty measures. Methods for assuring the consistency and correctness of probability estimates.

9. Detecting and eliminating bias. Some of the frequent sources of error in probability estimates.

10. Computer-assisted estimations of uncertainty. A description of some programs for estimating and aggregating uncertainties.

11. Communicating uncertainty. Methods for communicating uncertainty to the user.

12. Summary and recommendations.

Appendix A. Bibliography.

Appendix B. Training Manual (separately bound).

Appendix C. Computer-based systems for aggregating uncertainties.

Appendix D. Computer demonstration systems (separately bound).

SECTION 2

BACKGROUND

"Intelligence does not claim infallibility for its prophecies. Intelligence merely holds that the answer which it gives is its most deeply and objectively based and carefully considered estimate." (Sherman Kent, Strategic Intelligence for American World Policy.)

In common with current and basic intelligence, estimative intelligence attempts to discern the capabilities and attitudes of a potential adversary. Estimative intelligence differs from other forms of strategic intelligence in that it projects these capabilities into the future. Because future events cannot be predicted in detail - in part because of the unpredictable nature of human decisions and in part because of our human limitations upon processing the quantity of data that are potentially relevant to future events - a variable and unpredictable element of uncertainty is present in intelligence estimates.

Since these estimates nevertheless play an essential role in U.S. strategic planning, the degree of uncertainty in the estimates must be conveyed, in some fashion, to the eventual user. For this purpose, a large number of different techniques have been employed. For example, National Intelligence Estimates (NIE) and Defense Intelligence Estimates (DIE) have contained such words and phrases as these:

it is likely (or unlikely) that

it is possible

the possibility of

the likelihood of

probable

they probably believe

not likely

Such phrases, which appeared in an NIE in 1968, were intended to convey to the consumer the uncertain character of the projections contained in the estimate.

In 1976, more precise forms for expressing uncertainty were introduced into DIA/DE products, using such phrases as the following:

60 percent probability

80 percent likelihood

a 70 percent chance

Such expressions were applied to specific events, such as changes in military policy.

In some estimates issued during 1976, a colored sheet containing the following statement was inserted:

"[Numeric forms are used] to convey to the reader this degree of probability more precisely than is possible in the traditional verbal form. Our confidence in the supporting evidence is taken into account in making these quantifications. . . . All efforts at quantifying estimates are highly subjective, however, and should be treated with reserve."

At about the same time, a DIE contained the following notice:

"Completeness and Reliability of Evidence

"The evidence . . . is based on a wide variety of sources and is considered generally complete and reliable, although not necessarily definitive . . .

"There is as yet very little reliable evidence . . . The data base . . . is considered sufficiently reliable to support the judgments made . . ."

Throughout the estimate, both quantitative and verbal expressions were included:

there is a 75 percent chance that
except in the unlikely event that
it is possible that
it would probably react
we see no evidence of
it appears that
the possibility of . . . is ever present

[it] would probably

[it] could provide

we do not believe

In other words, both numeric and verbal qualifications frequently appeared in DIA estimates, even after the numeric forms had been introduced; DE felt it necessary to point out that the numeric qualifications were "highly subjective"; other caviats were included to assist the consumer in assessing the degree of confidence appropriate for the estimate.

The Defense Intelligence Projections for Planning (DIPP) have, for several years, provided another means for communicating uncertainty to the user. Numeric estimates frequently (but not invariably) include a "high," "low," and "best" value. These are selected in such a way that the true value will be found in the indicated range approximately 75 percent of the time. Although no attempt is made to specify the probability distribution over this range, the spread (from low to high) is intended to assist the user in determining the uncertainty present.

Several other methods for communicating uncertainty are reviewed in subsection 5.10. But while DIA has made an effort to communicate the degree of uncertainty to be attached to estimates, such problems as the following were reported to us by DE personnel:

- The numerical estimates of probabilities, generally reported as percentages, were indeed "highly subjective." No clearly-defined methods for obtaining the required numbers have been specified.

- There was little motivation to improve the quality of the probability figures, since they were generally ignored by intelligence consumers. Consumers appear to take the best estimates (or occasionally the high estimates) without any apparent regard for the uncertainty attached to them.

A vicious circle thus appears to have developed, in which neither the estimators nor the consumers take the probabilities very seriously. Estimators tend to regard them as mere guesses, and consumers tend to ignore them.

The goal of this project, then, was to review the methods that DIA-DE uses in preparing estimates, and to locate the types of uncertainties that enter into them. Methods for combining or aggregating these uncertainties were to be proposed. Finally, more effective methods for communicating uncertainties to intelligence consumers would be developed.

SECTION 3

THE USES OF UNCERTAINTY

"The main task for modern philosophy is to teach man to live without certainty and yet not to be paralyzed by hesitation." (Bertrand Russell, History of Western Philosophy.)

The first two sections have noted some of the difficulties that intelligence producers have experienced in communicating the uncertainty of their estimates. In this section we ask: Why is it important to communicate uncertainty to the intelligence consumer? What purpose does it serve -- or (since it is frequently ignored) what purpose should it serve?

The role of the intelligence consumer is a little like that of a traveler on a Mississippi riverboat, who finds himself involved in a poker game with a professional gambler. In this situation he is faced with uncertainties at many levels:

- Information may be concealed: the riverboat gambler may have cards up his sleeve.

- Information may be falsified: the gambler may be dealing from the bottom of the deck.

- Chance events may affect the outcome: even in a fair game, there is uncertainty concerning the deal.

- The gambler may change his strategy at any time: part of the essence of an effective combat posture is its unpredictability.

- All's fair in love, war, and cut-throat gambling: the gambler cannot be expected to abide by the rules of the game, or by any subsequent agreement, unless it is to his advantage to do so.

Any reasonable traveler would refuse to take part in a game like this. But in the context of world affairs, there is no alternative other than to play; the problem is learning how to play the game well. And in the world today, we are playing against professional gamblers, who may be expected to cheat, conceal, and lie whenever they think they can get away with it.

According to the theory of games and decisions, which serves as a background theory for much of the analysis presented in this report, two major factors enter into a rational person's decision: the probability of each expected outcome, and the gain or loss that can be expected for each outcome. The rational person will choose those actions which will, with greatest probability, maximize the chances of obtaining a gain or avoiding a loss. Such a person is willing to take very great risks to obtain a very large gain or to avoid a large loss; but if the outcome makes little or no difference, then the rational person will remain indifferent about choices.

In this context, the intelligence producer has the task of communicating information to the consumer that will serve as the basis for a rational decision. Such information can include estimates of the extent and deployment of enemy forces, assessments of their plans and goals, and other material which may be relevant to a decision. Because the adversary will attempt to conceal information or to circulate misleading information, the estimates of the intelligence producer are subject to error. In the case of estimates like those of DE, which may be projected for ten or more years into the future, the possibilities of major changes in enemy policy and in the technology of warfare introduce additional elements of uncertainty.

To develop more effective methods of communicating uncertainty to the consumer, it is essential to begin with the purposes that these methods are intended to serve. Throughout this report, we assume that the consumer will use information concerning uncertainty in such ways as the following:

- Strategic planning. DIA estimates for future military developments in foreign nations are essential for U.S. strategic planning. Realistic plans require some knowledge of the uncertainty to be attached to projections. For example, a decision concerning development of a major ABM system for the U.S. must include information concerning not only the nature of prospective Soviet missile systems, but the probabilities associated with each type of system. The intelligence consumer must make a judgment concerning the types of ABM systems, if any, to be developed, and such judgments cannot be made rationally without some estimate of the probability of

various Soviet missile postures. Uncertainty for such consumers should be communicated in the form of probabilities for discrete events, or of probability distributions for quantified estimates. In this way, the consumer can determine the need for U.S. actions which will respond to the most likely enemy posture, or to a posture which is somewhat unlikely but extremely threatening.

- Gaming and Simulation. Estimates may be used as inputs to gaming programs for training personnel and for simulating hypothetical military activities. In both types of activity, a numerical probability is required. For gaming, the probability might be used in conjunction with a randomizing procedure, such that the given event or weapon development would have a simulated occurrence with the specified probability, or with the specified probability distribution. In the simulation of hypothetical military activities or of major international developments, the probabilities would be used to determine the overall probabilities of various outcomes.

- Self-evaluation. Intelligence estimates are always subject to review, for the purpose of detecting problem areas and for improving the quality of estimative methods. Projections containing verbal phrases like "It is probable that" or "There is some probability that" are difficult to evaluate for quality control purposes, since no outcome could show them to be either true or false. Numerical probabilities, or probability distributions, however, can be evaluated and scored. Scoring rules, like those developed for the TEAMS

project, help to show precisely the degree to which a probabilistic projection is either right or wrong. Even if scoring rules are not applied, the use of numerical probabilities (rather than verbal phrases) permits an evaluation of projections to determine which of them were more or less correct. For example, if you had said that there is a 90 percent chance that System X would be developed and deployed by 1980, and if System X was deployed by that date, then you would be more correct than if you had said that there was a 60 percent or 20 percent chance.

In each of these applications - planning, gaming, and evaluation - we see that numerical statements of probabilities are more likely to be useful than verbal statements of uncertainty (like "It is probable that"). Numerical presentations of uncertainty can take various forms, such as:

- An 80 percent probability.
- A probability of 0.80.
- Odds of 8 to 2.

Since each of these forms can be reduced to the same numerical form for computation, there is no mathematical reason to choose among them. Experimental evidence suggests that the last form, using odds, may be somewhat easier for the non-expert to grasp. Throughout this report, however, we will use any of the above forms interchangeably, without attempting to resolve the subtle nuances of meaning that each conveys.

It is therefore important to use numerical probabilities in order to communicate the precise degree of uncertainty in estimates. To say, for example, "It is completely uncertain whether Primorye class intelligence ships will have facilities for processing of electronic signatures" could convey valuable information to the consumer: that these facilities may be installed on the ships, and that a long-range plan for countering them might appropriately be considered. Nevertheless, it is considerably more useful to the consumer to convey an accurate measure of your degree of credence by saying something like "There is a 50 percent probability that the Primorye class intelligence ships will have facilities for processing electronic signatures." This assists the consumer in assigning a degree of urgency to the long-range plan: a 50-percent probability suggests a greater degree of urgency than would a 10-percent or 20-percent probability. In addition, the numerical probability could be used in a rigorously-defined mathematical model for such purposes as cost-benefit analysis, while a purely verbal expression could not.

In recommending the use of numerical probabilities, however, this report does not make further recommendations concerning estimative methods. No new methods for developing intelligence estimates are proposed, since these would be beyond the scope of the research reported here. In particular, this report deals with the existing, rather informal methods now used by intelligence estimators. Sections 6 and 10 will, however, provide an introduction to statistical methods which have been proposed for use in producing and evaluating estimates.

In describing a computer-based study of the future of the Badger bomber, one person stated the central problem with which we will deal here:

"One would think that by this time everything about the Badger would be known. Unfortunately, this is not the case. All we know is what we have been able to physically measure by counting what we saw and heard. The causes of these phenomena are still unidentified. Hence, predictions of the future are fuzzy sets relying on the past. The heart of the problem seems to have been predictions. Even with the 'certainty' of the past, it has been an uncertain thing as to how one should apply it to the future." (Thomas H. Murray, "The Future of the Badger Bomber -- A Study in Information Science Techniques," 1974.)

Uncertainty concerning the future of the Badger thus includes (1) uncertainty concerning present deployment and capabilities of the Badger; (2) uncertainty concerning future deployments; and (3) uncertainty concerning the methods to be used in estimating the future.

These are just the uncertainties that the traveler faces in his conflict with the riverboat gambler. He does not know the cards in the gambler's hand or their order in the deck; he does not know how the gambler will bet or play; and he is not very sure of the way in which he should estimate the gambler's chances of winning. But there is one difference between U.S. intelligence consumers and the traveler: the traveler could always withdraw from the game.

SECTION 4

UNCERTAINTY AND PROBABILITY

". . . in each (situation) there is an uncontrollable random event inherent in the situation. The distinction between a risky situation and an uncertain situation is that in the former uncontrollable random event comes from a known probability distribution, whereas in the latter situation even the probability distribution is unknown." (Madansky, "System Analysis and Policy Planning Applications in Defense")

The proper definition of "probability" has been a subject of controversy from the time that probabilistic methods were first introduced in the Seventeenth Century. From a purely subjective point of view, it might be said to measure the degree of credence that we place in some hypothesis or other proposition. If we believe very strongly in something, then we ascribe a high probability to it -- from the subjective point of view.

Clearly, our subjective probability may be wrong. I may believe very strongly in my chances of winning in a game of poker against the riverboat gambler; but an objective observer would have to say that my chances are much smaller than I think they are. In general, as I look back over my lifetime, I can think of many times in which I believed very strongly in something, only to have it turn out to be false. For this reason, I generally look upon my own strong beliefs, and especially the strong beliefs of other persons, with a good deal of skepticism.

A major contribution of the Seventeenth Century probabilists was an alternative approach to the measurement of probabilities, which is objective in nature rather than subjective. It takes two forms, which we will label "analytic" (or "a priori") and "synthetic" (or "a posteriori").

The analytic approach is based upon the definitions of the objects or entities involved. For example, if we define a "fair die" as one which is equally likely to come up with any one of its six faces showing, then it follows logically and mathematically that the chance of any one face (such as the four) coming up is 1/6. This is a logical consequence of our definition of "fair die." If the chance of the four coming up were anything other than 1/6, then it wouldn't be a fair die.

We can, of course, test any specific die to find out whether it is fair. We can throw it a hundred times, and count how many times it comes up one, two, three, and so on. If, on every one of the hundred throws, it comes up six, then we can say, "It is not very likely that this is a fair die." And we can easily compute the likelihood that it is fair; this probability is $(1/6)^{100} = 1.5306 \times 10^{-78}$, which is a very small number. A die which came up six in every one of one hundred tosses, then, would not be likely to be a fair die.

The other approach to measuring probabilities is the synthetic approach. Instead of beginning with the definitions, it begins with a count of the proportions present in a population. Since this approach has been used by actuaries for determining insurance rates, it can also be called an "actuarial" approach.

Suppose that a person has received 10,000 Christmas tree lights for decorating the office. How many of them are faulty? Without attempting to test the entire lot of them, he or she decides to test 100 of them, to get some idea of the number of faulty bulbs to expect. Suppose that 10 of the bulbs refuse to light, or burn out immediately. Then the best guess concerning the entire lot of bulbs would be that the same proportion, or 10 percent, would be faulty.

Of course, this example is much too simple, because we would also want to know - to determine how many spare bulbs to order how likely it is that 15 percent of the total might be faulty, or 20 percent, or some other proportion. A statistician could easily provide a reply to these and many other questions.

Three approaches to the measurement of probability, then, are: (1) a subjective approach, measuring our degree of belief; (2) an analytic approach, based on the definitions of the entities involved; and (3) a synthetic approach, based on a statistical investigation of the behavior of similar entities in the past.

All three types of probabilities play a role in determining the uncertainty of intelligence projections, but this report will concentrate on the first type, "subjective" probabilities, because they are most useful for intelligence estimates; they are also most controversial, because they are difficult to measure, may differ from person to person, and are difficult to evaluate, to determine whether they are right or wrong. Where they are

available, mathematical and statistical probabilities (the second and third types) should certainly be used in measuring and communicating the uncertainty of intelligence estimates. For example:

- The known resolution accuracy of an aerial camera gives a precise range of error in the estimation of the length of a Soviet missile, photographed from a satellite at a known height. For any photograph, there is a statistical distribution of possible lengths of the object photographed. On the basis of this information, we could, for example, determine the probability that two photographs represent missiles of the same length, or missiles of two different lengths. This result would not be based upon a large-scale statistical survey of missile photographs, but upon the characteristics of the equipment involved. It would therefore represent a probability distribution which used the second, or analytic, approach.

- Barracks are under construction near a new Chinese factory. Previous (hypothetical) experience, including accurate counts of personnel at 100 other Chinese factories, indicates that the Chinese provide 20 square feet of floor space per person in their barracks. This factor may therefore be used in estimating the number of persons to be employed in the factory. In addition, previous experience has shown a degree of variability in the amount of floor space allotted; knowledge of this variability permits us to set upper and lower bounds on our estimate of the number of

persons to be employed. Use of this approach to the estimation of probabilities would represent the statistical, actuarial, or synthetic approach.

- Reports from a government agency indicate that the Soviets are developing a new type of radar specifically to detect and track U.S. cruise missiles. The agency estimates that 100 such radars will be deployed and operational by 1985. In checking their report, a large number of unanswered questions are found concerning the accuracy of their information and the validity of many of the inferences that they have drawn. We are willing to say only that there is some chance say 40 percent that the radars will actually be installed by the target date. We cannot, of course, base this figure on any largescale statistical study of radar installations. And we are not using any techniques of mathematical analysis to arrive at a welldefined number. Instead, we are saying that we think that there is some possibility that the installations will be completed, but we feel that there is something less than a fifty-fifty chance that they will be. We are stating, in short, a subjective probability.

Interviews with DIA estimators indicated that probabilities of of this type subjective probabilities were far more frequently used than probabilities of the other two types. For this reason, we will concentrate upon subjective probabilities in this report. Because the word "subjective"

carries connotations of guesswork, we generally use the term "probability assessments" to refer to them. As noted in Section 8, probability assessments can be calibrated in such a way as to permit their use in a consistent, well-founded manner.

Many of us are nevertheless hesitant about assigning probabilities to individual events, because it is difficult to determine precisely what is meant by such probabilities. Suppose, for example, that you yourself are playing solitaire with your own deck of cards. You shuffle it several times, cut the deck, and place the top card face-down on the table. What is the probability that the card is the ace of spades? Most of us would say that it is $1/52 = 0.019$. We have no serious problem in estimating this probability which is an "analytic" probability based on the definition of the card deck, and of a random draw from such a deck.

Next, suppose that the card comes from an unfamiliar deck, which belongs to a riverboat gambler. He has shuffled and cut the deck himself. He is wearing a baggy coat that could easily conceal some extra cards. And you stand to lose a substantial amount to him, if you fail to guess correctly. Now, what is the probability that your guess will be correct?

Obviously, the probability in the second example is much more difficult to estimate than the probability in the first example. There are an indefinitely large number of factors which may be relevant to the estimation, including unknown factors such as the possible presence of a conspirator

among the onlookers. A really clever opponent will be attempting to find exactly those ruses that you have neglected to identify.

It should be clear that the situation faced by the intelligence estimator is considerably closer to the second example than to the first. Our potential adversaries have absolutely no reason to play a "fair" game, if it is not to their advantage to do so. They may be expected to take advantage of every opportunity for concealment or misrepresentation of their capabilities and intention.

Because of the large number of elements that can serve to increase or decrease the probability of an intelligence estimate, it is rarely possible to rely on mathematical or statistical probabilities. Instead, the judgment of an experienced intelligence producer, who can take into account the many factors that may affect the probability of an estimate, must be used.

A subjective probability represents an assessment of the chance that a given proposition will be found to be correct. Like other probabilities, it is expressed as a proportion, in the range from 0.0 to 1.0. A subjective probability can be correct or incorrect, depending on the degree to which it is well-calibrated. Calibration is defined in terms of a statistical probability: over a large number of subjective probabilities, if we have assigned a probability of 0.70 to some propositions, then 70 percent of them should be found to be correct; and similarly for other probabilities. If these proportions hold, then we are said to be well-calibrated; if they do not hold, then

we are biased toward conservatism (if we underestimate probabilities) or toward anti-conservatism (if we overestimate them).

You are well-calibrated, then, if you do a good job of evaluating the quality of the information that you have, and if you have a realistic sense of your own ability to examine and to integrate this information.

Much of the remainder of this report is devoted to the development of methods for producing well-founded probability estimates. In particular, Section 7 presents an approach to the estimation of probabilities, Section 8 includes methods for calibrating probabilities, and Section 9 provides an approach to the elimination of bias.

In this section, we will suggest two general approaches to the quantification of uncertainty, the holistic and the compositional:

The holistic approach deals with wholes, the organic, inclusive structures of events. In artificial intelligence applications, these wholes are sometimes called frames, scripts, or scenarios; in Section 11 of this report, we will use the term "scenario" to refer to the inclusive structure of hypothesized future events, which fit together to form a consistent whole.

Using the holistic approach, the intelligence producer provides a probability for the scenario as a whole; based on this overall estimate, figures for the individual components can be derived. Several examples of such

scenarios are provided in Section 11. Here, we will use a somewhat more artificial example.

It is well known that many of the more hawkish forces within the Soviet Union believe that a large-scale nuclear war can be actually be fought and won. On the basis of this belief, they may be expected to emphasize those elements of the Soviet military structure that would seem to make such a war possible. Offensive missiles in concealed, hardened locations might be among the elements of this strategy. An increase in submarine forces might also be considered, with deployments which would permit effective launches of SLBMs against the U.S. continent at a moment's notice. Civil defense equipment would be maintained, and training would help to insure survival of the civilian population during a U.S. retaliatory strike.

A scenario would contain the details of this plan. Prepared by U.S. intelligence personnel, it would begin with the overall approach, and would contain the actions and developments that would be essential in carrying the Soviet plan into action. Since the scenario begins with the overall plan, we call this approach "top-down"; it begins at the most general level of a plan, and works down to the smaller details.

Probabilities are next assigned to the plan in a top-down fashion. Based on U.S. knowledge of the composition of leadership in the Soviet Union, and upon a general view of Soviet intentions, a probability figure is obtained for the total scenario. Next, probabilities can be assigned to each of the

major components of the scenario. For example, if the Soviets are preparing for a major offensive nuclear war, then the probability is very high that they will develop an effective civil defense structure.

Calculation of the probabilities for the elements of the scenario are straightforward. For example, suppose that the probability of an overall Soviet plan for aggressive nuclear war during the next five years is 0.35. Suppose that, if such a plan were implemented, then an increase in civil defense allocations carries a probability of 0.90. We may now calculate the unconditional probability of an increase in civil defense as 0.35 x 0.90 = 0.315.

Of course, we may know from other sources that civil defense is being emphasized in the Soviet Union. This means that the actual probability that we attach to this development is greater than 0.315. In its pure form, however, the top-down, holistic approach derives these probabilities only from the probabilities attached to the top-level scenarios, and the conditional (if-then) probabilities that are included in the scenarios.

The compositional approach, which will be described in Section 6, begins with individual events. It could be called a "bottom-up" approach, because it begins with the low-level individual developments and works upward to the most general ones.

Probabilities are estimated for specific events (such as a Soviet decision to develop a cruise missile), and a probability range is estimated for

a quantitative projection for a single weapon system. These probabilities are then combined to obtain higher-level probabilities -- obtaining, for example, a probability distribution for all offensive missiles, then for all missiles combined, and finally a figure indicating the total combined strength of all Soviet military resources.

The compositional approach is often used in decision analysis; because it frequently relies on Bayes' theorem, it is sometimes called "Bayesian analysis." This approach has been extensively studied, and it is supported by several computer-based systems. The Bayesian approach is briefly described in Section 6, and a related non-Bayesian approach is outlined in Section 10.

Other approaches to the quantification, aggregation, and communication of uncertainty are possible, of course. Information theory, for example, defines uncertainty in terms of an analogue of entropy, and provides statistical methods for combining uncertainties. (See Shannon, Claude E., and Weaver, Warren, The Mathematical Theory of Communication, Urbana: University of Illinois Press, 1949, pp. 51-53.) Briefly, our feeling has been that insofar as information theory is applicable to the problems addressed in this research, its methods are consistent with those of decision analysis. While it would certainly be interesting to work out the details of an information-theoretic approach (in which uncertainty = entropy, for example), it would be confusing to attempt to use this terminology in the final report.

In Section 3, we have attempted to show that some method for communicating uncertainty to intelligence consumers is required, and that this

method should be numerical, rather than verbal, in form. Furthermore, we believe that the most useful numerical form will be that of probabilities, which range from 0.0 to 1.0. In this range, 0.0 represents the subjective probability corresponding to total disbelief, absolutely no credence, while 1.0 represents total belief or credence.

SECTION 5

ESTIMATIVE INTELLIGENCE METHODS

"What is called 'foreknowledge' cannot be elicited from spirits, nor from gods, nor by analogy with past events, nor from calculations. It must be obtained from men who know the enemy situation." (Sun Tzu, Fifth century B.C. Chinese sage.)

This section reviews methods currently used by DIA-DE for the production of estimative intelligence, with emphasis on methods used for identifying and communicating uncertainty. It is based on interviews with DE estimators conducted on 1-3 November 1978 in connection with this project. Material is also drawn from interviews of 26 May 1976, which were conducted in connection with Contract No. F30602-76-C-0206 (TEAMS). Background material has been drawn from literature as cited.

It should be emphasized that the methods described in this section are intended to represent the approach which is actually in use at DE, not those methods which have been recommended by others. The development of new methods was specifically excluded from the scope of this project. However, the methods described here are synthesized and generalized from interviews with several different estimators who employ a variety of approaches; they therefore represent a somewhat idealized picture of the estimative process.

5.1. STRATEGIC INTELLIGENCE METHODS

The type of reasoning employed in strategic intelligence may be illustrated by this brief anecdote:

"While working on Eisenhower's scientific advisory committee in 1959 and 1960 I had to assess some of the early claims that the Russians were developing an ABM system. The Soviets, we knew from our intelligence, had a center for antiaircraft and antimissile work at Sary Shagan in Central Asia. Our U-2 planes observed there a large radar installation that might, it was thought, be a device for detecting incoming missiles. Our intelligence experts immediately linked this installation to the Soviet tests of medium-range ballistic missiles at Kapustin Yar, many hundreds of miles to the west. They conjectured that the Russians were putting together the combination of radar (to detect incoming missiles), computers (to track them), and interceptor missiles (to destroy them) that makes up an ABM system." (George B. Kistiakowsky, "False Alarm: The Story Behind Salt II," New York Review of Books, March 22, 1979, pp. 33-38.)

A specific hypothesis is required, which is suggested to the analyst by observations. In addition, inferences are needed to link the observations to the hypothesis. Finally, the hypothesis is verified by supporting evidence.

Note the crucial role of the hypothesis -- in this case, the claim that the USSR was developing an ABM system. It is this hypothesis that makes sense of the various observations, such as the simultaneous development of a radar

installation and tests of ballistic missiles. No mention is made in this anecdote of the verification (or validation) of the hypothesis. (Verification would mean the gathering of supporting evidence; validation means the gathering of conclusive evidence.) Verification and validation are not always possible for hypotheses in strategic intelligence.

Another approach to verification would be the development of alternative hypotheses. Can we find some other conjecture that would do an equally good job of explaining the evidence that we have? If so, are there any further tests that we can perform that would help us to choose between them? If no other hypothesis can explain the available data, then the given hypothesis may be regarded as verified (or validated, if conclusive evidence shows that other hypothesis is tenable).

Sources of uncertainty in this anecdote may be identified:

- The initial identification of the center at Sary Shagan as a center for antiaircraft and antimissile work is based on a train of reasoning which is here omitted; nevertheless, the identification may be incorrect.

- The identification of construction at this site as a radar installation is somewhat uncertain, although it appears to be highly likely; nevertheless a misidentification, due perhaps to Soviet deception, is possible.

- On the other hand, the surmise that the radar installation might be intended to detect incoming missiles plays a somewhat different logical role. It provides negative evidence for the hypothesis: there is no clear sign that the radar installation is going to be used for some other purpose.

- Another source of evidence, and of uncertainty, is in the identification of Soviet tests of medium-range missiles at Kapustin Yar. In this case, of course, the degree of uncertainty is minimal; it would be difficult to fake a missile test.

- The most significant source of uncertainty is in the completed hypothesis itself, that the Soviets were developing an ABM system. This claim must be shown to make sense of the various Soviet actions and installations that have been observed, and it will in the end depend on a comprehensive understanding of Soviet strategy, internal politics, economic structure, technical capabilities, and, in short, the total Soviet system. At the same time, if the given hypothesis can be shown to fit within the overall picture of Soviet aims and capabilities, and to be consistent with the specific observations, then it can be advanced with as high a degree of certainty as can be obtained in the social sciences. Thus, the hypothesis <u>makes sense</u> in the context of a larger set of hypotheses concerning the Soviet system.

5.2. THE NATURE OF ESTIMATIVE INTELLIGENCE

Like other forms of strategic intelligence, estimative intelligence is concerned with the determination of the capabilities and intentions of a potential adversary (and, in the case of the NATO nations, of an ally). Estimative intelligence differs from current and basic intelligence in that the capabilities and intentions must be projected into the future, where (at least for finite human beings) events are indeterminate. In practical terms, this means that the intentions of a potential adversary may change in unpredictable ways over the course of the next ten or twenty years. The rise of a Sadat or a Khomeini in the Middle East, for example, has brought about changes in the intentions of the nations that they represent. While it is possible, thanks to hindsight, to identify those forces within Egyptian or Iranian society that gave rise to the policies of Sadat and Khomeini, it would have been wildly speculative to have predicted them ten years in advance. Similarly, new inventions and discoveries may contribute to basic changes in the capabilities of an adversary. Thus, many technological developments which are operational today could not have been realistically foreseen ten or twenty years ago.

Projections over a limited time -- over perhaps as much as five years -- can be made with some assurance, on the basis of extrapolations of known technology, known production capabilities, and reasonable assumptions concerning intentions. Beyond this point, however, assurance drops off rapidly. It may not be known, for example, when a given weapon system will be regarded as obsolete, and major errors may occur in predicting this point. Because of the

uncertainty of projections of the future, estimative intelligence contains a degree of uncertainty which is not present in current and basic intelligence. Nevertheless, it is necessary to assess the intentions and capabilities of foreign nations in an effort to determine, as precisely as possible, what the future will bring.

5.3. THE NEED FOR CREDIBILITY

Another factor which affects the methods that DE uses is the need for estimates which represent a consensus of the intelligence community concerning the future development of foreign weapon systems. DE personnel are required to meet and consult with representatives of other intelligence agencies. The central role of these meetings is essential to an understanding of DE's approach.

Estimators must justify their projections in a dialogue with other intelligence personnel. It is not enough merely to hold and vote upon an estimate. To obtain a consensus, it is necessary to convince other representatives that an estimate is correct. A convincing line of argument must be developed and must be defended against contrary arguments.

In short, DE must not only produce estimates, but must produce defensible estimates. It is this need for justifying estimates that provides the incentive for developing well-founded methods in intelligence production. These methods are essentially those of the scientific method, based on hypothesis

generation and testing, as they have been applied to research in the social sciences.

In practice, this process would appear to work somewhat as follows. A hypothesis, such as a date for withdrawal of the Badger, is proposed. Arguments for or against the hypothesis are considered: the Soviet tendency to retain obsolescent equipment, the record of success of the aircraft, lack of evidence of new equipment to replace the Badger, the general need for aircraft with these capabilities within the Soviet defense system, outstanding orders from satellite nations. Alternative hypotheses are considered and evaluated. A selection is then made from among the competing hypotheses, which serves as the basis for a defensible projection.

Many discussions of the estimative process, particularly those that are critical of the intelligence community, appear to suffer from "hindsight bias." In Section 9, a full discussion of "hindsight bias" will be presented. Briefly, this is a form of second-guessing in which intelligence personnel are criticized for their failure to predict a given event or development. In some sense, it is said, they should have known that the Soviets would invade Czechoslovakia, or that the Shah would be overthrown. From a contemporary vantage point, it is possible to look back on the plethora of indicators that were available to intelligence officers, from which they should have been able to foresee the future.

As Roberta Wohlstetter has pointed out in her classic study, Pearl Harbor: Warning and Decision, actual prediction of a future event is far more

difficult than second-guessing after the event. Mistaken presuppositions concerning the enemy's intentions, erroneous information concerning their capabilities, and an overwhelming supply of "noise," or irrelevant information, all tend to distort and confuse our vision of the future. Hindsight bias, then, is the tendency to suppose that prediction is really much easier than it is.

In the context of this section, recognition of hindsight bias is needed to avoid the temptation to succumb to it -- to suppose that unusual or unexpected events can actually be predicted, and that the role of estimative intelligence is to prophesy the occurrence of unusual changes in the intentions or capabilities of our potential adversaries.

Estimators see their role, however, as the development of credible projections of the intentions and capabilities of the Soviet Union, the PRC, and other foreign powers. The credibility of their projections is essential; repeated attempts to predict exotic or unexpected happenings would soon result in a "cry wolf" response from intelligence consumers. After a few incorrect predictions of cataclysmic events, the consumers would tend to disregard further predictions.

The need to maintain credibility therefore introduces a healthy conservatism into DE's methods. Projections must be justified on the basis of known information and acceptable models (where the "models" are representations of the social structure, intentions, military forces, and other relevant elements of Soviet, Chinese, or other nations). The need for justification of a

projection thus determines the methodology used in the production of estimative intelligence.

5.4. THE ROLE OF UNCERTAINTY

To encourage a realistic understanding of its projections, DE has used several techniques to communicate the uncertainty present in them. Where the justification of a projection indicates some degree of doubt, this doubt must be communicated to the user. Later information should show that the more doubtful projections are less accurate, on the whole, than those that are stated with a greater degree of assurance. In this way, the credibility of the estimative process may be maintained.

In Section 3, we found that the role of uncertainty in DE's projections is more important than that of simply providing a hedge against possible errors, and thus maintaining credibility. It should be possible for intelligence consumers actually to make use of information concerning uncertainty, in such ways as the following:

- In the development of games and simulations, some estimate can be made of the probabilities of various alternative scenarios. For example, if it is uncertain whether a Soviet ABM capability will be developed (with specified characteristics), then the corresponding scenario will be equally uncertain.

- o In recommendations for R&D developments to meet a projected Soviet threat, some estimate of the uncertainty of the threat would help to determine the urgency of the recommendation.

- o Similarly, in recommendations for deployment of a given U.S. weapon system, it is essential to determine the degree of certainty to be attached to projected Soviet developments in related areas.

It is important, then, that some indication of the uncertainty to be attached to DE projections be provided for the guidance of intelligence consumers.

5.5. A DEFINITION OF UNCERTAINTY

In this report, the "uncertainty" of a predicted event has been defined as one minus the estimated probability that it will occur. In the case of quantitative predictions, uncertainty may be defined in terms of a confidence interval, in which the probability that the quantity will lie within the range has been estimated. As noted in Section 4, "estimated probability" is also called "subjective probability" or "personal probability." The determination of estimated probabilities will be treated more fully in Section 7 of this report.

We may, of course, be somewhat uncertain about the estimated probability; we may be uncertain about how uncertain we are. And we may be uncertain about <u>this</u> level of uncertainty, and so on, to any level of meta-uncertainty. These

cascading uncertainties could threaten any system for the measurement and communication of uncertainty. To avoid this difficulty, we will simply ignore it here; the uncertainty in a projection is whatever the estimator says it is. Later (in Section 7) we will want to explore this problem much more carefully.

In some of the Army intelligence literature, "uncertainty" is contrasted with "risk." Risk occurs when a decision is made, knowing the probabilities involved; uncertainty is present when we do not know the probabilities. In the context of this report, "uncertainty" will apply to both situations, since we may or may not be able to obtain an accurate estimate of the probability of various alternatives.

5.6. MISSING AND ERRONEOUS DATA

By "missing data" we mean those pieces of unknown information which would be relevant to a given projection if they were known. Like the missing pieces in a picture puzzle, they may be clearly identifiable as missing: Are these multiple warheads independently targetable, or aren't they? On the other hand, some data may be totally unknown: The Chinese are constructing an underground testing facility at point X, about which we know nothing whatever. Finally, some data may be unknown because they have not yet occurred, they are part of the future: A coup d'etat overthrows the Chinese regime and re-installs the radical policies of Mao Tse-tung, thus greatly modifying Chinese policies toward the U.S.

Our general understanding of the nation helps to provide a context which will limit the effect of missing data. Like a partially-completed picture puzzle, it provides a general picture of the policies and capabilities of a potential adversary. While we may not know the details of a specific meeting in the Kremlin, we can at least gain some idea of what would happen at such a meeting, based on our general knowledge of Soviet attitudes, combined with all the information that we do have concerning Soviet activities before and after the meeting.

In short, the scientific method requires that we account for all the available data within the context of a general hypothesis concerning the phenomena that we wish to investigate. As new data are obtained, they tend to verify our tentative hypothesis; or they may cause us to modify or reject it. The role of missing data, then, is to increase the uncertainty present in our general model; if all data were missing, uncertainty would be total, and if no data were missing, then there would be no uncertainty.

Another source of uncertainty in the initial data may derive from errors in the order of battle (OB), which serves as a base line for projections. This is a serious problem, since we cannot know the true level of forces in foreign nations, particularly in China, and as a result estimators cannot compare past projections with a true figure, but only with a figure which may be correct with a certain probability. In addition, when past figures (e.g., for 1971) are revised (e.g., in 1974) we have no assurance that the revision makes the newly revised figures more accurate. In fact, the revision may simply be the result of smoothing a trend line in one plausible direction or

the other. In addition, changes may reflect not only more accurate figures, but more effective collection systems. Specifically, when there is a sharp increase in (say) ICBM figures from 1971 to 1972, this may not mean that forces were significantly increased, but rather that satellite cameras were greatly improved at that time. As a result, the accuracy of earlier figures may be thrown into doubt, and earlier OB data are less credible than more recent data. No confidence levels are available for OB data.

5.7. DECEPTION

As in all forms of strategic intelligence, conscious deception by a potential adversary provides another source of uncertainty. Even our allies may habitually provide misleading figures concerning their capabilities. One nation, wishing to conceal the extent of its defenses, may produce figures which are much smaller than the actual figures; another nation, which wants to give an exaggerated vision of its capabilities, may provide artificially inflated figures.

DE estimators are aware of these deceptions, and may revise their projections toward more realistic numbers than those provided by the governments. A more difficult task is provided in some instances by those nations in which military planning is performed badly, and in which there are simply no realistic figures available to anyone. Under these conditions, DE's task is to develop realistic projections on the basis of whatever data may be available.

For the communist nations, the work of the estimators must be based on a comprehensive understanding of the nation. Every factor which might influence the development of a weapon system must be taken into consideration -- the national economy, domestic and international policy and goals, the location and capacity of production facilities, natural resources located within the nation or available through its allies, technological capabilities and the output of research laboratories, areas which are receiving special attention in research, the power base of the current regime and the likelihood that it will remain stable, the organization and leadership of the armed forces, military policies which have become traditional -- in short, many aspects which, together, form a "model," or rational intellectual picture, of the nation as a whole.

With a clearly-defined model of the nation, it is possible to develop reasonable estimates of its present and future capabilities in specific areas. For example, if we understand the importance of the five-year plans for Soviet resource development, and if we understand that the Soviets are rather slower than the Americans in disposing of obsolescent equipment, we can make some reasonable estimates of the dates by which a given weapon system will be replaced.

A clearly-defined and correct model of Soviet intentions and capabilities provides a basis for dealing with attempted deception. The deception itself fits into the pattern of overall Soviet strategy, which provides a rationale for the deceptive maneuver. Within obvious limits -- the model must itself be

tested against reality -- the use of a comprehensive model provides a defense against deception.

Deception does not merely introduce an element of uncertainty into the estimative process. The attempted deception must be motivated, and valuable information may be derived from the most deceptive material, if the underlying motive can be correctly identified. For example, two Soviet political scientists have prepared an article for a recent issue of Fortune magazine, which attempts to justify the USSR's massive expenditures for armaments within the context of peaceful Soviet intentions. The U.S. reader will not, of course, take these protestations at face value. Valuable information can nevertheless be obtained concerning Soviet intentions if we succeed in interpreting them correctly, since the article surely indicates what the Soviets want Americans to believe. More generally, the art of propaganda analysis attempts to derive information of value from deceptive material, not merely to reject it as false.

5.8. THE "PARADOX" OF INTELLIGENCE

The so-called "paradox of intelligence" is present in all forms of strategic intelligence. In one version, this says, "If you're right, then events will prove you wrong." In less paradoxical form, we note that the goal of all intelligence is to provide information for the use of military and other decision-makers, who may be expected to take action which will counteract any projected threat. For example, if we project the development of a substantial Soviet ICBM capability, the U.S. should be expected to respond in such a way

as to reduce or eliminate the threat that the ICBMs present. If the U.S. does successfully develop an effective counterforce, this could lead the Soviets to modify or abandon their ICBM development. If they were to do so, then our original projection will be "wrong." The Soviets would not have the ICBM force that we projected.

But in any reasonable sense, of course, the original projection was "right." The Soviets did indeed plan to develop an ICBM capability, but thanks to our timely response, they were forced to change their plans. We were correct in identifying the original Soviet intentions. Unfortunately, the format in which the projections are made does not clearly indicate that they represent intentions and capabilities; instead, they appear to represent firm predictions of the future. Thus, in an evaluation of the quality of the projections, they are judged "wrong."

This "paradox" indicates another source of uncertainty in DE's projections: the possibility that the intentions of a potential adversary may change as an eventual result of the projections themselves.

5.9. "INTUITION"

The word "intuition" was sometimes used by estimators to describe the process by which they arrived at projections. This is rather misleading, since it suggests that intelligence production is sometimes little more than guesswork.

The role of "intuition" becomes more significant if we recall that master-level checker players, who were questioned about their methods in connection with a checker-playing computer program, often said that they chose their most successful moves by "intuition." By this, they simply meant that there were no general rules guiding their choices; instead, they relied on their understanding of the game as a whole, their sense of the patterns present on the checker board, their choice of a strategy for this particular game and opponent, and so on.

Similarly, "intuition" for an intelligence estimator could include a global understanding of the nation as a whole, some insight into typical strategies employed, a recognition of specific capabilities, and a variety of "fringe" or ancillary factors that could influence a decision concerning weapon development, deployment, or withdrawal.

For example, an estimator may be considering Soviet ABMs. The current SALT agreement may permit 100 missile launchers, as a maximum, at Moscow. In fact, they have 64 launchers. What will they do? The estimator believes (let us suppose) that the Soviets are very concerned to protect Moscow, and that therefore they will increase the number of launchers around the city. The estimator thus draws on a general model or "picture" of Soviet goals and priorities, using it to predict a concrete action to be taken by them.

Traditionally, "intuition" has meant an immediate "seeing," as we see the truth of the formula "2 + 2 = 4," once we understand the meaning of the various symbols that make it up. The process does not represent a flippant substitute

for scientific thought, but acts as the basis for scientific thought, providing the first premises that are required for further analysis.

In strategic intelligence applications, "intuition" is the process by which the experienced analyst attempts to take relevant factors from a variety of sources into account, and to combine them to form a comprehensive pattern that "makes sense" of the observed phenomena.

In actual practice, however, projections are generated through the use of a limited number of estimation parameters, such as:

- Deployment rate
- Rate of change
- Retirement rate
- Estimates of ratios among weapon systems.

These more mundane figures provide the basis for determining how many weapon systems of a given type will be deployed at a specified future date. In the extreme case, an estimator who is pressed for time may simply use a straight-line extrapolation of current trends. Some uncertainty is present in the projections, since there is uncertainty concerning all these parameters -- particularly in the timing for introduction and phase-out of a weapon system.

A specific example may make this process clearer. In estimating future naval systems, the estimator knows that prototypes are planned five years in advance. Designs and requirements are specified, and, in the Soviet Union, the vessels are produced over a period of ten years. Naval vessels have a twenty-year life span. Using these figures, the estimator can construct a simple mathematical model of Soviet naval development.

A complicating factor is the "learning curve" exhibited in the development of a weapon system. Production begins slowly, as people in a factory are learning how to produce a system and as bugs in production and in design are located and eliminated. Then there is a period of rapid growth in numbers of weapons, as maximum production is obtained, for a period of years. Then this tapers off, as production is slowed and finally halted.

Although computer simulations have not actually been used in the production of estimates, it is clear that some portions of the estimative process might be automated, to the extent that models like these may be formalized.

5.10. COMMUNICATING UNCERTAINTY

Several methods are currently used for communicating the uncertainty present in intelligence reports. This subsection will include a brief description of such methods, together with commentary concerning their potential value for DIA-DE.

5.10.1. Kent Chart

Until recently, DE has used reporting methods proposed by Sherman Kent (Figure 5-1). Essentially, the Kent chart provides a translation from certain natural language phrases ("It is likely that . . . ") into numerical estimates of probability. Kent developed this approach following his observation that the natural-language phrases were subject to wide variations in interpretation, and that they served to conceal disagreements concerning the likelihood or uncertainty present in intelligence reports.

There has been some disagreement about the correctness of Kent's original observation; that is, it may be possible for humans to communicate in natural language with less ambiguity than he supposed, and it may be more difficult for humans to use numerical indications of probabilities.

Thus, in an unpublished paper, "Probability and Modality in the Lexicon," Valerie F. Reyna found that 34 adult volunteers substantially agreed on the ranking that they gave, based on the degree of uncertainty present, to the following modal words: 'impossible,' 'inconceivable,' 'unfeasible,' 'improbable,' 'unlikely,' 'uncertain,' 'indefinite,' 'unnecessary,' 'conceivable,' 'feasible,' 'possible,' 'probable,' 'likely,' 'necessary,' 'definite,' and 'certain.' Human beings are clearly capable of using and understanding modifiers like these, and they appear to agree concerning their relative force. Experimental results, however, do not invalidate Kent's claim that expressions like "It is likely" convey different meanings to different persons and can

This figure explains the terms most frequently used to describe the range of likelihood in the key judgment of DIA estimates.

Order of Likelihood	Synonyms	Chances in 10	Per Cent
Near Certainty	Virtually (almost) certain; we are convinced; highly probable, highly likely	9	99 90
Probable	Likely We Believe We estimate Chances are good It is probable that	8 7 6	60
Even Chance	Chances are slightly better than even Chances are about even Chances are slightly less than even	5 4	40
Improbable	Probably not Unlikely We believe ... not	3 2	10
Near Impossibility	Almost impossible Only a slight chance Highly doubtful	1	1

NOTE: Words such as "perhaps," "may," and "might" will be used to describe situations in the lower ranges of likelihood. The word "possible," when used without further modification, will generally be used only when a judgment is important but cannot be given an order of likelihood with any degree of precision.

Figure 5-1 Estimative Terms and Degrees of Probability

serve to conceal disagreements about the uncertainty present in an intelligence estimate.

The significant conclusion to be drawn from research in this area is that natural-language expressions like "It is likely" or "It is feasible" are used for communication of uncertainty in normal human discourse, where the vagueness or ambiguity of such expressions reflects the difficulty that ordinary people feel in estimating the uncertainty of their purported knowledge.

DE appears to have abandoned the use of the Kent chart in favor of the direct reporting of estimated probabilities (Subsection 5.10.3.).

5.10.2. Reliability-Accuracy Ratings

A widely used coding system assists in communicating the estimated reliability and accuracy of an intelligence report (Figure 5-2). According to Harry Howe Ransom, in The Intelligence Establishment, "The most frequent complaint from intelligence consumers is that this system is too mechanical; they would like to know more about the source of material as an aid in their own evaluation of the contents. Another complaint is that too frequently a middle-ground evaluation, such as 'undetermined' is given" (pp. 40-41).

A more intensive study of the usefulness of the reliability-accuracy ratings is provided in Michael G. Samet's "Quantitative Interpretation of Two Qualitative Scales Used to Rate Military Intelligence." Thirty-seven intelligence officers were tested to determine their subjective, quantitative

Figure 5-2

SOURCE RELIABILITY	INFORMATION ACCURACY
A -- Completely Reliable	1 -- Confirmed
B -- Usually Reliable	2 -- Probably True
C -- Fairly Reliable	3 -- Possibly True
D -- Not Usually Reliable	4 -- Doubtfully True
E -- Unreliable	5 -- Improbable
F -- Reliability Cannot Be Judged	6 -- Accuracy Cannot Be Judged

interpretations of the source reliability and information accuracy (plausibility) rating scales. In judging a report, they were influenced much more by the accuracy rating of the report's content than by the reliability rating of the report's source, when they assigned a numerical value to the likelihood that the report was true.

In summary, difficulties with the scaling system included the following:

- Accuracy and reliability could not be interpreted as independent factors.

- The system was under-used. Only 48% of spot reports in an Army field exercise were rated for both reliability and accuracy.

- Ratings were mostly confined to the high end of the scale. Category B2 alone contained 74% of all ratings.

- Ratings were inconsistent, even for experienced intelligence analysts, and could not be improved through training. For example, assigned probabilities for both "fairly reliable" and "probably true" ranged from 0.40 to 0.80.

It should be noted that Samet's results differ from those reported by Reyna and described in the preceding subsection. According to Samet: "Although intersubject agreement on the meaning of one rating relative to another is encouraging, the wide disparity in the absolute interpretation of each

rating raises doubts about the effectiveness of the qualitative rating scales to communicate specific levels of judgment" (p. 199).

Samet recommended that the dual rating system be modified to provide a single-dimensional quantitative scale, in which the probability or likelihood that a report was correct would be indicated by a number in the range from 0.00 to 1.00, by a percentage, or by odds. Thus, a report that was judged very likely to be correct might be rated 0.90, 90%, or 9 to 1 odds; or a fixed verbal phrase might be attached. "A specific rating could be based upon integration of all available information: the reliability of the source, confirming and nonconfirming reports from the same and other sources, the situation, etc. This likelihood rating could be associated with the report and used in subsequent data communication and processing" (pp. 200-201).

5.10.3. Probability Ratings

Some recent issues of the DIPP have included estimates of probabilities in numerical form. Thus, a statement concerning a specific event or development may be followed by "40 percent chance" in parentheses, reflecting an estimate of the probability or likelihood that the preceding statement is correct. This procedure is consistent with Samet's suggestions.

However, as Samet points out, "formidable problems can be expected with regard to whether data raters can reliably assign likelihoods that will be empirically valid (i.e., of all reports assigned a truth likelihood equivalent to an x% probability, x% of those should turn out to be true)." He recommends

the use of interactive computer aids which will elicit the appropriate likelihood judgment from the estimator.

In addition to the difficulty that estimators have experienced in determining such probabilities, they have also found that intelligence consumers tend to disregard them. Under these conditions, there is little motivation to improve the quality of the ratings assigned to intelligence estimates. Finally, no attempt has been made to validate the numerical ratings; there is no indication that estimates which are "70% probable" have ever been tested to determine that they were correct 70% of the time.

Still another type of problem that estimators found in the use of probabilities was the loss of information that occurs when only a probability is attached to an estimate. If we attach a particular percentage or probability figure to a projection, we are simultaneously ignoring the assumptions that combined to form that probability. These could include the probability attached to the figures for the overall forces, the probability of a particular mix of forces, the probability of deployment of a particular mix of forces, the probability of deployment by a certain date, the probability of replacement by another system, the probability of a response to U.S. postures, and so on. All of these probabilities can be aggregated in some way, but the aggregation process serves to mask the rather more sophisticated reasoning that has gone into a particular projection. The analyst may be entirely correct in a general claim that system X will be replaced by system Y; but if the date is somewhat in error, then the individual projections of numbers may be incorrect, with the result that the entire projection is regarded as an error. In short,

the use of a single probability figure may be an extremely crude way of representing what the estimator has actually thought.

5.10.4. Confidence Ranges

In addition to the probabilities assigned to specific events or developments, DE provides confidence ranges for numerical estimates of force levels for most countries. (A single estimate, not a range, is provided for Noncommunist Nations.)

Three estimates are provided: a high, a low, and a best. These are selected as roughly representing a three-out-of-four chance (75%) that the actual figure will lie between the high and the low estimate, and that the best estimate is most likely. The distribution of probabilities within the range is not specified, and no specific probability is assigned to the best estimate. In addition, the 0.75 probability is not an exact figure; at one point, it is suggested that one standard deviation from the best estimate, which would represent a 0.68 confidence interval, could equally well be used.

As with probability estimates, no attempt has been made to validate the confidence intervals; there is no indication that they have actually included the correct figure 75% of the time.

The high estimate does not always represent the largest number; and, conversely, the low estimate is not always the lowest figure. In these cases, the high estimate should be understood to represent a high level of effort

concentrated in a given area, with the result that an aging weapon system may be retired more rapidly in favor of a newer system. Similarly, a low effort would mean that the obsolescent system might be retained for a longer time, resulting in larger numbers for that system.

