AD/A-003 260

SUBJECTIVE INTERPRETATION OF RELIABILITY AND ACCURACY SCALES FOR EVALUATING MILITARY INTELLIGENCE

Michael G. Samet

Army Research Institute for the Behavioral and Social Sciences

Prepared for:

Army Training and Doctrine Command

January 1975

Technical Paper 260

SUBJECTIVE INTERPRETATION OF RELIABILITY AND ACCURACY SCALES FOR EVALUATING MILITARY INTELLIGENCE

Michael G. Samet

SYSTEMS INTEGRATION & COMMAND/CONTROL TECHNICAL AREA

ari

U. S. Army

Research Institute for the Behavioral and Social Sciences

January 1975

Approved for public release; distribution unlimited.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Paper 260	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) SUBJECTIVE INTERPRETATION OF RELIABILITY AND ACCURACY SCALES FOR EVALUATING MILITARY INTELLIGENCE		5. TYPE OF REPORT & PERIOD COVERED Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Michael G. Samet		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 1300 Wilson Blvd., Arlington, VA 22209		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 20162101A754
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Training & Doctrine Command, Ft Monroe, VA		12. REPORT DATE January 1975 13. NUMBER OF PAGES 38
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

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

18. SUPPLEMENTARY NOTES

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

reliability	information processing
accuracy	Tactical Operations System (TOS)
scales	quantitative
intelligence	
computer	

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

This report is part of a broader effort to research human performance in required judgmental tasks and to determine ways of improving such performance, particularly as it affects data input to tactical intelligence information processing. The research reported here assessed the adequacy of source-reliability and information-accuracy rating scales for conveying intelligence information.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

i

20. Intelligence officers completed an original set and a replication of paper and pencil tasks which measured their attitudes toward and knowledge about the scales; recorded their judgments as to which report in each of 100 pairings of reports with different joint accuracy and reliability ratings was more likely to be true; asked them to estimate the probability that a report with a specific reliability rating would also carry a specific accuracy rating, and vice versa; and had them assign numerical values representing the probable truth of reports with given levels of reliability, of accuracy, and reliability-accuracy combinations. About one-fourth of the subjects treated reliability and accuracy as independent dimensions; the majority treated reliability as highly correlated with accuracy, and their judgment of a report's truth was influenced more strongly by its accuracy rating. Numerical (probabilistic) interpretations of scale levels were consistent within individuals but varied widely between them. Development of a new scale is suggested. The scale should require the assignment of a quantitative value which would reflect the likelihood of a report's being true and be based on all available information including the reliability of the source.

ARI Research Reports and Technical Papers are intended for sponsors of R&D tasks and other research and military agencies. Any findings ready for implementation at the time of publication are presented in the latter part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

The Intelligence Systems Work Unit Area within the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is concerned with problems of advancing and exploiting man/computer technology for improved tactical intelligence information processing. A major objective is to determine basic capabilities and limitations of man as an information processor and to devise complementary and compensatory processing aids and techniques for these capabilities and limitations. A specific requirement under this objective is to provide research findings whereby human performance in required judgmental tasks can be enhanced. The entire research effort is responsive to requirements of RDTE Project 20162101A754, "Intelligence Information Processing," FY 1974 Work Program and to special requirements of the U.S. Army Training and Doctrine Command and the Project Manager's Office, Army Tactical Data Systems.

The U.S. Army currently is developing systems intended to provide computer-based support for command and staff functions on the battlefield (e.g., TOS). The effectiveness of such systems for intelligence functions will be determined in part by the characteristics of the input data. The present publication describes one effort which confirmed the need for improved evaluation procedures for intelligence data and identified the direction of follow-on efforts most likely to satisfy this need.

J. E. Uhlaner

J. E. UHLANER
Technical Director

SUBJECTIVE INTERPRETATION OF RELIABILITY AND ACCURACY SCALES FOR EVALUATING MILITARY INTELLIGENCE

BRIEF

Requirement:

To assess the adequacy of and the relationship between the reliability scale (levels A-F) and accuracy scale (levels 1-6) by determining the quantitative meaning attached by intelligence officers to the various rating levels of each scale.