At times in the past, when an unusually high degree of uncertainty was present, "spreads" were used for high, low, and best estimates. A spread was a range of figures, rather than a single figure, for the estimate. Thus, all three figures -- high, low, and best -- might be regarded as uncertain.

Information provided by these confidence intervals does not seem to be used by consumers of estimative intelligence. For many purposes, only the best estimate is used. At times, the worst case, represented by the high estimates, is used. The low estimates are generally ignored. This selective use of DE's products may indicate a lack of sensitivity to the information concerning uncertainty that DE provides. At the same time, it suggests that a more effective method for communicating uncertainty may be required.

5.10.5. General Assumptions

At the start of each DIPP, and in footnotes throughout the DIPP, assumptions are stated which assist in communicating some of the uncertainties present in the projections. Assumptions may include the observance or nonobservance of a treaty (such as SALT agreements), lack of major hostilities, continuation of the present regime, etc.

While no probabilities are attached to these assumptions, users of the DIPP are thereby notified that the published estimates are conditional upon them. Thus, they assist in conveying some degree of uncertainty for the reports to which they are attached.

The assumptions do not, however, seem to be used by consumers of DE's products. It is difficult to determine exactly how such information should be used, except to provide some sense of the uncertainty present in all estimates of future events. The assumptions do, however, provide some aid in modifying assessments of the quality of previous DE projections: if the assumptions upon which the earlier projections have been made are violated, then the projections no longer apply.

5.10.6. Lack of Consensus

Occasionally, no consensus is obtained among the agencies responsible for developing estimates. When this occurs, a footnote or appendix may be added, indicating the lack of agreement, together with some of the justification provided for each position.

When there is no agreement, such footnotes or appendices assist the consumer in evaluating the degree of uncertainty that may be present in a published projection.

5.11. SUMMARY

In this review of DE methods, several elements should be emphasized:

- DE must meet with other groups and justify estimates, with the eventual goal of presenting a consensus estimate for the intelligence community. This requires that estimates be defensible, that appropriate justification be available to support them.

- Justification takes place within the context of a comprehensive model of the nation under study, composed of defensible hypotheses concerning national goals, military and industrial capabilities, social structure, and other factors which might contribute to its military posture. A thorough understanding of all relevant characteristics of the nation is necessary to produce a justifiable projection concerning specific military developments.

- Projections are intended to represent verified scientific hypotheses concerning development and deployment of military capabilities, based on a general model of the nation, together with observed data concerning the specific capability.

- Uncertainty enters into the projection process in many ways, including missing data, conscious deception, errors in the models employed, and so forth.

- It is important to communicate this uncertainty to intelligence consumers, who must know the degree of uncertainty in a projection in order to make reasonable use of it.

- Attempts to communicate uncertainty have not been successful, since little use is made of any of the proposed measures of uncertainty. In addition, the need for calibration (validation) of measures of uncertainty has been neglected.

SECTION 6

STATISTICAL METHODS

". . . as we move further into the age of scientific achievement, the complicated machines and scientific detection devices require the greatest sophistication on the part of the operators and analysts. Without this our scientifically produced information as well as that furnished by the tools of espionage would be of little use. For it is the patient analyst who arranges, ponders, tries out alternate hypotheses and draws conclusions. What he is bringing to the task is the substantive background, the imagination and originality of the sound and careful scholar." (Allen Dulles, The Craft of Intelligence.)

This section provides an introduction to statistical methods for aggregating uncertainty. The quotation from Allen Dulles with which this section begins is intended as a caveat: statistical methods are designed to assist humans in the production of intelligence estimates, not to replace them. There is no substitute for the human being who has a broad understanding of the goals and structure of a foreign nation, of the technology required to support its goals, and of its ability to achieve them. As we shall see in Sections 7, 8, and 9, statistical methods can nevertheless help to detect biases in projections and to locate inconsistencies in estimates of uncertainty.

The problem for which we seek a solution here is that of determining precisely the degree of uncertainty to be attached to an estimate or projec-

tion, when the estimate is based on several sources of information, each of which is somewhat doubtful. For example, we know that information obtained from defectors and prisoners is likely to be self-serving and therefore faulty. However, if several prisoners independently agree upon a report, we are likely to give a much higher degree of credence to their combined story than we would to any one of them taken separately. Similarly, if several sensors (radars, photographs, infrared sensors) independently agree in identifying the missiles under test at a given location, we should attach a higher degree of certainty to the identification than we would to any one fallible report standing alone. The problem for consideration here, then, is the most effective, statistically correct method for combining measures of uncertainty attached to various reports, in order to obtain an aggregate measurement of uncertainty.

Several mathematical models, which have been developed as computer program designs, are described in Appendix C. Although the mathematics involved in these models is not particularly complex, it would burden the main text of this report to include it here. Instead, a general, non-mathematical overview will be presented.

Suppose that two reports of an event, such as a successful test of a new anti-satellite weapon, are received. Suppose also that these reports are independent. By "independent" we mean that the two reports came to us from two completely different sources; they are not simply two versions of the same report. If we know the probability that each of these reports

separately is correct, what is the likelihood that they are correct when taken together; that is, when they confirm one another?

A very simple probabilistic model for this problem might be the following. Let us call the first source of reports A, and the second source B. Suppose that 70 percent of the reports produced by A are correct, and 80 percent of those from B are correct. Since the two reports confirm one another, either both are correct, or both are wrong. What are the probabilities that (a) both are correct, or (b) both are wrong?

We select at random, out of the mass of reports produced by A and B, one report from each of them. Our chances for each combination of correct and incorrect reports would then be:

P(A correct and B correct) = 0.70 x 0.80 = 0.56

P(A correct and B wrong) = 0.70 x 0.20 = 0.14

P(A wrong and B correct) = 0.30 x 0.80 = 0.24

P(A wrong and B wrong) = 0.30 x 0.20 = 0.06

Then we calculate the probability that both A and B are right, given that they agree, as P(A and B, given agreement) = 0.56 / (0.56 + 0.06) = 0.903. Thus, there are somewhat better than 9 out of 10 chances that they are correct, given that they both agree in truth-value.

Unfortunately, this simple probability model is much too simple for use in aggregating the credibility of reports. It supposes that each source produces a large number of independent reports, as a bottle factory might produce bottles, and that some of these reports will randomly be found to be faulty. In point of fact, some events are far more likely than other events, thus some reports are more plausible than other reports.

When we say that one report is more "plausible" than another report, we mean that it is more likely to be true -- based on our general knowledge of the context in which it appears. Suppose that: (1) Soviet aircraft X is a rickety, ancient plane that is constantly in need of repair, and (2) aircraft Y is a sleek, new machine that performs effectively and reliably. Suppose that we receive two reports: (A) that X is being mothballed, and (B) that Y is being mothballed. Which report, A or B, is more likely to be correct? Obviously, assuming some degree of rationality on the part of the Soviets, A is more likely than B; we say that A is more plausible than B.

Two factors enter into the evaluation of the uncertainty of a report: (1) the record of reliability of the source, and (2) the initial plausibility of the event which it reports.

Bayesian methods are intended specifically for dealing with such situations. They represent a straightforward development of Bayes' theorem, which combines the initial plausibility measure with each piece of additional information to obtain the probability of a projected event. The following information is required: (1) the initial probability, before any new infor-

mation is received, of the event; (2) the probability of receiving a report of this type from this source; and (3) the probability, given the occurrence of the event, that a report from this source would be received. Bayes' theorem can be applied iteratively, using each new piece of information, to obtain the probability of the event based on current information.

A full description of various forms of Bayes' theorem is provided in Appendix C to this report.

Bayes' theorem is particularly effective in applications in which decisions are structured and repetitive. One application in which we have used it successfully is the identification of aircraft in the battlefield situation, in which there are (1) several different radars and other sensors, which can be used for identifying aircraft, and (2) a definite mix of friendly and enemy aircraft for identification. In addition, there is (3) a well-defined logical structure for the combination of reports from several sensors in such a way as to define the probability of a specific aircraft. This structure was experimentally determined in the research which preceded the system development.

We believe that an effective Bayesian system for aggregating uncertainties can be designed, and we have provided several examples of such systems in Appendix C, which is based on our successful work in pattern recognition. Because the level of mathematical detail is somewhat more complex than that of the main body of this report, the details have been relegated to the Appendix.

At the same time, we do not want to underestimate the difficulties that this approach presents. A Bayesian approach -- and probably any related approach -- would require such information as the following:

- The determination of prior probabilities for all events under study. These represent the initial plausibilities of the events, before current information is used. But one of the characteristics of estimative intelligence is the need to determine the probability of non-routine events -- that is, events for which prior probabilities may not be known. The application of Bayesian methods then becomes problematic.

- In any case, determination of prior probabilities for a Bayesian system may require as much work and be as subject to error as the assessment of probabilities without the use of an automated system.

- Experimental results indicate that the determination of conditional probabilities may be extremely difficult for intelligence personnel. In particular, Section 5 of this report indicates that the use of a numerical system for reporting plausibility and reliability in Army intelligence led to a situation in which most of the probability assessments lay within a narrow range of values, and in which many intelligence officers simply omitted one or the other of the values for plausibility and reliability.

Our conclusions are, then, that:

- Further research in Bayesian and related systems for decision analysis is needed. Operational testing of such systems designed specifically for estimative intelligence is essential.

- As we have noted in Section 4, the use of numerical probability assessments is required for the consumers of estimative intelligence. Since subjective probability assessments may be in error, there should be more feedback to the estimators to assist them in assessing uncertainty correctly.

- The aggregation of probability assessments is largely a bottom-up activity; that is, it begins with low-level judgments and combines these to obtain high-level assessments. An alternative approach, to be described in Section 11, would be to begin with general assessments on a more global level, and use these to obtain probabilities at the lower levels. The top-down approach more closely resembles the actual procedures employed by many estimators.

SECTION 7

PROBABILITY ASSESSMENTS

"It is a common view that belief and other psychological variables are not measurable, and if this is true, our inquiry will be vain; and so will the whole theory of probability conceived as a logic of partial belief; for if the phrase 'a belief two-thirds of certainty' is meaningless, a calculus whose sole object is to enjoin such beliefs will be meaningless also. Therefore unless we are prepared to give up the whole thing as a bad job we are bound to hold that beliefs can to some extent be measured." (F. P. Ramsey, "Truth and Probability.")

In this section, methods for determining probability assessments are described. This disscusion is intended to provide a clear answer to the question, "What is really wanted when intelligence consumers ask for measures of uncertainty?" Section 8 includes further methods for calibrating or insuring the accuracy of probability assessments, and Section 9 describes common errors in estimating uncertainty.

The measures of uncertainty to be described here are frequently called "subjective probabilities" or "personal probabilities" in the literature of decision analysis. These phrases, however (like "intuition"), suggest a type of irresponsible guessing that we would like to avoid. For this reason, we will call them "estimated probabilities" or simply "probabilities."

Roughly, these probabilities represent the "degree of belief" that we hold in a given proposition. They were first given an operational definition in papers written by F. P. Ramsey in 1927-1929 and collected in The Foundations of Mathematics: a person's degree of belief may be measured by offering him a series of bets. Each bet has a known probability of sucess. This probability can be compared with the person's degree of belief to obtain a well-defined numerical measure of belief.

Suppose that we ask someone, "What is the likelihood that Jerry Ford will be elected President in 1980?" He may not be able to give a numerical reply. We therefore offer him a choice between two bets: (1) the first will pay him $10.00 if Ford wins, and nothing otherwise; and (2) the second will pay him $10.00 (in November, 1980) if a coin comes up heads, and nothing if it comes up tails. Clearly, if he chooses the first bet, then he believes that Ford has better than a 50-50 chance of winning, which would make it more likely that he would obtain the $10.00. If he chooses the second bet, then he believes that Ford's chances are less than 50-50, since a flip of a coin would be more likely to obtain $10.00 for him.

This simple example scarcely does justice to the more elaborate bets and counter-bets that may be offered to obtain a precise, numerical characterization of a person's beliefs. Nor does it attempt to deal with the many irrelevant factors that might lead to an unrevealing choice of bets: a general dislike for gambling, a sense of loyalty to a political party, an unwillingness to reveal personal beliefs. Nevertheless, if these irrelevant factors can be identified and excluded, Ramsey's approach provides a clarification of

the meaning of "degree of belief" that can be used to define the type of probability that is required. When a measurement of uncertainty is needed, we can provide a number which indicates our degree of belief in the given proposition.

But we are left with many problems. For example, one person's probabilities may be better than another's, in the sense that one person may be better able to pick the correct bets, which more adequately represent his degree of belief, than some other person. This is the problem of calibration, or proper estimation of probabilities, which will be discussed in Section 8.

We may also hesitate to call these estimated probabilities "measures of uncertainty" (or, more properly, "measures of certainty"). "Uncertainty" seems to suggest a vagueness, an indecision, an ambiguity, which does not lend itself to a precise numerical characterization. For example, I am quite uncertain about the future of Israeli-Egyptian relations. I am, in fact, so uncertain that I would not know how to estimate the probability that the current peace treaty will last for the next five years. Its chances could be 20%, 50% or 80%, for all I know. I am thus uncertain about my uncertainty.

This is the problem of "cascading uncertainties," in which we are uncertain about our uncertainty about our uncertainty . . . and so on, with no obvious stopping point. Interestingly enough, Ramsey's method offers us a way of bypassing this problem, since it presents us with a choice, which must be accomplished within a given time. We may, Hamlet-like, continue to debate about the pros and cons, the ifs and the buts, but if we fail to make a

choice, then we lose any chance at the $10.00 -- or whatever else may be at stake. This is an entirely realistic picture of the human situation, in which the military officer or the corporation officer must make a definite choice within a limited span of time, even though serious doubts may remain concerning potentially relevant evidence. Uncertainty, then, enters into a decision at many different levels, but it may be summarized in a single number, which represents a degree of belief in the projected outcome.

7.1. SUBSTANTIVE VS. NORMATIVE JUDGMENTS

The use of probabilities in connection with intelligence estimates requires that two numbers be reported to the consumer: the estimate itself, together with a measure of uncertainty. The latter would represent the estimator's judgment concerning the credibility or accuracy of the estimate.

Two types of "goodness" are thus required of an estimate: substantive goodness, or the correctness of the estimate; and normative goodness, or the correctness of the attached probability. (Cf. Hogarth, "Cognitive Processes and the Assessment of Subjective Probability Distributions.")

The substantive quality of an estimate is, of course, of overriding importance, and it therefore receives the lion's share of the estimator's attention. It is to improve the substantive quality of projections that he reviews the current and past history of the nation, studies the capabilities of its weapon systems, and, in short, performs the work required to produce an estimate.

The normative quality of the estimated probability is also important, however. The eventual user must be able to distinguish a tenuous hypothesis from an established fact, if he is to make an intelligent decision. The development of a U.S. system may depend on whether there is a 20% or a 60% chance that the USSR will develop an opposing system; and the size and urgency of the U.S. development will depend in part on the probability that the Soviet development will attain a given size by a specified date.

It is important, therefore, to recognize that probability assessments play a critical role in the decisions underlying U.S. policy. Uncertainty plays a part in determining that policy. The converse of this claim may make its meaning clearer: if a particular level of accuracy of an estimate does not play a role in any U.S. decision, then it is not necessary to attain to that level of accuracy. If it doesn't really make any difference to U.S. policy if the Soviets have 103 destroyers rather than 105, then one projection is just as good as the other. (The word "good" here -- in "just as good" -- means the worth of the projection to the user. Obviously one projection may be closer to the correct figure than another. But it would not be worth spending any amount of money, or taking any risk whatever, to obtain the more accurate figure -- if, in the end, it <u>really</u> made no difference. In the sense in which "goodness" means "value to the user," then one of these projections is no better than the other.)

The estimated probabilities, which are used to communicate the uncertainty of a projection, are important to the extent that they can make a difference to U.S. decisions. The current practice, which reports probabilities in

increments of 10 (10 percent, 20 percent, and so on), thus seems appropriate, since it is not likely that any finer discrimination could make any significant difference to U.S. policy makers and other users of DE projections. In addition, it would be difficult to justify a finer increment, given the quality of data available to estimators.

The meaning of "goodness" for a probability assessment is discussed in more detail in the next section (Section 8). There is no absolute "right" or "wrong" for an individual probability assessment, but there are better and worse assessments. If we say, for example, that there is an 80 percent chance that the Soviets will abandon the Homer helicopter, and it is found that they do in fact abandon it, then obviously our 80-percent estimate is better than a 20-percent or 50-percent estimate. The more certain we are, the more often we should be wrong. But unless we (foolishly) claim to be 100 percent certain of a future event, there is no absolute sense in which we can be found right or wrong.

Over a large number of estimates, however, whenever we claim a probability of n percent for our figures, they should be found correct approximately n percent of the time. Probability estimates can then be said to be well-calibrated or accurate. The scoring rules described in Section 8 provide a means for measuring the normative goodness of the estimates (as well as their substantive goodness).

7.2. ELICITATION OF PROBABILITIES

Several techniques have been developed to assist in eliciting probabilities from the analyst. These generally resemble Ramsey's method, in that they assume that the primary task is to force a decision concerning the analyst's degree of belief in a given proposition. Examples of methods used for eliciting probabilities are:

- The analyst is shown a disk, in which a pie-shaped segment can be varied in size. Depending upon the area that it cuts out, the segment can represent any percentage of the whole area of the disk. This area is varied until the analyst agrees that it represents the probability or degree of belief that he holds.

- As in Ramsey's original suggestions, the analyst is offered a series of bets. Depending upon the bets that he accepts or rejects, his estimate of probability can be derived.

- Several experiments have been conducted to determine whether numerical probabilities (e.g., 0.70), percentages (70%), or odds (7 to 3) produce the most realistic estimates of uncertainty.

These methods of determining probabilities can be reviewed in the cited literature. They will not be further discussed here, however, because they do

not appear to deal with the more fundamental question of arriving at a realistic degree of belief. Given the need to produce a timely estimate, and given the evidence available to us, how strong should our belief in an estimate be? Sections 7, 8 and 9 of this report are intended to assist in answering this question.

7.3. ALTERNATIVE HYPOTHESES

Two major points were made in subsection 7.1.:

- The degree of confidence that an estimator reports is valid or well-calibrated under the following conditions: if he reports that his degree of belief is n% for a large number of estimates, then n% of the estimates should, in the long run, prove to be correct.

- The degree of confidence that he reports should be suitable for use in making a decision, such as the commitment of U.S. funds to a research and development program. The degree of confidence in the estimate will assist in determining the degree of risk in the decision.

Statistical methods, such as those described in Section 6, are available to assist in the decision process. Automated aids are discussed in Section 10; these might be used to combine various estimates in arriving at a decision.

Computer-based systems for decision analysis may thus be used to assist in developing a final decision.

Nevertheless, there is no substitute, in the production of estimates, or in the estimation of the credibility of those estimates, for a thorough knowledge of the policies and capabilities of the nation under study. The initial "subjective" probabilities must be based on the estimator's understanding of the factors that have entered into an estimate, which in turn is based on his study of available information concerning the subject. This "intuitive" process probably cannot be mechanized, since it involves the formulation of one or more reasonable hypotheses, and the testing of these hypotheses against all available data. Although research in computer-base artificial intelligence has succeeded in developing programs for performing such tasks, there is little likelihood that they can be generalized to include the broad range of data required for intelligence estimates. The problem therefore remains a task for well-informed, talented human beings, not machines. (Cf. Stuart E. Dreyfus, "Informal Models of Decision-Making," Forefront, Research in the College of Engineering, University of California, Berkeley, 1976-77.)

This approach requires that the estimator formulate an initial global hypothesis concerning the subject under study. For example, the hypothesis might be, "The N-6 missiles on Yankee class submarines will be replaced by N-8 missiles." It is important to note that without this initial hypothesis, the estimator would not know what data were relevant and what were not; he would not even know what data to look for in the flow of information available to him. However, <u>with</u> the hypothesis available, he can begin to formulate

additional supporting hypotheses: For example, "With the longer-range missile, it will be possible for missiles from Yankee class submarines to reach mid-continental areas in the U.S. without a close approach to U.S. coasts; they will therefore tend to stay further away from the coastline." A review of current and past submarine sightings will tend to confirm or disconfirm this hypothesis. Similarly, other subordinate hypotheses may be reviewed, and an estimate concerning their likelihood may be combined with that of the initial hypothesis (using the statistical methods described in Section 6 or the more informal suggestions of Section 9). A review of supporting evidence for the hypothesis will provide the estimator with a basis for roughly determining the uncertainty of the hypothesis.

A valuable next step will then be the exploration of alternative hypotheses. In the example, suppose that the N-6 missiles are not being replaced by N-8 missiles. What alternative courses of action could the Soviets take to solve the problem of providing a missile capability for reaching mid-continental North America? Several hypotheses may be generated and tested (including the possibility, of course, that the USSR does not plan to deploy a submarine-launched missile capable of reaching the mid-continent). Each alternative hypothesis is developed and reviewed, supporting data for each hypothesis are collected, and credibility estimates for the set of alternative hypotheses are prepared.

7.4. CONSISTENCY

An important constraint upon probability estimates is that they must be consistent among themselves. While it is possible to develop a full statistical calculus defining the meaning of "consistency" for probabilistic judgments, the informal approach described here does not seem to justify this formal a treatment. It is nevertheless important to notice what is meant by the consistency requirement. In a formal system of probabilities, the following rules must hold:

- Every probability must be greater than or equal to 0.
- The probability of the certain event is 1.
- If two events are mutually exclusive, then the probability that either of them will occur is the sum of their individual probabilities.

From these rules and the appropriate definitions, we can derive some additional rules, such as these:

- The probability of the impossible event is 0.
- The probability that an event will <u>not</u> happen is one minus the probability of the event itself.
- The probability of any event lies between 0 and 1, inclusive.

Another observation might be that, in a consistent system of beliefs:

- If the assessed probability of A is greater than that of B, and that of B is greater than that of C, then the probability of A must be greater than that of C. (This rule, which is obvious when we treat probabilities as numbers, may not be equally obvious when probabilities are reported in verbal terms, such as "very likely" or "somewhat likely.")

While the rules for consistency of a system of beliefs may seem intuitively obvious, to the point that they scarcely need mentioning, it is nevertheless difficult to maintain them in a very large system, like the computer systems reviewed in Appendix C. In general, human beings have a great many beliefs about a great many things, and they are rarely forced to integrate all these beliefs into a consistent system. Instead, when conflict among beliefs occurs, they tend to make ad hoc judgments concerning the best way of resolving the conflict. But this type of ad hoc resolution is not available when the conflicts occur as part of a set of estimates, such as those that appear in the DIPP. If a conflict were to occur in probability assessments attached to the DIPP, the consumer would have no way of knowing how to resolve them. It would be important, then, to insure some measure of consistency among the probability or uncertainty assessments incorporated into the DIPP.

Given the manual methods for preparation of DIPP estimates currently in use, there is probably no automatic way of locating and eliminating conflicts among the probability assessments. Instead, it would be most important to review the entire set of probability (or uncertainty) measures to determine that they meet the requirements stated here.

One obvious objection to this procedure is simply that the probabilities which are determined in such an informal manner are not rugged enough to be treated in any mathematically sophisticated way; we simply cannot take these "subjective" probabilities that seriously. Thus, if probabilities are simply "plucked out of the air," as some estimators have suggested that they are, they do not have the mathematical validity that would be required for combining them, aggregating them, or reporting them as correct to the fifth decimal point.

The reply is simply that a test for consistency may be used to test the rough estimates, in order to improve their quality. When an inconsistency is found, the estimator is challenged to determine how to resolve it. This will be particularly true in the discussion of calibration in Section 8. The purpose of such exercises is to improve the quality of probability estimates, not to treat the estimates with more respect than they deserve. In Section 11, we will describe a mathematical approach to the detection and resolution of inconsistencies.

Laboratory experiments have frequently given the impression that probability assessments are much more difficult than they are in practice. There are several reasons for this:

- One would be the use of college students in psychology courses for many of the experiments; as naive estimators, they cannot always be expected to behave like area experts with several years of experience. As will be noted in Section 8, experiments using experienced

weather forecasters, in their customary tasks, have shown them to be capable of producing accurate probability assessments. Since intelligence estimators are frequently faced with ill-defined, unique tasks, they may not perform as well as did these forecasters. But they may be able to perform better than inexperienced college students.

- A second reason for supposing that probability assessment is extremely difficult is that many of the examples are unrealistically broad in scope. Subjects were asked to respond to questions like: "What is the probability that Russia is about to attack the UAR?" (Barbara Heinrich Beach, "Expert Judgment About Uncertainty," p. 21). While it might be reasonable to ask for an assessment of this probability during 1979, it would be extremely difficult, even for an area expert, to project a probability over the next ten or twenty years. Radical changes in governmental policies in the USSR, to say nothing of those in Egypt, could rapidly make any such projections obsolete. But DE's projections are specifically designed to exclude broad, radical changes in governmental policies, and to deal with estimates under well-defined assumptions concerning attitudes and capabilities of the nation under study. Given this limitation of scope, DE's probability assessments can be considerably more accurate than those obtained in university laboratories. (Suggested by Beach's discussion, pp. 24-25.)

- Finally, the primary goal of probability estimates is to provide the basis for decisions concerning future U.S. actions. Thus, the probability estimates may be extremely rough, yet provide sufficient information for determining policy. For example, the confidence interval -- the range from "high" to "low" estimates -- is nominally set at a 75% level, a level which is not taken especially seriously by the estimators. In the opinion of estimators, the actual figure may range from 50% to 90%, depending on the accuracy of the data on which it is based and on the length of time over which it is projected. Nevertheless, for the purposes of the consumers, no greater accuracy may be required. In their present form, the ranges provide the basis for the necessary decisions -- a "worst case," a "best case," and a "most likely case" -- on which a realistic U.S. policy can be based. At the same time, any means which can be used for improving the quality of the probability estimates will provide the basis for better policy decisions.

7.5. A NOTE ON SCIENTIFIC METHOD

Much of the history of scientific method, from the time of Plato down to the present, has been that of a search for absolute certainty, of the type which is obtained in pure mathematics. Scientists have felt that it was their job to attain a level of proof which would allow no room for uncertainty. During the Seventeenth Century, for example, to say that something was "probably true" was regarded as an admission that a conclusive proof could not be found.

It was only during the second half of the Seventeenth Century, in fact, that the foundations of probability theory were being laid; until that time, no firm methods of proof using probabilities were available.

Much of our knowledge nevertheless is probable, not certain. While mathematics and symbolic logic may attain to absolute certainty, they do so only so long as they remain abstract; the moment that mathematical theorems are applied to the real world, their certainty vanishes -- we never find a perfect circle or a perfect triangle in the real world, and we are therefore never certain whether our pure geometry can be applied to the objects that we find there. All our measurements are inexact, and therefore applied mathematics is always inexact.

This lack of certainty is endemic to the social sciences, and particularly to estimative intelligence. Here it may be difficult or impossible to determine the intentions and choices of human beings, particularly when these intentions and choices are projected into the future. To this must be added the uncertaintly that is introduced through secrecy and deception; essential data are concealed from us, and apparent data are deliberately falsified.

The underlying logic of the scientific method, under conditions of uncertainty, nevertheless remains applicable. It begins with a problem to be solved. One or more hypotheses which may constitute solutions to that problem are formulated. Data are gathered which may tend to verify the hypothesis. In practical applications, definite time and cost constraints are applied during the data-gathering phase, since decisions must be made in real time,

and since resources are limited. On the basis of the available information, a choice is made among the hypotheses.

The final choice will be probabilistic in form. It is an answer to the question, "How likely is this hypothesis in competition with other available hypotheses?" The answer to this question provides an answer to the question, "How uncertain is this estimate?" An estimate is uncertain to the extent that other hypotheses can explain the available data.

In practical terms, consider the hypothesis, "The Soviets are constructing an ABM system." Data to support this hypothesis include the construction of a large radar installation at Sary Shagan, and tests of intermediate-range missiles at Kapustin Yar. The probability of the hypothesis will be a function of the probability of any competing hypothesis that can explain the same data. Specifically, the question to be asked is whether the radar installation might be used for some purpose other than the detection of incoming missiles, and whether the intermediate-range missiles might be used for something other than destruction of incoming missiles.

The estimator thus must act as a devil's advocate, either for himself or for other estimators, in formulating and defending alternative hypotheses. In the end, all competing hypotheses may be ranked in order of likelihood, and probability values assigned to each of them, on the basis of their ability to explain (or make sense of) the available data.

Note again that the data-gathering process is an active process; data are gathered for a specific purpose: the verification of one of the hypotheses under test.

7.6. UNCERTAINTY IN THE DATA

Within the informal model of the scientific method sketched in the preceding subsection, the data are properly treated as "givens," with essentially zero uncertainty. While this approach is appropriate for a laboratory situation, it clearly does not apply in most areas of historical research, and it certainly does not apply in estimative intelligence. The estimator must work with reports, estimates, and opinions.

DE does not usually work directly with raw intelligence -- with unscreened reports, uninterpreted radar data, or other direct sources of information. Occasionally an estimator may have the opportunity to tour an Eastern European nation, and actually to see some of their weapon systems at first hand; but while such first-hand observation helps to give a sense of realism to DE's projections, it probably is not as helpful as the detailed specifications prepared by experts in current and basic intelligence.

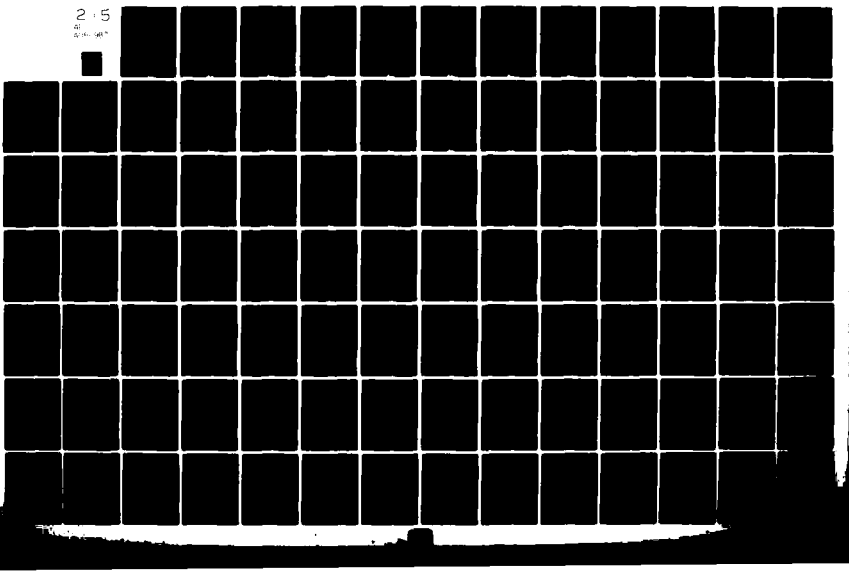
Firm, well-documented information is available concerning capabilities and numbers of most Soviet weapon systems. Data concerning naval systems may be somewhat better than data concerning smaller, better-concealed systems. Information concerning Chinese capabilities is considerably less certain, due in part to the lack of suitable observers in the PRC, to better concealment of

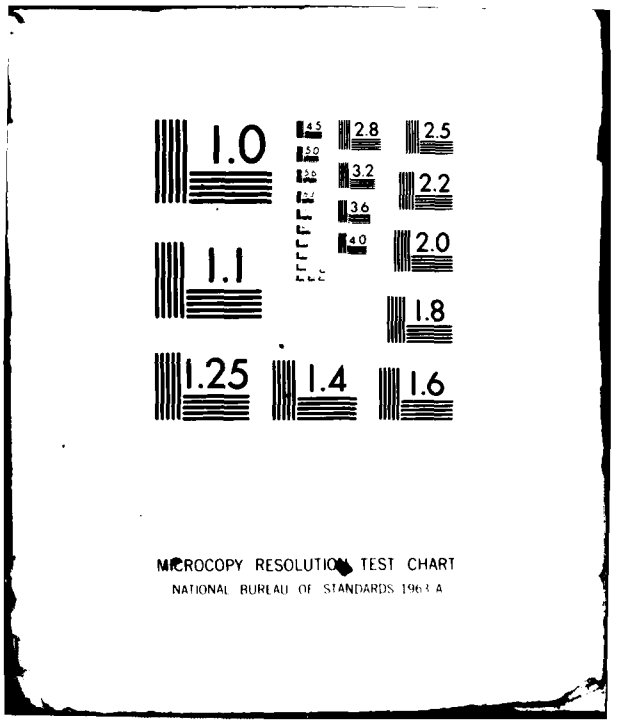
many of their installations, and to a general lack of knowledge concerning Chinese attitudes and methods. Information from some NATO nations, based on governmental reports, may show a consistent upward or downward bias. The initial data with which DE works nevertheless have generally passed through at least one stage of evaluation and can therefore be treated as givens; at worst, the initial data have a known degree of uncertainty.

Order of Battle (OB) information constitutes a special case. Since it represents an official estimate of current force levels, it is taken as the base-line upon which DE projections of future levels are founded. In addition, in evaluating past projections, the most recent OB estimate is taken as a "true" value; failure to predict the OB value represents an "error" in the projection.

The OB values nevertheless are estimates which are sometimes found to be in error themselves. For example, more accurate photographic equipment may lead to a substantial increase in the estimates of the number of ICBM installations, indicating that previous estimates were too low. As a result, projections which were based on the earlier OB values will be found to be incorrect, with an error equal (in numbers) to the error in the OB. Errors in the OB thus introduce a degree of uncertainty into projections which will be equal to the degree of uncertainty in the OB itself. Our information on OB errors is not precise enough to permit a quantitative measure of the uncertainty that such errors introduce. (Methods for obtaining such a measure are described in Section 8.) Nevertheless, experienced analysts should be familiar enough with such errors to permit a reasonable estimate of them. In particular, it has

AD-A086 987 PATTERN ANALYSIS AND RECOGNITION CORP ROME NY F/G 15/4
AGGREGATING AND COMMUNICATING UNCERTAINTY.(U)
APR 80 J M MORRIS, R J D'AMORE F30602-78-C-0291
UNCLASSIFIED PAR-79-1 RADC-TR-80-113 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

been found that OB values for the PRC are more likely to be erroneous than those for the USSR, and thus that the degree of uncertainty introduced into PRC projections will be greater. Again, the estimates of naval systems are much more likely to be accurate than those of other, more easily-concealed systems; there will therefore be less uncertainty in the projections of naval systems than in those of other systems.

Finally, projections must be based on opinions obtained from area experts. The quality of such opinions will vary widely; one of the tasks of the estimator is to seek informed, dependable opinions concerning future developments in the area under study.

7.7. UNCERTAINTY IN THE MODEL

It has been suggested in this report that the model used for DE projections is essentially that of scientific method as it has been traditionally applied in history and the social sciences. A major role is played in this predictive model by the intentions, as well as the capabilities, of the nation under study. It may be assumed, for example, that the Soviet Union intends to develop a capability for surviving a major nuclear war, and that this capability will require the existence of a suitable civil-defense system. Training programs and equipment will reflect this need. It will be possible to interpret the presence of a large number of troop transports as equipment for evacuating the civilian population: underground structures can be interpreted as shelters.

In short, the general assumptions concerning Soviet policy can assist in interpreting information about their facilities and actions. The data in the preceding example would have rather a different meaning if it were assumed that the Soviets were preparing for a land war against the Chinese with conventional weapons. The troop transports would become vehicles for transporting military forces, the underground excavations might be production facilities or storage areas.

Generalizations concerning Soviet intentions, therefore, function as hypotheses in an overall model. Such hypotheses may be supported or undermined by available information concerning Soviet capabilities and actions. Again, it is important to emphasize the necessity of some general hypotheses concerning the policies and intentions of the nation under study. Without such hypotheses, there is no way to determine which data may be relevant or irrelevant to the projection.

Since data concerning a potential adversary will always be incomplete and subject to error, the general hypotheses themselves will also be uncertain. It is not likely that they will ever be completely verified; and existing data may always be found which tend to undermine even the most plausible thesis. The suggested approach, therefore, will be to formulate alternative hypotheses concerning the adversary's intentions. In the example outlined above, the alternatives were: (1) an intensive civil-defense effort, and (2) a possible attack on the PRC. These two policies are, of course, neither mutually exclusive nor exhaustive--the USSR may choose to follow either, both, or neither--and the probabilities for them may therefore add up to a number which is

greater or less than one. These probabilities may be estimated for each of the competing hypotheses, based on the data available to the estimator.

7.8. SUMMARY

In this section, several approaches to the assessment of probabilities, which provide a quantification of the uncertainty present in intelligence estimates, have been described. While there is no purely mechanical way of determining the probability of a projected event or development, there are nevertheless several ways of improving the quality of probability assessments:

- Intelligence estimators can be assisted in visualizing the meaning of probabilities by describing various wagers. Flipping a coin, for example, helps to clarify the meaning of a 50 percent probability.

- A distinction between substantive and normative goodness in the assessment of probabilities has been drawn. Experimental evidence has suggested that development of substantive goodness (based on knowledge of the subject area) improves the quality of normative judgments (ability to assess probabilities correctly). Probability assessments can be evaluated through the use of a scoring rule, like those to be described in Section 8.

- A serious problem is the provision of feedback to the estimator, to permit him to see where errors have occurred in the past. The

Institutional Memory, to be described in Section 8, is one tool to assist in this process.

- Another approach to probability assessment is the use of alternative hypotheses, in which the comparative probabilities of various scenarios are assessed. A more detailed description of this procedure, together with examples of its use, will be found in Section 11.

- Several rules for determining the consistency of probability assessments were described. A discussion of methods for evaluating and resolving inconsistencies in probability assessments is included in Section 11.

SECTION 8

CALIBRATING UNCERTAINTY MEASURES

"A great part of the information obtained in war is contradictory, a greater part is false, and by far the greatest part, somewhat doubtful." -- Karl von Clausewitz

Measures of uncertainty are valuable insofar as they provide the consumer with guidance in the use of the estimates to which they are attached. Although a rough, verbal measure of uncertainty ("It is likely that . . . ") will be helpful to the consumer, a more precise statement ("There is a 70 percent probability that . . .") will be of greater value, for the reasons noted in Section 3: lack of ambiguity, usefulness in gaming and decision analysis, and a better basis for evaluating estimates.

Advantages such as these have encouraged the use of numerical measures of uncertainty in conjunction with projections and other estimates. To be effective, however, probabilities must be calibrated; this means that in the long run, if a probability of n% has been attached to a large set of estimates, then n% of those estimates should be found to be correct. Calibration is the process by which an estimator's probabilities are adjusted upward or downward to make them approximate the ideal more closely.

Calibration has been intensively studied in several areas other than estimative intelligence; weather forecasting is the most-studied area. We

should take care in applying the results of these studies to estimative intelligence, which differs from weather forecasting in the following ways:

- Studies of weather prediction have concentrated on a simple, dichotomous outcome: rain or no-rain. Intelligence estimates generally consider a much wider range of possible outcomes.

- Weather forecasts can be verified as correct or incorrect within a few days; feedback to the forecaster is very rapid. Feedback to the intelligence estimator may take several years. Thus, the opportunity for learning from one's mistakes is much smaller.

- The number of weather forecasts is much larger than the number of intelligence estimates. The statistical basis for calibration is thus much broader.

- Weather forecasts are repetitive, in the sense that the same types of data are used each day in preparing the same types of forecast. Intelligence estimates range over many different types of projection, using constantly changing sources of information.

- Intelligence estimates require the prediction of decisions made by foreign leaders, which may be unpredictable in principle, and which are certainly difficult to predict in practice. While factors which control the weather are sometimes difficult to predict, they are physical in nature, and thus are predictable in principle.

- There is an accepted methodology for weather forecasting. There is no comparable methodology for estimative intelligence, where the variety of problems encountered and the wide range of potentially relevant data make each problem unique.

It does not seem likely, then, that results obtained in studies of weather forecasting will be directly applicable to problems in strategic intelligence. Such a conclusion would apply with even greater force to the differences between estimates obtained from college students (or other similar groups used in psychological studies) and estimates produced by professional intelligence analysts.

Without attempting to ignore these caveats, however, it should be possible to learn a great deal from studies of weather forecasting. In particular, they will be used here to assist in developing a general approach to the calibration of probability estimates. In addition, laboratory studies will provide several suggestions, to be reviewed in Section 9, concerning frequent errors that are made by humans in assessing the probability of future events.

It is important to determine what constitutes a "good" or a "bad" estimate, and it is for this purpose that the earlier studies of weather forecasting and other types of prediction will be valuable. (In this report, we do not distinguish between "forecasting" and "prediction.")

We begin in subsection 8.1. with an initial discussion of uncertainty as contrasted with credibility, in an effort to determine which of these will be

found most valuable by consumers. The partially successful use of probabilistic predictions by weather forecasters is discussed in subsection 8.2., and the use of scoring rules is reviewed and criticized. Subsection 8.3. proposes an institutional memory to assist in calibrating estimates of uncertainty. Finally, subsection 8.4. describes techniques for assessing uncertainty in historical data.

8.1. UNCERTAINTY AND CREDIBILITY

We are always uncertain about the future. Whenever future events depend on human judgments, or when there is any risk of unexpected or chance events, our predictions can be falsified by the outcome. Intelligent planning and action nevertheless require that we predict the future as the basis for our decisions.

Practical decision-making requires several types of estimate:

- The likelihood that relevant events will occur -- that a given level of production will be achieved by a potential adversary, that a given weapon system will be deployed, etc.

- The degree of risk or benefit that will accrue if the event occurs or fails to occur.

- The time at which the event may occur, or the time-span over which a capability is available.

- The kind of counter-measures that may be required to achieve various levels of response.

- The time and cost required to develop counter-measures.

These and other factors enter into the decisions of U.S. and foreign leaders. Factors concerning probabilities or likelihoods clearly play an essential role in the decision process. The USSR, for example, is not likely to replace the Scarp Mod 5 missile if the cost of the replacement is greater than the benefits that are alleged for the new model, the risks involved in the use of liquid fuel appear to be too great, the estimated likelihood that the missile would be required against a threat from the U.S. is not high, and so on. Measurements of uncertainty thus enter into the Soviet decision. For example, uncertainty concerning U.S. intentions must be considered.

Conversely, U.S. intelligence estimators must ascribe a probability to the Soviet decision; such measurements of uncertainty emerge from the Soviet decision process. In the example, this becomes the estimate of the probability that the Soviets will replace the Scarp Mod 5 within a given time span. (The actual estimate will probably include a timetable for phasing out the Mod 5, with an estimated probability distribution for the numbers of missiles at each date.) The resulting estimates enter into the decision of U.S. intelligence consumers, where they are combined with information concerning U.S. policies and goals, estimated costs, etc. In short, estimates of uncertainty enter into both U.S. and Soviet decisions.

The purpose of the preceding discussion has been to motivate our choice of a measure of uncertainty. It requires a definition of "uncertainty" which is somewhat different from that of ordinary language. On the other hand, this definition is consistent with the terminology and methods of research in decision analysis, particularly with those studies which use Bayesian methods. It is a definition which fits neatly into a model of the decision process.

The meaning of a degree of belief can be partially formalized and clarified as follows. It is well-calibrated when, for all estimates to which it ascribes a value of n%, in the long run n% of them are found to be correct. If the estimated probability is too high, then we are said to be over-confident; if it is too low, then we are under-confident.

A degree of belief can be quantified in terms of real numbers, ranging from 0.0 to 1.0. When its value is near 1.0, this means that we are completely confident in the proposition to which it is attached. (Note the important difference in meaning from credible, as in "It is completely credible that the Soviets are planning to use nuclear weapons against the Chinese." The meaning of "We are completely confident . . ." is approximately the same as "It is completely certain . . ." in the sense that either phrase means that we would be willing to attach a high probability to the proposition which follows.)

When our degree of belief is close to 0.0, this means that we believe that the proposition is false, and we are completely confident in rejecting it. (This meaning is different from "It is completely incredible that . . ." since we may be confident in rejecting something without claiming that it is

incredible: I am quite confident that the Soviets are not now launching nuclear missiles against the Chinese capitol, but I do not think that some such action is incredible. The meaning is also different from "It is completely uncertain that . . ." To be completely uncertain is to have no idea whatever whether a proposition is true.)

Several rules will be listed, in verbal form, for the combination of estimated probabilities (where "estimated probability," "subjective probability," and "degree of belief" all are taken to have the same meaning). The rules could, of course, be translated immediately into symbolic formulas, but there seems little point in doing so at this time.

- The estimated probability of a proposition can be measured by real numbers in the range 0.0 to 1.0.

- As the estimated probability of a proposition increases, the estimated probability of its negation decreases; the estimated probability of the negation of a proposition is completely determined by the estimated probability of the proposition itself.

- The estimated probability of the conjunction of two propositions is less than or equal to the estimated probability of either of the two.

- The estimated probability of the disjunction of two propositions is greater than or equal to the estimated probability of either of the two.

- The estimated probability of the simultaneous truth of two propositions is a function of the estimated probability of the first, given the second, and of the estimated probability of either proposition taken separately (i.e., Bayes' theorem applies).

These rules could provide material for a formal treatment of measurements of uncertainty. There are, however, a number of informal constraints that do not appear to be amenable to formal treatment, but which may be of equal importance in the quantification of uncertainty. For example, we may make the following claims:

- New information will increase or decrease the estimated probability of a proposition, depending on (a) the estimated probability of the new information, (b) its independence from existing evidence (it can't simply repeat evidence that we already have considered), and (c) the degree to which it confirms or conflicts with the proposition.

- Cascading uncertainties may make the estimated probability of a proposition indeterminate. For example, we may not have sufficient data to determine the credibility of a source; thus the measurement of credibility will be uncertain.

- As new information becomes available, the estimated probabilities will vary over time. Maintaining up-to-date probability estimates

throughout a system of interacting estimates has proved to be extremely difficult.

- The estimated probability of an event or development may be changed by changes in policies or capabilities of a potential adversary. Such a change does not show that the earlier estimates were "wrong"; it shows that some of the assumptions on which they were based have changed.

While the informal rules will make it difficult to develop a completely formal model for the assessment of probabilities, the formal and informal rules will make it possible to review estimates for consistency.

For example, if we estimate that the probability of p is 50 percent, and the probability of q is 60 percent, we should know that the probability of p and q cannot be more than 50 percent. It could be less; it could even be zero, if p and q are mutually exclusive.

8.2. PROPER SCORING RULES

A proper scoring rule is a device for assisting forecasters in calibrating their estimated probabilities. Such rules have been intensively studied in connection with weather forecasting, in which predictions are often stated in terms of probabilities: "There is a 20 percent chance of rain tonight." Probabilistic predictions like these are neither completely right nor completely wrong, unless they are stated as "100 percent" or "0 percent."

Some probabilistic predictions are nevertheless better than others. They are better to the extent that they give high probabilities to the events which actually occur, and low probabilities to those which do not occur. We need a scoring rule to measure the degree to which one probabilistic forecast is better than another.

One simple -- and misleading -- scoring rule was used in earlier appraisals (1974-76) of DE projections. This can be called the "hit-or-miss" scoring rule. It simply counts the number of times that projections have been correct (the hits) and compares this number with the number of times that they have been incorrect (the misses). The result is stated as a percentage: "You've been wrong 70 percent of the time."

If the hit-or-miss scoring rule were taken seriously, it would have the effect of pressuring the estimator into hedging his bets by increasing the spread between High and Low estimates. The wider he makes these spreads, the higher his score. If he says, for example, "By 1984, the Soviets will have between 0 and 56 Foxtrot submarines," he is fairly certain to be right -- and to get a high score -- since only 56 Foxtrots were produced. But this heavily-hedged projection will not be of much help to the intelligence consumer, who really needs a less wide-ranging estimate. The hit-or-miss scoring rule is thus an improper scoring rule, since it encourages the estimator to produce a less-useful estimate.

Another type of improper scoring rule might be called the "direct" scoring rule. Suppose that we give the analyst a score of 70 every time he assigns

a probability of 70 percent to the actual outcome, a score of 80 every time he assigns a probability of 80, and so on. Thus, the higher the probability that he assigns to the actual outcome, the higher his score.

But the direct scoring rule is also improper, because it encourages the estimator to falsify his predictions. Specifically, he finds it possible to raise his score if he "goes for broke" -- that is, if he assigns a 100 percent probability to events that he believes likely, and a 0 percent probability to events that he believes unlikely, without attempting any of the finer shades of discrimination.

But this go-for-broke strategy would not be useful to the consumer, since it would encourage an untoward degree of overconfidence in the predictions. The consumer needs to know, with somewhat more precision, how likely the prediction or forecast is. For this reason, the "direct" scoring rule is improper, since it encourages the estimator to produce misleading estimates of probabilities.

A number of "proper" scoring rules have been developed to meet objections like these. They have the property of maximizing the analyst's score when his probability assessments are properly calibrated. A proper scoring rule will give the analyst a higher score when he assesses probabilities correctly.

Several scoring rules were described in the TEAMS Final Report and could be used in evaluating DE's performance. One of the simplest of the proper scoring rules is the logarithmic rule: this is simply

-log (p)

where p is the probability that the estimator has assigned to the event which actually occurred. For example, the estimator claims that there is a 70 percent likelihood that all Bear F aircraft will be withdrawn by 1980. In 1980, it is found that all Bear Fs have been withdrawn. He then gets a score of -log (0.70) = 0.15. On the other hand, suppose that some Bear Fs are still sighted in operation at the end of 1980. The analyst has allowed only a 30 percent chance (100 - 70) for this event. He then gets a score of -log (0.30) = 0.52. (Obviously, he is receiving a higher score for a worse estimate; the lower his score, the better.)

The primary advantage to the logarithmic scoring rules is that they encourage better calibration. There is a penalty for underestimating or overestimating the probability attached to a projection, and the analyst is thus rewarded for estimating the probability as correctly as possible. (It should be noted that substantively better projections also receive better scores. Two factors enter into the scoring rule: the accuracy of the projection and the calibration of the assigned probability.)

Present TEAMS designs include a scoring rule, which may be applied to the numerical projections in DIPPOLS. It would be possible to extend the use of scoring rules to include other data, such as the probability estimates included in various DE reports. Such an extension is not recommended at this time, however, for several reasons.

First, as noted in the introductory discussion in this section, the character of DE's estimates is quite different from that of weather forecasts, and there does not appear to be sufficient warrant -- for the reasons stated -- for extending the scoring rules now incorporated into TEAMS.

Second, several difficulties have been noted in the use of proper scoring rules, even in the weather forecasting applications for which they were designed. (Allan H. Murphy and Robert L. Winkler, "Forecasters and Probability Forecasts: Some Current Problems," Bulletin American Meteorological Society, April, 1971, pp. 239-247). Among the problems that Murphy and Winkler identify are these:

- Forecasters for both the National Weather Service and the Travelers Weather Service tended to hedge their forecasts (i.e., move them closer to 0.50), in the belief that they would receive higher scores in this way, even though the proper scoring rules were intended to discourage this kind of hedging.

- The go-for-broke effect can also occur, according to Murphy and Winkler, with a forecaster who is near the bottom of the status ladder. He has nothing to gain or lose from the production of mediocre predictions, but if he can score a few spectacular successes, he has some chance for advancement. He gambles, then, by exaggerating his estimated probabilities, in the hope of producing a few high scores. This is because he has a great deal to gain from a high score, and nothing much to lose from a low score.

Thus, the use of scoring rules has not proved wholly successful even in the field of weather forecasting. The primary problems for estimative intelligence, however, are:

- The great variety of types of estimate which are made, which rarely provide the kind of statistical basis needed for adequate calibration.

- The long time required for feedback: from five to ten years may be needed before an estimator knows whether his earlier projections were correct.

For these reasons, it is difficult to recommend the use of proper scoring rules for applications other than the numerical projections of the DIPP. In Section 2, an alternative approach is provided.

SECTION 9

DETECTING AND ELIMINATING BIAS

Humans do a poor job of estimating their own uncertainty. Generally, they are far more confident about their judgments than the evidence would warrant. At other times, they may be inclined to hedge -- to overstate their uncertainty in an effort to avoid the penalties for error.

In this section we shall review some of the research that has recently been directed at the question of human bias in estimates of uncertainty. We might expect that if people know something about the common biases that may occur, they may be able to correct them. As we shall see, however, this last expectation remains a rather forlorn hope; there is no clear evidence that people can correct their errors, even when they know that errors occur. What is needed is a better understanding of the estimative process itself; if estimators are skilled at developing estimates, they also tend to be better judges of the uncertainty in their estimates. The study of bias will be most valuable, then, in helping to clarify the nature of the estimative process.