Procedure:

Thirty-seven intelligence officers completed an original and a replication of paper and pencil tasks constructed to measure their interpretation of the reliability and accuracy scales. The tasks measured officers' attitudes toward and their knowledge about the scales; recorded their judgments as to which report in each of 100 pairings of reports with different joint accuracy and reliability ratings was more likely to be true; asked them to estimate the probability that a report carrying a specific reliability-of-source rating would also carry a specific accuracy-of-information rating, and vice versa; had them assign numerical values representing the probable truth of reports with given levels of reliability, given levels of accuracy, and given reliability-accuracy combinations.

Findings:

Approximately one-fourth of the subjects treated reliability and accuracy as independent dimensions; the other three-fourths of the subjects treated the reliability rating as highly correlated with the accuracy rating. A subject's estimate of a report's truth was influenced much more by its accuracy rating than by its reliability rating. Numerical (i.e., probabilistic) interpretations of scales were relatively consistent within individuals, but such interpretations varied widely between them. Group attitude toward the scales as measured on a 6-point continuum, which ranged from "very adequate" to "very inadequate," produced a mean rating of only "slightly adequate."

Utilization of Findings:

The findings point out several inadequacies in use of the reliability and accuracy scales that appear to result from their qualitative nature and from frequent interdependence. The basis is provided for research to design and properly validate a new, less ambiguous, more sensitive system for communicating evaluations of intelligence data. One approach derived from the findings could require that a report have a single quantitative value assigned which reflects its likelihood of being true. This value would be based on all available information including the empirical reliability of the source.

SUBJECTIVE INTERPRETATION OF RELIABILITY AND ACCURACY SCALES FOR EVALUATING MILITARY INTELLIGENCE

CONTENTS

TABLES

SUBJECTIVE INTERPRETATION OF RELIABILITY AND ACCURACY SCALES FOR EVALUATING MILITARY INTELLIGENCE

INTRODUCTION

An essential initial operation in the intelligence processing cycle is the evaluation of information. According to the Army Combat Intelligence Field Manual 30-5[1], evaluation includes the determination of the pertinence of the information, the reliability of the source through which the information was derived, and the plausibility or truth, termed "accuracy," of the information. Reliability is judged mainly from previous experience and represents an estimate of the relative frequency of times that reports from a given source turn out to be true. Accuracy refers to the content of a report; it is not a measure of whether the information was reported accurately, but rather of the probability that the reported fact is true in light of all other available information.

The importance of obtaining reliable and valid ratings for each of these evaluative dimensions has been emphasized by the development of such automated systems as the Tactical Operations System (TOS) and the Integrated Battlefield Control System concept (IBCS), which will enhance capability to process and utilize data ratings. The present investigation concerns ratings of source reliability and information accuracy which commonly appear together on an intelligence spot report.

The standard rating system that has been widely used by the Army and other organizations, including NATO, indicates source reliability on a six-point alphabetically coded scale and information accuracy on a six-point numerically coded scale. The verbal labels associated with these codes are as follows:

Reliability of source:

A -- Completely reliable
B -- Usually reliable
C -- Fairly reliable
D -- Not usually reliable
E -- Unreliable
F -- Reliability cannot be judged

Accuracy of an item of information:

1 -- Confirmed by other sources
2 -- Probably true
3 -- Possibly true
4 -- Doubtfully true
5 -- Improbable
6 -- Truth cannot be judged

[1] Department of the Army Field Manual No. 30-5, Combat Intelligence. February 1971.

For example, information from a Fairly reliable source considered Probably true would have a joint rating of C2. A detailed description of the scales, together with the standards to be used in assigning ratings, is available in FM 30-5.

Although these six-point scales have been used operationally for many years, certain basic questions can be raised with respect to their intrinsic meaning to raters and to users of intelligence reports. Within each scale, to what extent do the respective ratings represent separate, non-overlapping categories? To what extent do the first five categories of each scale reflect a strictly monotonic function of the level of reliability or accuracy? For example, does an accuracy rating of 1 - Confirmed by other sources always imply greater accuracy than a rating of 2 - Probably true? If so, what about the case in which a few sources confirm each other by reporting the same item of highly improbable information? To what extent are ratings of accuracy made and perceived as independent of ratings of reliability, as they should be according to FM 30-5? How consistent are individual analysts across different situations in the way they interpret the scales? How well do different analysts agree upon the absolute and relative quantitative meaning of each of the individual and joint ratings? For example, what kind of agreement would be obtained on just how reliable--in terms of an absolute judgment--is a Not usually reliable source and how much more reliable it would be than an Unreliable source, or, does a report rated A3 carry more weight than one rated C1? Previous experimental studies and the present one address aspects of these questions.

BACKGROUND

Using the method of magnitude estimation scaling, Meeland and Rhyne[2] measured the amount of relative confidence that would be placed in reports bearing various joint reliability and accuracy ratings. The results allow for comparisons among the mean weights assigned to each of the 36 joint ratings and among those derived for each of the individual reliability and accuracy ratings. For example, a B1 rating was assigned six times as much confidence as a rating of F3, or the confidence in a source rating of A was equivalent to the sum of the confidence in ratings of C, D, and E. For each source rating, a nearly constant ratio between successive accuracy ratings was obtained, e.g., B1:B2 = B3:B4. Although this research offers useful information as to the relative interpretation of the various ratings, it offers no direct measurements of the absolute interpretation of each.

[2] Meeland, T., and R. F. Rhyne. A confidence scale for intelligence reports: An application of magnitude estimation scaling. Menlo Park, Calif.: Stanford Research Institute Technical Note 4923-31. June 1967.

Baker, McKendry and Mace[3] investigated the rating of spot reports in an Army TOS field exercise. Of all reports processed, only 48% contained ratings of both source reliability and information accuracy. Those that contained ratings tended to have the reliability and accuracy rating correlated and were generally confined to the high end of the scale; in fact, A1, B2, C3, D4, E5, and F6 made up 87% of all ratings, A1 and B2 comprised 80% of all ratings, and B2 alone comprised 74% of all ratings.

In a follow-on study by Baker and Mace[4], ratings assigned by officers enrolled in an intelligence course were found to differ from the school solution about 15% of the time for reliability and 49% of the time for accuracy. Moreover, no improvement in ratings was observed when they were made with the aid of a decision flow diagram (i.e., a programmed sequence of simple questions) designed to reduce the complexity of the rating procedure and guide the rater to the appropriate rating.

The studies by Baker suggested that real difficulties accompany the use of the reliability (A-F) and accuracy (1-6) scales. However, whether these problems are mainly due to basic inadequacies in the scales themselves or the inability of intelligence analysts to use them properly remains to be determined. The present investigation examines the structure of the scales, with particular emphasis upon subjective quantitative interpretation of the various ratings and combinations.

OBJECTIVE

The objective of this study was to determine the adequacy of the existing qualitative reliability and accuracy scales by measuring the performance of intelligence officers in: 1) knowledge about and attitude toward the scales; 2) judging which report of a pair with different joint accuracy and reliability ratings is more likely to be true; 3) estimate the probability that a report will carry a second suggested rating when only the reliability of source rating or the accuracy of information rating is known (i.e., perception of relationship of ratings); 4) assigning numerical probabilities for the likelihood that reports with given levels of reliability, given levels of accuracy, and given reliability/accuracy combinations would be true; and 5) intrasubject and intersubject consistency in giving a numerical (i.e., probabilistic) interpretation to the scales.

[3] Baker, J. D., J. M. McKendry, and D. J. Mace. Certitude judgments in an operational environment. ARI Technical Research Note 200. November 1968. (AD 681 232).

[4] Baker, J. D., and D. J. Mace. Certitude judgments revisited. ARI Technical Paper. In preparation.

METHODOLOGY

Subjects

The suojects were 37 Army captains nearing completion of the Military Intelligence Officers Career Course. All subjects were generally familiar with the scales under study.

About 60 officers participated in each of the two experimental sessions, although only 49 officers participated in both sessions. Of these 49, 12 had to be disqualified--and their data excluded from all analyses--on the basis of one or more incoherent response patterns. A reasonable assumption is that the final group of subjects was motivated to provide responses that accurately reflected their true personal judgments.

Procedure

At the beginning of the experiment, subjects were asked how long they had used reliability/accuracy ratings prior to entering the Intelligence Officer's course. Sixteen subjects had no prior experience, and 21 subjects had experience ranging from half a year to 4 years with a median of 1.1 years.