The subsections present the various biases in the assessment of uncertainty that have been experimentally observed and studied. There is not yet sufficient information concerning DE's estimates of uncertainty to determine the direction and quantity of bias that may be present in them. A review of common biases will nevertheless provide some guidance in determining the types of errors that humans frequently make in their attempts to assess their own uncertainty.

Each of the subsections contains a discussion of a type of bias that has been observed and studied. The first of these, subsection 9.1., is of particular importance in this report, since it represents an error that frequently occurs in outside evaluations of estimative intelligence. In the relevant literature, it is called "hindsight bias"; more familiarly, it is "second-guessing," the "prediction" of events after they have already occurred. Other subsections will review typical errors that are made in the assessment of uncertainty, and will suggest applications to the probability measures that have been recommended to DE.

In the instructional manual (Appendix B), practical suggestions for identifying and eliminating biases will be presented.

9.1. HINDSIGHT BIAS

If someone were to ask, "How likely did it seem to you in 1977 that the Shah of Iran would be overthrown?," we would think back to the evidence that was available then: massive student protests among Iranian students in this country, the presence of secret police and the other paraphernalia of dictatorship, and perhaps other signs of a fragile, rigid regime. With this information, we certainly should have seen that the Shah was about to crumble. If we also had the special sources of information that were available to intelligence officers in 1977, then we certainly should have predicted the rebellion that was then immanent. Why, then, was it not predicted?

"Hindsight bias" is the claim that the future is fairly easy to predict. As we look at past events, we can see the chain of causes and conditions that made them possible, that may indeed have made them inevitable. For each of them, we ask: "Why didn't people see that it was coming?"

The answer to this question is complex, and it lies at the heart of the problem of estimative intelligence:

- Intelligence producers have too much information. They are overwhelmed with data that far exceed their capacity to ingest and digest them.

- The information that they have is not structured around the specific events that they are supposed to have predicted. While there may have been enough information to have enabled them to predict the overthrow of the Shah, for example, they did not have it neatly filed in a drawer marked "Evidence of forthcoming rebellion in Iran."

- Even if they did accumulate the proper information in time, there are too many chance factors that might intervene. For example, what would have happened to the rebellion if the aging Khomeni had become violently ill at the critical moment? What if the Shah had been more conciliatory? And so on, through an indefinite number of variables.

- o Over the long run, intelligence producers have to avoid the "cry wolf" response that comes when they predict dire events too often -- and the wolf fails to appear. Simply to maintain their credibility, they have to avoid premature and unnecessary warnings.

These are some of the reasons for failure to predict the future accurately. Hindsight bias is the claim that they should have predicted it far more accurately than they did. In terms of intelligence projections, it would demand an unwarranted increase in their degree of belief in specified future events. The reasoning would be this: "As we look at the past, we see that many events could have been predicted on the basis of available information. We therefore can attain a relatively high degree of certainty concerning future events. Therefore, our predictions of the future can have a high degree of certainty."

A second effect of hindsight bias would be the attempt to make predictions based on a simplistic view of the past. Since they can put together a reasonable scenario that would permit the prediction, say, of the overthrow of the Shah, this does not mean that they will be able to put together a scenario for the prediction of other anomalous events in the future. ("Anomalous" is used here in the sense that Thomas Kuhn uses it in The Structure of Scientific Revolutions: a violation of the laws that make up the currently accepted world-picture. The world-picture, for Americans in 1977, saw the Shah as an established, stable ruler. His overthrow violated this picture, and was thus anomalous.)

In contrast to hindsight bias, the point of view suggested here is conservative. On the one hand, it says that accurate, useful predictions can be made within the structure of widely-accepted assumptions concerning the goals and capabilities of the nations of the world. On the other hand, the attempt to predict major technological breakthroughs, or massive changes in leadership or policies, is far more difficult and problematic -- no matter what hindsight bias may tell us.

9.2. OVERCONFIDENCE

A person is overconfident when he or she gives too high an estimate of the probability of a projected future event. In quantitative estimates, overconfidence is reflected in too narrow a spread between the High and the Low estimates.

Research has shown a strong, consistent tendency toward overconfidence, both among subject-matter experts and among novices, in field situations as well as in the laboratory:

- Studies of Las Vegas casino patrons showed irrational preferences for certain bets.

- Studies of bankers and stock market experts in the prediction of closing prices for stocks showed exaggerated confidence in the accuracy of the predictions.

- Studies of military intelligence officers predicting a coup in a designated country, the shooting down of a reconnaissance plane, or an arms shipment from one country to another showed overconfidence in their predictions.

  (See Slovic, Fischhoff, and Lichtenstein, "Behavioral Decision Theory," p. 15, for references.)

As noted in the discussion of scoring rules in Section 8, there may also be some pressure toward underestimation, particularly if an estimator is rated as "right" or "wrong." Under these conditions, an estimator could be expected to hedge his projections as much as possible, by underestimating the probabilities of the more likely outcomes. For example, instead of saying, "There is an 80 percent chance that . . .," he says, "There is a 60 percent chance that . . .," which will (a) give him some credit if he's right, and (b) reduce the penalty if he's wrong. Similarly, he could increase the range from Low to High, to help insure that the actual value will lie within the range.

It would be pleasant to suppose that the pressure toward overconfidence (the go-for-broke strategy) is exactly balanced by the pressure toward underconfidence (the desire to avoid the cry-wolf effect). Obviously, the two opposite types of pressure operate in different contexts. To avoid overestimates and underestimates of probabilities, it is important to be aware of these pressures.

9.3. REPRESENTATIVENESS, AVAILABILITY, ADJUSTMENT

What specific techniques do humans use in dealing with uncertain information? In some imaginary world populated entirely by statistical geniuses, they would proceed like this:

- Define an experimental hypothesis for testing.
- Define the experimental population for which the hypothesis is to be tested.
- Employing standard statistical sampling techniques, obtain a representative sample of sufficient size and composition to achieve the required level of confidence.
- Under experimental conditions, perform a controlled experiment as required to test and validate the hypothesis.
- And so on, through the sequence of techniques developed by the experimental sciences.

An appropriate experimental design, following something like this sequence, should certainly be used in those situations in which available time and available information make it possible (and where the cost of the tests does not exceed the expected value of the results). Unfortunately, estimative intelligence -- and real life -- rarely makes it possible to carry out the

full sequence of experimental tests. Intelligence estimates face two major constraints:

- To be effective, estimates must be available for a decision within strict time limits, which may not permit a review of all available data.

- The nature of intelligence data collection is such that important pieces of information may never become available. Other pieces of information may be misleading or false.

Strategic intelligence thus represents, in somewhat exaggerated form, the situation that we all face in real life, where we never have enough time to investigate fully, and where much of the information that we must use is no more than rumor, hearsay, and fraud.

Humans have been found to use several shortcuts, or heuristics, in dealing with uncertainties in everyday decisions. These heuristics have a major virtue: they are fast and efficient. They also have a major vice: they are prone to errors.

Tversky and Kahneman have identified three frequently-used heuristics:

- Judgment by representativeness: a small sample is taken as representative of a large population. We judge the characteristics of a whole group on the basis of acquaintance with just a few of its members.

- Judgment by availability: an event is judged to be likely, if it is easy for us to imagine similar events. If, in our imagination, we can say, "That's just the sort of thing that would happen," then we tend to overestimate the probability that it will happen.

- Judgment by adjustment: when judging the numbers or sizes of things, we begin with a known value and adjust it upward or downward to obtain an estimate of the unknown value. In the process, we often fail to make a large enough adjustment.

Heuristics like these are likely to play a role in estimative intelligence, where the stringent time requirements and the lack of reliable data make it necessary to use short-cut techniques.

9.3.1. Representativeness

Judgment by representativeness will occur when it becomes necessary to make judgments concerning a total population on the basis of a small or non-representative sample.

Tversky and Kahneman have identified a "law of small numbers," which is a fallacious rule by which people tend to make judgments on the basis of a very small number of samples. For example, if we hear that two Ford Mustangs have had frequent brake failures, we are likely to generalize to the conclusion that all Mustangs are subject to brake failures. But our sample size is obviously much too small to make this sweeping a generalization.

Because information available to intelligence estimators can be limited to very small samples, it is important to recognize the high probability of error in generalizing to larger populations. For example, if we could obtain information only about the destroyer Bedovy, we might be tempted to generalize that all four Kildin class destroyers carry 45mm guns (with some high probability), when, in fact, the Bedovy is the only one that does so. Since information concerning Soviet naval vessels is very complete, estimators are not at all likely to make this particular error; but similar errors are possible wherever the information is skimpy and the time is short.

9.3.2. Availability

Judgment by availability is the tendency to use the information which is most easily available, but which may not adequately represent the population from which it is drawn.

Perhaps the best example of this fallacy was the poll undertaken by the Literary Digest magazine to determine the outcome of the 1936 American Presidential election. The poll indicated an overwhelming victory for Alfred M. Landon, the Republican nominee, over his opponent, Franklin D. Roosevelt. The magazine's prediction was, of course, badly mistaken. Its gross error was due to its use of a telephone survey to obtain its results, at a time when only the affluent could afford a private telephone. Since non-telephone households included the majority of voters, and since the vast majority of these voters favored Roosevelt, the poll gave badly misleading results. The magazine relied on data which were easily available, rather than making the greater

effort required to obtain data which were representative of the population from which they were drawn.

Psychological studies have indicated that people use this heuristic to shade their judgments upward or downward, depending on the ease with which they can recall similar objects or events. Such factors as familiarity, recency, and emotional saliency have been identified as affecting recall. Applying these results to estimative intelligence, we would expect that the following factors would affect judgments of uncertainty:

- Familiarity. If the estimator is familiar with a particular weapon system, capability, or other entity, he will be likely to overestimate its numbers, retention time, or other factors, in comparison with another system with which he is less familiar.

- Recency. A recent report, article, or briefing on a given Soviet weapon will tend to increase the importance of that weapon in the mind of the estimator. As a result, he is likely to overestimate the probabilities connected with that weapon, in comparison with other weapons, which may be equally important but which have been reviewed less recently.

- Emotional saliency. We are certainly likely to respond more readily to the more glamorous and more sophisticated weapons than we are to the dull, unglamorous ones. As a result, the estimator is more likely to overestimate the probability that the more glamorous

systems will be developed and deployed, ignoring the factors that would encourage development of the others.

Proper experimental design, then, requires that we take care to include data which are less "available" in the sense described here -- to include information concerning unfamiliar systems, older systems that we may have forgotten, and less-glamorous systems that may be overlooked.

9.3.3. Adjustment

Judgment by adjustment is a heuristic in which we begin with an existing estimate, and raise or lower it in response to new information. This process, called "anchoring and adjustment" by Tversky and Kahneman, is frequently insufficient.

This heuristic may have been partially responsible for underestimates of Soviet ICBM installations during the late 1960's. As Soviet policy changed in such a way as to dictate rapid expansion of ICBM facilities, U.S. estimates remained "anchored" to past estimates, and were not adjusted rapidly enough to take new Soviet policies into account. The result was a series of underestimates. (The underestimates of Soviet ICBMs have been widely publicized and discussed; this is obviously an oversimplification of the reasons for them.)

The use of this heuristic assumes the existence of a base rate, or commonly accepted level of development, production, deployment, and retirement.

Beginning with this base rate, the estimator makes adjustments upward or downward to take account of:

- Current political factors
- Shortages or surpluses of materials
- Difficulties or breakthroughs in production
- Changing economic conditions
- Responses to U.S. and other countermeasures
- Problems in training personnel
- Mechanical and other technological difficulties
- Availability of new technology
- Conservatism of Soviet policy

And any other factors that could influence the qualitative and quantitative projections that are required.

If the experimental evidence can be applied to intelligence estimates, it tells us that these adjustments will not be sufficient; human beings tend to be conservative in their use of the heuristic, retaining a bias in the direction of earlier estimates.

An alternative approach, then, would be the use of "zero-base" projections. Rather than beginning with existing estimates, the analyst would construct a new estimate entirely from scratch. Past projections would be ignored, and previous trends would not be used. Current information concerning foreign weapon systems would be used, to which information concerning

production and deployment rates would add appropriate numbers. Retirement rates could then be estimated, and the resulting figure would provide the final projection.

The essence of the zero-base approach would be its lack of assumptions; nothing would be taken for granted, and every projection would have to be justified. The zero-base approach differs from the anchoring-and-adjustment approach, which requires justification only for changes from existing projections.

The zero-base approach is not recommended here, primarily because there is no reason to suppose that it would produce improved projections. It is included simply to show the way in which anchoring-and-adjustment works, and to suggest a means for avoiding the bias that anchoring-and-adjustment introduces. Essentially, it says: look carefully at the assumptions that enter into projections, and make sure that these assumptions can be justified.

9.4. NEGLECT OF PRIOR INFORMATION

Another bias which has been widely studied seems almost the opposite of the anchoring-and-adjustment heuristic. While anchoring-and-adjustment is conservative, this is an anti-conservative bias, since it is the tendency to neglect prior information.

"Prior information" is represented by the prior probabilities discussed in connection with Bayesian methods in Section 6. It refers to the "base

rate," or the general information that we have concerning the population. In weather prediction, this would be the climatic information that we have. We may know, for example, that on any given day in August, the probability of a snowstorm in Arlington, Virginia, is 0.001. If we then received information concerning barometric pressure and wind direction that could indicate a snowstorm, we would nevertheless be very hesitant about predicting one. We are hesitant, because our prior knowledge makes such a storm very unlikely.

Unfortunately, we are not always hesitant enough in similar sorts of prediction, according to several experimental studies. Specific information takes precedence over the general information that we have. In one set of experiments, for example, an unreliable witness reports that a blue taxicab is involved in an accident. The witness's testimony is taken to have too high a probability, given the rarity of blue taxicabs in the overall population.

9.5. CAUSAL MODELS

". . . People predict and explain events by invoking their intuitive theories about underlying causal factors . . . In making predictions, people rely on information perceived to have a causal relation to the criterion while disregarding valid but noncausal information." (Icak Ajzen, "Intuitive Theories of Events and the Effects of Base-Rate Information on Prediction.")

In general, this approach suggests that the probability of an event is greater to the extent that we can find causal relationships between the hypothesized event and the data that we have. As Section 9 will point out,

this approach is regarded as fallacious, when it ignores the underlying probabilities of the events which it predicts. For example, if massive construction were observed in a Soviet shipyard, it might be hypothesized that the vessel under construction is an aircraft carrier. But this hypothesis would have to be tested not only against the question (1) is this construction appropriate for the building of an aircraft carrier, but also against the question (2) what is the probability that the Soviets feel the need for a balanced fleet capable of world-wide operations, as represented by their Kuril class aircraft carriers?

9.6. SUMMARY

This section has discussed several areas of bias which have been studied experimentally and which may be applicable to probability estimates in estimative intelligence. In general, the experimental evidence shows that many persons have trouble in assessing probabilities correctly, and that it is difficult to correct their biases.

One problem that may make the task especially difficult for untrained personnel, like those in college psychology classes, is simply that the questions deal with unfamiliar material ("Does the UAW have more or fewer members than the DAR?"), concerning which the subjects feel that they can make only random guesses about the accuracy of their replies. With a substantial fund of experience in the subject matter, professionals have been found to be more realistic in assessing the probability that their estimates are correct, or lie between specified limits.

There has nevertheless been little feedback to assist intelligence estimators in detecting biases and other errors. The long time span required for verification of their estimates, and the lack of an easily available data base containing past estimates, have meant that it is often impossible to tell whether a probability assessment has been satisfactory.

In addition, DIA's estimates are intended to represent a consensus of several agencies, rather than the opinion of a solitary estimator. This means that a major focus is upon obtaining an estimate which will be acceptable to members of the intelligence community. In this context, "better" and "worse" mean more or less acceptable to others in the community, rather than valid in some more global sense. The need to represent a consensus, then, seems to add a conservative bias to the estimates, which cannot easily be overcome.

If this sketch is correct, it would seem that the most valuable role that DE could play, during the initial stages of the preparation of estimates, would be that of a devil's advocate, introducing doubts and difficulties concerning the majority opinion. In this role, the estimator would collect as much evidence as possible for a minority point of view, and would provide a critique of arguments underlying the majority position. Such a role would help to reduce the overconfidence that has generally been found in probability assessments. In addition, it should help to provide better arguments to support the final estimates.

SECTION 10

COMPUTER-ASSISTED ESTIMATIONS OF UNCERTAINTY

The purpose of this section is to present an approach to computer-assisted methods for the aggregation of measures of uncertainty. A number of computer-based systems have been suggested for the support of intelligence analysis, and it is attractive to suppose that such a system could be used to increase the validity of DE's estimates of uncertainty, provide a rapid method for aggregating them, and assist in the production of outputs for communicating uncertainty to intelligence consumers.

Some of the systems to be discussed here represent computer implementations of the methods outlined in Section 6. One approach is essentially Bayesian -- that is, Bayes' theorem is used to combine known probabilities, in order to arrive at aggregated probability estimates, and to maintain the consistency of the estimates of uncertainty. Other systems incorporate logical structures of cause and effect, more closely resembling the approach described in Section 5.

The first subsection describes a relatively simple system, which may be taken as comparable to the Probabilistic Information Processor (PIP) system, which has been proposed as a tool for strategic intelligence analysis. The second subsection describes in detail some of the problems that have been observed in the actual operation of such systems. In the third subsection, some of the characteristics of a more adequate system are outlined. Appendix

C contains a more detailed technical description of the computer-based systems discussed here and in Section 6.

10.1. COMPUTER-BASED DECISION SYSTEMS

To provide concrete examples of computer-based systems which could be adapted to the production of uncertainty estimates, we shall refer to a class of medical diagnosis systems. Such systems accept a large body of information, often in probabilistic form, concerning the relationships between symptoms and diseases. Data concerning individual symptoms are then obtained from a patient, and a diagnosis is produced, indicating the probability that various diseases or other conditions are present.

The internal logic of a medical diagnosis system could incorporate statistical methods like those described in Section 6 of this report. The methods for determining the probability of a specific disease are similar to methods for determining the probability that a weapon system will be developed, or the likelihood that a given level of production will be reached. Both medical diagnosis and estimative intelligence are concerned with predicting future events through the use of available information about present events. The "symptoms" are the data available to the estimator; the "diagnostic procedures" are general hypotheses concerning the relationships between observations and the intentions and capabilities of the USSR, the PRC, and other nations of interest. The "diseases," of course, are their weapon systems.

MYCIN is one of the best-known of the computer-based systems currently available. (Cf. Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green, and Stanley N. Cohen, "Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System.") "It relies heavily upon artificial intelligence (AI) techniques that were originally developed for problem solving outside the environment of clinical medicine." It therefore represents a more complete demonstration of the strengths and weaknesses of the AI approach than any that have so far been applied to estimative intelligence. It should thus provide a reasonable test of AI methods in a practical setting.

The ultimate aim of the MYCIN project has been to develop a computer-based system to which physicians will refer for advice concerning antimicrobial therapy. It does not simply output a probable diagnosis, but includes the justification that underlies the diagnosis. This is important in communicating the rationale for a given response, and thus for communicating the uncertainty to be attached to it. In this way, the user can reject a given diagnosis if the rationale appears questionable, or if extra-systemic knowledge suggests an alternative response.

Conversely, MYCIN performs an instructional function; the diagnostician must follow through the chain of reasoning underlying each diagnosis, thereby increasing his skill in handling the relevant procedures. The implementation is interactive, permitting the physician to step through a diagnosis, obtaining and entering more data only as these are required to verify a potential

outcome. A specific therapy is recommended in connection with the final diagnosis.

Other systems, including many existing Bayesian systems, fail to provide an explanation of their decisions, thereby leaving the user with a take-it-or-leave-it outcome; it would be more helpful to include, as MYCIN does, some means for permitting the user to review the decision, to obtain grounds for accepting or rejecting it. The user simply types WHY or HOW to get a detailed explanation from the system of the type of conclusion it is trying to draw. The WHY and HOW commands may be repeated, to obtain a complete picture of the chain of reasoning employed by the system.

The MYCIN program therefore may serve as an initial model of the type of system that might be applied to quantifying, aggregating, and communicating uncertainty in estimative intelligence. It is particularly interesting in that it attaches scores to its various diagnostic rules, using these in such a way as to output a rough measure of the uncertainty of its conclusions.

MYCIN represents a "judgmental model," in which the relationships between symptoms and diseases (or observations and predictions, if it were to be used in certain intelligence applications) are represented in terms of a decision table. Entries in the table take the form, "If A and B and C are present, then D may be predicted (with probability p)."

It differs from a "statistical model," in which Bayesian, linear regression, or other statistical techniques are used to combine probabilities. In a

statistical model, entries in a table may consist of probabilities, in the form "the probability of symptom S, given disease D, is p." (Probabilities are stored in this form, rather than the converse, because these probabilities are less likely to be affected by epidemic conditions, changes in climate, etc.) The probabilities are combined to obtain the probabilities of diseases or other conditions which may be present.

The statistical models are particularly interesting, because they provide us with examples of systems which implement the statistical techniques which were described in Section 6. We will claim that if a statistical system can be made to work for estimative intelligence, for aggregating the quantitative estimates of uncertainty, then such a system should work for medical diagnosis. On the other hand, if there are problems in a system for diagnosis, then those same problems must be considered in a system for estimative intelligence.

Our reason for making this claim is that diagnosis is a simpler and more straightforward task than estimative intelligence, for the following reasons:

- Medical diagnosis is based on a highly developed taxonomy, with a well-documented basis in the prediction and verification of a very large number of cases. There is no developed taxonomy for estimative intelligence, the number of cases is much smaller, and experience has not been well documented.

- Diagnosis is generally a repetitive task, in which both reasoning and testing have been performed many times, using well-defined

procedures. For this reason, diagnosis is a process that seems particularly appropriate for computer modelling. While the production of estimative intelligence is often repetitive, the presence of factors that are described by estimators as "intuition" suggests that its procedures are not always well-defined.

- Medical diagnosis is seen as vitally important, playing a life-and-death role. It has therefore received a good deal of attention from agencies which fund medical research. There is no doubt of the essential role of estimative intelligence, which again may play a life-and-death role in the nation; but it is only recently that there has been support for research studies in its methods.

- There is a very large body of recognized experts in medical diagnosis, who can provide information essential for the development of a computer system. The number of experts in estimative intelligence is much smaller, and their methods are based upon a fund of personal experience which may be more difficult to communicate to a computer system designer.

Because the problem of medical diagnosis is similar to the problem of intelligence estimation, but because it is also a simpler, better-defined problem, it therefore can provide us with a test of computer-based decision systems to determine some potential strengths and weaknesses of such systems. More importantly, it can show the direction which the development of a computer-based system for aggregating and communicating uncertainty might take. In

the next subsection, we will review some shortcomings of the systems. The final subsection will suggest some of the ways in which the shortcomings may be overcome.

10.2. PROBLEMS WITH DECISION SYSTEMS

Four general approaches to the development of decision systems for aggregating uncertainty may be identified.

10.2.1. Regression Analysis

Regression analysis does not attempt to mimic the procedures of a diagnostician but uses the relations between symptoms and diseases to develop equations, in which various weights are assigned to the symptoms which may be relevant to each disease. This approach was found to produce predictions which were superior to those of experienced human diagnosticians.

Studies in regression analysis have generally shown that this approach is overkill for problems like diagnosis. Instead of precisely-determined coefficients to be attached to the various elements that enter into a diagnosis, a much simpler scheme, in which the coefficients are either zero or one, has often proved effective.

The interactions among various symptoms may also be important. For example, the presence of fever plus a flushed face may be significant for one medical condition, where the presence of either symptom alone would not be

significant. Or, in an intelligence application, the presence of extensive radar facilities, together with development of a site for launching intermediate-range missiles, could be taken to indicate the presence of an ABM capability, where either of these developments alone would not be significant for predicting the presence of an ABM system.

For reasons like this, regression analysis has not been widely used for medical diagnosis and is not recommended as an approach to computer-assisted intelligence estimations.

10.2.2. Bayesian Systems

In Section 6, a discussion of Bayesian methods for aggregating and communicating uncertainty was presented. This approach has been strongly urged as the basis for computer-assisted decision systems, since it provides an automated way of correcting some of the human errors in estimation that were discussed in Section 9 of this report. For example, clinicians have been found to be too conservative in their estimates of the uncertainty of a diagnosis; an automated Bayesian system might assist them in obtaining better estimates.

The problems involved in the development of an operational Bayesian system have, unfortunately, proved to be quite overwhelming. The need for obtaining a large number of prior and conditional probabilities for use in the system has been especially difficult to fulfill. And if this is true in systems for diagnosis, it would be even more difficult to obtain a satis-

factory base for estimative intelligence, where experience is more limited, where the number of cases is much smaller, and where the appropriate definitions and classifications -- the taxonomy -- are much less stable.

Another major problem for Bayesian systems has been the need for assumptions of independence and exhaustiveness among the conditions. For example, we might assume (a) that there is a 20% chance that the Soviets would attack Egypt, and (b) that there is a 30% chance that they would attack Cambodia. However, the chance that they would attack both Egypt and Cambodia, thus engaging in a two-front conflict, is much smaller than a simple statistical combination of the two probabilities. In fact, one would suspect that the probability of the USSR voluntarily engaging in a two-front war is near zero. Thus, the two actions are not independent, since one would very nearly preclude the other.

Unfortunately, many of the Bayesian systems must assume independence of the various probabilities, even though this requirement cannot be met in practical applications to medical diagnosis or to estimative intelligence.

Finally, many of the systems for Bayesian analysis in medical diagnosis must assume that the diseases are disjoint; but practical experience shows that two or more medical conditions can occur at the same time. Similarly, we can expect that two or more military developments may occur.

Like many applications of artificial intelligence research to practical problem areas, Bayesian systems have suffered from a combinatorial explosion --

that is, the number of combinations of relevant factors grows much more rapidly than the number of factors themselves, and thus cannot be handled effectively either by the system or by its human users.

10.2.3. Branching Logics

Some systems have included branching tree-like structures, essentially representing a Bayesian decision tree. The user steps through the tree, with branches chosen by the system, based on information obtained from the user.

Difficulties observed with such systems have included serious problems in obtaining the probabilities required by the Bayesian approach, the rigidity of the sequence of decisions that must be made as the user steps through the tree, and the degree of effort required to modify and update the decision tree.

10.2.4. Simulated Logics

The approach outlined in the discussion of MYCIN in Subsection 10.1. uses a combination of logic and probabilistic reasoning to mimic the intentions of skilled clinicians. For MYCIN, this is a simulation of inexact or probabilistic logics to obtain a "score" for the proposed diagnosis. Because of the difficulty in estimating prior probabilities for large numbers of correlated, overlapping events, this score must not be interpreted as a probability. Thus, it does not provide a measure of uncertainty as we have defined it in this report.

Without probabilities, systems like MYCIN are generally little more than relational data bases which provide some sort of ill-defined number which a doctor can use as he sees fit while making his differential diagnosis. For the consumers of estimative intelligence, such numbers would have little value, other than to provide a rank-ordering of alternative projections.

10.2.5. Problems in Existing Systems

The problems that have arisen in medical diagnosis systems will be summarized here, since they appear to be very similar to problems that may be expected to arise in any automated system to assist in estimative intelligence.

- The patient may have more than one disease.
- Symptoms are unrealistically considered independent or uncorrelated.
- The underlying distribution is not completely known.
- Most known probabilities are usually low-order conditional, marginal probabilities.
- Known probabilities may be obtained from different studies and therefore may be inconsistent.
- Subject matter knowledge, such as cause, effect, and "intuition" (i.e., broad understanding of the context), should be used to specify the distributions.

- The size of the distributions involved may not be known.

No existing medical diagnosis system deals with all seven of these problems. Since diagnosis, as noted in Subsection 10.1., is similar to, but less difficult than, estimative intelligence, it is likely that systems for intelligence applications may suffer from similar difficulties.

It is possible to develop parallels between the difficulties sketched here for diagnosis systems, and those which might be expected to occur in intelligence systems. For example, in parallel with the first problem, more than one military or political situation may be present. Secondly, the various pieces of information concerning such situations are not uncorrelated or independent. Underlying probability distributions are not known. Even when probabilities are known, they may be inconsistent or unreliable.

The sixth difficulty listed above is particularly significant, and this may account for some of the resistence to automated systems: most existing systems cannot incorporate cause-effect relationships and broader subject-area knowledge which is known to the estimator, but which does not fit the structure of the particular computer-based system.

Finally, the problem of sample size will be quite difficult for intelligence applications, since the available data may be severely limited. Under these conditions, it may be impossible to arrive at reasonable estimates for the probabilities required by the system.

10.3. A POTENTIAL SOLUTION

Other members of our staff have dealt with the problems discussed here in connection with systems for medical diagnosis and for other pattern-recognition applications. We believe that the problems which have arisen in attempts to apply academic models to real-world situations can be successfully overcome, if the intended application is known and the system is developed for effective use in the application.

A full discussion of the recommended approach requires a presentation which is somewhat beyond the mathematical level of the main sections of this report, and it is therefore included as Appendix C. The approach described there is, we believe, adequate to meet the difficulties which have been outlined in subsection 10.2.

Implementation of any system will require a review of the methods which have been discussed in Section 5 of this report. Since the body of knowledge underlying estimative intelligence methods is not as well understood as that of medical diagnosis (cf. Subsection 10.3.) or weather prediction (cf. Section 7), it would be premature to expect that a fully-automated system could be designed and programmed at once. Instead, a very few functions could be designed to provide assistance to the estimators, who would still rely heavily on background knowledge and informal reasoning processes to produce their projections and other estimates. As the supply of data and experience grows, additional functions could be added, as options, and used whenever estimators

found them valuable. There is absolutely no reason for using a system that does not actually assist the estimator in producing better and faster results.

Among the steps that might be found helpful in the development, in chronological order, are the following:

- Implementation of TEAMS, to provide a rapid method for locating trends, errors, biases, and uncertainties in previous projections. TEAMS is essentially a diagnostic device, for locating trouble spots in the estimative process.

- Development of the DIPPOLS data management system to provide a uniform method for accessing both current and past projections. This will assist in determining the degree of uncertainty which was actually present in earlier projections, and assist in determining uncertainties in current projections.

- Provide graphic outputs for DIPPOLS/TEAMS which will assist in showing the degree of uncertainty both in OB data and in DE projections. Graphic presentations will help DE users to visualize the uncertainty which has been present in their projections.

- Provide storage space in DIPPOLS for an institutional memory, which clearly indicates the justification for each projection. In the terms of this report, it will be the set of hypotheses which have led to the specific projection, in such a form that any future user

can readily determine the assumptions that entered into a projection. Thus, it will be possible not only to determine that a projection has gone wrong, but also why it has gone wrong.

- As more data concerning the formal character of the estimative process are gathered, it will be possible to consider automation of some aspects of the projection process. For example, the use of time-series analysis for more precise interpolations and extrapolations of trends in the data can be considered. Extrapolation of present trends can provide the estimator with a base-line against which to compare projections; trends may continue, or they may increase or decrease at some rate which can be supplied by the estimator. Computer-based methods can provide a more precise and automatic method for projecting trends.

- Artificial intelligence (AI) systems, like the diagnosis systems described in this section, can be tested on an experimental basis, to determine their potential usefulness to the estimator. As noted in Subsection 10.2., there are serious problems in existing systems. Problems which have been noted in systems for medical diagnosis are likely to be exacerbated in the more complex, less well-understood context of strategic intelligence production; nevertheless, it would seem possible that a more sophisticated system, like that described in Subsection 10.3., could provide a basis for further research in the automation of intelligence estimates.

SECTION 11

ASSESSING AND COMMUNICATING UNCERTAINTY

11.1. INTRODUCTION

In this section, computer-based methods for assessing and communicating uncertainty in DE forecasts are provided. The general procedure is focused on an intelligence databank called the "institutional memory" (IM). The IM provides a facility for automated documentation of DE projections and is an archive for past and current forecast descriptions. Individual estimators provide inputs to selected IM entries and, for each recorded projection, document important elements of the estimative process.

The IM is the core of a knowledge-intensive approach emphasizing storage of estimators' key insights and judgments. This differs from a data-intensive procedure where the emphasis is on retaining pure statistical data. Stored knowledge, in the IM, may include direct assessments of projection uncertainty as well as general descriptive information. The direct uncertainty assessments may be communicated through DE publications, such as the DIPP, or they may be limited to use within DE. The general descriptive information stored today may provide the basis for uncertainty assessments made tomorrow.

More specifically, the IM can be used to record the reason for numerical estimates (stored in the DIPPOLS/TEAMS data base) or to document nonnumeric (qualitative) projections. Qualitative projections, stored in the IM, can be reviewed as part of a calibration process allowing trends toward over- or

underconfidence to be located. Numeric projections can be calibrated using TEAMS routines as well as the Calibration Assessment Package (CAP).

Two procedures, supporting the IM, have been programmed for use on DE's Honeywell 6080 computer. One system, Subjective Probability Assessment (SPA) system, aids estimators in estimating consistent judgmental probabilities. The other system, Calibration Assessment Package (CAP), provides estimators with an automated procedure for calibrating projection data. A user manual is provided for each system.

In the remainder of this section, the following topics are covered:

1. a global examination of the IM and its integration into TEAMS,
2. the composition of IM entries (Section 11.3.),
3. a description of some of the analysis procedures supporting the IM (Sections 11.4., 11.5., and 11.6.).

11.2. THE IM: A GLOBAL PERSPECTIVE

In this section, the IM is examined from a global perspective. Integration of the IM into existing DE systems and procedures is discussed.

Integrating the IM into TEAMS

TEAMS provides a solid supporting base for the IM. As illustrated in Figure 11-1, the TEAMS design has been modified to incorporate the IM and some additional output channels. The visible additions include:

1. The IM data base (file) consisting of documented judgments and assessments with historical and current projections. The IM can store permanent and working records (much the same as TEAMS) which allows historical records to be preserved and current records to be updated as the forecast is developed.

2. The report generator which formats IM information for various output channels such as NIE and DIE reports as well as the DIPP.

In addition, changes have been made to existing DE modules such as:

1. The executive module, where the addition of file management and data retrieval routines provides general control over stored records. In addition, a text processing subsystem enables estimators to enter textual information.

2. The statistics package which includes additions such as the SPA system and CAP as well as several uncertainty statistics.

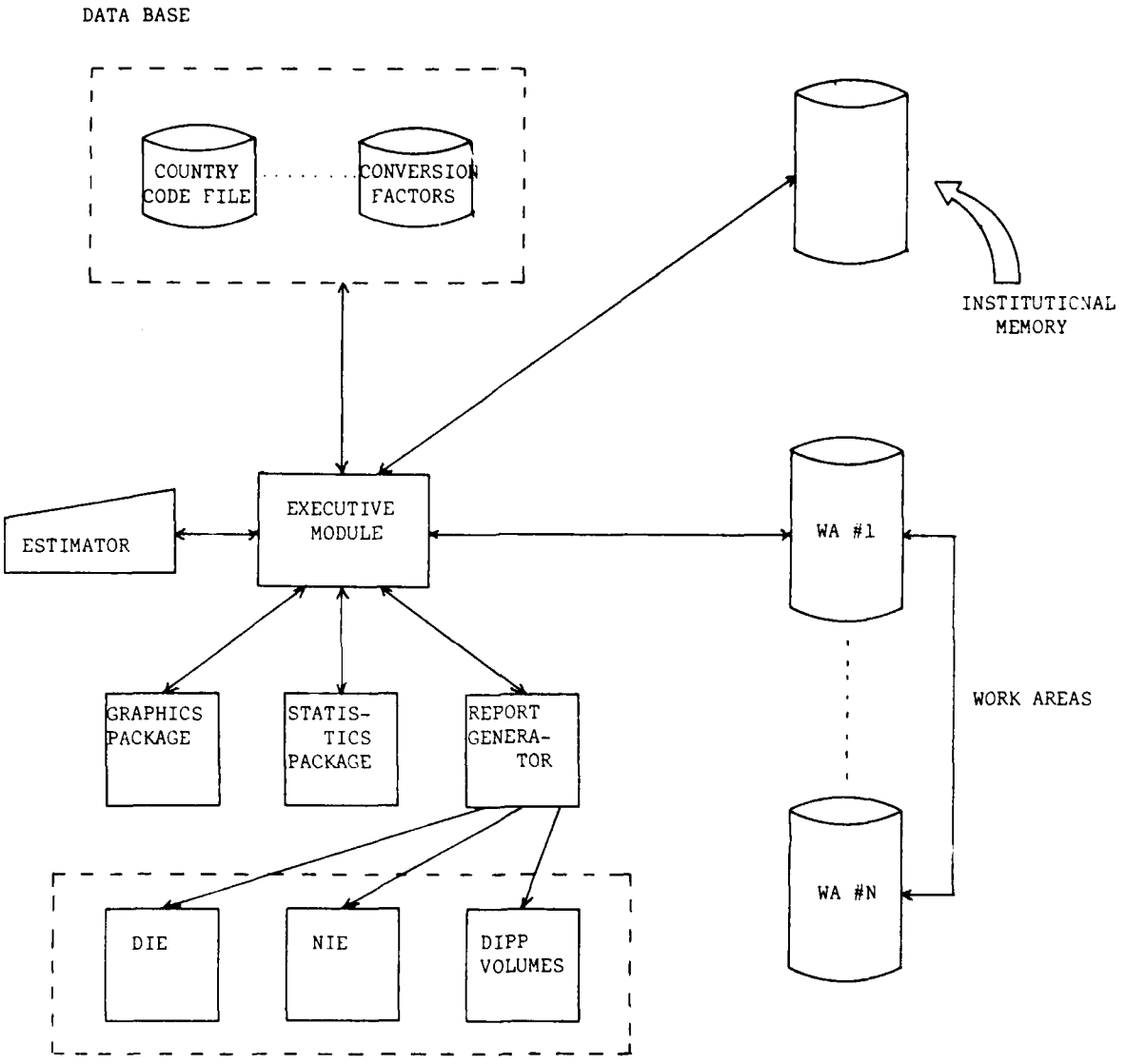


Figure 11-1 Modification of TEAMS to Incorporate the Institutional Memory

IM Procedures

The IM provides estimators with a number of general procedures which are quite similar to TEAMS operations. For example,

1. Maintenance of permanent (historical) and working (current) records. Estimators can access permanent records to perform historical analyses and working records to update documentation of current projections. The process of "permanitizing" working records (employed in TEAMS/DIPPOLS operations) could be expanded to include IM working records.

2. Inputting data to the IM. Many IM inputs can be generated independent of the IM and stored in the work area (WA). Refined data could then be directed to the appropriate IM entry.

3. Retrieving IM entries. IM entries can be indexed according to projection keys (descriptors) much the same as in the TEAMS/DIPPOLS data base. Therefore, IM records may be labelled according to Country Code, Force, System, and Projection Year. As a result of TEAMS, this retrieval mechanism is largely developed.

4. Using TEAMS statistics as basis for retrieval. TEAMS statistics such as, ERROR,BIAS,UNCERTAINTY, and SCORE could be used to locate anomalous projections. The reasons behind the numerical projections could be retrieved from the IM.

5. IM retrievals based on stored data. Estimators may identify documented information which correlates with specified levels of accuracy in forecasts. This information could then be used as a basis for retrieval. For example, all projections for Soviet ABM systems which are based on the United States deploying a cruise missile could be retrieved from the IM.

6. Generating reports. Specified elements of IM entries may be specially formatted for output to DE products.

11.3. COMPOSITION OF IM ENTRIES

The general format of IM entries is presented in this section. Two objectives of the format specifications are:

1. To provide a degree of structure in the framework of an entry. This facilitates computer storage and retrieval operations. In addition, it provides estimators with a file format decomposed into logically consistent information blocks.

2. To provide a degree of flexibility in each information block. For example, the format of information within a block is not critical. Each block entry might be written like text in a report or an entry in the DIPP.

In Figure 11-2, the components or blocks of a typical IM entry are presented. Before discussing each block in detail, note that the block labels and the block sequence correlate positively with the format in which projections are presented in the DIPP. To the left of each block, the block number is printed, to aid this discussion. To the right of each block, tools used to generate block information are listed.

11.3.1. Block 1 - File Identification Code

This block consists of a file code or index label which can be used to identify and locate an IM entry. In the computer this code would be represented as a number, but logically, to the estimator, the index consists of projection identification keys (possibly, Country Code, Force, System, and Projection Year).

11.3.2. Block 2 - General Assumptions

The general assumptions provide a "baseline" for forecasts. They may consist of elemental assumptions which are central to the validity of the estimates. For example, it may be assumed that there will be no war in the Middle East, NATO and other key alliances will not dissolve or substantially lose their effectiveness, and the Space Treaty will be in effect. Historical analyses of past projections would address the validity of the general assumptions and determine if those statements were in error.

BLOCK NUMBER	BLOCK DESCRIPTION	IM TOOLS
1	File Identification Code	
2	General Assumptions	Text Editor
3	Specific Assumptions/Factors	Text Editor
4	Script - Qualitative Forecast	
	o Most Likely Scenario (Detailed Description)	Text Editor
	o Alternative Scenarios (In Capsule Form)	
5	Sensitivity Analysis	
	o Qualitative description of forecast sensitivity to assumptions and factors	Text Editor
	o Quantitative description of forecast sensitivity to assumptions and factors (measures of uncertainty, sensitivity and robustness)	SPA Statistical Measures
6	Historical Analysis	Text Editor
	o OB Data Uncertainty Measure	OB Uncertainty Measure
	o Calibration Assessment	CAP

Figure 11-2 Information in IM Entries

Inputting assumptions to Block 2 could be nearly automatic when the TEAMS work area is used. For example, general assumptions can be named and stored in the work area. If the assumptions need to be edited, appropriate edit routines are called. If the assumptions are not edited they are simply transferred into the appropriate IM entry.

11.3.3. Block 3 - Specific Assumptions/Factors

Specific forecast assumptions are often more closely associated to one or a group of related projections than to the full spectrum of forecasts. They may not describe as far reaching events as the general assumptions, but instead point to factors which are most likely to influence selected forecasts. For example, assume that the Soviets are emphasizing production of air superiority fighter aircraft due to a perceived threat from the F-18. This assumption has direct impact on the production and deployment of the MiG-23, for example, and virtually no effect on production of the Mil Mi-8 (NATO "HIP") helicopter.

Essentially, then, the estimator documents assumptions, factors or forecast variables that are likely to influence weapon production and deployment rates. All of these elements may have been linked into a unifying scenario which describes various world event interactions. (This scenario will be discussed in Block 4.)

11.3.4. Block 4 - Script

The qualitative or nonnumeric forecast is documented in the script. This forecast is often communicated as a scenario and reflects expected critical events, anticipated adversary intentions, assumed political interactions, etc. In fact, DE products, such as NIE's, often present assessments in the form of scenarios. Scenarios may play a key role in communicating forecast uncertainty and therefore, the purpose of the script is twofold:

1. Provides documentation for the anticipated or most likely scenario

2. Supplies a means of stating alternative scenarios.

The two areas of documentation are useful from several perspectives; for example:

1. The most likely scenario is documented and therefore available for review. This scenario is useful in conveying the reasoning behind the numerical estimates.

2. The uncertainty associated with the most likely scenario is more readily assessed when alternative scenarios are available for comparison.

Although the most likely scenario may be quite explicit (in terms of communicating forecast factors) it may be difficult to assess the uncertainty

associated with the single scenario. (This is partly true because scenarios are often causally-related event sequences that are primarily composed of statements of fact or, in certain instances, uncertain events that are treated as facts. Scenarios are often not accommodating to statements of uncertainty.) A useful procedure may be to specify alternative scenarios in the IM and possibly in the DIPP and other DE products. A range of plausible outcomes, conveyed through alternative scenarios, aids in communicating the uncertainty associated with accepting the most likely scenario as well as numerical estimates that may be based on it.

The primary purpose of the script, then, is to document the full range of plausible events. In a qualitative sense, the degree of belief that the estimator has in the numerical forecast can be assessed, because alternatives, in a sense, convey uncertainty. Of course, it is likely to be difficult, if not impractical, to develop alternative scenarios with the level of detail found in the most likely scenario. The practical solution might be to provide alternative scenarios in capsule form, highlighting the primary factors and assessed events. In Block 5, a procedure for quantifying the uncertainty conveyed through alternative scenarios is given.

11.3.5. Block 5 - Sensitivity Analysis

Information stored in Block 5 is derived from sensitivity analyses. Procedures for conducting sensitivity analyses are discussed; however, it may be useful to first review the information content of Blocks 2 through 4. Blocks 2 and 3 serve to decompose the forecast into key assumptions and

factors. This procedure follows a decision-theoretic approach to decision making in which a prime objective is documentation of the important decision variables. In Block 4, a qualitative expression of the forecast is presented and generally unifies the information specified in Blocks 2 and 3. Although the forecast has been stated at the global and atomic levels, statements as to the uncertainty of the forecast have not been provided. This, then, is the role of the sensitivity analysis Block and its supporting routines; that is, to provide a range of procedures for assessing and communicating the uncertainty of the forecast. The remainder of the discussion on Block 5 first addresses the general procedure for conducting sensitivity analyses (defining the terms sensitivity and robustness). Next, individual approaches to the analysis are discussed and, following that, procedures for quantifying forecast uncertainty and robustness are presented. The quantification techniques are presented first conceptually and second through example. The last topic is qualitative procedures for conducting sensitivity analyses.

General Procedure and Definitions

Consider a numerical projection supported by stated assumptions and a descriptive scenario. If the numerical estimates are insensitive to deviations to stated assumptions, the projection is said to be robust with respect to those assumptions. If the projection is discounted should the assumptions be in error, the projection is said to be sensitive to departures in the stated assumptions. For example, assume that the production level for Soviet SS-16's is not strongly tied to a SALT agreement (even though SALT constrains the number of missiles that can be deployed.) Then, incorrect assumptions

concerning the existence of SALT may not affect the accuracy of the SS-16 forecast. The SS-16 is said to be robust to the assumption that SALT will exist. The approach to sensitivity analyses consists of identifying key forecast factors and assumptions and determining their impact on projection accuracy under changing world conditions. The following discussion addresses approaches to measuring the sensitivity and robustness of projections as well as the overall uncertainty.

Various Approaches to Sensitivity Analysis

There are many approaches that can be used in assessing the sensitivity or robustness of projections as well as the uncertainty. One is what is often called intuitive or holistic. In the holistic approach the estimator assesses the uncertainty, for example, using intuition, experience, or collective knowledge. Another approach may be termed rational or analytic. In the rational approach, the assessment of uncertainty is determined after having dissected the problem into its component features. In addition to the two approaches, the sensitivity analysis may be conducted on two bases. One basis is qualitative and the other is quantitative. Therefore, there are a number of approaches that can be employed. Selected procedures supported by the IM and affiliated routines are discussed next. The first approach is quantitative and employs the SPA system; the second approach is qualitative.

Quantifying Assessments of Uncertainty, Sensitivity and Robustness

Quantitative procedures for assessing forecast uncertainty, sensitivity and robustness are discussed first, in general, and then through example. The approach is holistic and focuses, first, on the script (Block 4). Assume that the most likely scenario is stated along with a number of alternative scenarios. To assist the estimator in conveying his degree of belief for each of the scenarios, a rank ordering of scenarios, based on their likelihood of occurrence, would be useful. As a first step, the ranking could be entirely qualitative such that the estimator would simply order the scenarios in descending fashion to convey the assessed decrease in the probability of occurrence. However, it is often not sufficient or even practical to simply order the alternatives based on their likelihood. In fact, significant information is often lost when only the ranks are supplied. For example, if along with the most likely scenario, call it A, three alternative scenarios are specified (B, C and D), it may be that a rank ordering would be: A, C, D, B. However, there is not any information supplied in the ranks alone which allows the degree to which one scenario is more likely than another to be inferred.

The ranking procedure, then, is suggested as a first step in structuring the probability assessment. To supplement the rank order procedure, a number of numerical assessment techniques are available for assigning probabilities to scenarios. For example:

1. Category methods which involve classifying the probability associated with a scenario into a fixed number of discrete categories.

2. Direct methods which involve the direct assessment of the probabilities. The estimator may be required to assign a probability to a scenario based on some internal assessment procedure (which often results in inconsistencies in the assessments).

3. Gamble methods which involve structuring wagers and then varying probabilities until the estimator is indifferent to the bets. The resulting probabilities reflect the estimator's beliefs.

The suggested approach to assigning probabilities is derived from a scaling procedure [Saaty], and is a variation of the direct estimation technique. The procedure employs compared probabilities or likelihood ratios. (As will be discussed later, this procedure is general in that the ratios can also be in units of value, sensitivity, importance or virtually any other comparison scale.) A compared probability, for example, consists of the ratio of the probability of one scenario to that of another. To illustrate, assume that an estimator has defined four scenarios as being plausible and that it is desirable to order or scale them as to their likelihood of occurrence. The estimator performs a pairwise comparison among the four scenarios. For example, the likelihood of scenario A compared to scenario B, L_{AB}, is by definition a ratio of their respective probabilities. However, the scaling method requires only that the ratios be assessed. The ratios of paired comparisons are stored in matrix form and solution of the subsequent eigenvalue problem results in the underlying probability assessments. (Note that a further description of the mathematical procedures involved in this method is provided in Section 11.4.)

The problem of extracting subjective or estimated probabilities from decision makers is difficult. Although the decision maker may determine which scenarios are more likely, it is difficult to quantify intuition. When the number of scenarios gets large, the difficulty of the problem increases. However, it is generally easier for the decision maker to compare the scenarios two at a time. The proposed scaling method is based on pairwise comparisons. The important feature of this method is that it allows estimators to measure the consistency of their assessments. The Subjective Probability Assessment (SPA) system, developed under this contract, can be used by estimators to estimate probabilities. It utilizes measures of consistency to direct the estimator to intuitively satisfying and mathematically consistent subjective probabilities.

The scaling method is illustrated. Assume that the basic task of an estimator is to predict the number of MiG-23's that the Soviet Union will deploy in the next 11 years. The estimator has developed a range of scenarios which capture the more plausible event sequences. (For this discussion, the scenarios are simplified.)

Scenario A: The MiG-23 will continue in its role as a fighter/interceptor aircraft. However, it is anticipated that the Soviets will take steps to counter the strike capability of the United States' cruise missile. The Soviets will likely develop a look-down radar capability for the MiG-23. This enhancement will give the MiG-23 the capability of detecting and destroying the cruise missile in limited weather conditions and terrains. MiG-23 development will increase significantly to satisfy this added role.

Scenario B: The Soviets will not modify the MiG-23 with look-down radar but instead will continue to utilize it in its current role. Instead, to cope with the cruise missile threat, ground-based radar and missile systems will be enhanced. Possibly the SA-9 (GASKIN) missile will be employed. Production of the MiG-23, under this scenario, is not affected by deployment of the cruise missile.

Scenario C: The Soviets will deemphasize the role of the MiG-23 as an interceptor aircraft and instead increase production of replacement aircraft, the MiG-27. The shift in emphasis will be slow, probably occurring over 3 or 4 years, the time necessary to develop production capabilities for the MiG-27.

Therefore, the stated scenarios consist of the most likely (scenario A), and the alternatives (B and C). These scenarios are documented in Block 4 of the IM while the underlying assumptions for scenario A are stored in Blocks 2 and 3.

The next step in the assessment is to perform a pairwise comparison among the three scenarios. In support of this, an assessment scale [Saaty] is used. (Note that use of this scale or any arbitrary yardstick is not crucial to the general technique, although it does encourage a consistent approach to its use. SPA allows the user to employ a stored scale or define another with only minor constraint on that definition.) Using the scale given in Figure 11-3, the following pairwise comparisons were performed and stored in matrix form as illustrated in Figure 11-4.