All subjects were required to complete paper-and-pencil forms, which will be referred to as Form 1, Form 2, etc., in each of two experimental sessions; the nature of each form and instructions for its use are described in separate sections. The initial session began with an introductory briefing (see Appendix), after which subjects completed Forms 1 and 2. Next, subjects were asked to review carefully a verbatim copy of the FM 30-5 section "Processing of Information, Evaluation," which concerns ratings of source reliability and information accuracy. They then proceeded to complete Forms 3, 4, 5, and 6, in that order. This order minimized the possibility that performance on one task would influence performance on a subsequent task. Subjects were permitted to refer to the evaluation section of FM 30-5 in responding to Forms 3, 4, 5, and 6. In the second experimental session, conducted two weeks later, subjects repeated Forms 3, 4, 5, and 6, in that order. They were then supplied with some feedback on group performance in the first session, and asked to repeat an attitude question (Form 1). A brief opinion questionnaire was completed by each subject at the experiment's conclusion.

Forms were distributed and collected one at a time. All subjects worked on identical forms simultaneously. However, Forms 4 and 6 consisted of two separate parts; in both sesssions, 19 subjects did part a first, then part b, while the 18 remaining subjects did part b first, then part a. Each session lasted a total of about two hours, including a ten-minute break between the administration of Forms 4 and 5.

As a measure of central tendency, both medians and means were computed for the data in Forms 4, 5, and 6. In all cases the medians closely approximated the means. Thus, for simplicity and because analyses of variance based upon the variance about the mean are reported, only the mean values are presented.

Form 1

Method. In Form 1 subjects were asked the following attitude question:

> Do you feel that the method used to rate information in terms of the reliability of the source (from A through F) and the accuracy of the information item (from 1 through 6) is (check one)
>
> ____________ very adequate
> ____________ moderately adequate
> ____________ slightly adequate
> ____________ slightly inadequate
> ____________ moderately inadequate
> ____________ very inadequate
>
> to meet Army intelligence-processing requirements?

Results. Responses were assigned scores of +3 for Very adequate to -3 for Very inadequate with the intermediate qualifiers receiving scores of +2, +1, -1, -2, respectively. Attitude scores were averaged separately for the groups with and without prior experience in using the scales. The mean score for the experienced group (1.29) was close to the mean of the inexperienced group (1.06). Subjects, on the average, considered the current scales to be only a little better than Slightly adequate.

Form 2

Method. Short quizzes on the meaning of the categories of the reliability and accuracy scales required subjects to fill in missing words from the FM 30-5 definitions of ratings. Form 2a covered reliability of source and agency as follows:

RELIABILITY
RATING   DESCRIPTION

A -- Completely   reliable

B -- ____________ ____________

C -- ____________ ____________

D -- Not ____________ ____________

E -- ____________

F -- ____________ cannot be ____________

In Form 2b, the scale concerning the accuracy of an item of information was presented to subjects as shown in the table here:

ACCURACY
RATING   DESCRIPTION

1 -- ______ by ______ ______

2 -- Probably true

3 -- ______ ______

4 -- ______ ______

5 -- ______

6 -- ______ cannot be ______

The subjects' task with respect to accuracy rating 1, for example, was to supply the three missing words: "confirmed," "other," "sources," respectively.

Results. Performance on both quizzes was combined to provide a single index of familiarity with the rating scales. For each completed blank, one point was scored if the word matched the word in the text exactly and one half point if it approximated the meaning of the textual word. Out of a total of 19 possible points, subjects averaged 12.22 or 64% correct responses. With the subjects divided according to experience in using the reliability and accuracy scales, the group with experience obtained a higher mean score (71%) than the group without experience (55%). Using the Kruskal-Wallis analysis of variance by ranks (i.e., H statistic),[5] the difference between these means was found to be statistically significant ($H(1) = 6.3$, $p < .02$). The results indicate that the subjects were relatively knowledgeable with regard to the verbal definitions associated with the scale categories; experienced subjects were slightly more knowledgeable than inexperienced subjects.

Form 3

Method. In Form 3, after review of the evaluation section of FM 30-5, subjects were required to make 100 comparative judgments of the type described in the following excerpt from the instructions.