The Scale and Its Description

Intensity of Importance	Definition	Explanation
1	Equal Importance	Two activities contribute equally to the objective
3	Weak importance of one over another	Experience and judgment slightly favor one activity over another
5	Essential or strong importance	Experience and judgment strongly favor one activity over another
7	Demonstrated importance	An activity is strongly favored and its dominance is demonstrated in practice.
9	Absolute importance	The evidence favoring one activity over another is of the highest possible order of affirmation
2,4,6,8	Intermediate values between the two adjacent judgments	When compromise is needed
Reciprocals of above nonzero	If activity i has one of the above nonzero numbers assigned to it when compared with activity j, then j has the reciprocal value when compared with i	
Rationals	Ratios arising from the scale	If consistency were to be forced by obtaining n numerical values to span the matrix

Figure 11-3 Assessment Scale

	A	B	C
A	1	6	4
B	1/6	1	1/2
C	1/4	2	1

Figure 11-4 Comparison Matrix

- Scenario A is strongly favored over B with intensity 6.
- Scenario A is slightly favored over C with intensity 4.
- Scenario C is weakly favored over B with intensity 2.

(Note even though the scale implies that it is discrete, it is acceptable to define it as continuous such that intensities such as 2.5 are permissible.) Therefore,

L_{AB} = the likelihood of A over B = 6

L_{AC} = the likelihood of A over C = 4

L_{BC} = the likelihood of B over C = 2

and by definition,

$$L_{BA} = 1/L_{AB} = 1/6, \text{ etc.},$$

and,

$$L_{AA} = L_{BB} = L_{CC} = 1.$$

Solution of the eigenvalue problem using the data in Figure 11-4 results in the following probabilities:

Probability of scenario A = .70

Probability of scenario B = .11

Probability of scenario C = .19

A consistency measure, μ, is computed and tested against a decision parameter, d. If μ is less than d, reasonable consistency is attained and the estimator may not wish to revise the original pairwise comparisons. In this case, the estimator is assumed satisfied with the results. (The reader may wish to consult the SPA user manual for a discussion of how consistent pairwise comparisons are attained.) Finally, the resulting probabilities can be graphically displayed as shown in Figure 11-5.

Using the generated probability for each scenario an uncertainty statistic may be defined and computed:

$$\text{Uncertainty} = U = 1 - P_{max}$$

where P_{max} is the probability associated with the most likely scenario, A. U, then, is an estimate of the uncertainty associated with accepting the most likely scenario, A. For example, if an intelligence consumer was to employ the expected scenario with associated numerical estimates in a war-gaming exercise, the measure, U, would give insight into the risk associated with basing decisions and actions only on the most likely scenario. For the previous example, concerned with estimates of MiG-23 production, the probability associated with the expected scenario is estimated to be .70. Therefore, the computed uncertainty associated with the most likely scenario is

$$U = 1-.70 = .30 = .11 + .19$$

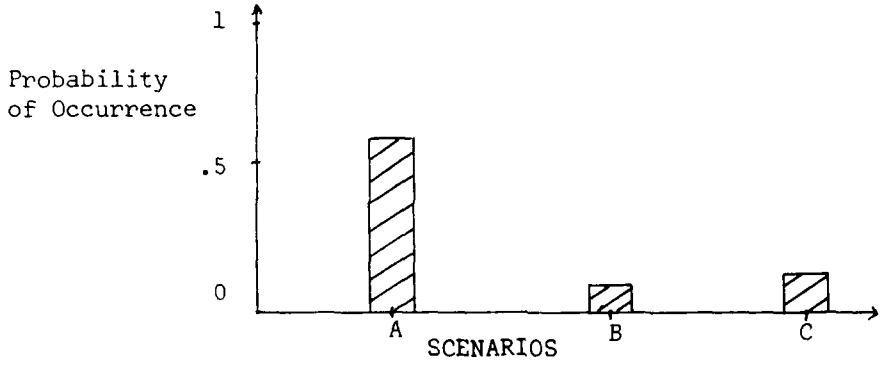


Figure 11-5 Assessed Probability of Occurrence

Uncertainty estimates can be computed in a like fashion for various weapon systems resulting in a straightforward procedure for assessing the relative degree of confidence that estimators have in a collection of projections. As illustrated in Figure 11-6, the uncertainty estimate for each weapon system in a group can be depicted graphically allowing recognition of particularly uncertain projections.

The assessed subjective probabilities for each scenario can be used as a base for the sensitivity analysis. In the procedure, the probabilities will be coupled with directly assessed measurements of sensitivity to generate a robustness measure. This quantitative measure of robustness can be used in conjunction with a qualitative assessment to provide a more complete appraisal of the overall confidence that the estimator has in a projection.

In computing a measure of robustness, a subjective measure of sensitivity for each alternative scenario is assessed by the estimator. This sensitivity measure, S, is defined to take on values between 0 and 1. The more sensitive a forecast is, with respect to a particular set of assumptions, the closer the value of S is to 1. As assessment that a forecast is insensitive would result in a sensitivity coefficient close to 0. Using sensitivity coefficients and probabilities assigned to alternative scenarios, an overall robustness measure, R, can be computed. The procedure for computing R will be described using an example. Referring to the previous example, a rank ordering and scaling of the scenarios, as to their likelihood, resulted in the following probabilities:

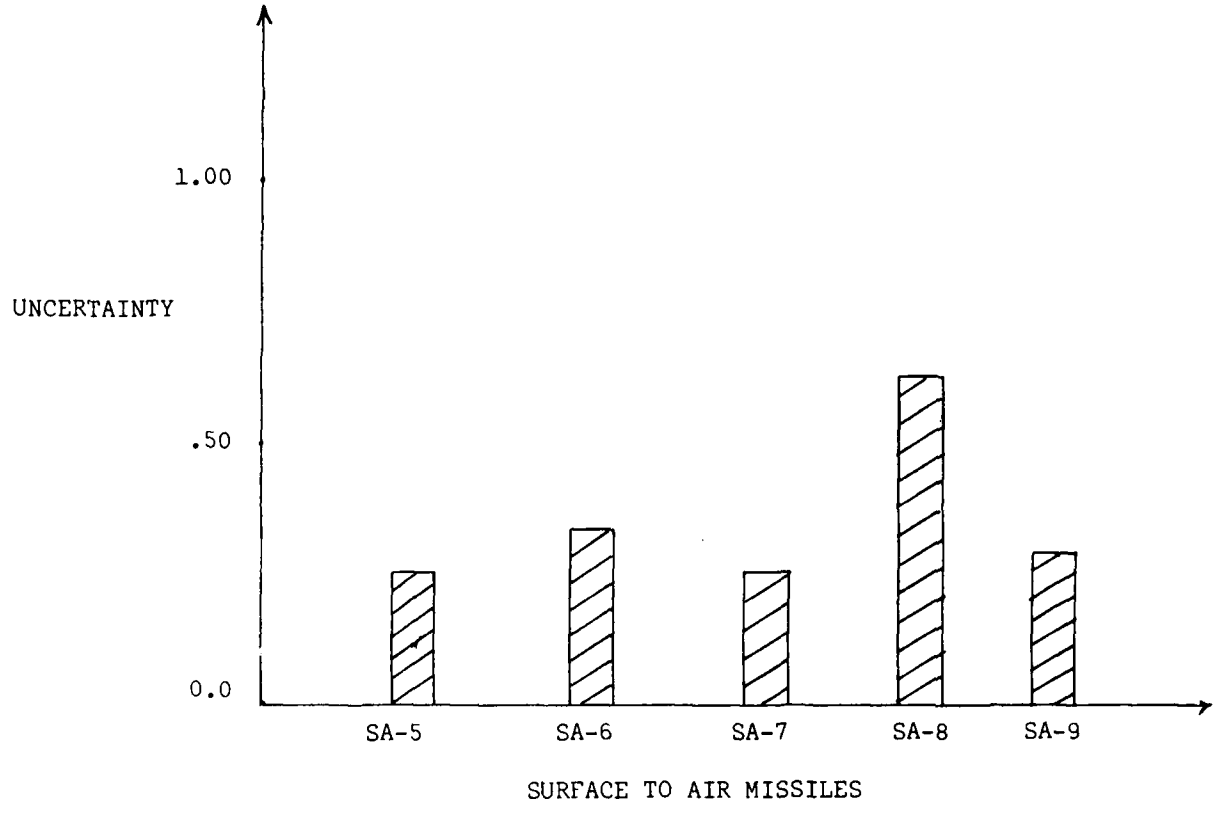


Figure 11-6 Assessed Uncertainty for a Group of Weapon Systems

- o Probability of scenario A = .70.....the most likely scenario
- o Probability of scenario B = .11.....the least likely alternative
- o Probability of scenario C = .19.....the most likely alternative.

Assume that the following sensitivity coefficients for each of the alternative scenarios were determined:

- o Sensitivity of the numerical estimates should B occur = .8
- o Sensitivity of the numerical estimates should C occur = .3

Note that the estimator has reflected the following in his quantitative assessment of uncertainty and sensitivity:

> Scenario A has a high likelihood of occurrence, .70. However, if A does not occur, scenario C is the next most likely; in fact, C is approximately 2 times as likely as the other alternative, B. Although scenario C is quite plausible, its realization would not seriously affect the validity or utility of the numerical estimates. Scenario B, should it occur, would essentially invalidate the numbers provided; however, its likelihood is quite small.

The measure of robustness aggregates the uncertainty and sensitivity assessments in the following way:

$$R = 1 - \sum_{i=1}^{n-1} \frac{S_i P_i}{(n-1)/n} = 1 - n \sum_{i=1}^{n-1} \frac{S_i P_i}{n-1} \qquad (1)$$

where n = the number of scenarios. Therefore, n-1 is the number of alternative scenarios.

S_i = the sensitivity coefficient for each of the n-1 alternatives for i=1 to n-1.

P_i = the probability of each alternative scenario for i=1 to n-1.

Note that in equation 1, the term in the denominator, (n-1)/n, is a normalizing factor which restricts R to the interval between 0 and 1. For the example cited, the following is computed:

$$R = 1 - 3\ (.8(.11) + .3(1.19))/2 = .78 \text{ (rounded)}$$

Therefore, the projection is assessed to be relatively robust, since the closer the value of R is to 1, the more robust the projection is assessed to be. The robustness measure, in addition to the uncertainty measure, provides another means for comparing a collection of projections. For example, as shown in Figure 11-7, robustness is plotted for a group of related weapon systems.

Given two measures of the quality of a projection, a scatter diagram could be generated for a group of related weapon systems. For example, in Figure 11-8, the scatter diagram aids in identifying those projections which are particularly uncertain or lacking in robustness or both. In the sample plot, the projection for the SS-18 is determined to be both significantly uncertain and sensitive to deviations from assumptions.

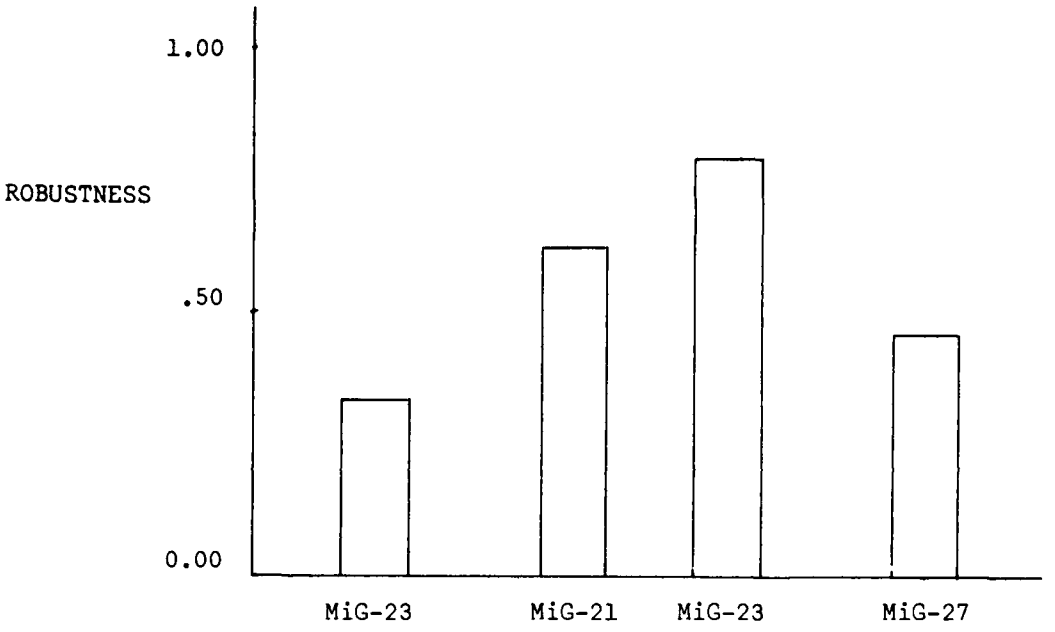


Figure 11-7 Comparing Robustness Measurements for a Group of Soviet Fighter/Interceptor Aircraft.

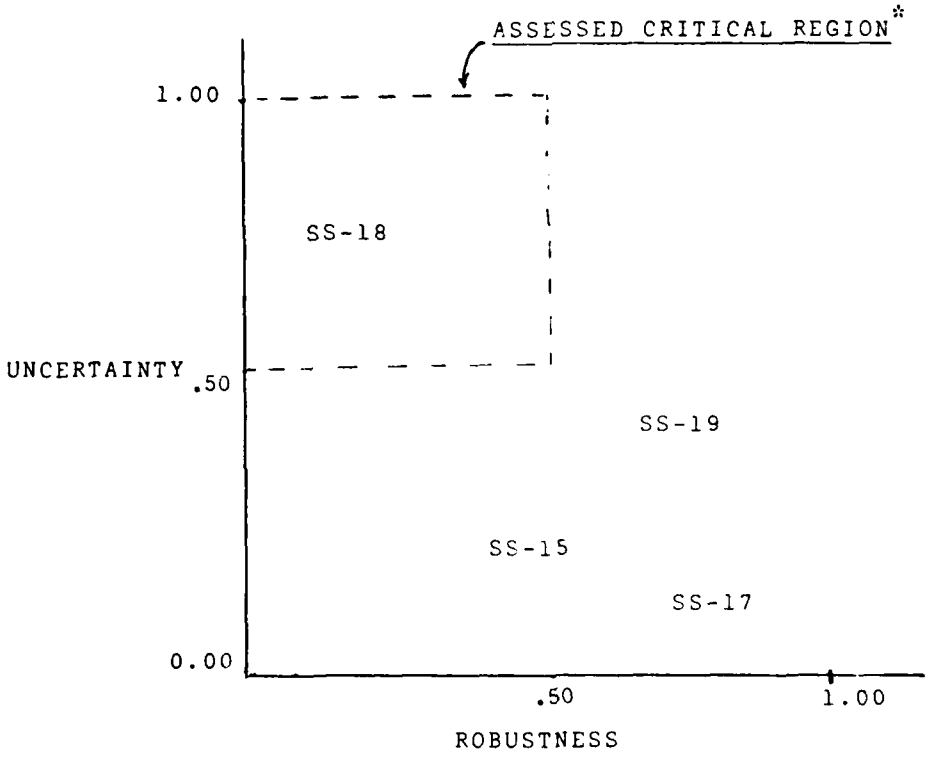


Figure 11-8 COMPARISON OF UNCERTAINTY AND ROBUSTNESS ASSESSMENTS FOR SELECTED SURFACE-SURFACE MISSILES.

* Projections falling in this region are determined to be particularly risky. A high degree of uncertainty coupled with a significant sensitivity to deviations from assumptions makes these projections questionable in terms of validity.

Thus, the robustness statistic, in a sense, summarizes the findings of the sensitivity analysis by providing an indicator of the reliability of the projection. It takes into account not only probability but the likely effect should alternative events occur. Some brief statements follow, concerning a qualitative approach to the sensitivity analysis.

Qualitative Assessments of Uncertainty, Sensitivity and Robustness

The approach to a qualitative sensitivity analysis can be rational or holistic. The holistic approach might consist of an overall appraisal of the uncertainty associated with a projection, pointing to various elements of the scenario which might be contributory. The analyst would most likely give a global view of the sensitivity or conversely the robustness of the projection under varying "world views". Another approach would be to dissect the qualitative expression of the forecast, assessing the contribution of each key assumption and factor to the overall uncertainty. This rational approach might involve a listing of the key projection factors accompanied by a qualitative assessment of their individual importance to the forecast. The following table is suggestive of the rational approach:

FACTOR	PARAMETER	FORECAST SENSITIVITY
Production rates for MiG-23 factories	Manpower	LO
	Raw Materials	LO
	Time after initial production	MED
Performance/Modifications for the aircraft	Aerodynamic	LO
	Avionics	HI

In this table key factors and associated parameters have been identified. The sensitivity of the forecast with respect to the stated parameters is given in the right hand column.

11.3.6. Block 6 - Historical Analysis

This Block is used to document the results of pertinent historical data analyses. The analysis results may have been significant inputs into the current projection and for that reason are given special consideration. Two historical data assessment procedures are proposed. The first is discussed in Section 11.5 and focuses on the accuracy of past OB data. Its primary purpose is to detect trends toward improvement or deterioration in the OB update sequence; in other words to determine if the accuracy of the OB varies significantly over time. The second procedure is employed in calibration assessments. The calibration assessment detects trends toward over- or underconfidence in past estimates. The Calibration Assessment Package (CAP) is provided as separate programmed package under this contract. It provides a statistically-based calibration procedure and is supported by tabular and graphic output.

11.4. THE MATHEMATICAL BASIS FOR ELICITING SUBJECTIVE PROBABILITIES

The Subjective Probability Assessment (SPA) system elicits comparison ratios from decision makers for defined items. The comparisons may consist of probability ratios, ratios of values, etc. For the purpose of this discussion,

we will mostly consider compared probabilities. From the compared probabilities, or ratios, the desired probabilities are obtained using a scaling procedure [Saaty]. The scaling procedure is outlined as follows.

Assume that the following set of items are of concern to the decision maker:

$A_1, A_2, \ldots\ldots, A_n$ and it is desirable to determine the set of probabilities

$P_1, P_2, \ldots\ldots, P_n$ associated with each item. (The items may be events or scenarios.)

It may be difficult to obtain the P_i's directly, however, it may be much easier for the decision maker to compare two items at a time, that is, determine the probability ratios. Consider the matrix of compared probabilities, M, where,

$$M = \begin{bmatrix} P_1/P_1 & P_1/P_2 & P_1/P_3 \ldots\ldots\ldots P_1/P_n \\ P_2/P_1 & P_2/P_2 & P_2/P_3 \ldots\ldots\ldots P_2/P_n \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ P_n/P_1 & P_nP_2 & P_n/P_3 \ldots\ldots\ldots P_n/P_n \end{bmatrix}$$

and where, $M_{ij} = P_i/P_j$, $M_{ji} = 1/M_{ij}$, and $M_{ii} = 1$

Saaty has shown that M is a matrix of rank 1, and it has n-1 eigenvalues which are zero and one that is n, if the P_i's are known exactly. Thus all the ratios, or elements of the matrix, are consistently defined. If the P_i's are not known, as is the case we are concerned with, the maximum eigenvalue will differ from n, and the remaining n-1 eigenvalues will differ from zero. Saaty has shown that the degree that the maximum eigenvalue deviates from n is proportional to the degree that the matrix of comparisons in inconsistent. By inconsistent, it is meant that the pairings do not obey a transitive rule.

The eigenvector associated with the maximum eigenvalue is the vector of desired probabilities. If the matrix is relatively consistent, then the probabilities should reflect the underlying belief system of the decision maker. If the matrix is inconsistent, then, the estimator may wish to alter the original paired comparisons. SPA provides a procedure for entering the paired comparisons, determining the resultant probabilities and checking for consistency. The consistency measure used in SPA differs from Saaty's; therefore it deserves mention.

Following the method of Saaty, the deviation between the maximum eigenvalue and n, the order of the matrix, can be used to determine if the matrix is consistent. An experiment, involving Monte Carlo testing, was conducted to determine how useful statistics based on the maximum eigenvalue would be in determining inconsistency. For various values of n, large numbers of random matrices were generated. For each, the statistic

$\mu = (\lambda_{max} - n) / n$ was computed. The resulting empirical distributions, for given n, were tabulated and the 1,5 and 10 percent fractiles were determined. It was found that small values for μ, occurring approximately 1 percent of the time, were in fact indicative of matrices that were on the aggregate consistent. Note the term "aggregate". Since P_i's are not known exactly, the resulting eigenvector can be considered to be the "true" or ideal eigenvector plus some perturbation. The perturbation results in localized errors in the elements of the eigenvector. It is difficult to assess the amount of error incurred on each eigenvector element. The statistic, μ, is an aggregate measure of inconsistency because it is sensitive to the summed error over the entire eigenvector. Therefore, the experimentation suggested that large local errors or inconsistencies could exist and still result in acceptable aggregate errors.

In SPA, a different approach to consistency was taken. An absolute error measure was calculated on a local basis; i.e. it sensed inconsistencies in each paired comparison. Using this, absolute error measurements on a local basis and variance measures at the aggregate level allowed SPA to present a total picture of the inconsistencies. Once the paired compariso . are judged consistent, the resulting eigenvector is scaled to provide the desired probabilities.

11.5. ANALYSIS OF HISTORICAL PROJECTION DATA

In this section, procedures are provided for analyzing historical projection data. One approach involves computing uncertainty measure for OB

data. This uncertainty measure is used to determine the overall reliability of past OB data for a particular system. The level of uncertainty in past OB data can be used as a basis for judging the accuracy of current OB data. A quantitative calibration procedure provides another approach to analyzing past projection data. This statistically-based calibration procedure complements the qualitatively-based approach available through the Institutional Memory (IM).

11.5.1. Uncertainty in Order-of-Battle (OB) Data

In this subsection, a technique for measuring uncertainty in OB data is presented. In this measure, uncertainty is attributed to variability (fluctuation) in the history of OB revisions. To begin, a description of the OB update process is presented. Next, the uncertainty measure is defined. This measure provides estimates of component uncertainties (yearly snapshots) and a basis for computing aggregate uncertainty (the overall assessment through time.) Finally, an example is provided.

The OB update procedure is defined. OB data are typically used as a basis for estimation. The OB may represent the best knowledge currently available on the trends in weapon system deployment over previous years. The OB is an estimate which may be revised as more information becomes available. Therefore, each year, for a period of four years, the OB for a given projection is updated. For the year 1972, one would expect four OB values to be available: the original given in 1973, and the updated values received in 1974 through 1976. The latest update, in 1976, is generally assumed to be the

most accurate estimate of weapon system levels in 1972. Variations or fluctuations in the OB over the four year period are a source of uncertainty whenever the OB is used as a basis for projections. Therefore, the reliability of the current OB, given the OB history, is an important consideration for analysts.

Graphically, updating OB values and computing component uncertainty estimates can be illustrated as shown in Figure 11-9. Note that to the right of each OB graph is an uncertainty measure, U, and each value of U is a component of the aggregate uncertainty, AU. The uncertainty measure, U, for each OB year can be defined as:

$$U_j = \left[\sum_{I=1}^{N-1} \frac{OB(N) - OB(I)}{N - 1} \right] \Big/ OB_{max}$$

where

U_j = the uncertainty associated with j^{th} OB year

N = the number of times the OB year is updated, generally 4.

OB(I) = the weapon system level for the I^{th} OB year.

OB_{max} = the maximum OB value over the N year period.

The aggregate uncertainty, AU, can be defined as:

$$AU = \sum_{j=1}^{NY} W_j U_j$$

where

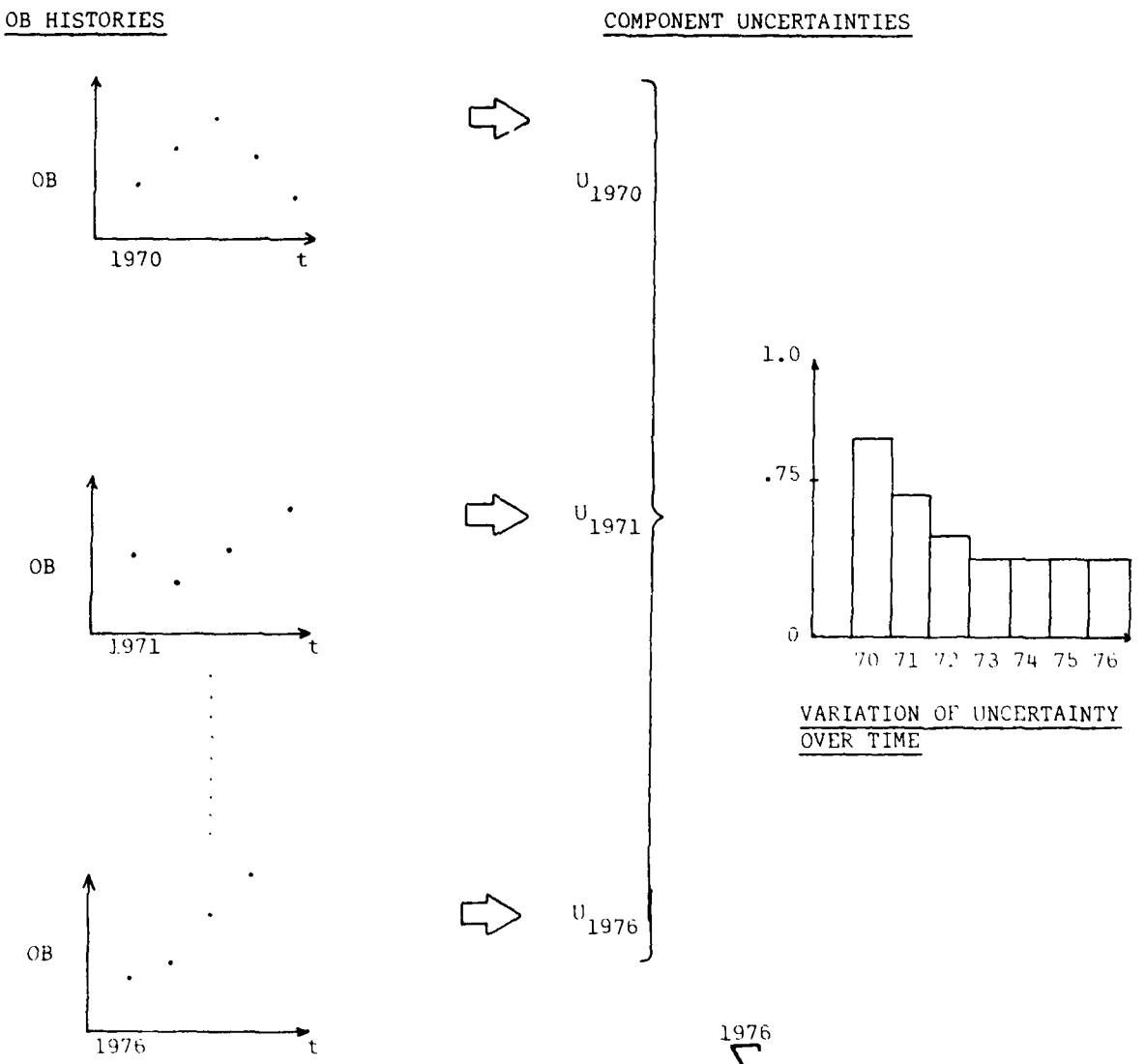


Figure II-9 Assessing Uncertainty in OB Data

W_j = the weight given to the uncertainty of the j^{th} OB year.

U_j = the uncertainty assessed for the j^{th} OB year.

NY = the number of OB years of history.

The following example illustrates the computation of the OB uncertainty measures U and AU. Assume that the OB history provided in Table 11-1 is stored in the historical data base. From examination of the component uncertainties in Table 11-2, it is evident that the OB values have not exhibited severe variation and that the uncertainty decreased through time. For example, in 1969 the uncertainty was assessed as .29 and the decreasing trend led to a value of .11 in 1978. The aggregate uncertainty, AU, can be computed using a weighting function, W. The estimator can determine whether uncertainties from the more recent past are more important than those further back in time. If they are, a weighting procedure other than uniform can be employed. As shown in Table 11-2, two weighting functions were used. Using the uniform weighting function, the aggregate uncertainty is computed to be .22, while the linear increasing weighting results in an aggregate of .17. In both cases the aggregate uncertainty is probably reasonable.

The linear weighting resulted in an aggregate uncertainty lower than that with the uniform weighting. This result reflects the fact that, in this hypothetical example, the OB variability was greater in the early 1970's than it was more recently. The uniform weighting emphasized early variability more than the linear weighting did, and the aggregate uncertainty was correspondingly higher.

PROJECTION YEAR	OB YEARS AND HISTORIES									
	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978
1970	10									
1971	10	10								
1972	12	12	14							
1973	14	14	16	16						
1974		18	18	18	20					
1975			22	22	24	26				
1976				28	28	30	30			
1977					32	34	36	38		
1978						36	38	40	40	
1979							38	44	46	50

Table 11-1 Hypothetical OB History For Soviet Non-Nuclear Icebreakers

OB YEAR	COMPONENT UNCERTAINTY	UNIFORM WT.	LINEAR
1969	.29	1/9	1/45
1970	.33	1/9	2/45
1971	.27	1/9	3/45
1972	.33	1/9	4/45
1973	.25	1/9	5/45
1974	.17	1/9	6/45
1975	.09	1/9	7/45
1976	.11	1/9	8/45
1977	.11	1/9	9/45
1978	*	*	*

* = Insufficient data

Table 11-2 Computation of Component Uncertainties

Uncertainty measurements computed on OB data may be quite useful in quantifying the inherent variability of that data. These uncertainty measurements may be used to communicate, within DE, the reliability of base rate data. If past projections are in serious error, for example, it may be of interest to conduct an analysis of the OB data to determine if OB variability or error was a significant factor in forecast degradation.

11.5.2. Calibration Procedures

In general, there is a basic procedure for processing probabilistic information; it involves refining, updating, calibrating and aggregating assessments. Therefore, judgmental probabilities often can be modified by analysts, updated with new information, and rectified by calibration. In this section, the role that calibration plays in probability-based assessments is discussed. It is shown that the prediction and calibration processes can be statistically modelled and that this formal approach results in precise assessments. More explicitly, the following will be discussed:

1. A definition of calibration that is useful to DE
2. The need and basis for a statistical model of the projection/calibration process
3. Some proposed calibration measures
4. The calibration procedure.

11.5.3. Definition of Calibration

Calibration assessments are used to determine the degree that stated probabilities agree with baseline or truth frequencies. Thus, the weather forecaster that assesses the probability of rain for April, in Washington, DC, to be 20 percent is well-calibrated if it rains approximately 20 percent of those days. If the forecaster is miscalibrated, calibration procedures can be used as a basis for removing systematic biases from uncertainty statements in forecasts. Then, if the forecaster was consistently high in assessing the likelihood of rain, this overstatement might be used to calibrate future forecasts.

Calibration procedures can be applied to DE projection data. Many projections consist of High, Low, and Best estimates of force levels. The High and Low define a confidence interval for the Best. A probability, p, or confidence coefficient is attached to the interval and states the likelihood that the true force level (OB) is contained within it. If a relatively large number of confidence intervals (High-Low's) and OB (actual) values were collected, the expectation would be that 100p percent of the OB's would be contained within the corresponding interval estimates. Therefore, if p equals .75, 100(.75) or 75 percent of the OB's would have been captured by respective High-Low intervals. Generally, 75 percent of the OB's will not be captured; the actual percentage (hit rate multiplied by 100) may be substantially higher or lower. A calibration assessment provides a procedure for testing whether deviations between 75 percent and the hit rate are significant. Given this definition of calibration, it is pertinent to discuss the need and basis for a statistical model of the calibration process.

11.5.4. A Statistical Model of Calibration

At this point in the discussion, a statistical approach to calibration assessments is described. The procedure involves stochastic modeling of the prediction/calibration cycle. This modeling yields probabilistic statements concerning the significance of deviations between the hit rate and p (.75). It is shown that the process in which the OB is compared to the High-Low range is analogous to flipping a biased coin (where the probability of a head is .75) and recording the number of heads. The analogy is reasonable because of the following:

- The process in which the OB is compared to the High-Low range has an uncertain outcome. The OB may fall within the interval, a hit, or it may not, a miss. The likelihood or certainty of a hit is .75. Flipping a coin also results in an uncertain outcome. The coin flip may result in a head (hit) or a tail (miss). The coin may be biased such that the probability of a head is .75.

- If each of the two processes is performed a number of times, the output from each will be the proportion of hits or hit rate. The coin tossing analogy and accompanying statistical model are often used to characterize processes where two outcomes are possible (a hit or miss).

Assume that 20 projections were assessed and the biased coin was flipped 20 times. In both cases the proportion of hits (hit rate) is .60 compared to

the desirable proportion .75. Are the projections well-calibrated? Is the coin consistent with the stated bias? The questions are equivalent and can be answered when considering the common probabilistic nature of each process.

Calibration procedures can result in summary statistics such as the hit rate or proportion of hits. Methods for treating proportions are now discussed. The treatment involves an important statistical distribution (the binomial) which, under certain assumptions, describes the way in which proportions vary. Knowledge of how proportions vary provides a basis for deciding whether projections might be miscalibrated. More explicitly, the binomial model provides the following information:

1. That calibration procedures based only on the hit rate are very likely to be misleading or at least incomplete. This is especially true when the number of projections is small.

2. That the hit rate and its confidence interval provide a more realistic basis for calibration. Miscalibration can be determined with known decision risk.

3. That the risk associated with deciding projections are calibrated is complex. This decision has uncertainty inversely proportional to the number of projections (amount of data).

The statistical model will be discussed using two approaches. The first will examine the binomial from an intuitive or informal perspective. The

discussion primarily employs logical examples instead of mathematical concepts in describing the model. The second approach is slightly more formal and uses explicit mathematical terminology in describing the binomial model. The approaches are identical in the sense that both describe the application of the binomial model to the calibration process. The major difference in the approaches is stylistic.

Intuitive Discussion of Calibration

A model of the calibration process is presented on intuitive grounds. It is shown that although the hit rate is a useful calibration measure, it is not sensitive to the inherent variability in analyst performance (as a probability assessor). Confidence intervals are useful in accounting for performance variability; therefore, calibration assessments are based on both the hit rate and its confidence interval. The hit rate and the confidence interval are used to determine whether projection data are calibrated or miscalibrated and to identify the risk or uncertainty associated with that decision.

The variability in the calibration process is discussed. Assume that a question and answer (Q/A) system is available to DE as a training tool. Estimators are required to answer a series of almanac-type questions. For each question, two answers are provided, one of which is correct. The analyst (estimator) is required to choose the correct answer as well as to assess the confidence he has in his choice. This confidence is expressed as a probability (that his answer is correct). For example, if an analyst was asked "Are potatoes indigenous to Ireland or Peru?" he might pick Peru (the correct

answer) and state a confidence of .7 in his answer. Over a large number of questions, responses are categorized according to the assessed probabilities such that all questions which have probability p attached to them are grouped together. Ideally, then, the hit rate for each group should be close to its assessed probability. Thus in the above example, all questions that had attached probabilities of .7 are grouped together, and, ideally, 70 percent of those questions are answered correctly.

The Q/A session is examined further. In particular, focus on the category p equal to .5. Assume that the estimator has answered 10 questions. If the estimator is well-calibrated with p = .5, how many questions would one expect to be answered correctly? This question is the same as asking: if a fair coin was tossed 10 times, how many heads would one expect to occur? Since the probability of a head is .5, the number of flips resulting in a head is expected to be 5 (i.e., the proportion of heads is expected to be .5). However, from experience, it would not be too shocking if 3, 4, 6, or even 7 heads resulted, even though the coin is fair. Similarly, for the Q/A system, given that p = .5, it is quite possible that the well-calibrated estimator would get 3, 4, 6, or 7 questions correct. Large deviations from the expected result are quite possible, especially when the number of questions is small. In light of the coin-tossing example alone, it is easy to be suspicious of the accuracy of calibration analyses based on hit rates computed over limited data.

Just how much the hit rate can vary for a well-calibrated estimator is now discussed; in doing so it is useful to define a confidence interval.

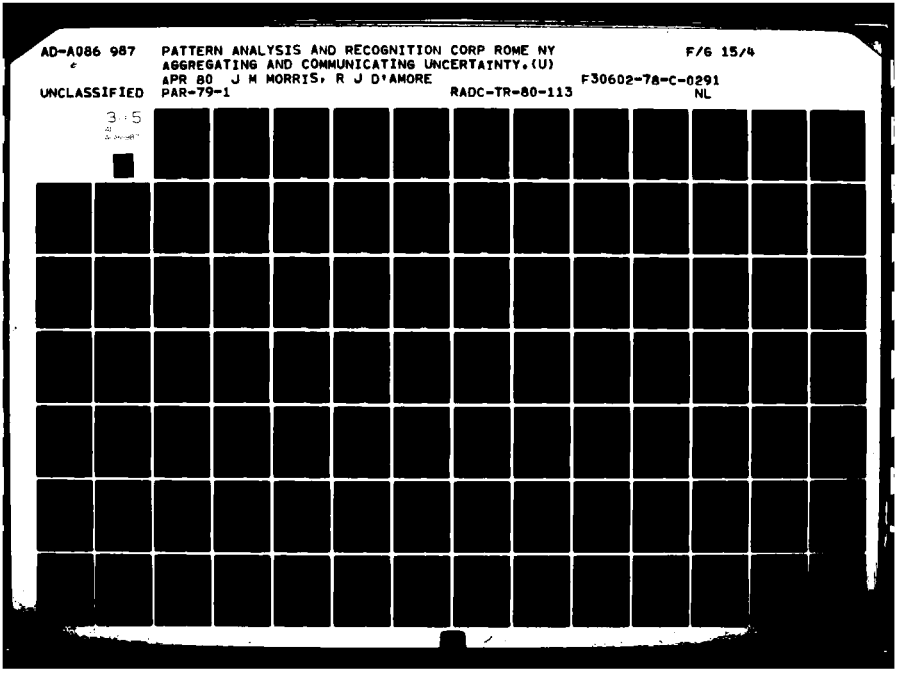
Confidence intervals generally define a range of plausible values; the range is usually centered about the expected value or mean. Confidence intervals based on two different values will be used in the following discussion; they are:

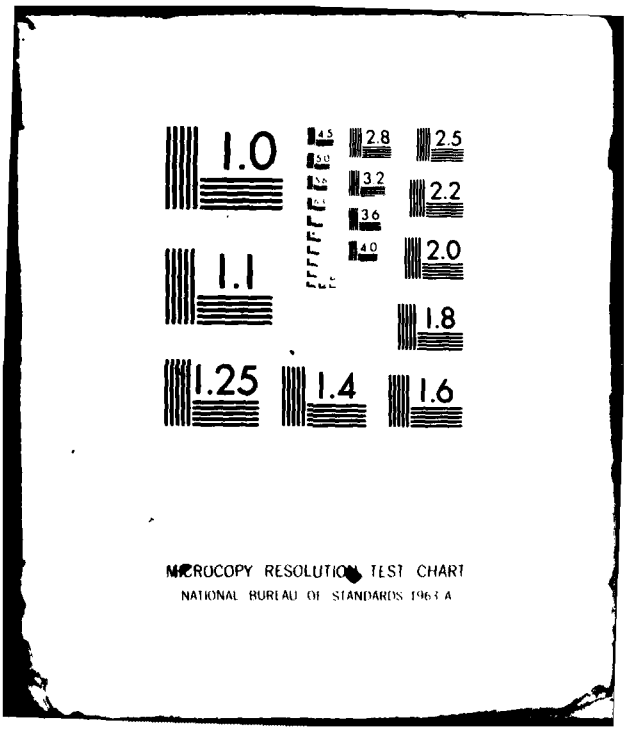
C1: A range of values that with some probability captures the true but unknown parameter value of interest.

C2: A range of values that captures some percentage of the possible values given that true parameter value is known.

For example, confidence interval C1 would result in the following type of statement: there is a 95 percent chance that the true unknown value of the range of the SX-19 missile lies between 6000 and 6500 miles. The basis for this statement might have been a number of observations made of actual SX-19 launchings. Suppose that it is known with confidence that the average range of the SX-19 is 6200 miles. Then a C2-type confidence interval could provide an interval estimate of the likely distance any SX-19 will travel. As an example, a launched SX-19 will, with 95 percent confidence, travel between 6000 and 6500 miles, with an expected range of 6200 miles. Both C1 and C2 confidence intervals will be used. C2 will be used to illustrate the amount of performance variability expected from the well-calibrated estimator; it will also be used to discuss the likelihood that miscalibrated estimators can perform like calibrated estimators, especially when data are limited. Miscalibrated and calibrated estimators may be difficult to identify over limited data just as it may be hard to separate the biased coin from the

AD-A086 987 PATTERN ANALYSIS AND RECOGNITION CORP ROME NY F/G 15/4
AGGREGATING AND COMMUNICATING UNCERTAINTY.(U)
APR 80 J M MORRIS, R J D'AMORE F30602-78-C-0291
UNCLASSIFIED PAR-79-1 RADC-TR-80-113 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

fair coin. C1 confidence intervals are more useful in computing performance estimates; they are used in the proposed calibration procedure discussed later in this section. It should be noted, however, that the two perform identically, in that the same decisions can be made with the same accuracy. The difference lies in the way calibration assessments would be conveyed.

The confidence interval, C2, provides a convenient tool for conveying the performance variability of the well-calibrated estimator. For the Q/A system if an estimator is calibrated for some p, say p = .7, then the 95 percent confidence interval is a range of values encompassing the "true" value, .7, such that the probability is .95 that the session hit rate will fall within the specified interval. Obviously, then, there is only a 5 percent chance that a hit rate from the calibrated estimator will fall outside this interval. There are two critical decisions to be made; they are that the estimator is:

1. Miscalibrated - If the hit rate falls outside the confidence interval, there is a 5 percent risk associated with deciding that the estimator is miscalibrated.

2. Calibrated - If the hit rate falls within the confidence interval, it is reasonable to state that the estimator is consistent with calibrated results but it cannot be necessarily stated that he is calibrated.

To illustrate the reasoning behind decision 1, miscalibrated, Figure 11-10 provides confidence intervals for p = .5 through 1.0, in increments of .1. The number of samples, N, is 8. The diagonal line is the calibration line and represents the desired result. For each assessed probability there is specified a range of hit rates or confidence intervals. If the hit rate falls inside the designated range, the assumption that the estimator is calibrated with a particular value, p, cannot be rejected. Note that for N = 8 and each p there is a wide range of acceptable hit rates. For example, for p = .7, the hit rate can range from .5 to 1.00; any value within this range is consistent with the assumption that the estimator is well-calibrated. Similarly, consider when p = .5. For 8 questions asked, the results can be considered consistent with the assumption that p = .5 if the hit rate is between .25 and .875, or if between 2 and 7 correct responses are recorded. The first conclusion to be drawn is that for small N, it is very difficult to reject, with small error, the possibility that the estimator is calibrated.

For decision 2, calibrated, the concern is whether or not the estimator is actually calibrated. If the hit rate falls within the prescribed confidence interval, the decision is not to reject the assumption that the estimator is consistent or calibrated. However, it may be that the estimator is not calibrated to the assessed probability but is in fact consistent with another. Reconsider the coin tossing example previously mentioned. It was claimed that even if the coin is fair, it would not be unusual for the proportion of heads to differ from the expected proportion, .5. Therefore, if the coin was flipped 8 times and 3 heads resulted, the original assumption that the coin was fair would not be rejected. However, it may be that the

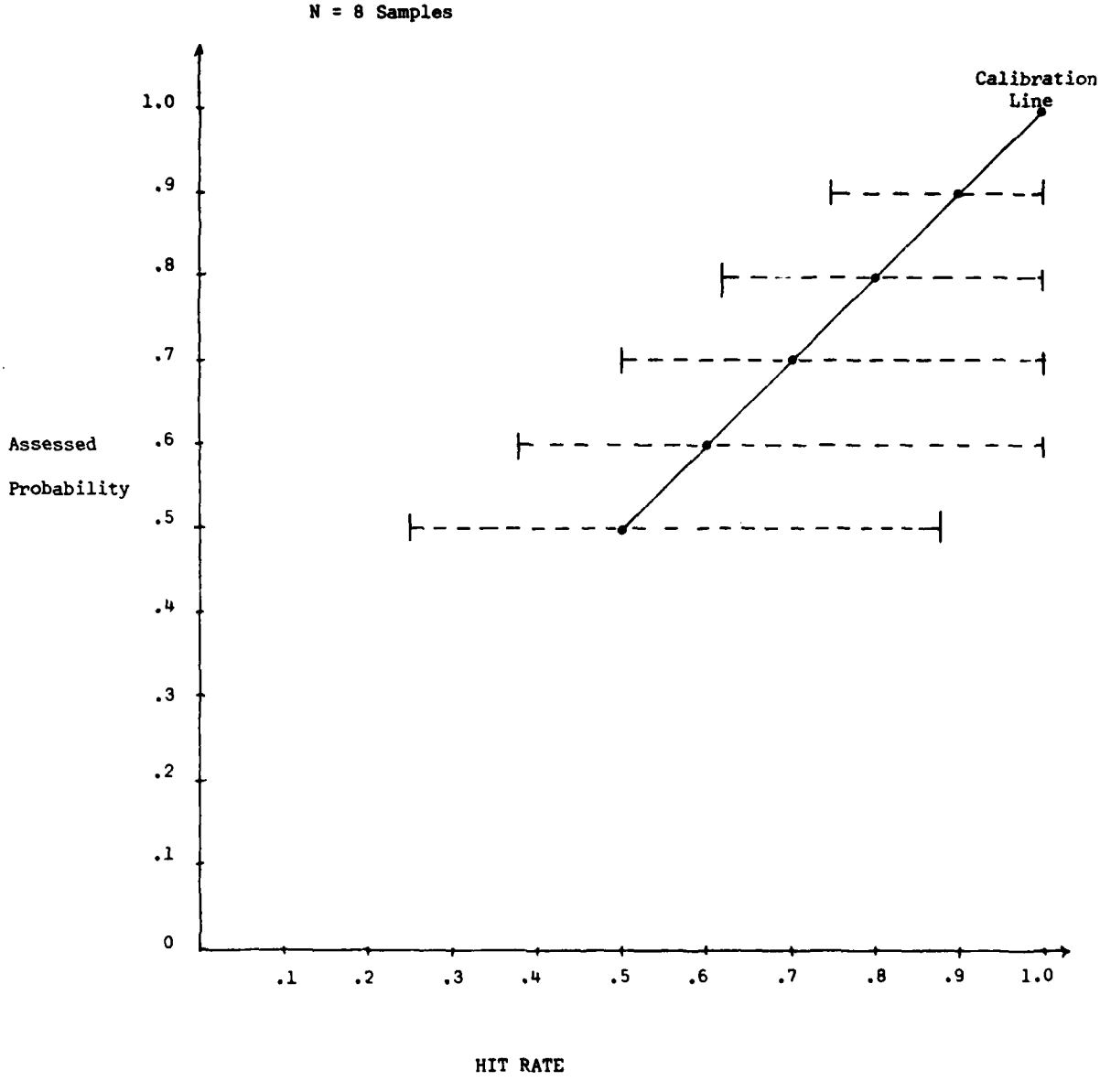


Figure 11-10 Ranges of Hit Rates Consistent with the Assessed Probability (95% Confidence Internal)

coin is biased, such that the probability of a head is .4. Flipping the biased coin 8 times may frequently result in a proportion of heads that would be identical to the proportion resulting from flipping the fair coin. Similarly, the estimator may be miscalibrated to some degree and still produce a hit rate that is acceptable. If the probability of incorrectly judging an estimator miscalibrated is fixed (say 5 percent), the ability of the test to distinguish those that are miscalibrated from those that are well-calibrated is dependent on N, the number of responses. The smaller N is the greater the possibility that the estimator will be considered well-calibrated when he is not. Therefore, a prime concern is that N be large.

In Table 11-3, the effect that increasing N has on the ability to distinguish those estimators that are miscalibrated from those that are calibrated as illustrated. For example, assume that the value of p is .7; that is, the estimator has assigned a probability of .7 to a number of responses. If the estimator is consistent with p, there is approximately a 90 percent chance that he will be judged so. However, assume that he is actually internally consistent with another probability, say .6. Then, for N = 8, there is an 81 percent chance that he would be consistent with p = .7. In other words, he will "look" consistent with p = .7. As N increases, however, the chances that he would be confused with being consistent with .7 decrease to .71 for N = 16 and to .6 for N = 32. Someone consistent with p = .2, would have little chance of being considered calibrated with p = .7, even when N = 8 (the probability is only .06). The general observation from Table 11-3 is that as N increases the ability to discriminate the miscalibrated from the calibrated increases. (In practice, the confidence in this discrimination

ASSESSED PROBABILITY	N=8	N=16	N=32
0.0	0.00	0.00	0.00
.1	.01	0.00	0.00
.2	.06	0.00	0.00
.3	.19	.02	0.00
.4	.41	.14	.02
.5	.63	.40	.19
.6	.81	.71	.60
Stated Probability .7	.89	.90	.88
.8	.82	.85	.64
.9	.57	.49	.09
1.0	0.00	0.00	0.00

Table 11-3 Probability of Being Labelled Calibrated with P=.7
Given true calibration to the "Assessed Probability"

would never be 100 percent, but "small" levels of miscalibration would be tolerable. For example, if it were desirable to calibrated to p = .7, miscalibration to neighboring values of p, such as .6 and .8 might very well be acceptable. From Table 11-3, it might be concluded that for N = 32 the two decisions, calibration and miscalibration, can be identified with confidence if calibration includes a range of p, .6 to .8, centered about the desirable value, .7.)

The above intuitive discussion of calibration is possible because of the statistical model it is based on (the binomial). Characterization of the calibration process makes clear the inherent variability and allows appropriate performance statistics to be developed.

A Mathematical Description of the Calibration Model

A mathematically oriented description of the statistically-based calibration model is given. The goal of this discussion is not mathematical rigor, but instead to provide a more explicit description of the statistical model. The treatment begins with a description of the binomial or calibration model. Given this model, the inherent variability associated with calibration data can be examined. Of special interest is the topic of Type I and Type II errors, which define the probability of incorrectly deciding estimators are miscalibrated and calibrated, respectively. Finally, the method of computing confidence intervals for the hit rate is presented. This estimation procedure is the basis for the calibration assessments described in Section 11.5.6.

The following data are considered:

1. The High-Low range (the DE projection confidence interval)

2. The associated Order-of-Battle (OB) data (the best estimate of the actual force level for past years).

Assume that there are N data items or trials, where each trial consists of the comparison of the OB to the High-Low range. There are two possible outcomes from each trial:

1. A hit - the OB value falls within the High-Low range
2. A miss - the OB falls outside the High-Low range.

Further, the assumption is made that the N trials are independent of each other; that is, that the probability of a hit is the same for each trial. This assumption is consistent with DE's assignment of a 75 percent confidence level to each High-Low estimate.

Consider one of the N trials. The outcome of the trial is coded as follows:

$x = 1$ when a hit occurs (the probability is $p = .75$)

$= 0$ when a miss occurs (the probability is $1 - p = .25$).

Thus the N trials form a Bernouli sequence where the outcome of any trial x, is said to be distributed like b(1,p). The distribution b(1,p) is binomial with parameters N = 1, and p. If a new random variable, X, is defined, such that,

$$X = \sum_{i=1}^{N} x_i$$

its distribution is given by the binomial b(N,p) where,

$$\Pr(X = k) = \binom{N}{k} p^k (1-p)^{N-k} \tag{2}$$

and

$$k = 0, 1, 2, \ldots, N; \quad 0 \leq p \leq 1.$$

The expected value and variance of the distribution are given as:

$$EX = Np \tag{3}$$

$$VAR(X) = Npq \tag{4}$$

where q = 1 - p. Thus the binomial distribution can be considered as the sum of N independent, identically distributed b(1,p) random variables. It should be noted that from an intuitive perspective, Eq. 2 relates to the following operation:

1. calculate the probability of k hits and N-k misses occurring in some specific order. This is $p^k q^{(N-k)}$.

2. compute the number of different orders of k hits and N-k misses; that is, $\binom{N}{k}$, the binomial coefficient.

3. multiply the results of (1) and (2) together to get Eq. 2, the desired probability.

The binomial model can be used to assess the inherent variability of the calibration process. To make this assessment the following hypotheses are considered:

H_o: $p = p_o$ (for projection data, $p_o = .75$)

H_1: $p \neq p_o$

A reasonable procedure for testing the above hypothesis would be to examine the number of hits (k), in N trials. H_o is rejected if $k \in C$, where C is the critical region for the test. H_o is not rejected when $K \in C^c$, where C^c is the complement of C. The following errors are important with respect with these decisions:

1. Type I error - the probability of incorrectly rejecting $H_o = \alpha$.
2. Type II error - the probability of incorrectly accepting $H_o = \beta$.

The probability of a Type I error is

$$\text{Pr (Type I error)} = \alpha = \text{Pr}_{p=p_o}(|K-Np_o| \varepsilon C) \tag{5}$$

The probability of a Type II error is

$$\text{Pr (Type II error)} = \beta = \text{Pr}_{p=p_i \neq p_o}(|K-Np_i| \varepsilon C^c) \tag{6}$$

where i = 1, 2, ... M, the number of alternatives.

The Type I and Type II errors are computed from Eqs. 5 and 6, respectively. Examination of Type I and II errors (for example, see Figure 11-10 and Table 11-3) illustrates the true variability of the process. In particular, the role that N, the number of trials, has on the variability is most obvious. Thus, the main conclusion is that N must be reasonably large before reasonable certainty can be attached to calibration decisions.

The proposed calibration procedure (Section 11.6.4) employs confidence interval estimates of the hit rate. The confidence intervals provide, with confidence r-α, a range of values which the estimator is consistent with. The development of the confidence interval follows from Rohatgi and will not be given in this report. The interval estimate for the hit rate is given as:

$$H_l = [K + Z_\alpha^2/2 - Z_\alpha \sqrt{\frac{K(N-K)}{N} + \frac{Z_\alpha^2}{4}}]/(N + Z_\alpha^2) \tag{7}$$

and

$$H_u = [K + Z_\alpha^2/2 + Z_\alpha \sqrt{\frac{K(N-K)}{N} + \frac{Z_\alpha^2}{4}}]/(N + Z_\alpha^2) \tag{8}$$

where Z_α = the $(1 - \alpha/2)$ point from N(0,1); e.g. $Z_{.10} = 1.64$. H_l is the lower limit of the range and H_u is the upper limit. The width of the confidence interval is given by $H_l - H_u$. This large sample approximation for the true confidence interval compares well with that given by the exact distribution when $N \geq 10$.

11.5.5. Calibration Measures

In this section, two calibration statistics, the hit rate and CAL measure, are presented. These measures are, in part, the basis for the calibration procedures described in Section 11.5.6. The hit rate is a measure of proportion while the CAL measure reflects the percent deviation between the computed hit rate and the assessed probability (i.e., the ideal hit rate).

Hit Rate

The hit rate will be defined as it applies to projection data. Assume that a number of historical projections and corresponding OB values are collected. Each projection consists of a set of High and Low estimates. Over some portion of each projection, OB values will be available. It is this data which can be used in the calibration assessments. Then the hit rate, HR, is defined as:

$$HR = \frac{\sum_{j=1}^{M} \sum_{i=1}^{N_j} H_{ij}}{\sum_{j=1}^{M} N_j} \qquad (9)$$

where, M = the number of projections over which the analysis is being performed.

N_j = the number of estimates (High and Low pairs) in the j^{th} projection, $1 \leq j \leq M$.

H_{ij} = 1, if the true value (OB) for the i^{th} estimate in the j^{th} projection falls within the High and Low, $1 \leq i \leq N_j$.

= 0, otherwise.

For the simple case in which Q/A data is being calibrated, the hit rate is defined:

HR = Number of questions correctly answered/Number of questions asked.

Thus, for question and answer data as well as projection data, the hit rate is a proportion.

CAL Measure

The CAL measure transforms the hit rate into a percent deviation measurement. The CAL measure determines the percentage difference between the assessed confidence or probability (the ideal value) and the computed hit rate. The CAL measure is defined as:

$$CAL = \frac{HR - P}{P} 100\% \qquad (10)$$

where HR = the hit rate

P = the assessed probability

Note that simple algebra allows the hit rate to be obtained directly from (10) when CAL is known; that is,

$$HR = P(1 + CAL/100) \qquad (11)$$

Both the CAL measure and the hit rate statistic are discussed more fully in Section 11.6.4. Some special comments about the CAL measure are, however, in order. For example, in Figure 11-11, hypothetical calibration results for four Soviet ICBM's are displayed. (Note that confidence intervals have not been employed since the purpose of this example is to focus on characteristics of the CAL measure.) Positive deviations from the correct result (p = .75) imply that the width of the projection confidence interval, the High-Low range, was on the average wider than it should have been, and that therefore the estimator was somewhat underconfident. For example, if the deviation for a particular system was 15 percent, then 15 percent more true values (OB's) fell within the High-Low intervals than should have. Note, however, that a deviation of 15 percent does not imply that the average distance between the High and Low should have been decreased by 15 percent. It may very well be that increasing the High-Low ranges by 15 percent would have little or no effect on the computed CAL or hit rate. Another point is that even if it was

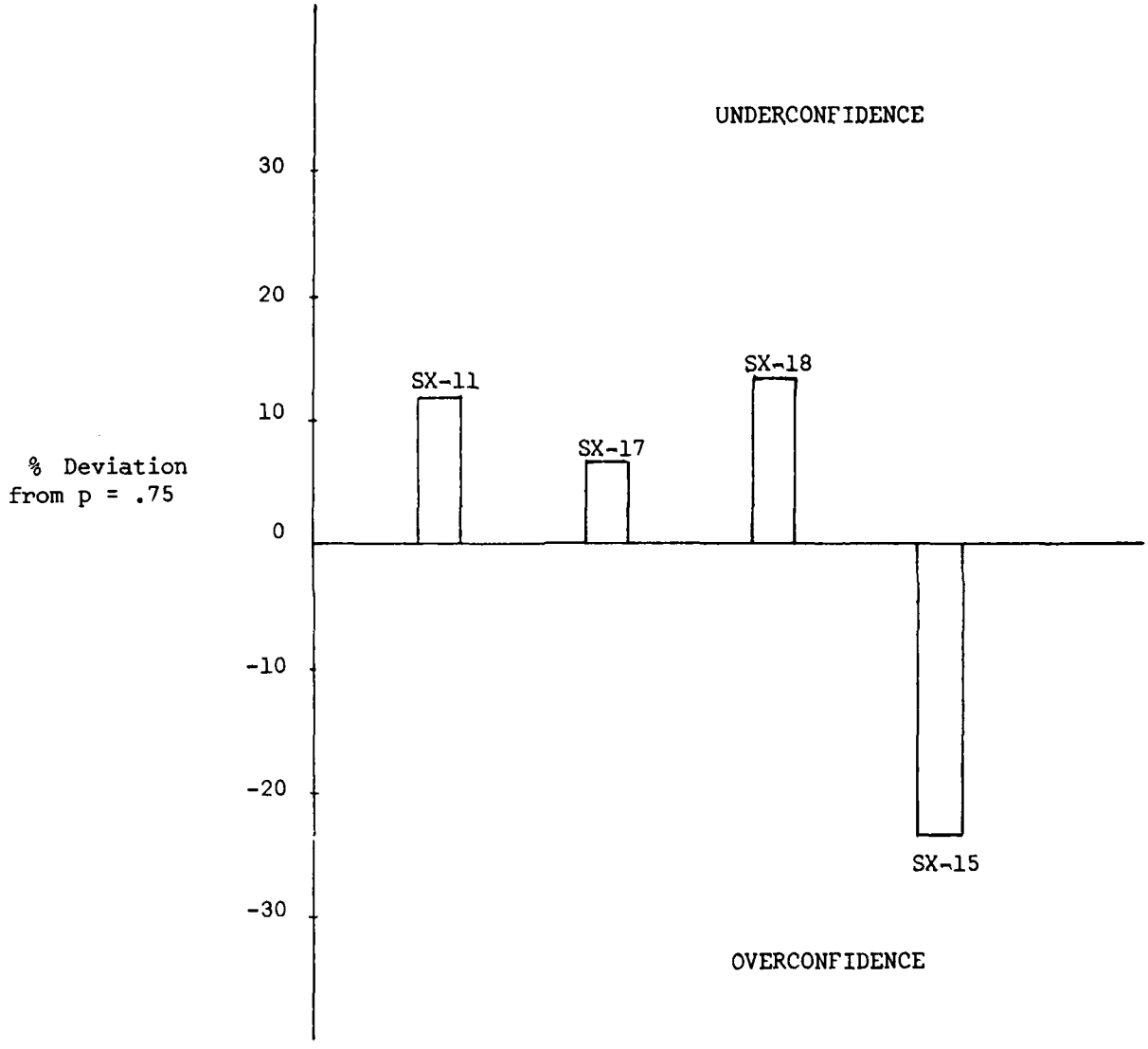


Figure 11-11 Calibration Analysis of Soviet ICBM's

possible to debias a group of projections by rescaling the High-Low interval, it may be that on the basis of a single projection, the approach is invalid. One reason is that the calibration analysis produces results from a composite of estimates; hence, what is effectively being investigated is an overall bias in High-Low assessments. Further, arbitrary expansion or contraction of the High-Low ranges may, in some cases, invalidate the assumptions and expectations on which the forecast is based. Therefore, the calibration analysis should be viewed as a global assessment procedure pointing towards trends in over- or underestimation.

11.5.6. The Calibration Procedure

The application of the binomial model to calibration assessments is illustrated. The calibration procedure employs point estimates, either the hit rate or CAL measure, and their respective confidence intervals. There are some key ideas associated with this procedure:

1. the estimator has assigned an assessed probability, p = .75, to each High-Low range.

2. the estimator is probably consistent with some probability, P_{true}, where P_{true} is not necessarily equal to p, .75.

3. the hit rate is the best estimate of P_{true}.

4. the hit rate and number of estimates can be used to compute a confidence interval for P_{true}; the probability attached to this interval is specified by the estimator. The confidence interval provides a range of plausible values for P_{true}.

5. the calibration decision rule is based on the relationship between the assessed probability, p, and the confidence interval. If the confidence interval encloses p, the decision is that the estimator is consistent with the assessed probability, i.e., possibly calibrated. If the confidence interval does not enclose the assessed probability, the decision is that the estimator is not calibrated (the confidence in this decision is that probability given in (4), above).

Assume that a number of estimates for the MiG-21, MiG-23 and MiG-25 have been collected, and sufficient OB information is available to conduct a calibration analysis. A summary of the data appears in Table 11-4. This data is input into the statistical model. Note that the following calibration analysis is primarily based on the hit rate statistic.

The following output is indicative of the type of information available in this statistically based calibration procedure. Included are a summary of computed hit rates, CAL measures, confidence intervals for the hit rate, and decisions as illustrated in Table 11-5. As supplement to this table, the confidence interval plot, illustrated in Figure 11-12 is provided.

WEAPON SYSTEM	NUMBER OF ESTIAMTES	NUMBER OF HITS	HIT RATE	ASSESSED PROBABILITY
MiG-21	35	25	.71	.75
MiG-23	40	20	.50	.75
MiG-25	10	9	.90	.75
Aggregate	85	54	.64	.75

Table 11-4: The data input to the calibration model

Table 11-5 Hit Rate-Based Calibration Assessments

CALIBRATION ANALYSIS

WEAPON SYSTEM	NUMBER OF ESTIMATES	NUMBER OF HITS	HIT RATE	RANGE OF HIT RATES LOW	HIGH	ASSESSED PROB.	GROUP HIT	CAL
MiG-21	35	25	.71	.58	.82	.75	YES	- 5.33
MiG-23	40	20	.50	.37	.63	.75	NO	-33.33
MiG-25	10	9	.90	.65	.98	.75	YES	20.00
Aggregate	85	54	.64	.55	.72	.75	NO	-14.67

Interpreting the Results

In Table 11-5, the hit rates and confidence intervals for each aircraft are displayed. In addition, data from each weapon system have been aggregated such that a composite analysis can be performed. Examine the individual aircraft estimates first. For the MiG-21, the confidence interval is given as the range .58 to .82. Therefore, there is a 90 percent chance that the true probability (the probability with which the estimator is actually consistent) is contained within the High-Low range. The best estimate of P_{true} is the hit rate, .71. Note that the confidence interval contains the value .75, the desired probability. This fact is verified in the column titled "GROUP HIT" in which "YES" is printed. Since .75 is included within the confidence interval, the estimator may be considered potentially calibrated. Note, also, that for the MiG-25, .75 is also contained within the corresponding confidence interval. However, since the number of estimates is only 10, as compared with 35 for the MiG-21, the result is less reliable.

If .75 is not included between the Low and High, then the data suggest that the estimator is miscalibrated. For example, the confidence interval for the MiG-23 is .37 to .63 and "NO" is noted in the "GROUP HIT" column. Therefore, the estimates for the MiG-23 can be considered miscalibrated; specifically, the estimator for this aircraft has been somewhat overconfident. Associated with this conclusion is a 10 percent risk of incorrectly labeling the estimator miscalibrated when in fact he is calibrated. In other words there is a 90 percent confidence in that decision.

A general assessment of the aggregate of the three aircraft leads to the conclusion that the estimates are somewhat miscalibrated. This is a fairly significant result as suggested by the number of samples and size of the confidence interval. In fact, since in the accompanying graph the confidence interval is entirely to the left of the vertical line denoting .75, it can be stated with reasonable certainty that a bias towards overconfidence may be present.

The calibration analysis, as specified, provides an estimator with a statistical assessment of the degree of consistency in his estimates. Through use of the confidence interval, the full range of probabilities of hit rates with which the estimator may be consistent is provided. This is one approach to quantifying the uncertainty of past probabilistic statements (i.e., uncertainty in projections). The computed hit rate is the best estimate of P_{true}, the probability with which the estimator is internally consistent. How well P_{true} compares with .75 is proportional to the degree to which the analyst is calibrated. Note that in the last column of Table 11-4, labelled "CAL," the percentage deviation between the computed hit rate and .75 is provided. The sign of CAL is useful in determining whether an overconfidence (-) or underconfidence (+) bias is in effect. The magnitude of CAL provides a percentage measure of the degree of bias. Finally, note that the information provided in Table 11-5 and Figure 11-12 can be transformed into CAL-based statistics as illustrated in Table 11-6 and Figure 11-13.

WEAPON SYSTEM	NUMBER OF ESTIMATES	NUMBER OF HITS	CAL%	RANGE OF CAL LOW	RANGE OF CAL HIGH	ASSESSED CAL	GROUP HIT
MiG-21	35	25	- 5.33	-22.67	9.33	0.00	YES
MiG-23	40	20	-33.33	-50.67	-16.00	0.00	NO
MiG-25	10	9	20.00	-13.33	30.66	0.00	YES
Aggregate	85	54	-14.67	-26.67	- 4.00	0.00	NO

Table 11-6 CAL-Based Calibration Assessments

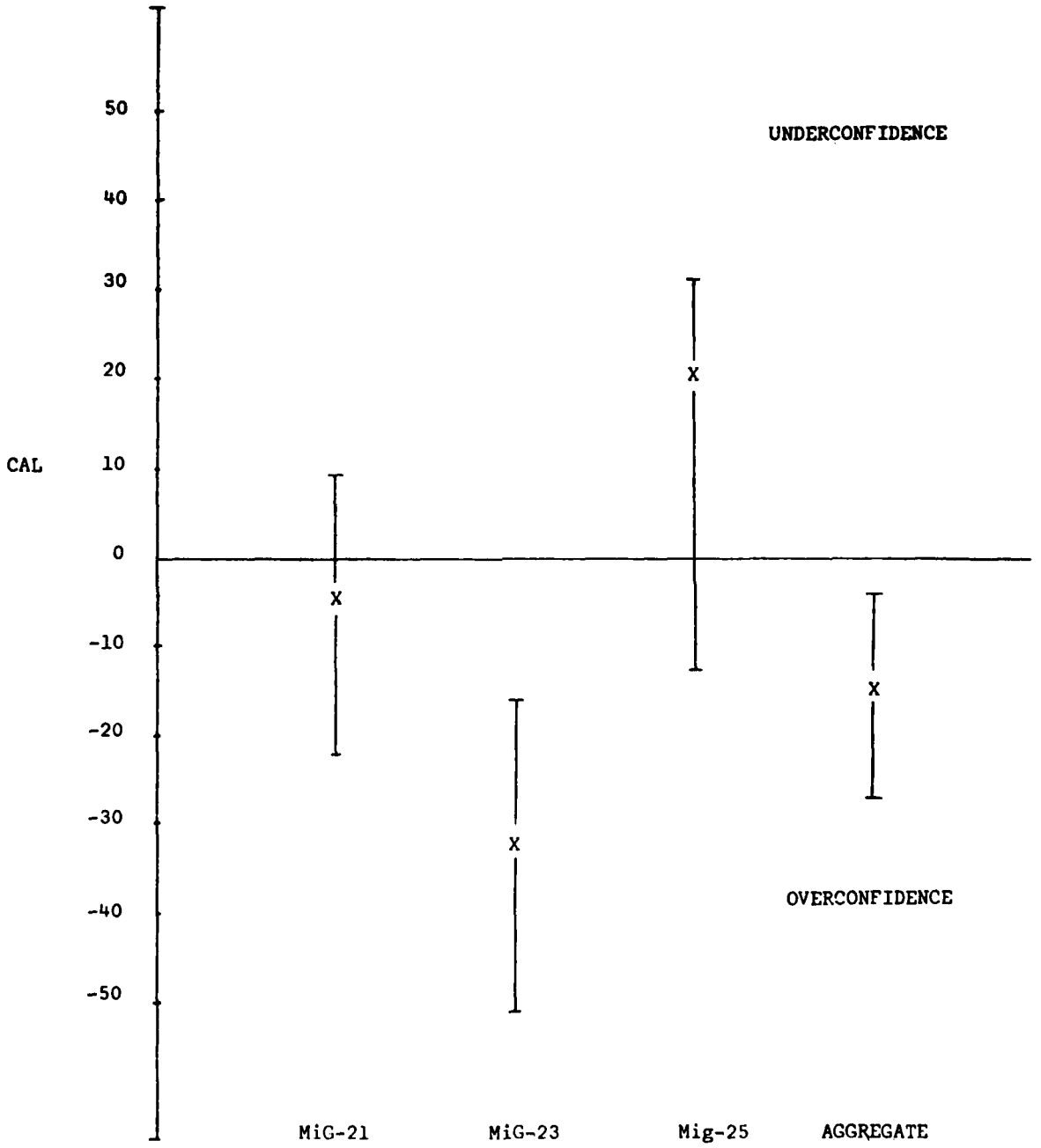


Figure 11-13. Point and Interval Estimates of the CAL Measure

SECTION 12

CONCLUSIONS AND RECOMMENDATIONS

The research reported here has identified two types of models of the estimative process:

- o The holistic, or top-down, approach, which begins with a broad, general picture of the nation under study.
- o The compositional, or bottom-up, approach, which begins with detailed studies of individual elements, such as weapon systems, combining these to arrive at a general estimate of military capabilities.

In practice, both methods are employed in a complementary manner for the production of intellignece estimates. They provide a check-and-balance approach:

- o The feasibility and credibility of a holistic evaluation are dependent upon evidence of intentions and capabilities at the lower levels. For example, the hypothesis that the Soviets expect a Chinese attack upon their Eastern border (high-level generalization) must be supported by evidence of appropriate defensive installations along the border (low-level specification). This in turn will be supported by photographs and other evidence of construction in the area.

- The detailed evidence of Soviet developments makes sense only in the context of a higher-level hypothesis concerning Soviet long-range goals. For example, construction along the Chinese border can be identified as defensive military developments only if we can go on to hypothesize that the Soviets regard the Chinese as potential adversaries. Otherwise, the construction could be intended as non-military production facilities, bases for offensive attacks against the U.S., or facilities for any other purpose.

Thus the top-down and the bottom-up approaches provide an interactive form of validation for one another. The high-level estimates are verified by the low-level observations, and the low-level observations are guided by the high-level hypotheses, which explain them and direct further investigations. These alternative approaches have motivated the description within this report of two alternative methods for the quantification, aggregation, and communication of uncertainty in estimative intelligence:

- From the top-down viewpoint, the subjective probabilities of alternative scenarios are obtained from specialists. Using the Institutional Memory described in Section 8, and statistical methods for insuring the consistency of the probabilities attached to these scenarios, as described in Section 11, the quality of general hypotheses can be evaluated and improved.

- From the bottom-up viewpoint, the probabilities of specific weapon system developments can be combined to obtain the joint probabilities of various mixes of weapon systems. Statistical methods for combining probabilities were described in Section 6, and computer-based decision systems were reviewed in Section 10.

To implement these two approaches to the aggregation and communication of uncertainty, two general types of computer systems have been suggested:

- For the top-down approach, a demonstration system written in FORTRAN for the HIS GCOS operating system has been provided. This provides a systematic method for the evaluation and resolution of inconsistencies in subjective probability assessments, in connection with high-level hypotheses or scenarios. Programs and documentation are being furnished separately from this report.

- For the bottom-up approach, descriptions of several system designs for combining assessments of uncertainty at the low level, to obtain consistent higher-level assessments, have been provided in Appendix C.

On the basis of research conducted for this project, several recommendations for future development are proposed here:

- Studies of the estimative process conducted during this project, and during the preceding TEAMS project, have indicated that estima-

tive methods are not clearly defined. Most estimators learn their tasks on the job. Although more sophisticated methods are recommended -- e.g. in courses conducted by Defense Intelligence Schools -- there is no indication that these methods are actually used. Among the estimators interviewed, for example, there was a strong tendency to rely on informal methods (called "intuition") and to regard assessments of uncertainty as little more than guesswork. Our interviews have suggested that the methods actually in use by the estimators represent the scientific paradigms ordinarily employed in the social sciences, and that they should not be regarded as unscientific or inadequate. It would, however, be desirable to review these methods in detail to provide estimators with better criteria for determining the quality of their work.

- While there have been several studies of the quality of past projections, these have been performed largely on an ad hoc basis, in response to specific requests from intelligence consumers. A continuing process of quality control is recommended, to provide DE with a facility for determining the existence of potential problem areas, and for making corrections where biases and other errors occur. TEAMS is one tool which was intended for this purpose. Implementation of TEAMS requires the development of an adequate data base, consisting of prior projections and the corresponding OB data. Such a data base would provide an objective record of the quality of estimative intelligence over the past several years, and

would permit a rapid response to requests for such information from intelligence consumers.

- A second type of historical data base has been described in this report. This is the Institutional Memory, which would consist of the higher-level assumptions and opinions that have entered into the projections. While it is important to know what the projections have been, it is also important to know why they have been made. It is recommended that the data base for the Institutional Memory described in Section 8 be developed and implemented concurrently with the TEAMS data base.

- Estimators are frequently forced to use manual methods for much of their work. While nearly all the estimators make use of hand calculators in their work, very few computer aids are used extensively -- the major exception being the tools provided by the DIPPOLS data base management system. Insofar as they are potentially helpful to the estimators, computer-based tools for interpolation, extrapolation, combination, and verification of projections should be developed. Tools like those designed for TEAMS for production of charts and graphs should be provided. It should be possible at any time for the intelligence producer to obtain past projections, identify trends, and develop further projections without the need for extensive manual preparation.

- Several of the estimators expressed an interest in more comprehensive computer systems for projecting future world developments, as well as for more limited projections of specific weapon systems. System Dynamics, as developed by Jay W. Forrester, is typical of these systems (Cf. Forrester, Principles of Systems, Cambridge: MIT Press, 1968). Our own research has not been sufficient to determine the potential usefulness of these systems in the specific context of DE's tasks. Further studies, including implementation of a demonstration system, would help to provide an answer to the questions that have been raised concerning this approach (H. Cole et al., Thinking About the Future: A Critique of the Limits to Growth, Chatto & Windus: Sussex University Press, 1973; David Berlinski, On Systems Analysis, Cambridge: MIT Press, 1976; etc.) One of the most striking outcomes of this study has been the discovery that much of the most effective work in estimative intelligence is based upon informal models -- that is, upon the process that estimators repeatedly called "intuition." Rather than attempting to criticize this approach, we have concentrated on the problem of making intuition more effective -- where the word "intuition" is taken to refer to a broad insight into the goals, technological capabilities, and habits of thought of a potential adversary. For this reason, our primary recommendations are for tools that will assist in making this process more effective. Specifically, there should be the means, like the Institutional Memory, for determining the informal reasoning that has underlain the past estimates, and which will assist in developing future estimates. Again, we have

described techniques for assuring the consistency of subjective probability assessments, where these assessments are based largely on informal models. Finally, we have suggested that the need to justify projections, through conferences with other members of the intelligence community, has provided one of the most effective means for identifying and communicating uncertainty.

APPENDIX A

BIBLIOGRAPHY

1. Ajzen, Icek; "Intuitive Theories of Events and the Effects of Base-Rate Information on Prediction," Journal of Personality and Social Psychology, 1977, Vol. 35, No. 5, p. 303-314.

2. Alpert, M. and Raiffa, H.; "A Progress Report on the Training of Probability Assessors," Unpublished Manuscript, Harvard University, 1969.

3. Alter, Steven; Drobnick, Richard and Enzer, Slewyn; "A Modelling Structure for Studying the Future," Center for Futures Research, Graduate School of Business Administration, University of Southern California, Los Angeles, CA 90007.

4. Amihud, Yakov; "The Effect of Uncertainty in Input Quantities on the Optimal Expected Input Combination," Management Science, Vol. 23, No. 9, May 1977, p. 957-962.

5. Anderson, N.H.: "Information Integration Theory: A Brief Survey," in D.H. Krantz, R.C. Atkinson, R.D. Luce, and P. Suppes (Eds), Contemporary Developments in Mathematical Psychology, (Vol. 2), San Francisco: W.H. Freeman and Co., 1974.

6. Ashton, Robert H.; "Cue Utilization and Expert Judgments: A Comparison of Independent Auditors with Other Judges," Journal of Applied Psychology, 1974, Vol. 59, No. 4, p. 437-444.

7. Bar-Hillel, Maya; "The Base Rate Fallacy in Probability Judgments," Hebrew University, Jerusalem, Decision Research Report 77-4, A Branch of Perceptronics, 1201 Oak Street, Eugene, OR 97401, April 1977.

8. Bar-Hillel, Maya; "On the Subjective Probability of Compound Events," Organizational Behavior and Human Performance, 9, 396-406 (1973).

9. Barclay, Scott; Ted H. Hazard, Rex V. Brown, Cameron R. Peterson, and Clint W. Kelly; "A Scoring Rule for Probability Assessment (Handbook for Decision Analysis)," Defense Intelligence School, September 1973.

10. Beach, Barbara Heinrich; "Expert Judgment About Uncertainty: Bayesian Decision Making in Realistic Settings," Organizational Behavior and Human Performance 14, 10-59 (1975).

11. Beach, L.R., and C.R. Peterson; "Man as an Intuitive Statistician," Psychological Bulletin, 1967, 68, 29-46.

12. Beach, Lee Roy and Phillips, Lawrence D.; "Subjective Probabilities Inferred for Estimates and Bets," Journal of Experimental Psychology, Vol. 25, No. 3, 1967, p. 354-359.

13. Birkmire, Deborah P. and Martuza, Victor R.; "Comparison of Selected Graphical/Tabular Displays of Quantitative Information," Department of Educational Foundations, University of Delaware, Newark, Delaware 19711.

14. Birnbaum, M.H., Wong, R. and Wong, L.; "Combining Information from Sources that Vary in Credibility," Memory and Cognition, 1976, Vol. 4, 330-336.

15. Blunt, C.R.; Luckie, P.T.; Mares, E.A. and Smith, D.E.; "The Role of Plausible Reasoning within Military Intelligence - An Application of Bayes Theorem as a Model for Problem Solving," HRB-Singer, Inc. - Navy Contract N00014-66-C0230, May 1967.

16. Box, G.E. and Tiao, G.C.; "Bayesian Inferences in Statistical Analysis," Addison-Wesley, 1973.

17. Brier, Glenn W.; "Verification of Forecasts Expressed in Terms of Probability," Monthly Weather Review, January 1950, p. 1-3.

18. Brown, Rex V.; Andrew S. Kahr and Cameron Peterson; "Decision Analysis for the Manager," Holt, Rinehart and Winston, New York, Chicago, San Francisco, Atlanta, Dallas, Montreal, Toronto, London, Sydney.

19. Brown, Thomas A.; "Admissible Scoring Systems for Continuous Distributions," Printed by the Rand Corporation, August 1974.

20. Brown, Thomas A.; "Probabilistic Forecasts and Reproducing Scoring Systems," Advanced Research Projects Agency, June 1970.

21. Brown, Thomas A., and Emir H. Shuford; "Quantifying Uncertainty into Numerical Probabilities for the Reporting of Intelligence," Defense Advanced Research Projects Agency, July 1973.

22. Brownell, H. and Caramazza, A.; "Categorizing with Fuzzy Categories," Submitted for Publication.

23. Brownell, H.; H.M Hersh, and A. Caramazza; "Influence of Population Distributions on the Form of Fuzzy Membership Functions," To Be Presented at the ORSA/TIMS Meeting, New York, May, 1978.

24. Brownell, H.; Caramazza, A. and Hersh, H.M.; "Reaction Time as a Measure of Imprecise Category Membership," In Preparation.

25. Campbell, Donald T.; "Reforms as Experiments," American Psychologist, Vol. 24, No. 4, April 1969, 409-429.

26. Capen, E.C.; "The Difficulty of Assessing Uncertainty," Journal of Petroleum Technology, p. 843-850, August 1976.

27. Caramazza, A., and H.M. Hersh; "A Fuzzy Set Approach to Modifiers and Vagueness in Natural Language," Journal of Experimental Psychology: General, 1976, 105, 254-276.

28. Caramazza, A., and H.M. Hersh; "Indeterminancy and the Integration of Relational Information," In Preparation.

29. Caramazza, A. and Hersh, H.M.; "Integrating Verbal Quantitative Information," Bulletin of the Psychonomic Society, 1976, Vol. 6, 589-591.

30. Caramazza, A. and Hersh, H.M.; "Quantification of Vague Concepts," Paper presented at the Spring Psychometric Society Meeting, Iowa City, 1975.

31. Charnetski, Johnnie R. and Richard M. Soland; "Multiple Attribute Decision Making with Partial Information III: The Use of Prior Decision Information," Paper still in preparation.

32. Collins, Allan; Warnock, Eleanor; Aiello, Nelleke and Miller, Mark; "Reasoning from Complete Knowledge," To appear in D.G. Bobrow & A.M. Collins, (Eds.) Representation and Understanding, New York: Academic Press, 1975.

33. Corbin, Ruth; "New Directions in Theories of Preferential Choice," Unpublished Master's Thesis, Department of Psychology, McGill University, 1973.

34. Corrigan, B. and Dawes, R.M.; "Linear Models in Decision Making," Psychological Bulletin, Vol. 81, No. 2, 1974, p. 95-106.

35. Crocker, Jennifer and Taylor, Shelley E.; "Theory Driven Processing and the Use of Complex Evidence," Paper presented at the American Psychological Association Annual Meeting, Toronto, Canada, August 1978 as part of a symposium, "Scripts and Schemas: Applications in Social Settings."

36. Crocker, Olga, L.K.; Mitchell, Terence R. and Beach, Lee Roy; "Sources of Judgement Uncertainty," Decision Making Research, Department of Psychology, University of Washington, Technical Report 77-11, September 1977.

37. Dawes, R.M.; "The Robust Beauty of Improper Linear Models in Decision Making," Paper presented at American Psychological Association Meeting, San Francisco, August 1977.

38. Dreyfus, Stuart E.; "Informal Models of Decision-Making," Industrial Engineering and Operations Research, Research Highlight, 46-48.

39. Duda, R.O.; Hart, P.E. and Nilsson, N.J.; "Subjective Bayesian Methods for Rule-Based Inference Systems," Artificial Intelligence Center, Technical Note 124, SRI Project 4763, January 1976.

40. Dulles, Allen; "The Craft of Intelligence," New York, Evanston, and London, Harper & Row, Publishers, 156-165.

41. Edwards, Ward; "Research on the Technology of Inference and Decision," Off. of Naval Research, Adv. Research Projects Agency, AD-A017 525, 31 August 1975.

42. Edwards, W.; Philips, P.D.; Hays, W.L. and Goodman, B.C.; "Probabilistic Information Processing Systems: Design and Evaluation," IEEE Transactions on System Science and Cybernetics, Vol. SSC-4, No. 3, September 1968, p. 248-265.

43. Edwards, Ward and Seaver, David A.; "Research on the Technology of Inference and Decision," Defense Advanced Research Projects Agency, Final Technical Report SSRI 76-7, October 1976.

44. Epstein, Edward S. and Allan H. Murphy; "Verification of Probabilistic Predictions: A Brief Review," Journal of Applied Meteorology, May 1967, p. 748-755.

45. Fain, Tyrus G.; "The Intelligence Community - History, Organization, and Issues," R.R. Bowker Company, New York and London, 1977.

46. Fallon, Richard; "Subjective Assessment of Uncertainty," Paper was originally prepared as part of the course requirements for "The Advisers" conducted by Dr. Herbert Goldhamer at the Rand Graduate Institute, January 1976.

47. Feeney, George J. and Sarah Lichtenstein; "The Importance of the Data-Generating Model in Probability Estimation," General Electric Company and Oregon Research Institute, Organizational Behavior and Human Performance, Vol. 3, No. 1, February 1968.

48. Fischer, Gregory W.; "An Experimental Study of Four Procedures for Aggregating Subjective Probability Assessments," Department of the Navy, Office of Naval Research, Technical Report 75-7, December 1975.

49. Fischhoff, Baruch; Sarah Lichtenstein, and Paul Slovic; "Behavioral Decision Theory," Annual Review of Psychology, 1977, p. 1-39.

50. Fischhoff, Baruch; L.D. Phillips and Sarah Lichtenstein; "Calibration of Probabilities: The State of the Art," In Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, ed. H. Jungermann, G. de Zeeuw, 1976, in press.

51. Fischhoff, Baruch and Sarah Lichtenstein; "Do Those Who Know More also Know More About How Much They Know?" Organizational Behavior and Human Performance 20, 159-183.

52. Fischhoff, Baruch; "Informal Use of Formal Models," Paper presented to session on "Formal vs. Informal Modes of Decision Making," May 1978 Joint Meeting of the Institute of Management Science and Operations Research Society of America.

53. Fischhoff, Baruch; Paul Slovic and Sarah Lichtenstein; "Knowing with Certainty: The Appropriateness of Extreme Confidence," Journal of Experimental Psychology: Human Perception and Performance 1977, Vol. 3, No. 4, 552-564.

54. Fischhoff, Baruch and Paul Slovic; "A Little Learning ...: Confidence in Multicue Judgment Tasks," Decision Research, a Branch of Perceptronics, Attention and Performance VIII, Princeton, NJ, August 1978.

55. Fischhoff, Baruch; "The Silly Certainty of Hindsight," Psychology Today, April 1975, p. 71-76.

56. Fischhoff, Baruch; Slovic, Paul and Lichtenstein, Sarah; "Subjective Sensitivity Analysis," Organizational Behavior and Human Performance, in press.

57. Fiske, Susan T. and Kinder, Donald R.; "Schemas and Political Information-Processing," Paper presented at the annual meeting of the American Psychological Association, August 1978, Toronto, Ontario Canada.

58. Fryback, Dennis G. and Kurt J. Snapper; "Inferences Based on Unreliable Reports," Journal of Experimental Psychology, Vol. 87, No. 3, p. 401-404, 1971.

59. Gettys, Charles; Charles Michel, and James H. Steiger; "Multiple-Stage Probabilistic Information Processing," Organizational Behavior and Human Performance 10, p. 374-387 (1973).

60. Glucksberg, S. and M. McCloskey; "Natural Categories: Well-Defined or Fuzzy Sets?" Paper presented at American Psychological Association Meeting, San Francisco, 1977.

61. Goldberg, Lewis R. and Nerella V. Ramanaiah; "Stylistic Components of Human Judgment: The Generality of Individual Differences," U of Oregon and S. Illinois University, Applied Psychological Measurement, Vol. 1, No. 1, Winter 1977, p. 23-39.

62. Gramann, Richard H.; "Synthesis of Physical Security Assessment Results," U.S. Nuclear Regulatory Commission, Washington, D.C. 20555.

63. Grayson, Anthony S., Capt. and Lanclos, Harold J., Capt.; "A Methodology for Subjective Assessment of Probability Distributions," U.S. Department of Commerce National Technical Information Service, AD-A032 536, September 1976.

64. Green, B.F.; "Descriptions and Explanations: A Comment on Papers by Hoffman and Edwards," B. Kleinmuntz (Ed.), Formal Representation of Human Judgment, New York, John Wiley & Sons, 1968.

65. Hamilos, Christopher A. and Pitz, Gordon F.; "The Encoding and Recognition of Probabilistic Information in a Decision Task," Organizational Behavior and Human Performance, 20, 1977, p. 184-202.

66. Hammerton, M.; "A Case of Radical Probability Estimation," Journal of Experimental Psychology, Vol. 101, No. 2, 1973, p. 252-254.

67. Hampton, J.M.; Moore, P.G. and Thomas, H.; "Subjective Probability and its Measurements," Journal of the Royal Statistical Society, Ser. A, 136 Part 1 (1973), 21-42.

68. Hazard, Ted H. and Peterson, Cameron R.; "Odds Versus Probabilities for Categorical Events," Office of Naval Research, 15 August 1973, AD-770 134.

69. Hersh, H.M.; "A Fuzzy Model of Human Reasoning," Paper presented at the ORSA/TIMS meeting, Atlanta, GA, 1977.

70. Hersh, H.M.; "Fuzzy Reasoning: The Integration of Vague Information," Unpublished doctoral dissertation, Department of Psychology, The Johns Hopkins University, 1976 (submitted for publication).

71. Hersh, H.M.; "A Fuzzy Set Theoretic Analysis of Age Terms," Submitted for publication.

72. Hersh, H.M., and J. Spiering; "How Old is Old?" Paper presented at the annual meeting of the Eastern Psychological Association, New York.

73. Hershman, R.L.; "A Rule for the Integration of Bayesian Opinions," Human Factor, 1971, 13 (3), 255-259.

74. Hogarth, Robin M.; "Cognitive Processes and the Assessment of Subjective Probability Distributions," Journal of the American Statistical

Association, June 1975, Volume 70, Number 350, Invited Paper, Applications Section, 271-289.

75. Hubert, G.P.; "Methods for Quantifying Subjective Probabilities and Multi-Attribute Utilities," Decision Sciences, 1974, Vol. 5, p. 430-458.

76. Jacques, J.S., and M. Norusis; "Diagnosis, I: Symptom Nonindependence Mathematical Models for Diagnosis," Computers and Biomedical Research, Vol. 8, No. 2, April 1975, p. 156-172.

77. Janis, Irving L. and Mann, Leon; "Coping with Decisional Conflict," American Scientist, Vol. 64, November-December 1976.

78. Jensen, Floyd A. and Cameron R. Peterson; "Psychological Effects of Proper Scoring Rules (Abstract)," Organizational Behavior and Human Performance 9, (1973) p. 307.

79. Johnson, Edgar M., Ph.D. and Halpin, Stanley M., Ph.D.; "Multistage Inference Models for Intelligence Analysis," U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington, VA AD 785 639.

80. Kahneman, Daniel, and Amos Tversky; "Availability: A Heuristic for Judging Frequency and Probability," Cognitive Psychology 5, p. 207-232 (1973).

81. Kahneman, Daniel and Tversky, Amos; "Causal Schemata in Judgments Under Uncertainty," Hebrew University, Jerusalem, Progress in Social Psychology, Hillsdale, NJ, 1977.

82. Kahneman, Daniel and Tversky, Amos; "Intuitive Prediction: Biases and Corrective Procedures," Defense Advanced Research Projects Agency, June 1977, Technical Report PTR-1042-77-6.

83. Kahneman, Daniel and Amos Tversky; "Judgment under Uncertainty: Heuristics and Biases," Science, Vol. 185, 1124-1131.

84. Kahneman, Daniel and Amos Tversky; "On the Psychology of Prediction," Psychological Review, Vol. 80, No. 4, July 1973, 237-251.

85. Kahneman, Daniel, and Amos Tversky; "Subjective Probability: A Judgment of Representativeness," Cognitive Psychology 3, 430-454 (1972).

86. Kauffman, A.; Introduction to the Theory of Fuzzy Subsets, Vol. 1, New York: Academic Press, 1975.

87. Kidd, John B.; "Scoring Rules for Subjective Assessments," Operational Research Quarterly, Vol. 26, No. 1, ii, p. 183-195.

88. Kirkpatrick, Lyman B., Jr.; "The U.S. Intelligence Community: Foreign Policy and Domestic Activities," Hill and Wang, New York, A division of Farrar, Straus and Giroux.

89. Kolata, Gina Bari; "Mathematics and Magic: Illumination and Illusion," Science, Vol. 198, October 1977, p. 282-283.

90. Lemmer, J.F.; "Algorithms for Incompletely Specified Distributions in a Generalized Graph Model of Medical Diagnosis," University of Maryland Dissertation, published University Microfilm, Ann Arbor, MI, 1976.

91. Lemmer, J.F.; "A Return to Probabilities in Computer Assisted Medical Diagnosis," Proceedings of the International Conference on Cybernetics and Society, September 19-21, 1977, p. 612-615.

92. Lichtenstein, S., and P. Slovic; "Comparison of Bayesian and Regression Approaches to the study of Information Processing in Judgment," Organizational Behavior and Human Performance, 1971, Vol. 6, 649-744.

93. Luckie, Peter T. and Smith, Dennis E.; "Research Applicable to Problems of Intelligence: Final Report," HRB-Singer, Inc., Science Park, Box 60, State College, PA 16801, July 1968.

94. Lyon, Don and Paul Slovic; "Dominance of Accuracy Information and Neglect of Base Rates in Probability Estimation," University of Oregon and Oregon Research Institute, Acta Psychologica 40 (1976), 287-298.

95. Mancini, L.; J. Meisner, and E. Singer; "A Look at the Use of Subjective Probabilities in an Industrial Environment," TIMS/ORSA Meeting, NY, May 1978.

96. Mayer, R.E.; Thinking and Problem Solving: An Introduction to Human Cognition and Learning, Glenview, Illinois: Scott Foresman and Co., 1977.

97. Murphy, ALH. and R.L. Winkler; "Credible Interval Temperature Forecasting: Some Experimental Results," Monthly Weather Review, 1974, Vol. 102, 784-794.

98. Murphy, A.H.; C.R. Peterson, and K.J. Snapper; "Credible Interval Temperature Forecasts," Bulletin of the American Meteorological Society, 1972, Vol. 53, 966-970.

99. Murphy, Allan H.; "Hedging and Skill Scores for Probability Forecasts," Journal of Applied Meteorology, October 1972, p. 215-223.

100. Murphy, Allan H.; "A Sample Skill Score for Probability Forecasts," Monthly Weather Review, December 1973, p. 48-55.

101. Murphy, Allan H. and Robert L. Winkler; "Forecasters and Probability Forecasts: Some Current Problems," Bulletin American Meteorological Society, Vol. 52, No. 4, April 1971, p. 239-247.

102. Murphy, Allan H. and Robert L. Winkler; "'Good' Probability Assessors," Journal of Applied Meteorology, Vol. 7, p. 751-758, October 1968.

103. Newell, A. and H.A. Simon; Human Problem Solving, Englewood Cliffs, NJ: Prentice-Hall, Inc., 1972.

104. Newsted, Peter R. and Wynne, Bayard E.; "Augmenting Man's Judgment with Interactive Computer Systems," Int. J. Man-Machine Studies, 8, 1976, p. 29-59.

105. Nisbett, Richard E., and Wilson, Timothy Decamp; "Telling more Then We Can Know: Verbal Reports on Mental Processes," Psychological Review, Vol. 84, May 1977, p. 231-259.

106. Offir, Carole Wade; "Seven Quick Ways to Kid Yourself," Psychology Today, April 1975, 66-68.

107. Olson, Chester L.; "Some Apparent Violations of the Representativeness Heuristic in Human Judgment," Journal of Experimental Psychology: Human Perception and Performance, 1976, Vol. 2, No. 4, 599-608.

108. Peterson, Cameron R.; "Judgments of Probability and Utility for Decision-Making," The Department of the Navy, Office of Naval Research, Contract NR 107-014, 30 September 1971.

109. Peterson, Cameron R. and Swensson, Richard G.; "Intuitive Statistical Inferences About Diffuse Hypotheses," Organizational Behavior and Human Performance, Vol. 3, No. 1, February 1968.

110. Pitz, Gordon F.; "A Model for Judgments Based on Uncertain Knowledge," Proceedings of IEEE International Conference on Cybernetics and Society, 1977.

111. Pitz, Gordon F.; "A Structural Theory of Uncertain Knowledge," Probability and Human Decision Making, 1975, p. 163-176.

112. Pitz, G.F.; "Subjective Probability Distributions for Imperfectly Known Quantities," in L.W. GREGG (Ed.), Knowledge and Cognition, Potomac, MD: Lawrence Erlbaum Associates, 1974.

113. Posner, M.I.; Cognition: An Introduction, Glenview, Illinois: Scott Foresman and Co., 1973.

114. Raiffa, Howard; "Decision Analysis," Addison-Wesley, 1970.

115. Ransom, Harry Howe; "The Intelligence Establishment," Harvard University Press, Cambridge, Massachusetts, 1970.

116. Reyna, Valerie F.; "Probability and Modality in the Lexicon," Paper presented at the forty-ninth Annual Meeting of the Eastern Psychological Association, March 29 - April 1, 1978.

117. Rohatgi, U.K.; An Introduction to Probability Theory and Mathematical Statistics, John Wiley, 1976.

118. Saaty, Thomas L., "A Scaling Method for Priorities in Hierarchical Structures," Journal of Mathematical Psychology, Vol. 15, No. 3, June 1977.

119. Samet, Michael G.; "Quantitative Interpretation of Two Qualitative Scales Used to Rate Military Intelligence," Human Factor, 1975, 17(2), p. 192-202.

120. Savage, L.J.; "Elicitation of Personal Probabilities and Expectations," Journal of the American Statistical Association, 1971, Vol. 66, 783-801.

121. Seaver, D.A.; "How Groups Can Assess Uncertainty: Human Interaction Versus Mathematical Models," 1977 Proceedings of the International Conference on Cybernetics and Society, Washington, D.C., September 1977.

122. Slovic, Paul; "From Shakespeare to Simon: Speculations - And Some Evidence - About Man's Ability to Process Information," Oregon Research Institute, Research Bulletin, Vol. 12, No. 2, April 1972, 1-29.

123. Slovic, Paul; "Toward Understanding and Improving Decisions," Decision Research, A Branch of Perceptronics, Eugene, Oregon, 1-47.

124. Slovic, Paul and Amos Tversky; "Who Accepts Savage's Axiom?" Behavioral Science, Vol. 19, p. 368-373, 1974.

125. Spetzler, Carl S. and Carl-Axel S. Stael Von Holstein; "Probability Encoding in Decision Analysis," Management Science, Vol. 22, No. 3, November 1975.

126. Stael Von Holstein, Carl-Axel S.; "An Experiment in Probabilistic Weather Forecasting," Journal of Applied Meteorology, Vol. 10, p. 635-645.

127. Stael Von Holstein, C.-A.S.; Assessment and Evaluation of Subjective Probability Distributions, Stockholm: The Economic Research Inst., 1970.

128. Stael Von Holstein, Carl-Axel S.; "A Bibliography on Encoding of Subjective Probability Distributions," Stanford Research Institute, July 5, 1973.

129. Stael Von Holstein, C.-A.S.; "Probabilistic Forecasting: An Experiment Related to the Stock Market," Organizational Behavior and Human Performance, Vol. 8, 1972, 139-158.

130. Stael Von Holstein, C.-A.S.; "Some Problems in the Practical Application of Bayesian Decision Theory," Behavioral Approaches to Management, Gothenburg: The Graduate School of Economics and Business Administration, 1970.

131. Stael Von Holstein, Carl-Axel S.; "A Tutorial in Decision Analysis," Paper presented at the Third Research Conference on Subjective Probability, Utility, and Decision Making, London, September 7-9, 1971.

132. Tversky, A.; "Assessing Uncertainty," Journal of the Royal Statistical Society, 1974, Ser. B, 36, 148-159.

133. Van Orden, M.D. USN (Ret.); "Management by Decision," Signal, September 1978, 35-39.

134. Wainer, H.; "Estimating Coefficients in Linear Models: It Don't Make No Nevermind," Psychological Bulletin, Vol. 83, No. 2, 1976, 213-217.

135. Winkler, R.L.; "The Assessment of Prior Distributions in Bayesian Analysis," Journal of the American Statistical Association, 1967, Vol. 62, 776-800.

136. Winkler, Robert L.; "Probabilistic Prediction: Some Experimental Results," Journal of the American Statistical Association, December 1971, p. 675-685.

137. Winkler, Robert L.; "Scoring Rules and the Evaluation of Probability Assessors," American Statistical Association Journal, September 1960.

138. Wyer, Robert S., Jr.; "An Investigation of the Relations Among Probability Estimates," Organizational Behavior and Human Performance 15, 1-18 (1976) p. 1-18.

139. Zadeh, L.A.; "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning - I," Information Sciences, 1975, Vol. 8, 199-249.

140. Zadeh, L.A.; "Fuzzy Sets," Information and Control, 1965, 8, 338-353.

141. Zadeh, L.A.; "Fuzzy Sets as a Basis for a Theory of Possibility," Fuzzy Sets and Systems, 1978, Vol. 1, 3-28.

142. Zadeh, L.A.; "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," IEEE Transactions on Systems, Man, and Cybernetics, 1973, SMC-3, 28-44.

143. Zeleny, Milan; "Adaptive Displacement of Preferences in Decision Making," TIMS Studies in the Management Sciences 6 (1977), 147-157.

APPENDIX B

INTRODUCTION

This Training Manual forms a part of the Final Report for Contract #F30602-78-C-0291, Aggregating and Communicating Uncertainty.

It consists of ten units which take the form of half-hour briefings on methods for quantifying, aggregating, and communicating uncertainty in estimative intelligence. Although the approach is specifically designed for use by DIA/DE, most of the suggestions included here should also be of value to the production of other forms of strategic intelligence.

The material included in this Training Manual is based upon research which is described in the main body of the Final Report. References and additional background information may be found there. This manual is not mathematical or statistical in orientation. A mathematical presentation of methods for aggregating uncertainty is included as an appendix to the Final Report.

Topics to be included in this manual are as follows:

1. Introduction and Background.
2. Intelligence as a Science.
3. The Uses of Uncertainty.

4. Probability Assessments

5. Combining Probabilities

6. Hindsight and Second-Guessing

7. Further Errors in Assessing and Combining Probabilities

8. Scoring Rules

9. Communicating Uncertainty

10. Conclusion

UNIT 1

INTRODUCTION AND BACKGROUND

Throughout this manual, we refer to two people; the intelligence producer and the intelligence consumer. The producer is the person who gathers information, analyzes it, summarizes it, tabulates it, and eventually produces a report of some kind. The consumer is the person who receives an intelligence report. The consumer may, in turn, become a producer if he subjects the information to further processing; on the other hand, he may take some other kind of action, such as developing recommendation for a military budget.

By its very nature, the information contained in an intelligence report is uncertain. It represents the best estimate that can be made, given the resources, time, and information available. Since it deals with the projected actions of a foreign, possibly hostile power, an intelligence report will often depend upon sources of information which contain errors, gaps, and misconstruals of the nation's intentions. In addition, the intelligence producer himself is fallible: his understanding of the material may not be complete, or his own biases may have led him to exaggerate an estimate.

It is important for the producer to be able to communicate the uncertainty present in an intelligence report, and DIA has long recognized the need for communicating this uncertainty to the consumers of its products. In 1968, for example, an NIE contained phrases like these:

it is likely that

it is unlikely that

the possibility of

the likelihood of

it is probable that

they probably believe

Such phrases were intended to let the consumer know that the estimate was not gospel. It contained valuable information, but it was not free of potential errors.

In 1976, more precise forms for expressing uncertainty were introduced into DIA/DE products, using such phrases as the following:

60 percent probability

80 percent likelihood

a 70 percent chance

Such expressions were applied to specific events, such as a projected military policy.