> Suppose that you know that one of two camps, X or Y, is definitely going to be attacked by the enemy. Now suppose that you have two intelligence reports, one saying that camp X will be attacked and the other saying that camp Y will be attacked. The reports differ only in their respective ratings for the

[5] Siegel, S. Nonparametric statistics. New York: McGraw-Hill, 1956.

reliability of the report's source and the accuracy of the report's information. Assume further that the given reliability and accuracy ratings for each report are correct assessments of their actual reliability and accuracy. On the basis of this information alone, your task is to decide whether it is more likely that camp X will be attacked or that camp Y will be attacked.

Let us take an example: suppose that in the first problem the report that camp X will be attacked was assigned A3 and the report that camp Y will be attacked was assigned C1. On your sheet the problem would appear as follows:

	Report camp X		Report camp Y
(1)	A3	vs	C1

The subject indicated his decision response by circling the joint rating of one of the two reports (either A3 or C1 in the example).

If one of the joint ratings was assigned a higher reliability value, then the other joint rating had a higher accuracy value, or vice versa. This arrangement was true for each of the 100 problems, which were generated in the following way: Reliability ratings of A through E, designated R_1 through R_5 respectively, were combined in all possible ways with accuracy ratings of 1 through 5, designated A_1 through A_5 respectively, to yield 25 joint ratings of the form $R_i\ A_j$. Reliability rating F and accuracy rating 6 were excluded from this task because they do not specify a judgmental level of rating which can be meaningfully compared with the other ratings. Further, the assumption was made for the purposes of this task that the reliability ratings A through E and the accuracy ratings 1 through 5 each represent a strictly monotonic decreasing function with respect to reliability/accuracy; that is, if r_i represents the value assigned to R_i, and a_j, the value assigned to A_j, then $r_1>r_2>r_3>r_4>r_5$ and $a_1>a_2>a_3>a_4>a_5$. Thus, any pair $R_i\ A_j$ vs $R_k\ A_l$ was excluded if any of the following relationships held: (1) $i=k$; (2) $j=l$; (3) $i<k$ and $j<l$; (4) $i>k$ and $j>l$. Each of the 100 problems, therefore, involved a comparison of the form $R_i\ A_j$ vs $R_k A_l$ where either $i<k$ and $j>l$, or $i>k$ and $j<l$.

The 100 problems were presented in each of the two sessions in a five-page booklet compiled from computer printout sheets, with 20 problems to a page. Each subject received the 100 problems in a different random order, but this same order was maintained in both sessions. Before printing each problem, the computer determined randomly whether the joint rating on the left side of the page would have the higher reliability (between R_i and R_k) or the higher accuracy (between A_j and A_l); the set of left-right orientations was different for each subject but it was maintained for each across sessions.

Results. On each problem, it was determined whether the subject made his decision on the basis of a rating of higher reliability or higher accuracy. If the subject indicated that the event predicted by a report rated $R_i A_j$ was more likely to occur than the event rated $R_k A_l$, then one of the two following possibilities held: $i > k$ and the subject decided in favor of higher reliability, or $j > l$ and the subject decided in favor of higher accuracy.

For each subject, the percentage of decisions in each session based on a higher accuracy rating was computed. An average of 72.1% [standard deviation (SD) = 18.8] and 78.0% (SD = 18.7) of decisions were based on higher accuracy in the first and second sessions, respectively. The Spearman rank order correlation coefficient[6] between the percentage of decisions based on higher accuracy in the first and second sessions was statistically significant ($r_s = .74$, $t(35) = 6.5$, $p < .001$), indicating that subjects were reasonably consistent with respect to how often each decided on the basis of higher accuracy or reliability. Only one subject used higher reliability as a basis for responding to a majority of problems in both sessions. To summarize: in determining which of two reports was more likely to correctly predict an event, subjects showed a clear and consistent preference in favor of the report assigned a higher accuracy rating.

Form 4

Method. In Form 4, subjects were given the specific reliability (or accuracy) rating carried by a report and were asked to estimate the probability that the report should carry a specific accuracy (or reliability) rating. The subjects had the option to respond as if the two scales were independent or as if they were not. The task was constructed around a hypothetical situation as described in the following excerpt from the instructions for Form 4a.