In some estimates issued during 1976, a colored sheet containing the following statement was included:

"[Numeric forms are used] to convey to the reader this degree of probability more precisely than is possible in the traditional verbal form. Our confidence in the supporting evidence is taken into account in making these quantifications. . . . All efforts at quantifying estimates are highly subjective, however, and should be treated with reserve."

At about the same time, a DIE contained the following notice:

"Completeness and Reliability of Evidence.

"The evidence . . . is based on a wide variety of sources and is considered generally complete and reliable, although not necessarily definitive . . . There is as yet very little reliable evidence . . . The data base . . . is considered sufficiently reliable to support the judgments made . . ."

Another measure of uncertainty has become familiar through its use in the DIPP volumes. Quantitative estimates frequently (but not invariably) include a "high," "low," and "best" value. These are selected in such a way that the true value should be found within the indicated range approximately 75 percent of the time. The spread from low to high, and occasional spreads within the high and low values, have been intended to assist the consumer in determining the uncertainty of the estimate.

In spite of DE's continuing effort to provide consumers with some indication of the uncertainty present in its products, there is little evidence that the consumers have actually been using this information. Instead, a number of problems, for both the producers and the consumers, have arisen:

- The consumers frequently take the "best" estimate as though it were an unqualified projection, without noting the range of uncertainty in the set of projections.

- In spite of warnings, some consumers appear to be taking the "high" estimates as representing the "worst case" against which U.S. forces must prepare. This is particularly unfortunate, since the high projections from various sections of the DIPP cannot be combined. (This would imply, for example, that the Soviets were placing a heavy emphasis on two different systems simultaneously; the estimator intended to show that either one system or the other could be emphasized -- but not both at once.)

- Intelligence producers apparently also had trouble in justifying their assessments of uncertainty. The numerical estimates of probabilities, generally reported as percentages, were regarded as "highly subjective." No clearly-defined methods of obtaining the required numbers have been specified.

- o There was little motivation to improve the quality of the probability figures. They were plucked out of the air, representing (in the view of some estimators) little more than guesswork or "intuition.")

A vicious circle thus appears to have developed, in which neither the producers nor the consumers take the probabilities very seriously. Producers tend to regard them as mere guesses, and consumers tend to ignore them.

In this manual, we will generally refer to the formulation of a measure of uncertainty as an assessment, to distinguish it from an estimate of foreign military capbilities. A primary goal of the manual will be to show that the assessments of uncertainty can be as important for the consumer as the estimates themselves. In fact, we have come to believe that probability assessments are essential, if the consumer is to make effective use of estimates.

The goal of estimative intelligence, like that of other forms of strategic intelligence, is the determination of the goals and capabilities of the adversary. This would mean that an "error" could only be a failure to determine those goals and capabilities correctly. Any other definition would cast the intelligence producer in the role of fortune teller, attempting to determine the future without the aid of empirical evidence. This definition of "error" may be somewhat difficult to enforce, since the only evidence that we may have of the enemy's intentions is what he actually does; an estimate is probably correct if it predicts what he does. But this

definition does help to take the role of estimation out of the realm of the mystical, and into the realm of the empirical sciences. Our point of view is one that says that estimative intelligence is a form of social science.

One word that has had mystical connotations has been "intuition." The word "intuition" was sometimes used by estimators to describe the process by which they arrive at projections. This is rather misleading, since it suggests that intelligence production is sometimes little more than guesswork.

The role of "intuition" becomes more significant if we recall that master-level checker players, who were questioned about their methods in connection with a checker-playing computer program, often said that they chose their most successful moves by "intuition." By this, they simply meant that there were no general rules guiding their choices; instead, they relied on their understanding of the game as a whole, their sense of the patterns present on the checker board, their choice of a strategy for this particular game, and so on.

Similarly, "intuition" for an intelligence producer could include a global understanding of the nation as a whole, some insight into typical strategies employed, a recognition of specific capabilities, and a variety of "fringe" or ancillary factors that could influence a decision concerning weapon development, deployment, or withdrawal.

For example, an estimator may be considering Soviet ABMs. The current SALT agreement may permit 100 missile launchers, as a maximum, at Moscow. In fact, they have 64 launchers. What will they do? The estimator believes (let us suppose) that the Soviets are very concerned to protect Moscow, and that therefore they will increase the number of launchers around the city. The estimator thus draws on a general model or "picture" of Soviet goals and priorities, using it to predict a concrete action to be taken by them.

In strategic intelligence production, "intuition" is the process by which the experienced producer attempts to take relevant factors from a variety of sources into account, and to combine them to form a comprehensive pattern that "makes sense" of the observed phenomena.

Of course, intuition is not a substitute for hard work. In actual practice, projections are most frequently generated through the use of such estimation parameters as deployment rates, rates of change, retirement rates, estimates of ratios among weapon systems, and so on. These more mundane figures provide the basis for determining how many weapon systems of a given type will be deployed at a specified future date. In the extreme case, an estimator who is pressed for time may simply use a straight-line extrapolation of current trends. In any case, some uncertainty is present in the projections, since there is uncertainty concerning all these parameters -- particularly in the timing for introduction and phase-out of a weapon system.

A specific example may make this process clearer. In estimating future naval systems, the estimator knows that prototypes are planned five years in advance. Designs and requirements are specified, and, in the Soviet Union, the vessels are produced over a period of ten years. Naval vessels have a twenty-year life span. Using these figures, the estimator can construct a simple mathematical model of Soviet naval development.

A complicating factor is the "learning curve" exhibited in the development of a weapon system. Production begins slowly, as people in a factory are learning how to produce a system and as bugs in production and in design are located and eliminated. Then there is a period of rapid growth in numbers of weapons, as maximum production is obtained, for a period of years. Then this tapers off, as production is slowed and finally halted.

Both "intuition" and somewhat more formal models, like the one just described, are used in the development of projections. The next important task will be a more complete description of the estimative process, to locate precisely those points at which uncertainty enters.

UNIT 2

INTELLIGENCE AS A SCIENCE

Strategic intelligence is a social science. It draws upon such specialized sciences as history, economics, and political science. In addition, strategic intelligence draws upon military science and the technology of weapon systems as other sources of information.

More importantly, strategic intelligence methods most closely resemble the methods used by working social scientists, which are often quite informal in the way in which hypotheses are developed and presented.

Most important of all, intelligence estimates must represent a consensus of the intelligence community. They not only draw upon the combined resources of the community, but they must also be justified to the community. The process of justification requires evidence and argument; it is not simply a process of taking a vote among delegates.

This point should be emphasized. Meetings with other representatives of the intelligence community are not merely vote-taking sessions. The majority opinion does not prevail (or should not prevail) if the majority has no evidence for its opinion, no logically-justified set of arguments. A minority opinion, even that of a minority of one, should be able to win the required consensus, if that opinion is properly buttressed with relevant evidence and argument.

It is the need for scientifically respectable arguments that motivates this study of estimative intelligence methods. If DE's task were simply that of taking an opinion poll among various specialists, then its job would be a good deal simpler than it is. Instead, DE must formulate and evaluate those opinions, selecting the viewpoint that is most plausible, in the light of current evidence. It is in this sense that DE's work can be called a form of social science.

Here is an example of the type of reasoning that strategic intelligence can employ:

"While working on Eisenhower's scientific advisory committee in 1959 and 1960 I had to assess some of the early claims that the Russians were developing an ABM system. The Soviets, we knew from our intelligence, had a center for antiaircraft and antimissile work at Sary Shagan in Central Asia. Our U-2 planes observed there a large radar installation that might, it was thought, be a device for detecting incoming missiles. Our intelligence experts immediately linked this installation to the Soviet tests of medium-range ballistic missiles at Kapustin Yar, many hundreds of miles to the west. They conjectured that the Russians were putting together the combination of radar (to detect incoming missiles), computers (to track them), and interceptor missiles (to destroy them) that makes up an ABM system." (George B. Kistiakowsky, "False Alarm: The Story Behind Salt II," New York Review of Books, March 22, 1979, pp. 33-38.)

Several elements make up this conjecture:

- o A general hypothesis has been proposed: that the USSR is developing an ABM system, which will include installations at Sary Shagan and Kapustin Yar.

- o Supporting evidence for this hypothesis: the observation of a radar installation and of Soviet tests of medium-range missiles.

- o A set of inferences that link the hypothesis with the observations: the fact that an ABM installation requires the existence of radars, computers, and missiles.

As more supporting evidence is found, the probability of the hypothesis increases; disconfirming evidence will tend to decrease this probability.

As in all sciences, these three elements are crucial. There must be a general hypothesis; without it, you would hardly know what counted as evidence and what didn't. You have to know what your are trying to prove.

Evidence is essential to every empirical science. Strategic intelligence -- more than any other science -- is based on observation; it requires knowledge of the enemy's actions and capabilities for action.

The third element is a set of inferences that tell us that if the hypothesis is true, then there will be certain observable results. The inferences provide a structure within which the scientific method operates.

The essential point is this. You must have a hypothesis. Otherwise, you don't know what facts to collect, where to collect them, or why you are collecting them. The general hypothesis makes sense out of a disparate collection of data.

Evidence for the hypothesis is provided by those observations which it implies:

- Example 1. You hypothesize the development of an ABM facility. This implies that radar equipment and IRBMs will be installed. Observation of the radars and IRBM installations will provide evidence for the hypothesis.

- Example 2. You hypothesize increasing emphasis on civil defense in the USSR. This implies that air raid shelters will be constructed and a civilian warning system will be installed. Observation of the shelters and the warning system provides evidence for the hypothesis.

The connection between hypothesis and conclusion is not logically tight, for several reasons. These points -- where there are several logical possibilities available -- are the points at which uncertainty enters.

Let's see whether we can find some of the logical loopholes in the argument that was presented as an example:

- The initial identification of the center at Sary Shagan as a center for antiaircraft and antimissile work is based on someone's train of reasoning. We don't know who he was, or what made him think that the center had those purposes. Without this information, we don't know whether this was just a wild guess or a firmly founded conclusion. The degree of credibility in this item is strictly unknown.

- The identification of construction at this site as a radar installation is also uncertain. This time, we may have an idea of who it was that made the identification. If we do, then we probably know enough of his past record of similar identifications. Since everybody makes mistakes, this record may not be perfect, but we will at least know how uncertain the identification was.

- Is the radar installation intended to detect incoming missiles? We have no idea. This identification is derived from our initial hypothesis. The example shows how the hypothesis makes sense out of the radar installation; if the hypothesis is true, then we know what the radar installation is for; otherwise, we would only be guessing at its purpose.

You should be able to find several additional sources of uncertainty in this example, which is a good deal simpler than most of the examples you work with in the real world. Unlike a mathematician, the intelligence producer rarely obtains a logically tight argument; his arguments always contain loopholes and missing, but essential, pieces of evidence. Much the same can be said, of course, of the historian. Uncertainty is an inevitable part of history, as of strategic intelligence. The task is to determine how much uncertainty there is.

Like other forms of strategic intelligence, estimative intelligence is concerned with the determination of the capabilities and intentions of a potential adversary (and, in the case of the NATO nations and Japan, of an ally). Estimative intelligence differs from current and basic intelligence in that the capabilities and intentions must be projected into the future, where (at least for finite human beings) events are indeterminate.

In practical terms this means that the intentions of a potential adversary may change in unpredictable ways over the course of the next ten or twenty years. The rise of a Sadat or a Khomeini in the Middle East, for example, has brought about changes in the intentions of the nations that they represent. While it is possible, thanks to hindsight, to identify those forces within Egyptian or Iranian society that gave rise to the policies of Sadat and Khomeini, it would have been wildly speculative to have predicted them ten years in advance. Similarly, new inventions and discoveries may contribute to basic changes in the capabilities of an adversary. Thus,

many technological developments which are operational today could not have been realistically foreseen ten or twenty years ago.

In short, the loose logical structure of the arguments used in estimative intelligence, together with the indeterminate quality of much of the evidence that supports them, means that a degree of uncertainty is present in all of the finished intelligence that it produces.

At th present time, DE uses several techniques for communicating uncertainty in its estimates, for the use of intelligence consumers. Since you are already familiar with these techniques, they will not be described in great detail here. They are listed only to indicate the ways in which DE has, in the past, attempted to communicate uncertainty:

- Kent Charts. Until recently, DE has used reporting methods proposed by Sherman Kent. Essentially, the Kent chart provides a translation from certain natural language phrases ("It is likely that . . . ") into numerical estimates of probability. Kent developed this approach following his observation that the natural-language phrases were subject to wide variations in interpretation, and that they served to conceal disagreements concerning the likelihood or uncertainty present in intelligence reports.

- Reliability-Accuracy Codes. A coding system which is widely used for qualifying intelligence reports (but which does not

appear to have been used by DE) indicates the estimated reliability and accuracy of the report as follows:

SOURCE RELIABILITY	INFORMATION ACCURACY
A. Completely reliable	1. Confirmed
B. Usually reliable	2. Probably true
C. Fairly reliable	3. Possibly true
D. Not usually reliable	4. Doubtfully true
E. Unreliable	5. Improbable
F. Reliability cannot be judged	6. Accuracy cannot be judged

Complaints concerning this system are interesting, because they indicate the sort of information that your consumers may be looking for in your assessments of uncertainty. *In general*, the reliability-accuracy rating system was found to be:

oo Too mechanical. Consumers would like to know the reasons behind the uncertainty in a report, rather than a brief code. They would like to know more about the source of material as an aid in their own evaluation of the contents.

oo Too few categories actually used. Too often, the rating would be given as "undetermined" or would be given some middle-ground evaluation. Consumers would like more discrimination among the levels of uncertainty.

oo Accuracy and reliability not separated. The accuracy and reliability ratings were not independent factors, to judge from the way in which they were used. If two different ratings are to be given, they should not simply be duplicates of one another. Apparently, neither the producers nor the consumers were able to keep them separate.

oo System under-used. Only 48 percent of spot reports in an Army field exercise were rated for both reliability and accuracy. This suggests that tactical intelligence personnel simply did not have the information necessary to provide reliability and accuracy ratings. It also suggests that if they were forced to provide such ratings, they would probably be guessing. Simple guesswork would not be useful for the consumers.

oo Ratings mostly the same. In the Army exercises, category B2 alone contained 74 percent of all ratings. This indicates that the ratings were not of much help to the consumers in distinguishing levels of uncertainty among the reports.

oo Ratings inconsistent. Even experienced intelligence analysts gave conflicting ratings to the reports, and these inconsistencies could not be removed through training. Obviously,

intelligence personnel had difficulty in using this rating system. Consumers would have difficulty in taking the ratings seriously, if different intelligence analysts could not agree on the ratings to be assigned to the same reports.

It should be noted that these difficulties were observed in the assignment of ratings to tactical intelligence reports in Army exercises, using a system which has not been employed by DE. The same difficulties can nevertheless be expected in the use of almost any rating system, in strategic intelligence applications as well as in tactical intelligence. A more effective approach is needed.

- Numerical Assessments. In 1976, a system of numerical assessments of the probability of estimates was introduced into DE products. Typically, these include parathetical statements of the form "(70 percent probable)" or "(60 percent likely)." These are much more precise and less ambiguous than phrases like "It is somewhat likely that . . .," but for this reason they may be harder to prepare. An estimator who was willing to say "It is somewhat probable that . . ." might not be willing to commit himself to a definite number. The numbers are just too precise to be determined with any degree of confidence. Another problem has been that there is little feedback. Was a figure like "80 percent probable" warranted? There is no way of finding out, because there have been no detailed studies of the success of projections of this type.

- Confidence Ranges. In addition to the probabilities assigned to specific events or developments, DE provides confidence ranges for numerical estimates of force levels for most countries. (A single estimate, not a range, is provided for noncommunist nations.) Roughly 75 percent of the actual values will be found between the High and the Low values of the range. Since there has been little effort to verify this figure, it can simply be taken as a general indication of the uncertainty of the projected values. Both the producers and the consumers tend to ignore the "75 percent" figure, and to assume that the High figures reflect an all-out effort, while the Low figures represent the minimum plausible effort.

- General Assumptions. At the start of each DIPP, and in footnotes throughout the DIPP, assumptions are stated which assist in communicating some of the uncertainties present in the projections. Assumptions may include the observance or nonobservance of a treaty (such as SALT agreements), lack of major hostilities, continuation of the present regime, etc. While no probabilities are attached to these assumptions, users of the DIPP are thereby warned that the published estimates are conditional upon them. Thus, they assist in conveying some degree of uncertainty for the reports to which they are attached.

- Lack of Consensus. Occasionally, no consensus is obtained among the agencies responsible for developing estimates. When this occurs, a footnote or appendix may be added, indicating the lack of agreement, together with some of the justification provided for the dissenting position. When there is no agreement, such footnotes or appendixes assist the consumer in evaluating the degree of uncertainty that may be present in a published projection.

Three conclusions from this survey are particularly impressive:

- DE has repeatedly attempted to convey the degree of uncertainty present in its estimates, using a variety of techniques.

- Intelligence producers find it difficult to assess the degree of uncertainty precisely.

- Intelligence consumers are not satisfied with current methods of reporting uncertainty, as evidenced by the fact that they rarely make use of this information.

With these preliminaries out of the way, we can turn now to the major question: Is there some more effective way of assessing and communicating the uncertainty of an estimate?

Projections over a limited time -- over perhaps as many as five years -- can be made with some assurance, on the basis of extrapolations of known technology, known production capabilities, and reasonable assumptions concerning intentions. Beyond this point, however, assurance drops off rapidly. It may not be known, for example, when a given weapon system will be regarded as obsolete, and errors of several years may occur in predicting this point. Even though you may correctly predict that a system will be abandoned in, say, the five-to-ten-year time frame, an error in predicting the exact date may result in serious numerical errors during the years over which your error extends.

For example, you may believe correctly that an aircraft type will be abandoned at some time between 1985 and 1990. Your projections indicate that this data will be 1987. As it happens, the retirement date for this aircraft turns out to be 1989. This minor error will mean that the actual numbers of this aircraft will be considerably higher than your projections for the years 1987-1988.

This type of uncertainty is difficult to handle under the current reporting system, which indicates a high and a low projection for each year along the vertical axis. In this case, however, the uncertainty lies along the horizontal axis; it is an uncertainty concerning the time of retirement, rather than an uncertainty concerning the number of aircraft. To indicate this uncertainty, a horizontal line, labelled "year of retirement," could be added to graphical presentations of the projections, in addition to the vertical line representing the number of aircraft for each year.

In addition, notes to the projection could indicate the reasoning that went into it. These notes permit the consumer to evaluate the degree of uncertainty to be attached to the specific numbers.

For example, a hypothesis, such as a date for withdrawal of the Badger, is proposed. Arguments for or against the hypothesis are considered: The Soviet tendency to retain obsolescent equipment, the record of success of the aircraft, lack of evidence of new equipment to replace the Badger, the general need for aircraft with these capabilities within the Soviet defense system, outstanding orders from satellite nations. Alternative hypotheses are considered and evaluated. A selection is then made from among the competing hypotheses, a probability is attached to indicate the degree of uncertainty that it contains, and the most-likely hypothesis then serves as the basis for a defensible projection.

This last example suggests the need for a flexible system for reporting uncertainty, rather than a mechanized or stereotyped system. If measures of uncertainty are presented in a variety of ways, to fit the data and the subject matter, they may do a better job of convincing the consumer that he ought to take them seriously.

UNIT 3

THE USES OF UNCERTAINTY

In a sense, reporting uncertainty provides you with a hedge. If you say, "There's a fifty-fifty chance of X," then you can't be all wrong, no matter what happens. But excessive hedging makes it more difficult for the person who eventually has to use your report, the intelligence consumer. He is not going to be happy with a report that gives him no basis for his decisions or other actions.

In this unit, we will argue first for the use of numerical expressions of uncertainty. While you can never, of course, be completely certain of the truth of your predictions -- this would imply that you were some kind of clairvoyant -- you can at least be precise about the degree of uncertainty that you believe that they contain. Numerical expressions give you this kind of precision.

Numerical expressions ("There is a 60 percent chance that . . .") are useful to several types of consumers, and to the producers themselves:

- In the development of games and simulations, it is important to include the probabilities connected with each scenario. These probabilities form the basis for evaluating the plausibility of various outcomes. In a training situation, for example, the trainee should be faced with a realistic mix of weapons and forces,

which can be derived from the probability distributions supplied as inputs to the program. These probabilities -- these numbers -- have to be obtained from intelligence estimates. If the estimates are vague -- if they use expressions like "It is somewhat probable that . . ." -- then the programmers of the games and simulations will simply have to guess at the actual probabilities involved.

Guesses about the meaning of "It is somewhat probable that . . ." are not likely to be very reliable. Experiments have shown that people can guess anything from 0.40 to 0.80 as the actual probability intended by this phrase. What is worse, the estimator may not even have known what probability he intended by it -- he may only have been hedging his estimate.

Providing a precise assessment of uncertainty in numerical form thus helps one group of users, by providing them with the numbers that they need for games and simulations. In addition, it forces you to <u>think</u> about the degree of uncertainty that is present in your estimate, and not merely to use a verbal expression as a hedge.

- A second major area in which consumers require numerical assessments of uncertainty is that in which recommendations must be made for research and development of U.S. forces and systems. It is important not only to know the specific projected strengths and

weaknesses of Soviet, Chinese, and other foreign military powers; it is also important to know the probability that various levels can be attained.

If the U.S. had unlimited monetary and material resources capable of developing overwhelming capacity to respond to every conceivable threat, there would be no need for assessing uncertainty in DE's projections. All U.S. capabilities would be developed to their maximum.

It is hardly necessary to discuss the reasons that this approach to military strategy could not possibly be considered. In any case, intelligent military strategy requires that available resources be used in the places in which they will do the most good. Under these conditions, it is important to know, with some precision, exactly what probability to attach to various potential threats. In this way, limited resources can be directed to exactly those areas in which they are most likely to be needed.

Numerical assessments of uncertainty can assist by providing a clear-cut basis for recommendations for U.S. weapon system development, required by U.S. strategic planning.

- Not only research and development, but production and deployment of U.S. weapon systems will depend on the projected defense posture of the USSR and other potential adversaries. The probabilities

associated with a projected Soviet development will have an effect on the U.S. defense posture, which must be prepared to counter the most probable enemy threats.

- Numerical assessments of probability are also useful to intelligence producers. They need an accurate method for determining when their work in the past has been correct. Vague expressions like "There is some likelihood that . . . " cannot be adequately evaluated, in the light of later developments. On the other hand, a numerical expression, like "There is a 30 percent probability that . . ." can be given a well-defined score, which will assist the intelligence producer in determining exactly how good his work has been, and where the trouble spots are.

Four potential applications for a measure of uncertainty have been suggested; the next step is to describe precisely what this measure does. What are we measuring?

In a sense, any consistent set of words or numbers could be used to represent your assessment of the uncertainty of an estimate. You could use "A" to represent complete certainty, and "E" to represent complete uncertainty, for example. You could use the numbers 1 to 100, or any other set of numbers, provided that the consumer knew what the letters or numbers meant.

For the types of application that we have suggested here, however, the most useful set of numbers would be those that represent the probability, likelihood, or possibility of the event, or the plausibility of the projected figures or other estimates. Such a probability is assessed in the light of the information available today. If the quality of the information available to you is good, then the degree of certainty will be high, and the probability will be close to 1.0. If your information is doubtful, and the hypotheses that you build upon it are speculative, then the degree of uncertainty will be greater, and the probability will go down.

Although you will probably never want to claim that you are absolutely certain of anything, at the same time you have a supply of information and assumptions which are not really doubtful. The numbers, dimensions, capabilities, and deployment of a variety of Soviet equipments are known, for example, to the extent that it would be wasteful of your time and your reader's time to attempt to deal with probabilities in connection with them. Well-documented, thoroughly confirmed information can be given a tentative probability rating of 1.0, indicating that there is no good reason to doubt it.

At the other end of the scale, there are situations about which we seem to know nothing whatever. It is hard to imagine a projected event about which we actually know nothing; if this were the case, we would not even know how to describe the event, at least not with any understanding. For example, suppose that I have thrown a pair of dice, and have covered them with my hand. Has a seven come up? If you were totally uncertain about the

outcome, you would nevertheless have to know that each die has six sides, with the numbers 1-6, and that there is no reason to believe that any one number is more likely than any other. This is background information that you need simply to understand what it means to throw a pair of dice. Thus, you could say that, if you were totally uncertain about the outcome of a roll of the dice, then the probability that it is a seven is 6/36, because that represents all the possible ways that it could come up seven, divided by all the possible ways that it could come up with any total.

In other words, total uncertainty is represented by the probability of the event, based on whatever background knowledge is available concerning events of this general type.

For example, if you were forecasting the weather and, thanks to a freakish electrical connection, you could gather no data concerning tommorrow's weather, you could produce a forecast based on your general background knowledge of the climate. If it generally rained on 40 percent of the days in your area in October, then you could forecast a 40 percent chance of rain -- a reflection of your total uncertainty about any specific day.

To assess a condition in which your uncertainty is complete, then, you assign a probability based on your general background knowledge of events of this type.

Note that, in general, the probability associated with total uncertainty is not "fifty-fifty." A 50 percent probability occurs only in those situations, like a flip of a coin, in which your general background knowledge suggests that the two possibilities are equally likely, and that no other events are possible. Your total uncertainty concerning the outcome of a flip of a coin can be expressed by your assessment of a 50 percent probability that a head (or a tail) will come up.

There are two extremes, then:

- Total certainty, with an assessed probability of 1.0, which is claimed for that information which is so well-confirmed that it can no longer be called into question.

- Total uncertainty, with an assessed probability that depends on your general background knowledge concerning events of this type, which is applicable to information for which you have no confirming or disconfirming evidence.

From a value for total uncertainty, each new piece of information raises or lowers the probability of a projected event. For example if the probability of rain is 40 percent, then your observation of a rising barometer will suggest that this probability should be decreased; a falling barometer will suggest that it be increased. An increase or decrease in relative humidity, a change in wind direction, an observation of cloud patterns, and more global information obtained from satellite photos, for example -- all of

these sources of new information will affect the initial (or prior) probability that you have assigned. They combine to form the final (or posterior) probability, which is a measure of the uncertainty of the event, on the basis of both old and new knowledge.

A familiar situation in intelligence work is that in which some of the information is missing. There is no way of finding out -- short of invisibly attending a meeting at the Kremlin -- exactly what the Soviets plan to do concerning a certain type of equipment.

This is the problem of "missing data," and it is a familiar one in all the social sciences including and especially history: for large sections of history, essential pieces of information have been lost forever and can never be recovered.

The effect of missing data is to increase the uncertainty of the estimate. This means that the estimate is pushed closer to the prior probability than it would be if the data were available. For example, construction work is observed at Factory X. If you could obtain information concerning the purpose of this construction, it would increase or decrease the probability that helicopter Y is eventually going to be produced. But you do not know the purpose of the new construction at the factory. Therefore, the uncertainty concerning the helicopter remains.

Suppose that we hypothesize that the Soviets have a requirement for equipment capable of transporting heavy munitions quickly to otherwise inaccessible areas along the Eastern front. Large helicopters would be ideal for this purpose, if the engineering problems could be solved. Given your general background knowledge concerning Soviet requirements, policies, and technology, you assess a 40 percent chance that the Soviets will produce and deploy several large helicopters of the Homer class by 1985. Next, you receive information concerning construction of Factory X. If you could obtain information concerning the purpose of Factory X, it might increase or decrease your confidence in your projection of the production of large helicopters. Unfortunately, you can obtain no information, and you can only hypothesize that Factory X could be used to produce Homers. It is consistent with your hypothesis, but it neither supports nor undermines it. Until you have additional information concerning Factory X, the probability remains at 40 percent.

In other words, missing data simply represent the source of some of your uncertainty concerning an estimate. Since some data will always be missing in any interesting problem in strategic intelligence, the primary problem is the development of methods to communicate the resulting uncertainty to the consumer.

Your general understanding of the nation helps to provide a context which will limit the effect of missing data. Like a partially-completed picture puzzle, it provides a general picture of the policies and capabilities of a potential adversary. While you may not know the details of a specific

meeting in the Kremlin, you can at least gain some idea of what would happen at such a meeting, based on your general knowledge of Soviet attitudes, combined with all the information that you do have concerning Soviet activities before and after the meeting.

In short, the scientific method requires that you account for all the available data within the context of a general hypothesis concerning the phenomena that you want to investigate. As new data are obtained, they tend to verify your tentative hypothesis; or they may cause you to modify or reject it. The role of missing data, then, is to increase the uncertainty present in your general picture of the nation; if all data were missing, uncertainty would be total, and if no data were missing, then there would be no uncertainty.

Another source of uncertainty in the initial data may derive from errors in such sources of information as the various orders of battle (OBs), which provide the base line for projections. If the initial figure for a projection is incorrect, then this error will be propagated throughout the projection, resulting in an underestimate or overestimate over the entire period.

Like all products of the intelligence community, the OB represents an estimate; it is an estimate of current force levels, and is, presumably, an authoritative source of information. Since it is an estimate, however, it is subject to correction. For example, more accurate sensor systems or more effective methods of analysis may produce better estimates of enemy force levels.

The variability of the OB over a period of time will introduce a degree of uncertainty into your projections, to the extent that projections are based upon the OB. The variability, and therefore the uncertainty, of your projections are increased by the amount of variability present in the OB. It may be somewhat difficult to introduce this degree of uncertainty into projections, since the OB is usually taken as a given, a unique figure for the year. But the variability of the OB can be taken as the minimum variability that is present in your projections. That is, you cannot expect to do better than the OB in eliminating uncertainty from your projections of the future. A review of errors in the OB, then, will help to give some sense of the types and amounts of errors to be expected in estimates of future forces and capabilities.

An additional source of uncertainty, which is present in all intelligence work, is conscious deception on the part of the foreign power under study. Even our allies may habitually provide misleading figures concerning their capabilities. One nation, wishing to conceal the extent of its defenses, may produce figures which are much smaller than the actual figures; another nation, which wants to give an exaggerated vision of its capabilities, may provide artificially inflated figures.

Intelligence producers are aware of deceptions like these, and may revise their projections toward more realistic numbers than those provided by the governments themselves. Even when there is no conscious deception, military leaders in a small nation may have an exaggerated (or excessively modest) view of their own country's financial or technical capabilities. A

more difficult task is provided in some instances by those nations in which there are simply no realistic figures available to anyone. Under these conditions, the task of the intelligence producer is to develop realistic projections on the basis of whatever data may be available.

What is required is a comprehensive understanding of the goals and capabilities of the nation as a whole. Typically, you work from your knowledge of the country's history, its aspirations, the specific goals of its leaders; these are combined with your understanding of its technological research capabilities, its production capacities, its economic resources. No political leaders have a completely free hand to accomplish whatever they want; on the contrary, some of the most tyrannical leaders have demonstrated most clearly their inability to move the nation in the directions that they seemed to want. The political realities, then, will also shape the country's future.

In developing a picture of the nation, every factor which might influence the development of a weapon system can be taken into consideration -- the national economy, domestic and international policy and goals, the location and capacity of production facilities, natural resources located within the nation or available through its allies, technological capabilities and the output of research laboratories, areas which are receiving special attention in research, the power base of the current regime and the likelihood that it will remain stable, the organization and leadership of the armed forces, military policies which have become traditional -- in short, many aspects which, together, form a "model," or rational intellectual picture,

of the nation as a whole.

With a clearly-defined model of the nation, it is possible to develop reasonable estimates of its present and future capabilities in specific areas. For example, since you understand the importance of the five-year plans for Soviet resource development, and since you understand that the Soviets are rather slower than the Americans in disposing of obsolescent equipment, you can make some reasonable estimates of the dates by which a given weapon system will be replaced.

A clearly-defined and correct model of Soviet intentions and capabilities provides a basis for dealing with the uncertainty that is introduced through deception. The deception itself fits into the pattern of overall Soviet strategy, which provides a rationale for the deceptive maneuver. Within obvious limits -- the model itself must be tested against reality -- the use of a comprehensive model provides a defense against deception.

Several sources of uncertainty have thus been seen to enter into the intelligence production process: missing data, errors in the OB, and deception on the part of friends and enemies. It is important to recognize and report these sources of uncertainty. At the same time, a comprehensive understanding of the nation can assist in reducing the degree of uncertainty that sources like these can introduce.

UNIT 4

PROBABILITY ASSESSMENTS

The proper definition of "probability" has been a subject of controversy from the time that probabilistic methods were first introduced in the Seventeenth Century. From a purely subjective point of view, probability might be said to measure the degree of credence that we place in some hypothesis or other proposition. If we believe very strongly in something, then we ascribe a high probability to it -- from the subjective point of view.

Clearly, our subjective probability may be wrong. I may believe very strongly in my chances of winning in a game of poker against a riverboat gambler; but an objective observer would have to say that my chances are much smaller that I think they are. In general, as I look back over my lifetime, I can think of many times in which I believed very strongly in something, only to have it turn out to be false. For this reason, I generally look upon my own strong beliefs, and especially the strong beliefs of other persons, with a good deal of skepticism.

A major contribution of the Seventeenth Century probabilists was an alternative approach to the measurement of probabilities, which is objective in nature rather than subjective. It takes two forms, which we will label "analytic" (or "a priori") and "synthetic" (or "a posteriori").

The analytic approach is based upon the definitions of the objects or entities involved. For example, if we define a "fair die" as one which is equally likely to come up with any one of its six faces showing, then it follows logically and mathematically that the chance of any one face coming up (such as the four) is 1/6. This is a logical consequence of our definition of "fair die." If the chance of the four coming up were anything other than 1/6, then it wouldn't be a fair die.

We can, of course, test any specific die to find out whether it is fair. We can throw it a hundred times, and count how many times it comes up one, two, three, and so on. If, on every one of the hundred throws, it comes up six, then we can say, "It is not very likely that this is a fair die." And we can easily compute the likelihood that it is fair; this probability is $(1/6)^{100} = 1.5306 \times 10^{-78}$, which is a very small number. A die which came up six in every one of one hundred tosses, then, would not be likely to be a fair die.

The other approach to measuring probabilities is the synthetic approach. Instead of beginning with the definitions, it begins with a count of the proportions present in a population. Since this approach has been used by actuaries for determining insurance rates, it can also be called an "actuarial" approach.

Suppose that you have received 10,000 Christmas tree lights for decorating your offices. How many of them are faulty? Without attempting to test the entire lot of them, you decide to test 100 of them, to get some idea of the number of faulty bulbs to expect. Suppose that 10 of the bulbs refuse to light, or burn out immediately. Then your best guess concerning the entire lot of bulbs would be that the same proportion, or 10 percent, would be faulty.

Of course, this example is much too simple, because you would also want to know - to determine how many spare bulbs to order - how likely it is that 15 percent of the total might be faulty, or 20 percent, or some other proportion. A statistician could easily provide a reply to these and many other questions.

Three approaches to the measurement of probability, then, are (1) a subjective approach, measuring our degree of belief; (2) an analytic approach, based on the definitions of the entities involved; and (3) a synthetic approach, based on a statistical investigation of the behavior of similar entities in the past.

All three types of probabilities play a role in determining the uncertainty of intelligence projections, but this manual will concentrate on the first type, "subjective" probabilities, because they are most useful for intelligence estimates; they are also most controversial, because they are difficult to measure, may differ from person to person, and are difficult to evaluate. Where they are available, mathematical and statistical probabili-

ties (the second and third types) should certainly be used in measuring and communicating the uncertainty of intelligence estimates. For example:

- The known resolution accuracy of an aerial camera gives a precise range of error in the estimation of the length of a Soviet missile, photographed from a satellite at a known height. For any photograph, there is statistical distribution of possible lengths of the object photographed. On the basis of this information, you could, for example, determine the probability that two photographs represent missiles of the same length, or missiles of two different lengths. This result would not be based upon a large-scale statistical survey of missile photographs, but upon the characteristics of the equipment involved. It would therefore represent a probability distribution which used the second, or analytic, approach.

- Barracks are under construction near a new Chinese factory. Previous experience, including accurate counts of personnel at 100 other Chinese factories, indicates that the Chinese provide 20 square feet of floor space per person in their barracks. This factor may therefore be used in estimating the number of persons to be employed in the factory. In addition, previous experience has shown a degree of variability in the amount of floor space allotted; knowledge of this variability permits you to set upper

and lower bounds on your estimate of the number of persons to be employed. Use of this approach to the estimation of probabilities would represent the statistical, actuarial, or synthetic approach.

- Reports from a government agency indicate that the Soviets are developing a new type of radar specifically to detect and track U.S. cruise missiles. The agency estimates that 100 such radars will be deployed and operational by 1985. In checking their report, you find a large number of unanswered questions concerning the accuracy of their information and the validity of many of the inferences that they have drawn. You are willing to say only that there is some chance -- say 40 percent -- that the radars will actually be installed by the target date. You do not, of course, base this figure on any large-scale statistical study of radar installations. And you are not using any techniques of mathematical analysis to arrive at a well-defined number. Instead, you are saying that you think that there is some possibility that the installations will be completed, but you feel that there is something less than a fifty-fifty chance that they will be. You are stating, in short, a subjective probability.

Interviews with DIA estimators have indicated that probabilities of this type -- subjective probabilities -- were far more frequently used than probabilities of the other two types. For this reason, we will concentrate

upon subjective probabilities in this report. Because the word "subjective" carries connotations of guesswork, we generally use the term "probability assessments" to refer to them. Probability assessments can be calibrated in such a way as to permit their use in a consistent, well-founded manner.

Many of us are nevertheless hesitant about assigning probabilities to individual events, because it is difficult to determine precisely what is meant by such probabilities. Suppose, for example, that you yourself are playing solitaire with your own deck of cards. You shuffle it several times, cut the deck, and place the top card face-down on the table. What is the probability that the card is the ace of spades? Most of us would say that is $1/52 \approx 0.019$. We have no serious problem in estimating this probability -- which is an "analytic" probability based on the definition of the card deck, and of a random draw from such a deck.

Next, suppose that the card comes from an unfamiliar deck, which belongs to a riverboat gambler. He has shuffled and cut the deck himself. He is wearing a baggy coat that could easily conceal some extra cards. And you stand to lose a substantial amount to him, if you fail to guess correctly. Now, what is the probability that your guess will be correct?

Obviously, the probability in the second example is much more difficult to estimate than the probability in the first example. There are an indefinitely large number of factors which may be relevant to the estimation,

including unknown factors -- such as the possible presence of a conspirator among the onlookers. A really clever opponent will be attempting to find exactly those ruses that you have neglected to identify.

It should be clear that the situation faced by the intelligence estimator is considerably closer to the second example than to the first. Our potential adversaries have absolutely no reason to play a "fair" game, if it is not to their advantage to do so. They may be expected to take advantage of every opportunity for concealment or misrepresentation of their capabilities and intentions.

Because of the large number of elements that can serve to increase or decrease the probability of an intelligence estimate, it is rarely possible to rely on mathematical or statistical probabilites. Instead, the judgment of an experienceed intelligence producer, who can take into account the many factors that may affect the probability of an estimate, must be used.

A subjective probability represents your estimate of chance that a given proposition will be found to be correct. Like other probabilities, it is expressed as a proportion, in the range from 0.0 to 1.0. A subjective probability can be correct or incorrect, depending on the degree to which it is well-calibrated. Calibration is defined in terms of a statistical probability: over a large number of subjective probabilities, if you have assigned a probability of 0.70 to some propositions, then 70 percent of them

should be found to be correct; and similarly for other probabilities. If these proportions hold, then you are said to be well-calibrated; if they do not hold, then you are biased toward conservatism (if you underestimate probabilities) or toward anticonservatism (if you overestimate them).

You are well-calibrated, then, if you do a good job of evaluating the quality of the information that you have, and if you have a realistic sense of your own ability to examine and to integrate this information.

Much of the remainder of this manual is devoted to the development of methods for producing well-founded probability estimates. In this unit, we will suggest two general approaches to the quantification of uncertainty, the holistic and the compositional:

The holistic approach deals with wholes, the organic, inclusive structures of events. In artificial intelligence applications, these wholes are sometimes called frames, scripts, or scenarios. We will use the term "scenario" to refer to the inclusive structure of hypothesized future events, which fit together to form a consistent whole.

Using the holistic approach, the intelligence producer provides a probability for the scenario as a whole; based on this overall estimate, figures for the individual components can be derived. Here is a somewhat artifical example:

It is well known that many of the more hawkish forces within the Soviet Union believe that a large-scale nuclear war can actually be fought and won. On the basis of this belief, they may be expected to emphasize those elements of the Soviet military structure that would make such a war possible. Offensive missiles in concealed, hardened locations might be among the elements of this strategy. An increase in submarine forces might also be considered, with deployments which would permit effective launches of SLBMs against the U.S. continent at a moment's notice. Civil defense equipment would be maintained, and training would help to insure survival of the civilian population during a U.S. retaliatory strike.

A scenario would contain the details of this plan. Prepared by U.S. intelligence personnel, it would begin with the overall approach, and would contain the actions and developments that would be essential in carrying the Soviet plan into action. Since the scenario begins with the overall plan, we call this approach "top-down"; it begins at the top of a plan, and works down to the smaller details.

Probabilities are next assigned to the plan in a top-down fashion. Based on U.S. knowledge of the composition of leadership in the Soviet Union, and upon a general view of Soviet intentions, a probability figure is obtained for the total scenario. Next, probabilities can be assigned to each of the major components of the scenario. For example, <u>if</u> the Soviets are preparing for a major offensive nuclear war, <u>then</u> the probability is very high that

they will develop an effective civil defense structure.

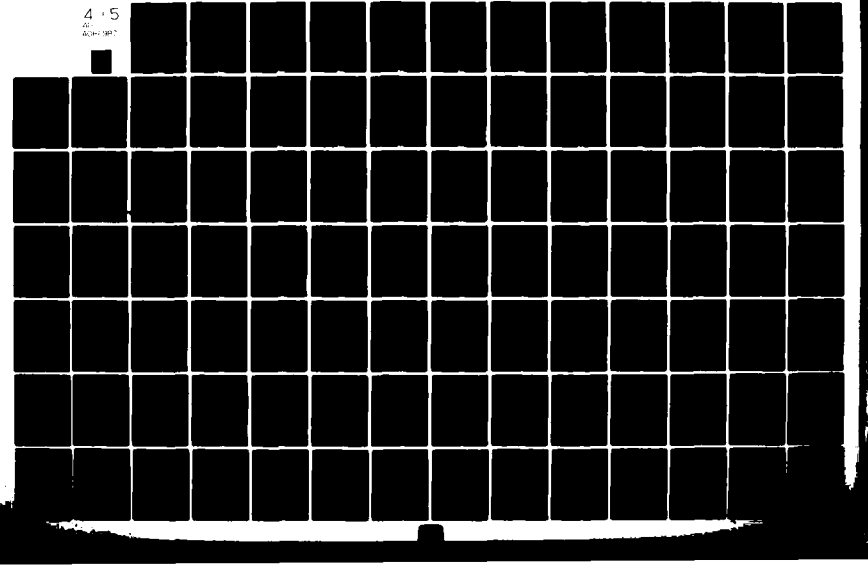
Calculation of the probabilities for the elements of the scenario are straightforward. For example, suppose that the probability of an overall Soviet plan for aggressive nuclear war during the next five years is 0.35. Suppose that, if such a plan were implemented, then an increase in civil defense allocations carries a probability of 0.90. We may now calculate the unconditional probability of an increase in civil defense as 0.35 x 0.90 = 0.315.

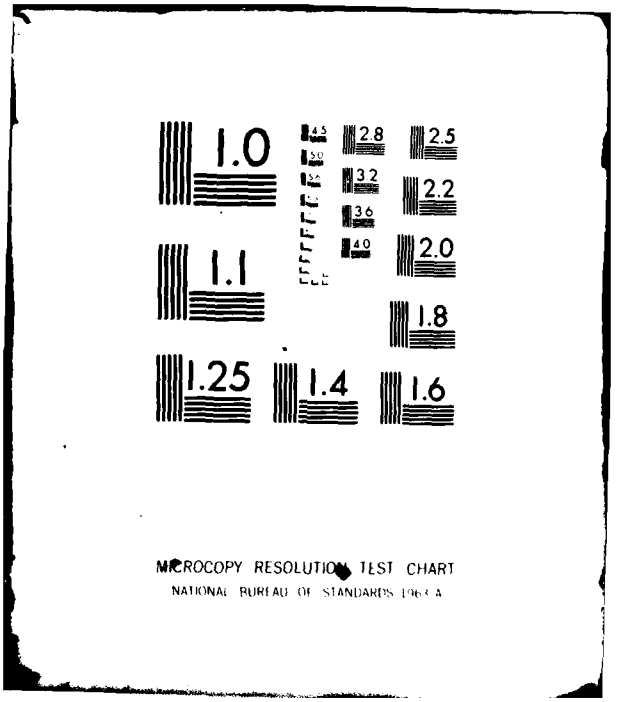
Of course, we may know from other sources that civil defense is being emphasized in the Soviet Union. This means that the actual probability that we attach to this development is greater than 0.315. In its pure form, however, the top-down, holistic approach derives these probabilites only from the probabilities attached to the top-level scenarios, and the conditional (if-then) probabilities that are included in the scenarios.

The compositional approach begins with individual events. It could be called a "bottom-up" approach, because it begins with the low-level individual developments and works upward to the most general ones.

Probabilities are assessed for specific events (such as a Soviet decision to develop a cruise missile), and a probability range is assessed for a quantitative projection for a single weapon system. These probabilities are then combined to obtain higher-level probabilities -- obtaining, for

AD-A086 987 PATTERN ANALYSIS AND RECOGNITION CORP ROME NY F/G 15/4
AGGREGATING AND COMMUNICATING UNCERTAINTY.(U)
APR 80 J M MORRIS, R J D'AMORE F30602-78-C-0291
UNCLASSIFIED PAR-79-1 RADC-TR-80-113 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

example, a probability distribution for all offensive missiles, then a distribution for all missiles combined, and finally a probability distribution indicating the total combined strength of all Soviet military resources.

The compositional approach is often used in decision analysis; because it frequently relies on Bayes' theorem, it is sometimes called "Bayesian analysis." This approach has been extensively studied, and it is supported by several computer-based systems.

These two approaches, then, the top-down and the bottom-up, give us a pair of methods for aggregating uncertainties -- for going in a consistent way from the probability of one event or development to the probabilities of related developments.

UNIT 5

COMBINING PROBABILITIES

Every report that you receive is somewhat uncertain. No source of information is totally free of errors. Nevertheless, by combining reports from several different sources, you can often obtain a composite picture that is more certain than any of the reports that make it up.

This unit provides a brief overview of methods for combining probabilities. In keeping with the non-mathematical tone of this manual, there is no attempt to provide a full introduction to statistical methods; Howard Raiffa's Decision Analysis: Introductory Lectures on Choices under Uncertainty (Reading: Addison-Wesley, 1970) is a particularly good textbook in this field, and the mathematics are elementary. Here, we will use mathematics only when it is impossible to avoid them.

The problem that we want to solve is this: How can you determine precisely the degree of uncertainty to be attached to an estimate or projection, when the estimate is based on several sources of information, each of which is somewhat doubtful? For example, we know that information obtained from defectors and prisoners is likely to be self-serving and therefore faulty. However, if several prisoners independently agree on a report, we are much

more likely to believe their combined story than we would be to believe any one of them taken separately.

In the same way, if several radar reports, photographs, and infrared sensors independently agree in identifying a group of missiles under test at one location, you would attach a higher degree of certainty to the identification than you would to any one fallible report standing alone.

The problem, then, is the most effective, statistically correct method for combining measures of uncertainty attached to various reports, in order to get an aggregate measurement of uncertainty.

Before continuing, we should emphasize three things:

- Statistical methods are designed to assist humans in the production of intelligence estimates, not to replace them. There is no substitute for the human being who has a broad understanding of the goals and structure of a foreign nation, of the technology required to support its goals, and of its ability to achieve them. In practical terms, this means that your own good sense would be used to correct the results of a statistical analysis -- not because there is something wrong with statistics, but because the statistical model may not be taking all factors into consideration.

- In all cases, we are assuming that reports are "fallible" or "somewhat uncertain." Obviously, in this imperfect world, no one is ever infallible. At the same time, we do have a substantial body of confirmed knowledge concerning Soviet military capabilities (and, to a lesser extent, those of the PRC and other nations of interest). It would be a waste of your time to attempt to develop elaborate statistical procedures for handling the uncertainty of this body of knowledge. It would be wasteful to assign, say, a 95 percent probability to the fact that the USSR now has precisely 2 Moskva-class helicopter cruisers, if you have a substantial body of evidence to support this figure, and no reason whatever to doubt it. The "95 percent" figure will simply require additional work for you in subsequent computation, and cause needless worry for the consumer who takes it seriously. If this is a substantially documented piece of information, then there is no need to treat it as uncertain at all; you can simply take it as a fact.

- The need for "independence" among the reports is based on a desire to keep the statistics simple. In practical terms, it means that reports have come through distinct channels; for example, a defector's report is confirmed by aerial photography. Obviously, not all reports are independent. For example, a story concerning Soviet aircraft may appear in both Pravda and Izvetsia. Since

both papers represent the views of the Soviet government, their articles do not tend to confirm one another; they are simply two versions of a single report. Thus, if the probability of a story in Pravda is 20 percent, the appearance of a corresponding story in Izvetsia does not increase this probability.

Suppose now that two reports of an event, such as a successful test of a new anti-satellite weapon, are received. Suppose also that these reports are independent. If we know the probability that each of these reports separately is correct, what is the likelihood that they are correct when taken together; that is, when they confirm one another?

A very simple probabilistic model for this problem might be the following. We call the first source of reports A, and the second source B. Suppose that 70 percent of the reports produced by A are correct, and 80 percent of those from B are correct. Since the two reports confirm one another, either both are correct, or both are wrong. What are the probabilities, then, that (a) both are correct, and (b) both are wrong?

Out of the mass of reports produced by A and B, we select at random one report from each of them. Our chances for each combination of correct and incorrect reports would then be:

P(A correct and B correct) = 0.70 x 0.80 = 0.56

P(A correct and B wrong) = 0.70 x 0.20 = 0.14

P(A wrong and B wrong) = 0.30 x 0.80 = 0.24

P(A wrong and B wrong) = 0.30 x 0.20 = 0.06

Next we calculate the probability that both A and B are right, on the assumption that they both agree:

P(A correct and B correct, given that they agree)

= 0.56 / (0.56 + 0.06) = 0.903

Thus, there are something better than 9 out of 10 chances that they are correct, given that they both agree in truth-value.

Although this probability model is very simple, notice that it tells us a good many things that our intuitions say are correct. For example, suppose that we have two independent reports which are such that (a) they tend to confirm each other, and (b) they are both fairly reliable. Then the probability that they are correct is higher than it would be for either of them taken separately.

In general, the more independent reports that you have which tend to confirm one another, the more certain you can be of their truth.

In the same way, notice how <u>disconfirmation</u> lowers the probability of a report. What is the likelihood that A is true, given that we have received report B which disconfirms it? Using the same set of numbers as before:

P(A correct and B wrong, given that they disagree)

= 0.14 / (0.14 + 0.24) = 0.368

The probability that A is correct, then, drops from 0.70 to 0.368, if a disconfirming report B is received, which has a probability of 0.80.

Similarly, suppose that we have received report B first. What happens when we receive report A, which disconfirms it?

P(B wrong and A correct, given that they disagree)

= 0.24 / (0.24 + 0.14) = 0.632

The probability that B is correct has dropped, then, from 0.80 to 0.632, when the disconfirming report A is received. While the probability of B has dropped, it is still likely, since its initial probability (0.80) was quite high. If additional confirming evidence were received, it would tend to outweigh the shakier disconfirming report A.

Finally, suppose that we have received both A and B, and that both agree upon their story. What is the probability that they are both wrong?

P(A wrong and B wrong, given that they agree)

= 0.06 / (0.06 + 0.56) = 0.097

Thus there is some possibility that both reports are wrong, even though they confirm one another. This probability, 9.7 percent, may be too high to

permit a critical decision to be made which assumes that they are correct; the possibility of error is too great. Additional information may be needed.

This very simple probability model, then, gives you some sense of the way in which uncertainties in various reports can be combined to obtain an aggregate measure of uncertainty.

Unfortunately, this simple probability model is much too simple for practical use in aggregating the credibility of reports. It supposes that each source produces a large number of independent reports, as a bottle factory might produce bottles. Then it supposes that some of these reports will randomly be found to be faulty. In point of fact, some events are far more likely to be reported than other events. Thus some reports are more plausible than other reports.

When we say that one report is more "plausible," we mean that it is more likely to be true -- based on our general knowledge of the context in which it appears.

Suppose that: (1) Soviet aircraft X is a rickety, ancient plane that is constantly in need of repair, and (2) aircraft Y is a sleek, new machine that performs effectively and reliably. Suppose that we receive two reports: (A) that X is being mothballed, and (b) that Y is being mothballed. Which report, A or B, is more likely to be correct? Obviously, assuming some degree of rationality on the part of the Soviets, A is more likely to be correct than B; we say that A is more plausible than B.

Two factors enter into the evaluation of the uncertainty of a report: (1) the record of reliability of the source, and (2) the initial plausibility of the event which it reports. We call this initial plausibility the prior probability (or a priori probability). It represents your best guess concerning the likelihood of the event before you have received the report.

Bayesian methods are intended specifically for dealing with situations like this. They represent a straightforward development of Bayes' theorem, which combines the initial plausibility measure with each piece of additional information to obtain the probability of a projected event.

The following information is required:

- The initial probability, before any new information is received, that you assign to the event.
- The probability of receiving a report of this type from this source.
- The probability, given the occurrence of the event, that a report concerning it from this source would be received.

Bayes' theorem produces, on the basis of this information, the probability that the event occurred. In its simplest form, it says:

$$P(E/R) = P(E) \times P(R/E) / P(R),$$

where P(E/R) is the probability that the event occurred, given that you have received the report; P(E) is the initial plausibility or prior probability of the event; P(R/E) is the probability that the report would have been produced, if the event occurred; and P(R) is the probability of reports from this source.

Notice that the theorem can be applied repeatedly. Each time, as a new report is considered, the plausibility of an event increases or decreases. In this way, on the basis of several reports, it becomes possible to assign a measure of the uncertainty of the event reported.

Bayes' theorem is particularly effective in applications in which decisions are structured and repetitive. Consider for example the identification of aircraft in a battlefield situation, in which there are (1) several different radars and other sensors, which can be used for identifying aircraft, (2) a definite mix of friendly and hostile aircraft for identification, and (3) a well-defined logical structure for the combination of reports from several sensors in such a way as to define the probability of specific aircraft.

Well-defined, repetitive tasks like this occur repeatedly in tactical situations. You may find that Bayesian methods are also useful in strategic intelligence applications, particularly where the tasks are highly structured and repetitive. They are probably most useful when a computer system is available to perform the routine computational tasks.

Some of the problems of the Bayesian approach, which make it difficult to use without computer support, are these:

- You will have to determine the prior probabilities of all the events under study. These represent the initial plausibilities of the events, before current information is used. This task can become quite onerous when probabilities for a large number of events must be assessed, or where the events are nonroutine. We can estimate, on the basis of past experience, the probability that Soviets will launch a new satellite next month, or that they will initiate troop maneuvers; but the probability that they will withdraw troops from Cuba - a nonroutine event - is much more difficult to assess.

- Determining the prior probabilities as inputs into the Bayesian system may be a time-consuming task. You may feel that they are too unreliable - that they contain too much guesswork - to make it worthwhile to subject them to Bayesian processing. Under these conditions, you may prefer to assess the final probabilities directly, without using a Bayesian approach.

- In any case, preparing probability assessments is a difficult task, if the assessments are to be of value to the consumers. In other areas, such as tactical intelligence, experiments have shown that intelligence analysis sometimes guess at the probabilities, give the same "average" rating to almost all information, or omit the probability assessments entirely. This suggests that strategic intelligence producers will also find it difficult to assign detailed prior probabilities.

The study of Bayesian methods, in spite of these caveats, is nevertheless useful in estimative intelligence production, for several reasons:

- It gives a good sense of the meaning of probabilistic information, and how such information functions in decisions.

- It helps to show how probabilities can be properly combined.

- In particular, it emphasizes the importance of the prior probabilities - the essential background information that must be used in assessing the plausibility of a report.

Work with statistical methods will, in short, give you a much better sense of the way in which uncertainty affects and afflicts all knowledge.

UNIT 6

HINDSIGHT AND SECOND-GUESSING

Subjective probabilities are most frequently used for reporting the uncertainty of an estimate. These represent your assessment of the likelihood that the projected event will occur, or that the numbers will lie within a given range.

These probabilities will, of course, be as accurate as you can make them, based on your assessment of all the factors that entered into your estimate. They are "subjective" or "intuitive" in the sense that each estimate is based on a different collection of information which, in your opinion, is relevant to the problem. Your sense of the overall policies and capabilities of the nation, as well as your detailed knowledge of specific weapon systems, enters into your assessment of the probability of an estimate.

Because this assessment is subjective, it is subject to the biases and distortions that enter into all human judgments. Experiments have shown that more experienced personnel are less subject to bias than are inexperienced persons. Specifically, experienced weather forecasters, working in the areas of their own expertise, did not show the same biases that were shown by college students in psychology courses, who are most frequently used in experiments of this type. It will be helpful, nevertheless, to see what some of the common biases are.

In this unit, we review several kinds of bias that have been found in assessments of probabilities. In addition, we explore some of the interesting side effects of these studies. We will find, for example, that some of the critics of the intelligence community have been more guilty of one type of bias -- hindsight bias -- than the intelligence producers themselves. But our main goal will be to learn something about the nature of bias in subjective judgments of uncertainty.

Humans generally do a poor job of assessing their own uncertainty. They are far more confident about their judgments than the evidence would warrant. At other times, they may be inclined to hedge -- to overstate their uncertainty in an effort to avoid the penalties for error.

We might expect that if people knew something about the common biases that may occur, they might be able to correct them. Experimentation has shown, however, that this last expectation remains a rather forlorn hope; there is no clear evidence that people can correct their errors, even when they know that errors occur. What is needed is a better understanding of the estimative process itself; if estimators are skilled at developing estimates, then they also tend to be better judges of the uncertainty in their estimates. This study of bias is therefore intended to help you understand the estimative process better -- in the expectation that this understanding will help you to assess your own uncertainty a little more accurately.

The first type of bias, hindsight bias or second-guessing, is particularly interesting, because it represents an error that frequently

occurs in outside evaluations of estimative intelligence. It is a bias that is based on the "prediction" of events that have already occurred.

If someone were to ask, "How likely did it seem to you in 1977 that the Shah of Iran would be overthrown?" You might think back to the evidence that was available then: massive student protests among Iranian students in this country, the presence of secret police and the other paraphernalia of dictatorship, and perhaps other signs of a fragile, rigid regime. With this information, you certainly should have seen that the Shah was about to crumble. If you also had the special sources of information that were available to intelligence officers in 1977, then you certainly should have predicted the rebellion that was then imminent. Why, then, was it not predicted?

The answer to this question is complex, and it lies at the heart of the problem of estimative intelligence:

- There is too much information. Intelligence producers are overwhelmed with data that far exceed their capacity to ingest and digest them.

- The information that you have is not structured around the specific events that are to be predicted. While there may have been enough information to have enabled you to predict the overthrow of the Shah, for example, you did not have it neatly filed in a drawer marked "Evidence of forthcoming rebellion in Iran."

- Even if the proper information had been accumulated in time, there were too many chance factors that might have intervened. For example, what would have happened to the rebellion if the aging Khomeni had become violently ill at the critical moment? What if the Shah had been more conciliatory? And so on, through an indefinite number of variables.

- Over the long run, intelligence producers try to avoid the "cry wolf" response that comes when they have predicted dire events too often -- when the wolf failed to appear. Simply to maintain their credibility, they have to avoid premature and unnecessary warnings.

Hindsight bias ignores these problems. It is the claim that someone could have, and should have, predicted the future far more accurately than he did. Of the intelligence community, for example, it says:

- The Pearl Harbor attack should have been predicted more accurately and effectively, given the vast amount of information in U.S. hands concerning Japanese plans and naval maneuvers, including information that was made available after the Japanese codes were broken.

- The Soviet invasion of Czechoslovakia should have been predicted. Specific messages were available which could have revealed Soviet plans. In addition, careful tracking of the levels of tension

between the two nations would have shown increasing hostility, reaching the ignition point at the time of the invasion.

- The overthrow of the Shah of Iran should have been predicted. It was generally assumed that the regime was stable, and that the increasing Westernization of the country had broad support. A better understanding of the nation's attitudes, together with specific information concerning anti-Shah movements, would have shown that this assumption was not warranted.

In general, the reasoning is this: "As we look at the past, we see that many events could have been predicted on the basis of available information. We therefore can attain a relatively high degree of certainty concerning future events. Therefore, predictions of the future should be given a high degree of certainty."

Hindsight bias reflects a simplistic view of the past. Since we can put together a reasonable scenario that would permit the prediction, say, of the overthrow of the Shah, this does not mean that we will be able to put together a scenario for the prediction of other anomalous events in the future. ("Anomalous" is used here in the sense that Thomas Kuhn uses it in The Structure of Scientific Revolutions: a violation of the laws that make up the currently accepted world-picture. The world-picture, for Americans in 1977, saw the Shah as an established, stable ruler. His overthrow violated this picture, and was thus anomalous.)

This simplistic view of the past has been called "creeping determinism." It says that there is enough information available about future events to permit us to predict them. Thus, the only problem, according to this view, would be to get more information to the intelligence producer, and to get him to do his job better. If the deterministic view is correct, then he can predict future events.

This point of view produces two types of problems for the intelligence producer:

- First, since it is a plausible viewpoint, it increases the level of criticism of the intelligence community. It leads to demands for performance which cannot be met. The most serious result, then, is a loss of confidence in the work of the intelligence community, and a tendency to rely on other sources of information -- such as, hunches, guesswork, and the popular press.

- Second, experimental evidence has shown that people tend to overestimate their own knowledge of the future. When asked to assess the probability of future events, they frequently assign too high a probability. Moreover, when the event which they predicted has occurred, they claim to have been more certain than they actually were. As they look back on the past, it seems to them that they had been able to foresee the future, with more assurance than they claimed at the time.

In practical terms, it is important to remember that the demand for more accurate predictions of future events can lead to over-assessments of the probability to be attached to your estimates. In addition, this type of criticism tends to focus on anomalous events, which are precisely those events which are most difficult to predict. (If we take "anomalous" seriously, they are the events which cannot be predicted at all.) The goal of strategic intelligence is the identification of the intentions and capabilities of the nation under study, not the prediction of the unpredictable.

Even without the presence of hindsight bias, however, people tend to be overconfident about the extent of their own knowledge. A person is overconfident when he or she gives too high an estimate of the probability of a projected future event. In quantitative estimates, overconfidence is reflected in too narrow a spread between the High and the Low estimates.

Research has shown a strong, consistent tendency toward overconfidence, both among subject-matter experts and among novices, in field situations as well as in the laboratory:

- Studies of Las Vegas casino patrons showed irrational preferences for certain bets.

- Studies of bankers and stock market experts in the prediction of closing prices for stocks showed exaggerated confidence in the accuracy of the predictions.

- Studies of military intelligence officers predicting a coup d'etat in a designated country, the shooting down of a reconnaissance plane, or an arms shipment from one country to another showed overconfidence in their predictions.

"Overconfidence" means that the probabilities that they assigned were too high. Over the long run, when they assigned a 70 percent probability to a series of estimates, then 70 percent of those estimates should have been correct. In fact, the actual percentage of correct estimates was found to be consistently lower than the percentages which these experts assessed. This was true of experts in such widely scattered fields as gambling, the stock market, and military intelligence.

There is no easy cure for overconfidence. All of these experts obviously had a good deal at stake in the accuracy of their assessments; their overconfidence was costing them money or prestige. Apparently, they were unaware of their bias. Thus, the first step toward a cure is to become aware of the accuracy of your past estimates. A continuing review of your past work and that of other people working on similar estimates will help to show where there are areas of overconfidence and underconfidence. Such a review will also have a beneficial side effect -- it will help to locate precisely those factors that contribute to errors in the estimates themselves.

In addition to problems with overconfidence, there also seems to be some pressure toward underconfidence -- that is, toward excessive hedging of

estimates. The motivation for underassessing the probability of your result is not difficult to locate. Suppose that, instead of saying, "There is an 80 percent chance that ...," you say, "There is a 60 percent chance that..." This will (a) give you some credit if you're right, and (b) reduce the penalty if you're wrong. Similarly, you could increase the range from Low to High, to help insure that the actual figure lies within the range that you have projected.

More generally, there is a pressure toward conservatism in the production of intelligence estimates, which is likely to counteract the general human tendency toward overconfidence. On the one hand, the go-for-broke strategy encourages overconfidence: you produce wildly daring estimates in the hope that one of them will be right, winning you glory and renown throughout the grateful nation. On the other hand, the don't-cry-wolf strategy encourages conservatism: if you repeatedly offer exaggerated assessments, then your credibility drops off sharply. Since these two tendencies can lead to serious errors in probability assessments, it would be a good idea to define them more clearly.

According to the go-for-broke strategy, there will be little opportunity for personal recognition for the person who produces dull, routine estimates. Even if these estimates are mostly correct, the consumers are scarcely likely to notice them. To gain that kind of notice, you have to go for broke -- that is, you have to indicate a high degree of confidence in a very striking, unexpected event. If you win, then you have gained the kind of recognition that you were looking for. What happens if you lose? Generally, the person

that tries this strategy doesn't have all that much to lose. It is a strategy for the person who likes to gamble for high stakes.

The don't-cry-wolf approach uses the opposite strategy. It says that you can gain credibility only through a chain of successes, and this means that you never speculate on the long shots. In practical terms, it means that probability assessments are kept low, and that the spreads from Low to High are kept wide. In this way, you are likely to be correct most often.

Both of these approaches place the emphasis in the wrong place. Instead of asking, What assessment will be of greatest value to the consumer? they are asking, What will most benefit the producer? But even here, these strategies are wrong. Clearly, in the long run, both producer and consumer are benefitted most by assessments which correctly indicate the degree of uncertainty that is present in an estimate. Such an assessment maximizes the amount of information transmitted to the consumer. In addition, accurate assessments of uncertainty will also maximize the credibility of the intelligence producer.

UNIT 7

FURTHER ERRORS IN ASSESSING AND COMBINING PROBABILITIES

What specific techniques do humans use in dealing with uncertain information? In some imaginary world populated entirely by statistical geniuses, they would proceed like this:

- Define an experimental hypothesis for testing.
- Define the experimental population for which the hypothesis is to be tested.
- Employing standard statistical sampling techniques, obtain a representative sample of sufficient size and proper composition to achieve the required level of confidence.
- Under well-defined conditions, perform a controlled experiment as required to test and validate the hypothesis.
- An so on, through the sequence of techniques developed by the experimental sciences.

An appropriate experimental design, following something like this sequence, should certainly be used in those situations in which available time and available information make it possible (and where the cost of the tests

does not exceed the expected value of the result). Unfortunately, estimative intelligence -- and real life -- rarely finds it possible to carry out the full sequence of experimental tests. Intelligence estimates face two major constraints:

- To be effective, estimates must be available for a decision within strict time limits, which may not permit a review of all available data.

- The nature of intelligence data collection is such that important pieces of information may never become available. Other pieces of information may be misleading or false.

Strategic intelligence thus represents, in somewhat exaggerated form, the situation that we all face in real life, where we never have enough time to investigate fully, and where much of the information that we must use is no more than rumor, hearsay, and fraud.

Humans have been found to use several shortcuts, or heuristics, in dealing with uncertainties in everyday decisions. These heuristics have a major virtue: they are fast and efficient. They also have a major vice: they are prone to errors.

Three frequently-used heuristics have been identified:

- Judgment by representativeness: a small sample is taken as representative of a large population. We judge the characteristics of a whole group on the basis of acquaintance with just a few of its members.

- Judgment by availability: an event is judged to be likely, if it is easy for us to imagine similar events. If, in our imagination, we can say, "That's just the sort of thing that would happen," then we tend to overestimate the probability that it will happen.

- Judgment by adjustment: when judging the numbers or sizes of things, we begin with a known value and adjust it upward or downward to obtain an estimate of the unknown value. In the process, we often fail to make a large enough adjustment.

Heuristics like these are likely to play a role in estimative intelligence, where the stringent time requirements and the lack of reliable data make it necessary to use short-cut techniques. In more detail, the role of these heuristics is this:

Representativeness

Judgment by representativeness will occur when it becomes necessary to make judgments concerning a total population on the basis of a small or nonrepresentative sample.

Researchers have identified a "law of small numbers," which is a fallacious rule by which people tend to make judgments on the basis of a very small number of samples. For example, if we hear that two Ford Mustangs have had frequent brake failures, we are likely to generalize to the conclusion that all Mustangs are subject to brake failures. But our sample size is much too small to make this sweeping a generalization.

Because information available to intelligence producers can be limited to very small samples, it is important to recognize the high probability of error in generalizing to larger populations. For example, if you could obtain information only about the destroyer Bedovy, you might be tempted to generalize that all four Kildin class destroyers carry 45 mm guns (with some high probability), when, in fact, the Bedovy is the only one that does so. Since information concerning Soviet naval vessels is very complete, estimators are not at all likely to make this particular error; but similar errors are possible wherever the information is skimpy and the time is short.

Availability

Judgment by availability is the tendency to use the information which is most easily available, but which may not adequately represent the population from which it is drawn.

Perhaps the best example of this fallacy was the poll undertaken by the Literary Digest magazine to determine the outcome of the 1936 American Presidential election. The poll indicated an overwhelming victory for

Alfred M. Landon, the Republican nominee, over his opponent, Franklin D. Roosevelt. The magazine's prediction was, of course, badly mistaken. Its gross error was due to its use of a telephone survey together with a poll of its readers to obtain its results, at a time when only the affluent could afford a private telephone. Since nontelephone households included the majority of voters, and since the vast majority of these voters favored Roosevelt, the poll gave grossly misleading results. The magazine relied on data which were easily available, rather than making the greater effort required to obtain data which were representative of the population from which they were drawn.

Psychological studies have indicated that people use this heuristic to shade their judgments upward or downward, depending on the ease with which they can recall similar objects or events. Such factors as familiarity, recency, and emotional saliency have been identified as affecting recall. Applying these results to estimative intelligence, we could expect that the following factors would affect judgments of uncertainty:

- Familiarity. If the estimator is familiar with a particular weapon system, capability, or other entity, he will be likely to overestimate its numbers, retention time, or other factors, in comparison with another system with which he is less familiar.

- Recency. A recent report, article, or briefing on a given Soviet weapon will tend to increase the importance of that weapon in the mind of the estimator. As a result, he is likely to overestimate

the probabilities connected with that weapon, in comparison with other weapons, which may be equally important but which have been reviewed less recently.

- o Emotional Saliency. We are certainly likely to respond more readily to the more glamorous and more sophisticated weapons than we are to the dull, unglamorous ones. As a result, the estimator is more likely to overestimate the probability that the more glamorous systems will be developed and deployed, ignoring the factors that would encourage development of the others.

Proper experimental design, then, requires that you take care to include data which are less "available" in the sense described here -- to include information concerning unfamiliar systems, older systems that you may have forgotten, and less-glamorous systems that you may have overlooked.

Adjustment

Judgment by adjustment is a heuristic in which you begin with an existing estimate, and raise or lower it in response to new information. This process, called "anchoring and adjustment," is frequently insufficient to account for the new data.

This heuristic may have been partially responsible for underestimates of Soviet ICBM installations during the late 1960's. As Soviet policy changed in such a way as to dictate rapid expansion of ICBM facilities, U.S. estimates

remained "anchored" to past estimates, and were not adjusted rapidly enough to take new Soviet policies into account. The result was a series of underestimates. (The underestimates of Soviet ICBMs have been widely publicized and discussed; this is obviously an oversimplification of the reasons for them.)

The use of the heuristic assumes the existence of a base rate, or commonly accepted level of development, production, deployment, and retirement. Beginning with this base rate, the estimator makes adjustments upward or downward to take account of:

- Current political factors
- Shortages or surpluses of materials
- Difficulties or breakthroughs in production
- Changing economic conditions
- Responses to U.S. and other countermeasures
- Problems in training personnel
- Mechanical and other technological difficulties
- Availability of new technology
- Conservatism of Soviet policy

And any other factors that could influence the qualitative and quantitative projections that are required.

If the experimental evidence can be applied to intelligence estimates, it tells us that these adjustments will not be sufficient; human beings

tend to be conservative in their use of the heuristic, retaining a bias in the direction of earlier estimates.

An alternative approach, then, would be the use of "zero-base" projections. Rather than beginning with existing estimates, the analyst would construct a new estimate entirely from scratch. Past projections would be ignored, and previous trends would not be used. Current information concerning foreign weapon systems would be used, to which information concerning production and deployment rates would add appropriate numbers. Retirement rates could then be estimated, and the resulting figure would provide the final projection.

The essence of the zero-base approach would be its lack of assumptions; nothing would be taken for granted, and every projection would have to be justified. The zero-base approach differs from the anchoring-and-adjustment approach, which required justification only for changes from existing projections.

The zero-base approach is not recommended here, primarily because there is no reason to suppose that it would produce improved projections. It is included simply to show the way in which anchoring-and-adjustment works, and to suggest a means for avoiding the bias that anchoring-and-adjustment introduces. Essentially, it says: Look carefully at the assumptions that enter into projections, and make sure that these assumptions can be justified.

Another bias which has been widely studied seems almost the opposite of the anchoring-and-adjustment heuristic. While anchoring-and-adjustment is conservative, this is an anti-conservative bias, since it is the tendency to neglect prior information.

"Prior information" is represented by the prior probabilities discussed earlier in connection with Bayesian methods. It refers to the "base rate," or the general information that we have. In weather prediction, this would be the climatic information that we have. We may know, for example, that on any given day in August, the probability of a snowstorm in Arlington, Virginia, is 0.001. If we then received information concerning barometric pressure and wind direction that could indicate a snowstorm, we would nevertheless be very hesitant about predicting one. We would be hesitant, because our prior knowledge makes such a storm very unlikely. Bayesian statistical methods make it possible to state this probability precisely.

Unfortunately, human beings are not always hesitant enough in similar sorts of prediction, according to several experimental studies. Specific information takes precedence over the general information that we have.

Another all-pervasive source of bias is the tendency of human beings to try to make sense out of the information that they have. They look for causal relationships. ". . . People predict and explain events by invoking their intuitive theories about underlying causal factors. . . . In making predictions, people rely on information perceived to have a causal relation

to the criterion while disregarding valid but noncausal information." (Icak Ajzen, "Intuitive Theories of Events and the Effects of Base-Rate Information on Prediction.")

In other words, we tend to regard an event as more probable when we can find causal relationships between the event and the data that we have. But this approach is fallacious, because it ignores the underlying probabilities of the event which it predicts.

For example, suppose that massive construction is observed in a Soviet shipyard. It could be hypothesized that the vessel under construction is an aircraft carrier. But this hypothesis would have to be tested not only against the question (1) Is this construction appropriate for the building of an aircraft carrier? but also against the question (2) What is the probability that the Soviets feel the need for a balanced fleet capable of world-wide operations, as represented by their Kuril-class aircraft carriers?

In other words, correcting for this source of bias would involve an investigation of <u>all</u> the factors that might influence the probability of an estimate, and not merely those which have a causal relationship with it.

This source of bias, like the other sources of error that have been discussed in the last two units, is intended to suggest an approach to the detection of potential biases in your own estimates. Some of the errors that have been observed in experimental situations tend to cancel each other out. We have noticed that:

- Overestimation in some situations may be balanced by underestimation in others.

- An overemphasis on the base rate, or on prior probabilities, can be balanced against a tendency to ignore prior knowledge.

These contrasting sources of bias tend to suggest that there is no simple advice that can be given to the intelligence producer who wants to eliminate bias. Instead, your goal should be to review past estimates against later information to determine where your errors, if any, are most likely to occur.

This review of your past work should not be confused with a test that you might take to illustrate the meaning of various types of biases. Experimental evidence has not shown that the biases that appear in laboratory tests are similar to the biases that might occur in actual production of estimates in your area of specialization.

This continuing review of your past work, together with an awareness that biases appear in <u>all</u> assessments of probabilities, will assist in located and reducing the errors in your probability assessments.

UNIT 8

SCORING RULES

". . . as we move further into the age of scientific achievement, the complicated machines and scientific detection devices require the greatest sophistication on the part of the operators and analysts. Without this our scientifically produced information as well as that furnished by the tools of espionage would be of little use. For it is the patient analyst who arranges, ponders, tries out alternate hypotheses and draws conclusions. What he is bringing to the task is the substantive background, the imagination and originality of the sound and careful scholar." (Allen Dulles, The Craft of Intelligence.)

In this manual, we have emphasized the role of subjective probabilities, based on "intuition" -- a comprehensive understanding of the nation, its goals, and its capabilities. This approach reflects Allen Dulles' view of the intelligence producer as a "sound and careful scholar."

We have also emphasized the view that this approach is not to be confused with irresponsible guessing. On the contrary, the characteristic that links the intelligence producer most closely to the scholar and the scientist is the requirement that they must justify their conclusions. They must be able to make their evidence explicit, showing that it supports their claims.

This requirement holds not only for the intelligence products that you produce, such as estimates, but also for the assessments of uncertainty that

are attached to them. When you say that an estimate holds with "70 percent probability," this assessment should be something that you can defend and justify. It is not a wild guess, plucked out of the air.

In this unit we will describe scoring rules, which are used for rating the quality of probability assessments. First, however, we should review the meaning of these assessments. Before we can rate them for quality, we ought to know what it is that we are rating.

The notion of a subjective probability can be traced back to the work of the brilliant young philosopher-mathematician F. P. Ramsey, who in 1926-28 worked out a logic of "partial belief." He noted that our beliefs are generally not all-or-nothing affairs, and that we frequently believe things partially -- we believe them in some sense, but we have doubts about them.

Since the time of Plato, philosophers had scorned the kind of knowledge that was less than completely certain. Then felt that a genuine science, like mathematics, had to be based upon absolute, unchanging truths. Unfortunately, this traditional view could not be applied to the real world in which we live, in which our information is always partial, and in which we can never really be sure about anything -- at least not in the ultimate, unswerving sense that Plato demanded. What was needed was some effective way of dealing with our partial knowledge, consisting of information about which we could have some doubts.

Among Ramsey's many contributions to this problem was an effective way of measuring partial beliefs, which are expressed in terms of the subjective probabilities that we have been discussing in this manual. Ramsey's methods are essentially an operational definition, describing what we mean by a partial belief or a subjective probability.

Suppose that someone claims that Jerry Ford will be elected President in 1980. On questioning, we find that he is fairly sure of this opinion; but he recognizes, like any sane person, that he could be wrong. Our job is to find out how strong his conviction really is.

We therefore offer him a choice between bets. Call these bets A and B. If he chooses A, then he will receive $10 on November 15, 1980, if Ford wins the Presidential election, and nothing otherwise. If he chooses B, then we will flip a coin, freshly obtained from the bank, on November 15, 1980, and will give him $10 if it comes up heads, and nothing otherwise. Which of these two bets will he choose, A or B?

Clearly, since he stands to gain $10 from one or the other of the bets, and will lose nothing, he ought to choose one or the other of them. If he chooses B, then he has a fifty-fifty chance of winning the money. He should therefore choose A if he thinks that Ford's chances of winning are better than fifty-fifty, and he should choose B if he thinks that they are worse.

We can continue this process, by offering him additional bets (perhaps using dice or a roulette wheel rather than a coin), in which the odds are

60-40, 70-30, and so on. At some point, he should reach a state of indifference, in which he doesn't know which of the two bets he prefers. Suppose that he can't choose between bet A and a bet in which the odds are 70-30. Then we can say that his degree of belief in Ford's election is 70 percent. This figure is the subjective probability that he ascribes to it.

Ramsey and his successors developed and refined this process, with the goal of obtaining much more accurate discriminations among various degrees of belief, but the basic idea is clear: we can obtain an objective measure of a person's beliefs, assigning a definite number to them, by comparing them with the bets that he is willing to make.

There are, of course, objections to the specific method. Some people hate to take chances, especially when large sums of money are involved; they are sometimes called "risk-averse" people. Other people may make mistakes in evaluating the bets correctly, especially when they are complicated or obscure. Some people may have moral or esthetic objections to gambling, while others may enjoy the thrill of betting on a long shot.

But none of the objections really undermines Ramsey's basic point: that there is a logic of partial belief, which we can treat in a rigorous, mathematical way. This point constitutes the basis and justification that we have for our treatment of subjective probabilities throughout this manual.

There are two different senses in which a subjective probability can be said to be correct:

- It can <u>really</u> represent the estimator's internal feelings about a projected event. If offered a series of bets, like those of Ramsey's method, this is the percentage or probability that he would <u>really</u> choose -- not some other.

- It can represent a realistic assessment of the evidence available, the value of conflicting evidence, general background knowledge, and his own strengths and weaknesses. In this sense, it represents the likelihood that the estimate is correct.

We will advocate the second point of view, which says that your probability assessments are not <u>merely</u> subjective. If they were only subjective, then a wild-eyed fanatic could correctly assess a greater certainty for his claims than could the careful, scientific investigator. But the opinions of the fanatic are worthless. This means that his subjective probabilities are wrong. And if they are wrong, then there should be some objective way of measuring them, to tell <u>how</u> <u>much</u> they are wrong.

It is for this purpose that a number of "proper scoring rules" have been developed. A proper scoring rule is a device for assisting forecasters in calibrating their probability assessments. An assessment is said to be "calibrated" when, over a large number of cases in which an n percent probability has been assessed, approximately n percent of them have been true.

Proper scoring rules have been intensively studied in connection with weather forecasting, in which predictions are often stated in terms of

probabilities: "There is a 20 percent chance of rain tonight." Probabilistic predictions like these are neither completely right nor completely wrong, unless they are stated as "100 percent" or "0 percent."

Still, some probabilistic predictions are better than others. They are better when they give high probabilities to the events which actually occur, and low probabilities to those which do not occur. We need a scoring rule to measure how much one probabilistic forecast is better than another.

One simple -- and misleading -- scoring rule was used in earlier appraisals (1974-76) of DE projections. This can be called the "hit-or-miss" scoring rule. It counts the number of times that projections have been correct (the hits) and compares this number to the number of times that they have been incorrect (the misses). The result is stated as a percentage: "You've been wrong 70 percent of the time."

If the hit-or-miss scoring rule were taken seriously, it would have the effect of pressuring the intelligence producer into hedging his bets by increasing the spread between High and Low estimates. The wider he makes these spreads, the higher his score. If he says, for example, "By 1984 the Soviets will have between 0 and 56 Foxtrot submarines," he is fairly certain to be right -- and to get a high score -- since only 56 Foxtrots were produced. But this heavily-hedged projection will not be of much help to the intelligence consumer, who really needs a less wide-ranging estimate. The hit-or-miss scoring rule is thus an improper scoring rule, since it encourages the estimator to produce a less-useful assessment of uncertainty.

Another type of improper scoring rule might be called the "direct" scoring rule. Suppose that you get a score of 70 every time you assess a probability of 70 percent to an event which later occurs, a score of 80 every time you assess a probability of 80 percent, and so on. This is a more plausible scoring rule, because it gives you a higher score when you assess a higher probability to the actual outcome, and a lower score when you assess a lower probability.

But the direct scoring rule is also improper, because it encourages the estimator to falsify his assessments. Specifically, he finds it possible to raise his score if he "goes for broke" -- that is, if he assigns a 100 percent probability to events that he believes likely, and a 0 percent probability to events that he believes unlikely, without attempting any of the finer shades of discrimination.

(To see this, suppose that you have issued 100 estimates, and that 70 percent of them have been correct. Suppose that you have, realistically, assessed a 70 percent probability for each of them. Using the direct scoring rule, you would get 70 points for each of the 70 estimates that were right, and 30 points for each of the 30 estimates that were wrong (since you allowed a 30 percent probability for this negative outcome). Then your total score would be 70 x 70 + 30 x 30 = 5800. Next, suppose that you had chosen a go-for-broke strategy. Instead of 70 percent, you said that all your estimates had a 100 percent chance of being right. Only 70 of them were actually right, however, so that you get a score of 70 x 100 + 30 x 0 = 7000.

This is a higher score than the 5800 you got when you made the correct assessments. Thus the direct scoring rule encourages the estimator to go-for-broke -- to exaggerate his probability assessments. For this reason, it is an improper scoring rule.)

The go-for-broke strategy would not be useful to the consumer, since it would encourage an untoward degree of confidence in the estimates. The consumer needs to know, with somewhat more precision, how likely the prediction or estimate is. The direct scoring rule is improper, because it encourages the estimator to produce misleading probability assessments.

A number of "proper" scoring rules have been developed to meet objections like these. They have the property of maximizing your score when your estimates are properly calibrated. For example, if 75 percent of the DIPP projections are alleged to fall within the Low-High range, then they are properly calibrated if and only if 75 percent of them do fall within the stated range. A proper scoring rule will give you a higher score when you assess probabilities correctly, without encouraging underassessment or overassessment.

One of the simplest of the proper scoring rules is the logarithmic rule; this is -log (p), where p is the probability that you have assigned to the event which actually occurred. For example, you claim that there is a 70 percent likelihood that all Bear F aircraft will be withdrawn by 1980. In 1980, it is found that all Bear Fs have been withdrawn. You then get a score of -log (0.70) = 0.15.

The log scoring rule gives you a lower numerical score when you are right; this means that you aim for the lowest possible score. Suppose that some Bear Fs are still sighted in operation in July, 1980. You have allowed only a 30 percent chance (100 - 70) for this possibility. You then get a score of -log (0.30) = 0.52.

Using the log scoring rule, you get the best possible score when you assess a probability of 1.00 to an event which actually occurs; this gives you a score of -log (1.0) = 0. On the other hand, using the same scoring rule, if you were unwise enough to assess a probability of zero to the event which actually occurred, you get a score of -log (0) = infinity, which makes it a very bad score. (The log scoring rule thus reflects that fact that we should never claim to be absolutely sure of anything.) Other proper scoring rules are less drastic in their treatment of completely wrong assesssments.

The primary advantage of a proper scoring rule is that it rewards accurate assessments. There is no pressure toward overassessment or underassessment -- neither a go-for-broke strategy nor a strategy that relies on hedging your bets will get as high a score as one which assesses probabilities correctly.

The reason for including a discussion of scoring rules in this manual has been to give you a sense of what it means to provide correct probability assessments. In general -- because of the great variety of products that

DE produces and because of the different sorts of information that must be considered for them -- the types of scoring rules that have been developed for weather forecasting cannot be easily adapted to intelligence production.

Instead, a more qualitative approach is needed. This suggests that the most effective way of evaluating past assessments is to review them regularly. For this purpose, you need extensive notes concerning your reasons for making a probability assessment -- the methods by which you have justified a particular assessment, and the assumptions that went into the estimate itself. Then, when errors are found in an estimate or a probability assessment, the response will be qualitative, rather than quantitative. It will be an explanation, in verbal terms, of what assumptions went into the judgment which were not justified, and of what later events conspired to undermine the accuracy of your product. This kind of information will be of more value to you, or to other people who must use your work, than a simple, numerical score.

The important point, however, is that your assessments of uncertainty are more than a statement of your subjective feelings about the estimate. They represent your careful assessment of the value of the information that went into your estimate, the relative value of other information available to you, and the understanding that you have of the larger context in which the estimate appears.

UNIT 9

COMMUNICATING UNCERTAINTY

The final task is the communication to the consumer of the uncertainty contained in your estimate. Several observations about this task can be drawn from earlier parts of this manual:

- Numerical forms should be used. This is especially true of those estimates which will be used in games and simulations, and in formal decision models. But all consumers will benefit from the use of precise, unequivocal numerical expressions.

- For consistency with other work in decision analysis and probability theory, the degree of uncertainty should be expressed as a probability. It should lie in the range from 0.0 to 1.0, where 0.0 represents your assessment that the event cannot possibly occur, while 1.0 represents your assessment that the event is completely certain to occur.

- Much information is sufficiently well-known that it can, in effect, be given a rating of 1.0. This does not necessarily indicate that there is absolute certainty concerning the event, but simply that it would make no sense in the current context to treat it as doubtful.

- Several different types of uncertainty may enter into an estimate. These can include uncertainty concerning the source data, possible deceptive material, missing information concerning aspects of the estimate, inclusion of speculative or inferential material, and possible changes in the political structure of the nation. Because these differing sources of uncertainty will require different types of responses from the consumer, it will be desirable to report them in different ways, rather than combining them all into a single measure of uncertainty.

- Different dimensions, as well as different forms of reporting, are required. For example, uncertainty concerning the numbers of aircraft can be measured along the vertical axis, while uncertainty concerning the date of retirement of the aircraft can be measured along the horizontal axis. These will require different formats when they are reported to the consumer.

Each of these points will be described in more detail in this unit.

In general, DIA has experimented with several different forms for reporting uncertainty, ranging from rather vague verbal expressions ("We believe that . . .") to more precise numerical assessments ("There is a 40 percent probability that . . .") Intelligence consumers have not made extensive use of these assessments, either ignoring them or taking the best estimate as though it were an unqualified projection. We believe that the appropriate response is not so much to change the formats in which uncertainty

is reported, but to improve the quality of these assessments and thus to make them more useful to the consumer.

In other words, while it might be possible to display your probability assessment in blinking red lights at the top of every report, the consumer will not pay attention to it unless (a) he thinks that the assessment is meaningful and correct, and (b) he knows how to make use of it. Your primary task, then, is to make sure that the probability assessment is meaningful, correct, and in a form that the consumer can use.

Several steps can be taken to support these goals:

- A uniform numerical method should be used for reporting assessments of uncertainty. The present use of the phrase "a 70 percent chance" or "an 80 percent probability" is immediately meaningful and should be retained. (Purists will object to calling a percentage a "probability," since probabilities are reported in a range from zero to one. Since current usage is completely meaningful, however, and since it can easily be translated into the required form for computations, there is no reason to discard it.)

- Other forms for reporting uncertainty may also be used. Experimentation has shown that the use of odds sometimes makes more sense to the nonmathematical reader. For this reason, you should try using phrases like "a fifty-fifty chance," or "the odds are six to one," when reporting the uncertainty of an estimate.

- Finally, the use of probabilities in their normal form is undoubtedly meaningful to the consumer. Thus, you can report "a probability of 0.60" when it makes sense in the context. This form can be used immediately by the consumer who is working with mathematical models.

The probability of the broad, general assumptions which underlie estimates may be much more difficult to quantify. For example, while you undoubtedly have some opinions concerning the likelihood that the Soviets will honor the SALT II agreements -- if, indeed, they are ever ratified by the U.S. -- you may not find it appropriate to include this opinion in a DIA estimate. The role of these general assumptions should nevertheless be made clear, in such a way that the consumer can, if necessary, insert his own assessment of the probability to be attached to them.

The most important next step, however, is to provide the basis for increasing the credibility of DE's assessments of uncertainty. There are several ways in which this goal can be achieved:

- First, if the credibility of DE's past performance is to be maintained, there must be some record of that performance. Specifically, there should be an accurate record of the probabilities that you have assessed for various events, together with records of the spreads that have been included in the DIPP. These should be checked against later information to determine how accurate they have been.

- While this information undoubtedly exists somewhere in the files, it is not in a form that permits rapid retrieval and evaluation. Small parts of the record can be retrieved, to assist in researching some specified topic, but these do not provide an overall batting average for the Directorate.

- For this reason, we have recommended the design and development of an Institutional Memory, which will contain information concerning past probability assessments, in a form which will permit them to be checked rapidly. Lacking this facility, it would still be important to have full information concerning the probability assessments that are currently being developed, for use in later documentation of DE's record.

- Second, to assist in developing better probability assessments, you need to know the rationale behind each assessment. Frequently, this information is known to the estimator but is lost when he moves to another assignment. While it would probably be inappropriate to include all of this information in a published estimate, nevertheless it would be useful if there were, somewhere, a more complete explanation of the chain of reasoning that led to a particular assessment.

- For this reason, we have recommended that the Institutional Memory include information concerning the rationale behind an estimate and the accompanying assessment of its credibility. This information

could be included in a file or a set of background papers accompanying an estimate. It could be used whenever it was necessary to review a particular estimate.

- Third, this information is not likely to be of much value unless it is used. This means that you should plan to spend at least some time in looking at past assessments of uncertainty, your own and others, simply to find those areas in which there was trouble. Without this kind of feedback, assessments of uncertainty are likely to remain mere guesses.

Another approach which will contribute to improving the usefulness of uncertainty assessments will be to attempt to determine, for each assessment, exactly how the consumer expects to use it. If this information is of no use at all to the consumer, then it doesn't matter whether you say that there is a "20 percent probability" or an "80 percent probability;" it doesn't really make any difference. Consumers have, however, requested this information, and the real task is to get it to them in the form that will be most useful.

For example, in estimates which are to be used in simulations, the probability assessments should be clearly attached to the estimates in a form that can be quickly entered into a machine-readable data base.

On the other hand, estimates prepared for briefings cannot rely on numerical addenda or footnotes to communicate uncertainty to the user.

Charts and graphs should clearly indicate the areas of uncertainty surrounding each projected figure.

Further techniques for aggregating and communicating uncertainty are included in the computer programs, together with the user's manuals, which accompany the final report for this project. These provide more practice in the use of the methods described here, together with other suggestions for communicating uncertainty.

The remainder of this unit will be concerned with specific problems in determining uncertainty.

You may, of course, be uncertain about the probabilities that you assess; you may be uncertain about how uncertain you are. And you may be uncertain about this level of uncertainty, and so on, to any level of meta-uncertainty. These cascading uncertainties could threaten any system for the measurement and communication of uncertainty.

The approach suggested earlier, in which you are asked to choose between bets, in order to clarify your own subjective probabilities, provides the answer to this difficulty. Even though you may not be sure that you are choosing the best bet, you are forced to choose some probability or another. Of course, this reflects the situation that we all face in real life, in which we have to make choices on the basis of partial evidence.

In other words, there is no one who is better qualified that you are to assess the degree of uncertainty in your estimates. You should be aware of your own biases, your tendencies to overassess or underassess your level of certainty, you own conservatism or anti-conservatism, and so on. But no one else can tell you how certain you are about your estimates.

How does missing data affect your assessment of uncertainty? By "missing data" we mean those pieces of unknown information which would be relevant to your estimate if they were known. Some missing data are like the missing pieces in a picture puzzle; they may be clearly identifiable as missing. Others may be totally unknown to you, like the blank areas in a puzzle about which you know nothing at all. Still others lie in the future, where human decisions may reverse or undercut an earlier projection: if a coup d'etat were to overthrow the current Chinese regime and re-install the radical policies of Mao Tse-tung, this could have a pervasive effect upon the Chinese military posture. But coups d'etat are difficult to predict, and the specific directions to be taken by a new government are even more difficult to prophesy; this is undoubtedly the reason that such cataclysmic events are explicitly excluded from DE's projections.

Your general understanding of the nation helps to provide a context which will limit the effect of missing data. Like the partially-completed picture puzzle, it provides a general outline of the policies and capabilities of a potential adversary. While you may not know the details of a specific meeting in the Kremlin, you can at least gain some idea of what would happen at such a meeting, based on your general knowledge of Soviet attitudes,

combined with all the information that you do have concerning Soviet activities before and after the meeting.

The scientific method requires that you account for all the available data within the context of a general hypothesis concerning the phenomena that you wish to investigate. As new data are obtained, they may tend to verify your tentative hypothesis, or they may lead you to modify or reject it. The role of missing data, then, is to increase the uncertainty present in your general model: if all data were missing, uncertainty would be total, and if no data were missing, then there would be no uncertainty (or at least no uncertainty concerning current goals and capabilities of the nation; data concerning future events may be unobtainable in principle).

Another source of uncertainty in the initial data may derive from errors in the order of battle (OB), which serves as a base line for projections. This is a serious problem, since it means that you cannot know the true level of forces in a foreign nation, particularly in China, and as a result you cannot build projections upon a firm base of knowledge. In addition, errors in the OB may make it difficult to determine whether your earlier projections have been in error; the error may be in the OB figure, rather than in your projection. Moreover, when past figures (e.g., for 1971) are revised (e.g., in 1974), there is no assurance that the revision makes the newly revised figures more accurate. In fact, the revision may simply be the result of smoothing a trend line in one plausible direction or the other. Another source of changes in the OB may be the use of a more effective collection system. Suppose, for example, that there is a sharp increase in ICBM

figures from 1971 to 1972; this may not mean that forces were significantly increased, but rather that satellite cameras were greatly improved at that time. As a result, the accuracy of earlier figures may be thrown into doubt.

The computer programs that accompany this report provide an approach to the assessment of uncertainty in the OB figures. Specifically, they show how to measure the variance in these figures, and the ways in which it affects your assessments of the probability of later estimates.

Another source of uncertainty, conscious deception by a potential adversary, afflicts all forms of strategic intelligence. Even allied nations may habitually provide misleading figures concerning their capabilities.

You are aware of these deceptions and may revise your projections toward more realistic numbers than those claimed by foreign nations. In some instances, there is a more difficult problem, when military planning is performed badly, and the nation has no clearly identified military goals or policies (perhaps because of internal political conflict). Under these conditions, your task is to develop realistic projections based on whatever data you can get.

In the final unit of this manual, this problem will be re-attacked at a more global level.

UNIT 10

CONCLUSION

Your estimates are based on your comprehensive understanding of the nation. Every factor which might influence the development of a weapon system, military action, or capability must be taken into consideration -- the national economy, domestic and international policy and goals, the location and capacity of production facilities, natural resources located within the nation or available through its allies, technological capabilities and the output of research laboratories, areas which are receiving special attention in research, the power base of the current regime and the likelihood that it will remain stable, the organization and leadership of the armed forces, military policies which have become traditional -- in short, many aspects which, together, form a "model," or rational intellectual picture, of the nation as a whole.

With a clearly-defined model of the nation, it is possible to develop reasonable estimates of its present and future capabilities in specific areas. For example, if the estimator understands the importance of the five-year plans for Soviet resource development, and if he understands that the Soviets are rather slower than the Americans in disposing of obsolescent equipment, he can make some reasonable estimates of the dates by which a given weapon system will be replaced.

A clearly-defined and correct model of Soviet intentions and capabilities provides a basis for dealing with errors, deception, and missing data. The

deception itself fits into the pattern of overall Soviet strategy, which provides a rationale for the deceptive maneuver. Within obvious limits -- the model itself must be tested against reality -- the use of a comprehensive model provides a defense against deception.

Deception does not merely introduce an element of uncertainty into the estimative process. The attempted deception must be motivated, and valuable information may be derived from the most deceptive material, if the underlying motive can be correctly identified. For example, two Soviet political scientists have prepared an article for a recent issue of Fortune magazine, which attempts to justify the USSR's massive expenditures for armaments within the context of peaceful Soviet intentions. The U.S. reader will not, of course, take these protestations at face value. Valuable information can nevertheless be obtained concerning Soviet intentions if we succeed in interpreting them correctly, since the article surely indicates what the Soviets want Americans to believe.

More generally, the art of propaganda analysis attempts to derive information of value from deceptive material, not merely to reject it as false. In practical terms, this means that you formulate a generalized hypothesis concerning the goals of the nation. This hypothesis provides you with a context in which to interpret the specific information that you receive, including erroneous and deceptive information. The general hypothesis must, of course, be reviewed frequently against available, well-documented information, lest it become a form of paranoia -- in which assumptions are never

really tested. Nevertheless, it is this general viewpoint that helps to provide you with a context in which to interpret data of dubious value.

There is, in addition, a "paradox" of intelligence which has frequently been described. It is more apparent in tactical and current intelligence than in estimative intelligence, but it appears in all intelligence work to some degree. In one version, it says, "If you're right, then events will prove you wrong." In less paradoxical form, it simply calls attention to the fact that the goal of all U.S. intelligence is to provide information which will assist in determining U.S. policies and actions. In the dynamic context of world events, this means that the U.S. responds in such a way as to counteract or avoid the projected threat. The foreign power therefore must abandon or redirect its effort, and the threatened action does not take place. Since the threat does not materialize, the intelligence report was "wrong."

Such a sequence is not, of course, a paradox in any real sense. It serves primarily to illustrate a point that we have repeated several times: that the goal of intelligence is not to prophesy the future, but to determine the plans and capabilities of a nation. For example, if you were to project the development of a substantial Soviet ICBM capability, the U.S. should be expected to respond in such a way as to reduce or eliminate the threat that the ICBMs present. If the U.S. does successfully develop an effective counterforce, this could lead the Soviets to modify or abandon their ICBM development. If they were to do so, then your original projection would be "wrong." The Soviets would <u>not</u> have the ICBM force that you projected.

But in any reasonable sense, of course, the original projection was "right." The Soviets did indeed plan to develop an ICBM capability, but thanks to your timely projection and the U.S. response, they were forced to change their plans. You were correct in identifying the original Soviet intentions. Unfortunately, the format in which the projections are made does not clearly indicate that they represent intentions and capabilities; instead, they appear to represent firm predictions of the future. Thus, in an evaluation of the quality of the projections, they are judged "wrong."

The dynamic character of the world situation in which estimative intelligence must work therefore creates an irremovable source of uncertainty.

In summary, several points have been emphasized in this manual:

- DE must meet with other groups and justify estimates, with the eventual goal of presenting a consensus estimate for the intelligence community. This requires that estimates be defensible, that appropriate justification be available to support them.

- Justification takes place within the context of a comprehensive model of the nation under study, composed of defensible hypotheses concerning national goals, military and industrial capabilities, social structure, and other factors which might contribute to its military posture. A thorough understanding of all relevant characteristics of the nation is necessary to produce a justifiable projection concerning specific military developments.

- Projections are intended to represent verified scientific hypotheses concerning development and deployment of military capabilities, based on a general model of the nation, together with observed data concerning the specific capability.

- Uncertainty enters into the projection process in many ways, including missing data, conscious deception, errors in the models employed, etc.

- It is important to communicate this uncertainty to intelligence consumers, who must know the degree of uncertainty in a projection in order to make reasonable use of it.

- Attempts to communicate uncertainty have not been successful, since little use is made of any of the proposed measures of uncertainty. In addition, the need for calibration (validation) of measures of uncertainty has been neglected.

The primary goal of this manual has been to suggest some effective means of dealing with the problem of assessing uncertainty correctly, and of communicating this uncertainty to the intelligence consumer.

APPENDIX C

COMPUTER-BASED SYSTEMS FOR AGGREGATING UNCERTAINTIES

This appendix will include descriptions of computer-based systems for the aggregation of measures of uncertainty. They represent successively more sophisticated developments of a basically Bayesian approach.

The systems chosen for inclusion in this section were designed by Dr. Edward R. Hogan and Dr. John F. Lemmer of our staff; the descriptions have been adapted here to the specific task of aggregating uncertainties in strategic intelligence.

C.1. INTRODUCTION

In the production of strategic intelligence, there will be a number of sources of information available, including reports from the same source over a period of time. To facilitate exposition, we will refer to all pieces of information as "reports," regardless of the specific format or source. Some of these reports can be taken as completely accurate, beyond all reasonable doubt; a verified engineering report on a piece of captured Soviet equipment might qualify as this kind of information. Other reports will be doubtful, inaccurate, or subject to confirmation in varying degrees; unverified information obtained from prisoners, defectors, or casual tourists would fit this category. Since some of these reports will be wrong part of the time, it is desirable to combine the information from several reports in such a way that

a higher level of confidence may be obtained from the pooled data than from the independent responses.

We will consider several mathematical models that perform this function. In all cases, we will assume that the reports from the various sources are independent. We do this to make the model tractable. Of course, not all reports will actually be independent. For example, several prisoners may have been briefed prior to capture concerning the appropriate responses to make under interrogation. In this instance, their stories cannot be taken as confirming one another, since they are all essentially the same story, told by various persons. In such cases, the actual levels of confidence will be lower than those given by all of the models. In this sense, the first three models will tend to be optimistic. In Subsection C.6., another approach is described which does not require the assumption of independence.

In evaluating the worth of the various methods, including those used in Subsection C.6., the degree of confidence or uncertainty of the projection is important, of course, but so is the speed with which the projection can be produced. While the time factor for estimative intelligence is not as critical as that for current intelligence, it is nevertheless necessary to produce estimates in a timely fashion, to permit appropriate responses by intelligence consumers. Of equal importance is the cost of the estimate; with limited resources available, the comparative costs of various techniques must be taken into consideration. Thus, a system which made modest demands upon the computing system would be preferable to one which required very extensive computer resources.

The Bayesian approach to be described here uses the prior probabilities (or a priori probabilities) of the various events and developments for which estimates are made. These are the probabilities that the intelligence producer attaches to a specific event before receiving or reading a report; Bayesian techniques tell us how much those probabilities must be changed when new information is taken into consideration. It is also important to use a method which will take into account the fact that an error in determining some estimates is more costly than an error in determining others. This information is essential in determining the degree of effort to be expended to achieve a given level of confidence in an estimate. (In practice, however, it may be necessary to assume that all estimates are equally valuable.)

C.2. CLASSICAL BAYES DECISION MODEL

Classical Bayesian decision theory assumes that there exist n possible states of nature denoted by $\theta_1, \ldots, \theta_n$. These are assumed to be exhaustive and mutually exclusive. In terms of intelligence projections, these could be possible types of missiles selected for defense of an urban area during the next ten years. The specific typology would depend on the breakdown used for reports and projections. In particular, possible classifications might include: "all others," "no missiles used for this purpose," and so on (Figure C.1.).

To each possible state of nature, as identified by the analyst, there corresponds an action, which will be denoted by $a_1, \ldots, a_m$. For the

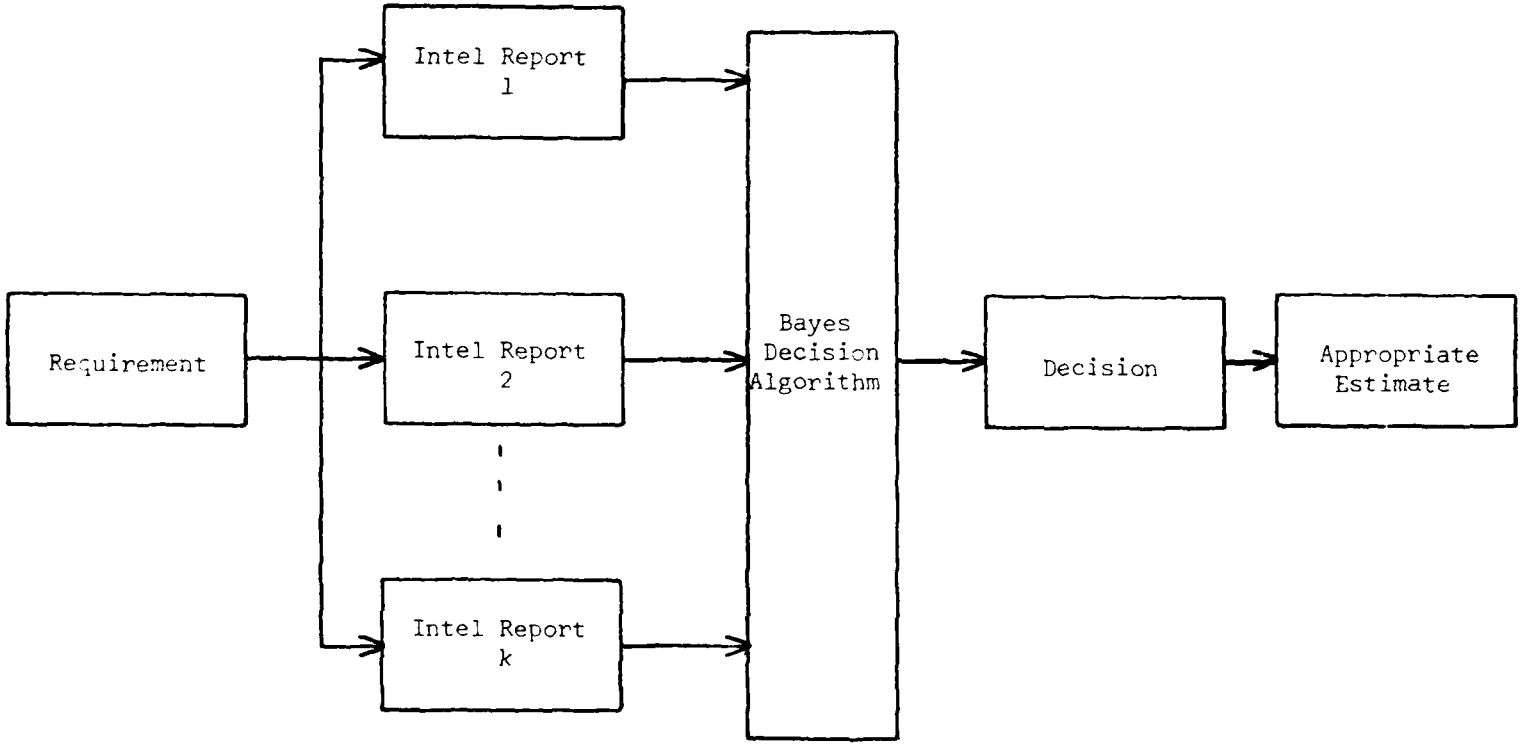


Figure C-1 Classical Bayes Decision Model

intelligence estimator, each action is simply the identification of a future state of nature; for this reason, every a_i corresponds to one of the θ_j, and consists simply of the action of stating that the state of nature will be θ_j on some future date. Thus, in our application, each of the a_i will correspond to one and only one of the θ_j.

Bayesian theory also defines a loss function $L(\theta_i, a_j)$ that gives a quantitative measure of taking action a_j when the true state of nature is θ_i. This is essentially the cost to the user of each estimate -- the cost or loss will be positive, in general, when the estimate is wrong, and will be negative (representing a gain) when the estimate is right. The introduction of a loss function, while it is a standard component of Bayesian analysis, is rather interesting in connection with intelligence estimates. It says, briefly, that you should know how valuable this information is to the user. If it has no value--that is, if it doesn't matter whether the projection is right or wrong, in terms of gain or loss for the user--then it doesn't much matter what figures are included in a projection. The more valuable the projection is for the user, the more care should be expended in collecting information.

The Bayesian approach presupposes that some estimate of the prior probabilities is available. In general, these estimates would come from prior probabilities would consist of earlier estimates of the probability that each type of missile system would be deployed in the given location at the specified date. This determination may be expressed as a probability distribution and is denoted by $\pi(\theta)$.

We will denote the various reports which are to be used to update the probabilities by $x_1, \ldots, x_k$, and the k-tuple $(x_1, \ldots, x_k)$ by X or X_k. Each report x_i will correspond to one of the states of nature, θ_j. (That is, each report tells you that a particular missile system is to be deployed; or, more generally, it says that a particular situation will obtain.) By prior testing, estimation, or a mathematical model, the probabilities of the form:

$$P(x_i \mid \theta_j) = \text{Prob}\{\text{report says that } x_i \mid \text{given that the true situation is } \theta_j\}$$

may be obtained. From the above, we may calculate the probabilities $P(x_1, \ldots, x_k \mid \theta_j) = \prod_{i=1}^{k} P(x_i \mid \theta_j) =$

prob (reports $x_1, \ldots, x_k$ given that the true state of nature is θ_j). The equation holds if we assume that the various reports are independent.

For a fixed θ_j we define the risk function:

$$R(\Theta, d) = E_\Theta\,(L(\Theta, d(X)),$$

where d (X) is the decision function, a statistic that takes values in $\{a_1, \ldots, a_m\}$, and the expectation, E, is taken with respect to the distribution of X.

Bayes' theorem now provides us with a means of going from the prior probabilities, using the conditional probabilities, to obtain the posterior probabilities -- that is, the probabilities of each of the possible states of

nature, given the reports that you have about them:

$$P(\theta_j \mid x_1, \ldots, x_k) =$$

$$\frac{\pi(\theta_j)\, P(x_1, \ldots, x_k \mid \theta_j)}{\sum_{i=1}^{n} \pi(\theta_1, \ldots, x_k \quad \theta_i)} =$$

$$P_{post}(\theta_j \; X).$$

Given the prior distribution (θ), we define the Bayes Risk function to be:

$$R(\pi, d) = E_{\pi} R(\theta, d).$$

Your task now is to decide which *projection* to make, using a decision rule d* that minimizes the Bayes risk.

If we examine each of the conditional risks,

$$R(a_j|X) = \sum_{i=1}^{n} p_{post}(\theta_i|X)\, L(\theta_i, a_j,)$$

then the decision rule that will minimize the expected loss is the rule that chooses the action that gives minimum conditional risk. Thus if

$$R(a_j \mid X) = \min R(a_i \mid X),$$
$$\text{for } 1 \leq i \leq m,$$

then the decision rule $d^*(X) = a_j$. In the event of a tie, any convenient tie-breaking procedure may be used.

The Bayesian approach is designed to be used iteratively. When you have calculated a set of posterior probabilities, these may be used as the prior probabilities with a new set of reports, and a new computation. In practical terms, this means that with every new set of estimates, the probabilities attached to the old estimates are combined, using Bayes' theorem, with a new set of reports, to obtain the new probability estimates. Over a period of time, the probability estimates will tend more and more to approach the actual probabilities to be attached to each potential weapon system development. This is an essential feature of the Bayesian approach: regardless of the initial estimated probability distribution, new information will bring the probability values closer and closer to the actual values as more reports are incorporated in the estimate. The trade-off, of course, is that the estimates become more costly and less timely as more reports are required.

C.2.1. Confidence Level

Our purpose here is to evaluate the expected probability that a given set of reports will produce the correct combined estimate. This will provide a measure of the aggregated uncertainty of the estimate.

As noted above, there is a unique action associated with each state of nature. In our application, this action is simply to state that a particular state of nature, θ_j, will occur. For example, your action might be to say, "Fifty SS-20 missiles will be deployed by 1985." The state of nature is

the Soviet intention to deploy 50 SS-20 missiles; the action is your projection that they will do so or that they intend to do so.

To simplify computational form, we will assign an action a_j to each state of nature θ_j. Suppose, temporarily, that we assume that some projection must be made; we omit the possibility that "no decision" can be one of our possible actions, $a_1, \ldots, a_m$. It will be useful to calculate our expected probability of making a correct decision with any combination of reports. This will simply be:

$$P \text{ (projection is correct)}$$

$$= \sum_{j=1}^{n} \pi(\theta_j) \left[\sum_{X} P(d(x) = a_j | \theta_j) \right]$$

$$= \sum_{j=1}^{n} \pi(\theta_j) \, P_{\substack{\text{combined} \\ \text{projection}}} (a_j | \theta_j).$$

where

$$P_{\substack{\text{combined} \\ \text{projection}}} (a_j | \theta_i)$$

$$= \sum_{X} P(d(X) = a_j | \theta_i),$$

and where the sum is taken over all possible X's.

Now suppose that we wish to extend the above formula to the case in which an action "no choice possible" is allowed, and where this action is denoted by a_{n+1}. Then we wish to determine:

P (projection is correct given that choice was made)

$$= \frac{\text{P (choice was made and is correct)}}{\text{P (choice was made)}}$$

$$= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} E_i \; P_{proj.}(a_j|\theta_j)}{\sum_{j=1}^{n+1} \sum_{k=1}^{n} E_k \; P_{proj.}(a_j|\theta_k)}$$

In a similar manner, we can calculate the confidence level for any subset of the total set of weapon systems. For example,

P (projection is correct for IRBM)

= P (actual installation is IRBM | projection is for IRBM)

$$= \frac{\text{P (actual installation is IRBM and projection is correct)}}{\text{P (projection is for IRBM)}}$$

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} E_i P_{proj.}(a_j|\theta_i)}{i \text{ is IRBM } j \text{ is IRBM}}$$

$$= \sum_{j=1}^{n}\sum_{k-1}^{n} E_k P_{proj.}(a_j|\theta_k)$$

$$j \text{ is IRBM}$$

C.2.2. Example

We will compute the confidence level of a projection made up of several reports concerning missile installations to be installed by 1985 around a hypothetical Soviet city. Two sources of information are available concerning Soviet plans for the projected missile installation, which we will simply call Source 1 and Source 2.

The probability matrices indicate our estimates, based on prior experience with similar sources, that Sources 1 and 2 are correct or incorrect. The upper left-hand entry is the probability that Source 1 or 2 will report an IRBM installation, given that the actual Soviet plan calls for IRBM; the upper right-hand entry is the probability that the source will report (erroneously) that the installation will include IRBM when Soviet plans call for non-IRBM; and so on.

Source 1

a_j \ θ_i	IRBM	NON-IRBM
IRBM	0.89	0.10
NON-IRBM	0.11	0.90

Source 2

a_j \ θ_i	IRBM	NON-IRBM
IRBM	0.95	0.04
NON-IRBM	0.05	0.96

The loss matrix, indicating the cost to the U.S. of an incorrect identification and the corresponding gain, or negative loss, for a correct identification could, for this hypothetical example, be:

a_j \ θ_i	IRBM	NON-IRBM
IRBM	-1	1
NON-IRBM	0.5	-1

The entries in this matrix are values for $L(\theta_i, a_j)$. Thus, the relative cost of a projection which states that the missiles are <u>not</u> IRBM, when they actually <u>are</u> IRBM (according to Soviet plans), is 0.5. The costs need not represent actual dollar amounts, but simply the relative loss or gain (where negative values represent gains).

You assume also, on the basis of general statistical information concerning Soviet plans, that 1 out of every 5 Soviet missile installations is IRBM. Thus, your prior probability figure tells you that any missile installation selected at random has 1 chance out of 5, or a probability of 0.20, of being an IRBM installation, and a probability of 0.80 of being non-IRBM.

We let $X = (x_1, x_2)$, where x_1 is the identification provided by Source 1, and x_2 is the identification provided by Source 2. Since we are assuming that Source 1 is independent of Source 2, letting I abbreviate IRBM, and NI abbreviate non-IRBM, we have:

$P(I,I|I) = P_1(I|I)\, P_2(I|I) = .89 \times .95 = .8455$, which is the probability that both sources will identify the proposed installation as IRBM, given that the Soviets actually plan to deploy IRBMs at this site. Similarly, the other combined probabilities can be calculated. (E.g.: What is the probability that the first source will report IRBM and the second will report non-IRBM, given that the Soviets actually plan a non-IRBM deployment? Answer: 0.10 x 0.96 = 0.096.)

These results can be summarized in the following table.

Source Reports X	$P(X/\theta_j)$ IRBM	NON-IRBM	prob. of X
(I,I)	0.8455	0.004	.8455x .2 + .004x .8 = .1723
(I,NI)	0.0445	0.096	.0857
(NI,I)	0.1045	0.036	.0497
(NI,NI)	0.0055	0.864	.6923
$\pi(\theta)$	0.2	0.8	

Now the posterior probabilities are calculated using Bayes' theorem, e.g.:

$$P(I|I, I) = \frac{0.2 \times 0.8455}{.1723} = 0.98143$$

Summarizing, you get the following values of $P(\theta|X)$:

	θ	
X	IRBM	NON-IRBM
(I,I)	0.98143	0.01857
(I,NI)	0.10385	0.89615
(NI,I)	0.42052	0.57948
(NI,NI)	0.00159	0.99841

You now need to find d(X), the decision to be made concerning the projection, for each of the four possibilities. Using the risk function:

$$R(a_j|X) = \sum_{i=1}^{n} P(\theta_i|X)\, L(\theta_i, a_j)$$

You obtain:

X	$R(a_j, X)$		d(X)
	I	NI	
(I,I)	-.96286	.47215	I
(I,NI)	.7923	-.84423	NI
(NI,I)	.15898	-.36922	NI
(NI,NI)	.99682	-.99762	NI

The decision function, in the right hand column, simply reflects your choice of the outcome that carries the smallest risk, or the largest expected gain (negative risk).

Using this information for the combined sources of information you get:

$P(I|I) = 0.8455$

$P(NI|I) = 0.1545$

$P(I|NI) = 0.004$

$P(NI|NI) = 0.996$

Thus the overall confidence level is

$C = 0.2 \times 0.8455 + 0.8 \times 0.996 = 0.9659$

This represents the confidence you obtain through the use of two sources of information, rather than one. Your confidence in either of the sources of information separately is:

C (Source 1) $= 0.98 \times 0.2 + 0.90 \times 0.8 = 0.898$

C (Source 2) $= 0.95 \times 0.2 + 0.96 \times 0.8 = 0.958$

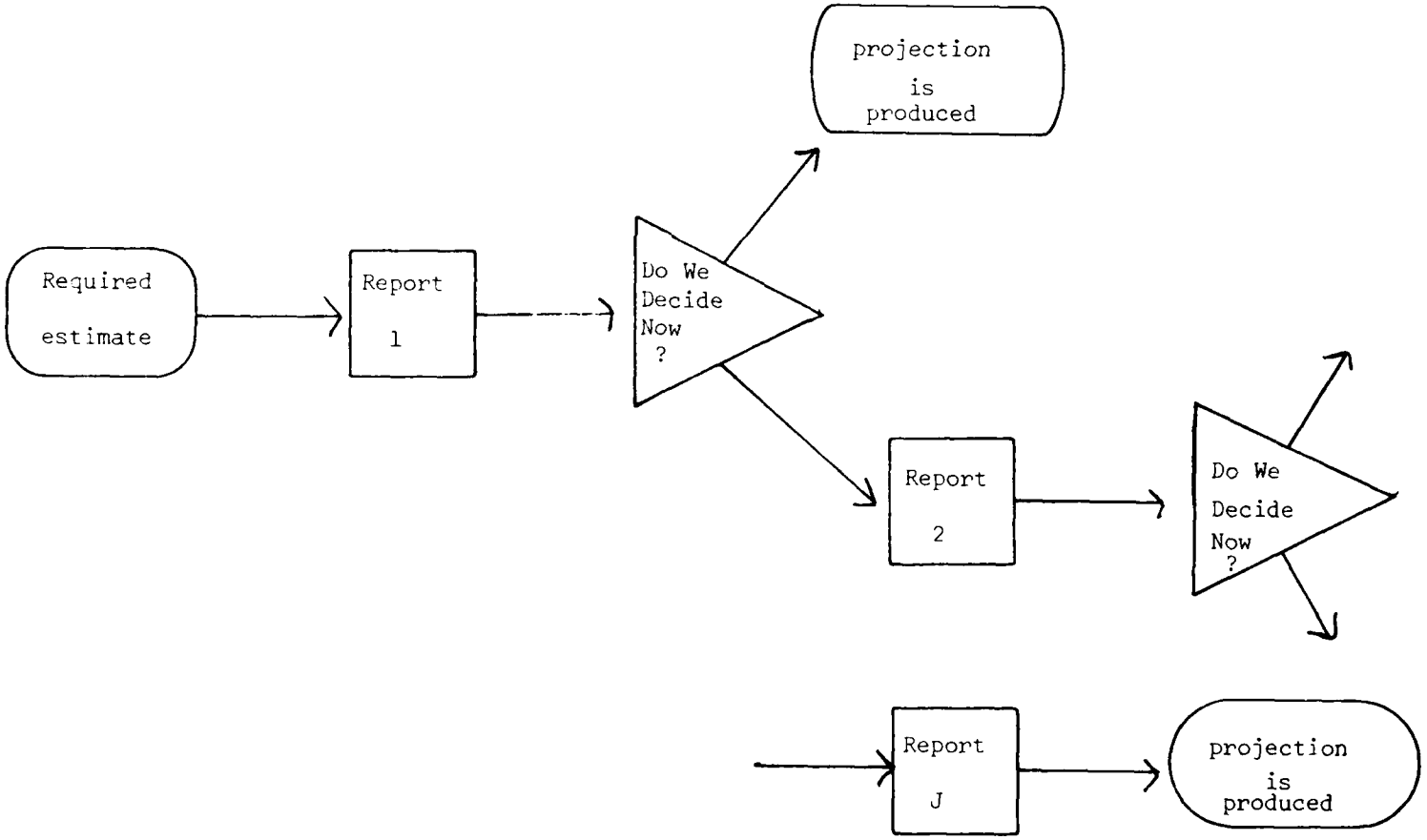
C.3. SEQUENTIAL BAYES DECISION MODEL

The preceding model may be adequate under some conditions, especially when the number of sources of information is fairly small, say around six, and they are similar in character -- if, for example, they are all reports based on prisoner interrogations. That is, if you can expect to be getting information from all of a small number of sources at approximately the same time, then the classical Bayes model should be ideal. But because of their combinatorial nature, the Bayes rules become computationally long and expensive if reports from many varied sources are combined.

In addition, the preceding model ignores one of the fundamental rules of intelligence production: that intelligence rapidly loses its value through time, and that intelligence is valueless if it does not arrive soon enough for timely action. If you wait until you are absolutely certain of a possible event, then it will be too late for anyone to take action to prevent that event. On the other hand, if you act prematurely, before you are sufficiently sure of the enemy's intentions, you are subject to disasters of a different nature. You must achieve the appropriate balance between too-hasty and too-tardy action. The methods to be described in this subsection are intended to identify that point of balance.

The following model takes into account the cost of using additional sources of information. It also allows for repeated use of the same source of information over a period of time (Figure C.2.).

Figure C-2 Sequential Decision Model



Suppose that you now take intelligence reports $X = (x_1, x_2, \ldots)$ sequentially, and, as in the preceding model, the distribution for each finite segment $X_n = (x_1, \ldots, x_n)$ is assumed to be discrete and known, given a state of nature θ_j. Since you are assuming independence of the reports,

$$\phi_n(x_1, \ldots, x_n|\theta_j) =$$

$$f_1(x_1|\theta_j) \ldots f_n(x_n|\theta_j).$$

You may now assume, quite reasonably, that there is a cost, or many costs, associated with obtaining each report. Determining the precise cost of reports is, of course, a very serious problem. However, in the context of estimative intelligence, an overriding cost is the amount of time spent in obtaining the report and the consequent delay in producing the estimate (together with the delay which is introduced in the production of other estimates). It might be possible, and desirable, to devise a fairly complex cost function based on the anticipated time required to obtain a particular report, together with the time necessary to combine that report with other reports. However, the time required to obtain n reports should be roughly proportional to n. That is, we can devise a rough cost function by defining

$$C(X_n) = \text{cost of obtaining } X_n = Kn$$

where K is a constant representing the time required to obtain and integrate one report.

We will also assume that it is possible to obtain only a finite number of reports, J. In other words, your sequential decision problem will be truncated at J. Again, J may be estimated simply by dividing the time available for preparation of the estimate by the amount of time required for obtaining and integrating each report. In any case, there will come a time at which you must produce an estimate, if the estimate is to be useful to the consumer; and J is a measure of the amount of information that can be obtained and digested within that time limit.

We will define a stopping rule s to determine when to stop sampling (obtaining reports) and a terminal decision rule d to determine what action to take when you have stopped sampling. The terminal decision rule may be viewed as a vector of decision rules, one for each sample size:

$$d = [d_0, d_1(x_1), d_2(x_1, x_2), \ldots],$$

where d_0 is one of the available actions and $d_n(x_1, \ldots, x_n)$ assigns an action to the vector or sequence of reports $(x_1, \ldots, x_n)$.

A stopping rule is determined by a family of functions of the form

$$\phi = [\phi_0, \phi_1(x_1), \phi_2(x_1, x_2), \ldots]$$

where

$$\phi_0 = \begin{cases} 0, \text{ if d tells you to take the first observation} \\ 1, \text{ if d tells you to take no observations} \end{cases}$$

and

$$\phi_n (x_1, \ldots, c_n) = \begin{cases} 0, \text{ if d says take an additional observation, given } x_n \\ 1, \text{ if d says take no further observations, given } x_n. \end{cases}$$

The functions ϕ_k completely determine the point at which to stop, for any given stream of information. If the stopping rule is given in another form, the functions ϕ_k may always be determined from the formulation.

Dr. Hogan's original study provided a derivation of the stopping rule, which is somewhat complex and will be omitted here. Essentially, a stopping rule will determine that point at which it would be more costly -- in terms of expected loss -- to continue gathering information, than to stop. For example, if your current information indicates, with a 60 percent probability, that a given weapon system will be deployed, then it may not be worth your time to continue collecting reports which might raise this probability to 70 percent. The need to produce a timely projection, the cost of obtaining additional reports, and the need to work on other projections are among the factors that would influence your decision to stop collecting information. You therefore issue the projection with the 60 percent probability attached to it.

As the stopping rule is formulated and combined with a decision rule, indicating the projection to be made, it minimizes the Bayes risk, B (s,d), over all decision and stopping rules.

This model, the sequential Bayes decision model, allows for a tradeoff between timeliness and confidence in a projection. But the model is even more complex computationally than the first model. It might also be extremely difficult to devise a cost function that would be general enough to fit a wide enough set of circumstances to make the model useful. In addition, it might be difficult or costly to determine some of the other parameters, such as the truncation point, at which no further reports will be included.

The sequential model is quite sensitive to the cost function. In the example given in subsection C.2.2., if the cost of new information is high enough, an estimator might simply decide to issue a projection stating that the installation is non-IRBM, given that 80 percent of all Soviet missile installations are non-IRBM. (This would be comparable to issuing a weather forecast in Rome, New York, stating that there will be snow on January 14, based only on the forecaster's prior knowledge that 80 percent of the days in Rome in January are snowy.) For a randomly selected missile installation, the estimator would have an 80 percent chance of being right.

C.4. A MODIFIED SEQUENTIAL DECISION MODEL

Unlike the first model, which takes reports from all available sources of information before making a decision as to the proper projection, the second model uses only enough reports to assure a given level of confidence in the projection (with the additional requirement that the number of available reports is limited to some finite number, J).

If an estimator were required only to track and report a single weapon system, then it might be possible to use a fixed number of different sources of information, which would provide periodic updates, and which could be combined using the first Bayes decision model, to obtain projections as required. Since the appropriate level of confidence for each source of information could be obtained over time, the overall level of confidence could be readily computed, using Bayes' theorem. A minimum of time would be required to make the best possible decision -- that is, to issue a projection with the highest possible level of confidence. On the other hand, if reports became available at different times, with varying levels of confidence; if there were a priority on using a small number of reports (in order to reduce the cost of obtaining them); and if the computations were not prohibitively long; then the second model might best fit the problem.

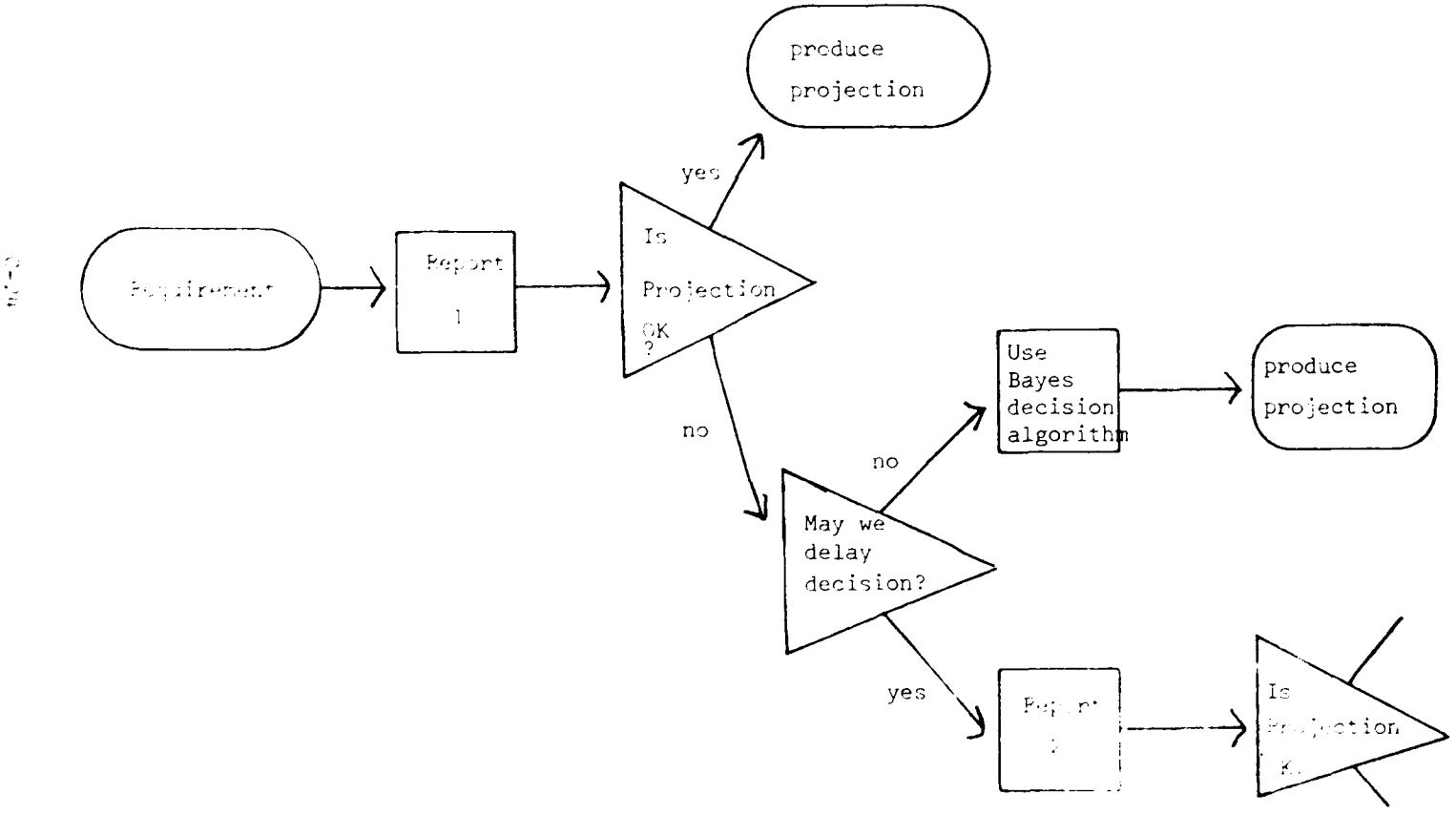
But the actual production of intelligence estimates is not as orderly as these models suggest. Many different types of projections must be produced to meet the requirements of consumers, and the required reports may not be avail-

able when they are needed. In such a situation, the first model may be critically costly, in that it requires a complete set of reports before it can produce a projection; in such a situation, the projection may not be completed in time to be of value to the consumer. And since the second model may potentially require even more computations than the first, the second model may take too long to arrive at a decision.

The critical problem with the second model is its lack of flexibility. Although the preliminary computations are lengthy, they need to be done only once. Then the algorithm is implemented by a simple lookup table. However, all of the required reports must be available for the lookup method to function. Such a situation is not likely in the actual context in which estimates are produced. The heterogeneous nature of the reports used in estimative intelligence means that some sources of information may be unavailable in time to be of use. Under such circumstances, either tables would have to be available to account for all possibilities, or recalculations would be necessary, perhaps several times, during the production of an estimate. The classical Bayesian sequential model would then become unwieldy computationally.

The following sequential statistical technique may be found desirable for actual estimates, under the conditions of heavy personnel loading and fixed deadlines which normally obtain (Figure C.3.). Using this approach, more than one projection can be produced, and no more reports need to be utilized than will be necessary to produce a satisfactory projection.

Figure C-3 Modified Sequential Model



We will again use a Bayesian approach and assume nonparametric probability distributions. Suppose that a request for a projection arrives, and that your current information indicates several possible projections -- that is, several types of weapon systems which could be developed during the indicated time span -- of m possible types, $T_1, \ldots, T_m$, with the following discrete probability distribution:

$f_1(S_1|T_j)$ = probability (of having report S_1 | given that actual system is T_j)

The m possible classifications of weapon systems would be predetermined types that had been assessed to be likely or possible for the required time period. Of course, as in the previous models, there are many other possible calssifications suitable for use with this model; for example, a classification of "system unknown" could be added, or generic classifications, like ICBM, IRBM, etc., could be used.

Generally, one of the $f_1(S_1|T_j)$'s would be larger than 0.5 and considerably larger than most of the other $f_1(S_1|T_i)$'s. Otherwise the report would represent a source of questionable value. Indeed, many of the $f_1(S_1|T_i)$'s should be equal to ze -- indicating that there is no likelihood that a source of information would make this particular error. The required probabilities may be estimated as indicated in the discussion of the first model.

Let the prior distribution be denoted by $\pi_0(T_i)$. Then, using Bayes' theorem, we can calculate the posterior distribution:

$$h_1(T_i|S_1) = \frac{\pi_0(T_i)\, f_1(S_1\, T_i)}{\sum_{k=1}^{m} \pi_0(T_k)\, f_1(S_1|T_k)}$$

For each weapon system, or other category, predetermined probability levels, A_i and B_i, would be selected to satisfy the following conditions: If $h_1(T_i|S_1) < A_i$, it would be sufficiently improbable that the system would fall into category T_i; hence, that category could be eliminated from further consideration. On the other hand, if $h_1(T_i|S_1) > B_i$, then a sufficiently confident determination would have been made, and the correct weapon system would be considered identified.

If $A_i < h_i(T_i|S_1) < B_i$, then the $h_i(T_i|S_1)$'s would serve as the new prior distribution and an additional report would be used which would give probabilities $f_2(S_2|T_1), \ldots, f_2(S_2|T_m)$. These would be combined with the new prior distribution using Bayes' theorem, and the above decision algorithm would be applied again. This procedure could be continued until an identification could be determined. If there is a deadline for production of the estimate, and no decision is reached by the time that this deadline is immanent, it would be natural to use the classical Bayes decision algorithm to force a decision.

Since all of the necessary probabilities would have been calculated, this model could be easily implemented.

In this context it is interesting to observe that if the posterior distribution is used as a prior distribution for a subsequent computation, the resulting new posterior distribution based on the outcome of the new computation is the same as the posterior distribution that would have been obtained based on the combined data from all the observations and the initial prior distribution.

Thus if we let $f(S_1, S_2 | T_j) = f_1(S_1 | T_j) f_2(S_2 | T_j)$ (since we are assuming independence, which is not a necessary condition for this computation), then

$$h_1(T_j | S_1) = \frac{\pi_0(T_j) f_1(S_1 | T_j)}{\sum_{k=1}^{m} \pi_0(T_k) f_1(S_1 | T_k)}$$

Now, using h_1 as a prior distribution combined with the data S_2, we get the posterior distribution:

$$h_2(T_j | S_1, S_2) = \frac{h_1(T_j | S_1) f_2(S_2 | T_j)}{\sum_{k=1}^{m} h_1(T_k | S_1 f_2(S_2 | T_k)}$$

$$= \frac{\dfrac{\pi_0(T_j)f_1(S_1|T_j)f_2(S_2|T_j)}{\sum_{k=1}^{m} \pi_0(T_k)f_1(S_1|T_k)}}{\sum_{k=1}^{m} \dfrac{\pi_0(T_k)f_1(S_1|T_k)f_2(S_2|T_k)}{\sum_{j=1}^{m} \pi_0(T_j)f_1(S_1|T_j)}}$$

$$= \frac{\pi_0(T_j)f(S_1,S_2|T_j)}{\sum_{k=1}^{m} \pi_0(T_k)f(S_1,S_2|T_k)}$$

which is what we would have obtained by using the combined data from the two reports and the initial prior distribution.

Perhaps we should note that the different levels of confidence A_i and B_i for different types of weapon systems allow the model to reflect the fact that an error in determining the identity of one type of weapon system may be more costly than errors for other systems. These variable confidence levels fulfill a purpose that corresponds to the loss function in the other models.

For any mathematical system to be valuable, it must have the confidence of the user. If the intelligence consumers consistently believe that a system is giving unreliable information, they are not apt to make use of it.

The conceptual simplicity of the third model is of particular value in this regard. The other two models might work extremely well, yet because they seem to make consistently incorrect identifications, they might lose the confidence of an operator who was not well acquainted with the subtleties of decision theory. For example, suppose that the cost of mistakenly projecting a Foxbat development is considerably greater than the cost of mistakenly projecting a Flogger development -- perhaps because the effort required to counter the Foxbat is greater than that which would be required to counter the Flogger. Under these circumstances it would be desirable to design a loss function for the first two models such that any reasonable doubt as to which of these two aircraft would be developed would result in a projection of a Flogger development. While this procedure might be desirable, a user would soon come to the conclusion that the system simply did not work. But the concept of different confidence levels for different weapon systems, used in the third model, should be easier for the user to grasp.

One other obvious algorithm for production of uncertainty estimates is simply to eliminate possibilities as reports are received, in the hope that only one outcome will be left after processing a small number of reports. This is essentially the method used in the third model, since weapon systems with zero or sufficiently low probabilities are automatically eliminated from consideration.

C.5. POSSIBLE PROBLEMS WITH THE PRIOR DISTRIBUTION

Although the use of an a priori distribution seems generally appropriate, there are several problems that might arise in connection with the distribution that should be noted.

Prior estimations of the mix of enemy weapon systems certainly provide an important starting point for new estimates. It would be foolish to ignore this information, and all the models include the prior distributions in their computations. That is, it would be foolish not to take into account the fact that a given weapon system was very unlikely, if the source of the report concerning that system was known to be somewhat uncertain. Additional support for the use in the models of prior distributions from earlier estimates is the fact that as long as the prior probabilities are not too near to 0 or 1, they will not significantly affect the eventual outcome of the final probabilities.

Very small prior probabilities will, however, have a significant effect on the final result. Suppose, for example, that the prior probability of Soviet development of a particular type of aircraft is 0.01, and that two reports are received which independently indicate that the Soviets are indeed developing this type. Suppose further that both reports have a credibility of 0.95. Then $P(A|E_1) = 0.1610$ (the posterior probability after receiving the first report) and $P(A|E_2) = 0.7848$ (the posterior probability after receiving both reports). Even here, the model does not give a seriously

wrong result, but intuition tells us that two almost certain reports should indicate an almost sure projection.

Now suppose we use the first probability as the prior distribution for the second report, and discard the original prior. Then, after the above two reports, $P(A|E_2) = 0.9972$, which is closer to our intuitions concerning the probability that the projection is correct.

The essential point here is that the enemy is likely to do something unusual. It should be to their advantage to do so. Thus, if the prior probabilities are low, it might be well to use the probabilities from the first report as the prior distribution, or to run the calculations using both values, until confirmation or disconfirmation is obtained.

We noted above that, using the classical Bayes decision model, after a posterior distribution is calculated, this distribution is used as the new prior distribution for the next identification. In this way, the prior distribution would constantly be updated and reflect the actual mixture of weapons that could be expected during the period of the projection. This mixture, however, can be expected to change through time, as weapon systems become obsolete and are retired. The prior probabilities must be modified to reflect this process, in particular to reflect the need for constant revision of estimates of the mix of Soviet and other forces.

In Figure C.4., a summary of the advantages and disavantages of the three models is presented.

Figure 6-5 Comparison of the Models

	Bayes	Sequential	Modified Sequential
Advantages	o Very Good for Small number of reports o Good if data come simultaneously o Mathematically optimal	o Efficient in use of reports o Good if reports arrive at different times o Mathematically optimal	o Efficient in use of reports o Computationally simple o Good if reports arrive at different times o Conceptually simple o Shortest expected computational time
Disadvantages	o Computationally prohibitive for large numbers of reports o Must use all reports -- inefficient in use of reports	o Has additional parameters which may be hard to estimate o Highly inflexible	o May have to resort to Bayesian decision

C.6. AN ALTERNATIVE APPROACH

This subsection will present a detailed description of an alternative computer system for aggregating uncertainties. Material in this subsection has been adapted from John F. Lemmer, "A Return to Probabilities in Computer Assisted Medical Diagnosis," PAR, 1977.

We believe that the problems which have arisen in the development of systems for the aggregation of uncertainties can be successfully overcome, if the intended application is known and the system is developed for effective use in the application. The procedures to be described here are based on a probability model in which each node of a graph (Figure C-4) can be interpreted as an event, with the node at the top being the universal event. The sample space may be considered to be a set of weapon systems, with the random experiment to be the selection of a weapon system for which a projection is required.

The basic concept underlying these procedures is that of sequentially "extending" a marginal distribution to include additional variables. The complete distribution underlying the graph of Figure C-4 is a joint distribution over the set of events defined by the nodes of the graph, G = a, b, c, d, e, f, g, h, i, j. We shall denote this distribution by D(G). In extension, we proceed from a distribution having as a variable only the universal event to build gradually larger component marginal distributions (CMD) until the last distribution is D(G) itself.

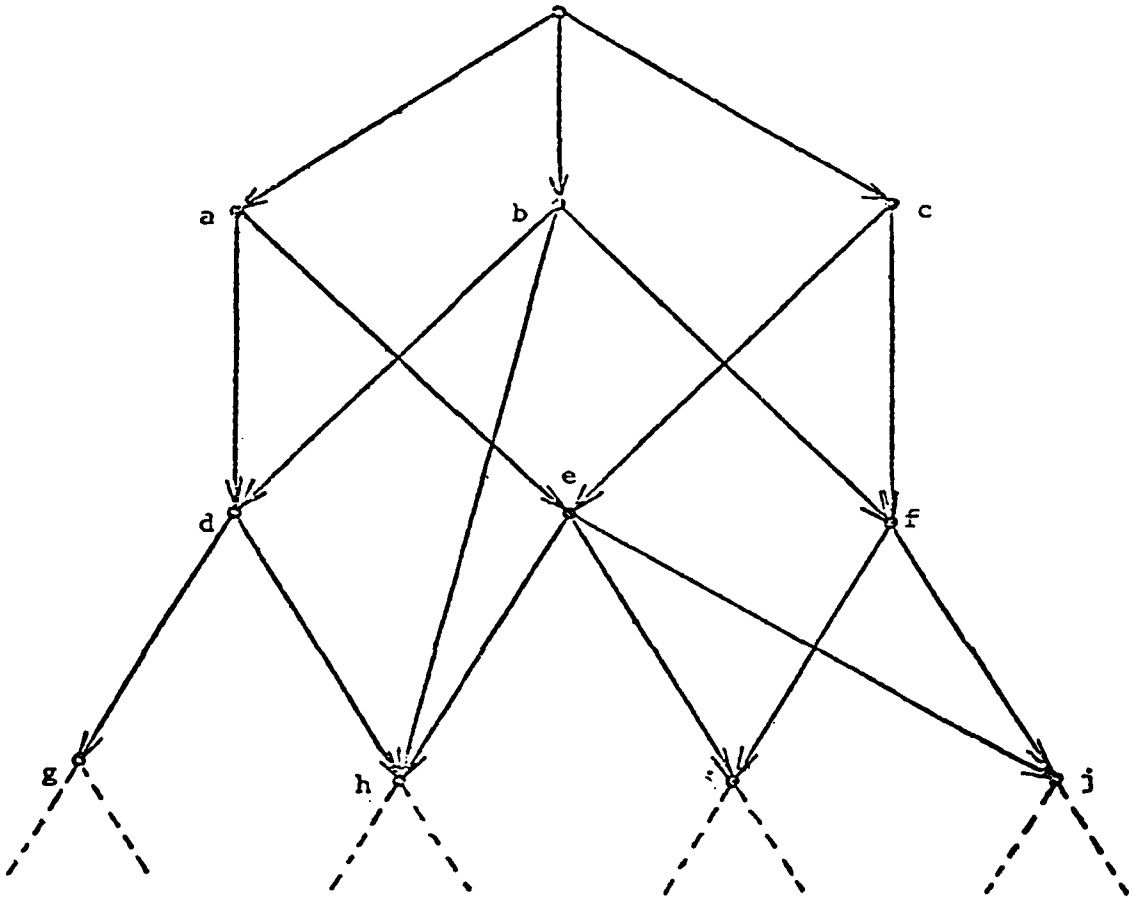


Figure C-4 Graph of Probability Model

We can start extension by including any one of the nodes in the first level of the graph. For example, assume that the value of p(a) is known. Then CMD ({a}) = [p($\overline{ab}$), p($a\overline{b}$), p($\overline{a}b$), p(a)].

We then continue in the sequence to CMD ({a, b, c}), CMD ({a, b, c, d}), etc. Difficulty arises if we do not know some of the values in the vector needed to specify the CMD, but we will return to this point later.

By this extensions approach we overcome the problem that most known probabilities are conditional probabilities. To see this, consider that in a cycleless graph, such as Figure C-4, each node lies at some maximum depth. For example, node h lies at maximum depth 3, although there is one path to it that has only length 2. By not extending a CMD to include any node at maximum depth ℓ until all nodes lying maximum depth ℓ-1 have been included in the CMD, we can assure that the conditional probabilities for nodes at depth ℓ can be converted to unconditional probabilities. For example, suppose for Figure C-4 we know p(i/e), p(i/f), and p(i/ef). Nodes e and f lie at a shallower maximum depth than i. Thus p(e), p(f), and p(ef) can be computed from CMD ({...e, f, ...}) and the conditioning removed by multiplication.

It remains to be shown how the extension process can be carried out. The method for extension will overcome the difficulties of both missing probability information and inconsistent probability information. The method also allows a subject matter specialist's intuition to be incorporated into the distribution. Thus the extension procedure overcomes three of the problems listed in Section 10.

Extension of a CMD to include another variable can be easily understood in terms of a linear programming problem though actual implementation by this method is too expensive to consider. Suppose we wish to extend CMD ({a,b}) to CMD ({a, b, c}). Suppose we know CMD ({a,b}) in terms of the set of four joint events $\{\bar{a}\bar{b}, \bar{a}b, a\bar{b}, ab\}$. Suppose $p(\bar{a}\bar{b}) = \nu_0$, $p(a\bar{b}) = \nu_1$, $p(\bar{a}b) = \nu_2$ and $p(ab) = \nu_3$. When we include event c, the set of joint events grows to 8 elements, namely $\{a\bar{b}\bar{c}, a\bar{b}c, \bar{a}\bar{b}\bar{c}, ab\bar{c}, \bar{a}\bar{b}c, a\bar{b}c, \bar{a}bc, abc\}$. The property that we desire in the extended CMD is that

$$
\begin{aligned}
p(\bar{a}\bar{b}) &= p(\bar{a}\bar{b}\bar{c}) + p(\bar{a}\bar{b}c) \\
p(a\bar{b}) &= p(a\bar{b}\bar{c}) + p(\bar{a}bc) \\
p(\bar{a}b) &= p(\bar{a}b\bar{c}) + p(\bar{a}bc) \\
p(ab) &= p(ab\bar{c}) + p(abc)
\end{aligned}
\qquad (1)
$$

If we consider the values of the probabilities for the joint events of the extended distribution to be the variables of linear programming problem, (1) is equivalent to the constraint set:

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{matrix}
p(\bar{a}\bar{b}\bar{c}) \\
p(a\bar{b}\bar{c}) \\
p(\bar{a}\bar{b}\bar{c}) \\
p(ab\bar{c}) \\
p(\bar{a}\bar{b}c) \\
p(a\bar{b}c) \\
p(\bar{a}bc) \\
p(abc)
\end{matrix}
=
\begin{matrix}
p(\bar{a}\bar{b}) \\
p(a\bar{b}) \\
p(\bar{a}b) \\
p(ab)
\end{matrix}
\qquad (2)
$$

If we know any marginal probability involving the event c, it will be a linear function of the vector on the left side of (2). For example, if we knew a value for p(ac), we would have

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{matrix} \qquad \begin{matrix} p(\bar{a}\bar{b}c) \\ p(a\bar{b}c) \\ p(\bar{a}b\bar{c}) \\ p(ab\bar{c}) \\ p(\bar{a}bc) \\ p(a\bar{b}c) \\ p(\bar{a}bc) \\ p(abc) \end{matrix} \quad = \quad p(ac) \qquad (3)$$

If there is a possibility that p(ac) is inconsistent with the rest of our data, or a subject matter specialist wants to guess at its value, we could treat (3) as the criterion function for the set of linear constraints (2). By maximizing and minimizing this criterion, we could determine the feasible range for (3). From this feasible range the expert could select one, or we could choose a value near our questionable value of p(ac). Then linear programming techniques allow us to find a feasible solution satisfying the initial constraints and giving the selected value for p(ac).

We can then repeat this step for each piece of probability information concerning c. Each time we repeat this step the old criterion function becomes the new member of the constraint set. Thus we eventually arrive at

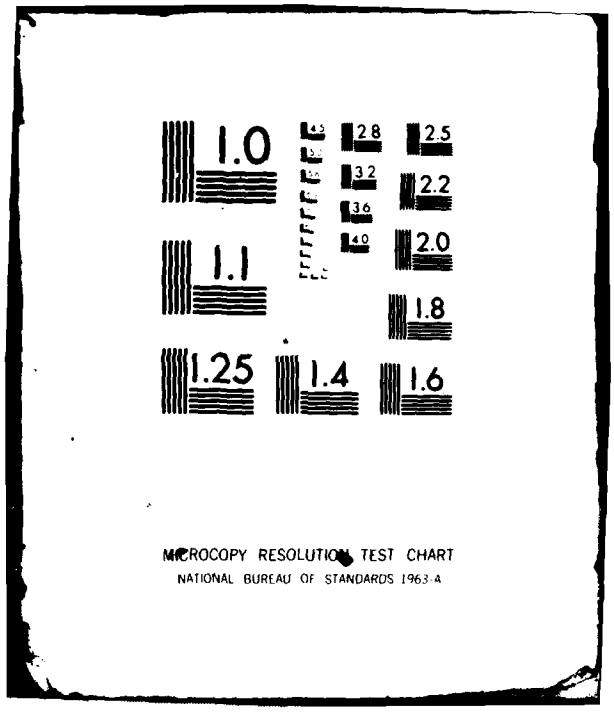
a feasible solution for CMD ({a, b, c}); that is probability values for $p(\overline{abc})$, $p(a\,\overline{bc})$, $p(\overline{a}b\overline{c})$, $p(ab\overline{c})$, $p(\overline{a}\overline{b}c)$, $p(a\overline{b}c)$, $p(\overline{a}bc)$, and p(abc). This solution staisfies all supplied probability information.

Thus by the above procedures we have used marginal information concerning c, which was originally conditional information, and which could have included intuition of a subject matter specialist. There has been no need to assume the reports independent. Since all weapon systems are in the joint distribution after updating, we have probabilities for multiple systems as well as single systems. And even though the distribution was incompletely specified, we have succeeded in estimating it. (The case where no information, even intuition, is available to choose a value for some constraints is covered fully by Dr. Lemmer's full presentation.) No specific amount of probability information is required. In fact, the methods are most applicable when some variables are not contained in <u>any</u> completely known component marginal distribution.

The only remaining problem is the size of the resulting distribution for real problems and the amount of computational effort required to estimate it. We have developed partial answers in this regard. Indeed the major portion of Dr. Lemmer's work is devoted to this.

He has shown that under certain general conditions there is no advantage to computing the complete underlying distribution. Thus, it is shown that, for Figure C-4, estimating CMD ({a, b, c, d, e, f}) and extending CMD ({d, e, f}) to CMD ({d, e, f, g, h, i, j}) is equivalent to estimating D

AD-A086 987 PATTERN ANALYSIS AND RECOGNITION CORP ROME NY F/G 15/4
AGGREGATING AND COMMUNICATING UNCERTAINTY.(U)
APR 80 J M MORRIS, R J D'AMORE F30602-78-C-0291
UNCLASSIFIED PAR-79-1 RADC-TR-80-113 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

({a, b, c, d, e, f, g, h, i}). Remembering that these are binary variables, this means that only $(2^6-1) + (2^7-2^3-1) = 182$ parameters must be estimated instead of 511. Such savings grow exponentially with the number of variables and the number of layers in the graph.

Efficient techniques have also been developed to replace the linear programming procedures discussed above. If N is the number of variables needed to specify the CMD, that is $N = 2^n$ where n is the number of events in a CMD, linear programming techniques can be shown to be approximately N^3 in computational complexity. We have developed techniques for solving the particular extension problems which are about NlogN in complexity. The price paid for this speedup is a fixed order of adding constraints during the extension process.

Thus, we have developed techniques overcoming all the original problems listed in Section 10. The actual procedures are too detailed and non-intuitive to be presented here, but they are fully explained in Dr. Lemmer's papers. They are quite easy to program and have been implemented, in part, by John Franko of Rutgers University. With some further development, we believe these techniques can be used to develop a computer-based system for aggregating uncertainties.