> Suppose that you are one of the links in a real-time computerized intelligence processing system. You are stationed at a display terminal in an intelligence center and receive periodic spot reports from the field. With each report you are supposed to receive a source reliability rating and information accuracy rating. However, there is a bug in the computer system and only the reliability rating is coming through. On each problem you will be given the reliability rating carried by the report; in light of this rating your task is to estimate the probability that the report carried a given accuracy rating. You are to state the probability as any

[6] Siegel, 1956, op. cit.

> whole number between 0 and 100, inclusive: the higher the number you assign the greater you feel the probability is. This number can be thought of as your estimate of the number of chances in 100 that, given the known reliability rating, the report carried the given accuracy rating...
>
> A sample problem is the following: "Given that the reliability of the source was rated B, what is the probability that the accuracy of the information was rated 4?"...A report has come through whose source reliability was rated B. The information accuracy was also rated, but the rating has been lost in the processing system and you do not know what it is. What we are asking for is your estimate of the probability that the accuracy rating was actually 4 in light of the fact that the reliability rating was B.

In Form 4b, reliability and accuracy were interchanged, and the corresponding question was: "Given that the accuracy of the information was rated 4, what is the probability that the reliability of the source was rated B?"

Each of the six reliability ratings (A-F) was paired with each of the six accuracy ratings (1-6) to generate 36 problems each for Form 4a and 4b. Subjects were not required in either form to normalize their probabilities (P) across a given accuracy or reliability rating; that is, $\sum_{j=1}^{6} P(A_j/R_i)$ for any i did not have to equal 1 in Form 4a, and $\sum_{i=1}^{6} P(R_i/A_j)$ for any j did not have to equal 1 in Form 4b. The problems in each form were presented in a nine-page booklet compiled from computer print-out sheets, with four problems to a page. For each subject, the 36 pairs of ratings were put in a different random order; problems were presented according to that order in both Form 4a and 4b.

Results. Ten of the 37 subjects consistently responded as if ratings of reliability and accuracy were statistically independent, and all 10 assumed that each of the six reliability and accuracy ratings was equally likely to be observed. Half of the 10 subjects had experience using the scales. The independence assumption strictly conforms to the description of scale utilization in FM 30-5. The response strategy translates mathematically to: $P(A_j/R_i) = P(A_j) = 1/6$ and $P(R_i/A_j = P(R_i) = 1/6$ for any i and j. In other words, to the 10 subjects, knowledge of the rating assigned to one dimension of the report (either reliability or accuracy) provided no information with respect to its rating on the other dimension. Subjects responded this way on both forms or not at all. Thus, 27% of the 37 subjects in the study followed a procedure that supports independence of the accuracy and reliability scales.

The remaining 27 subjects (73%) responded to the task in a way that supports interdependence of the scales. None of these subjects normalized their response probabilities across a given reliability or accuracy rating. A Reliability (6) x Accuracy (6) x Form (2) x Session (2) x Subject (27) analysis of variance was performed on the probability data to assess the effects of form and session and the Reliability x Accuracy interaction. Neither the main effects of form and session nor any of the interactions with form and/or session as components were significant. These results suggest that subjects were consistent over sessions in assigning probabiiities, and were not significantly affected by the conditional probability orientation; i.e., for any i and j, the mean response for $P(A_j/R_i)$ was not significantly different from the mean response for $P(R_i/A_j)$.

The Reliability x Accuracy interaction effect proved highly significant, $F(25,650) = 60.1$, $p < .001$. Data were pooled over sessions and subjects to obtain mean values (each based on 54 data points) of the form $\overline{P}(A_j/R_i)$ and $\overline{P}(R_i/A_j)$ for Forms 4a and 4b, respectively. Each mean was then normalized by dividing by $\sum_{j=1}^{6} \overline{P}(A_j/R_i)$ in Form 4a and by $\sum_{i=1}^{6}$ $\overline{P}(R_i/A_j)$ in Form 4b. The normalized means are shown in Table 1. The nature of the Reliability x Accuracy interaction is clearly shown by the diagonal position of the modal normalized mean probability in each row. In other words, for any i, $\overline{P}(A_j/R_i)$ is highest when $j=i$; similarly, for any j, $\overline{P}(R_i/A_j)$ is highest when $i=j$. Furthermore, in departing from the diagonal (i.e., as the difference between i and j increases) within any row, the means fall off monotonically. For example, given that a report carried a reliability rating of C, subjects considered it most likely also to carry an accuracy rating of 3, and considered it more likely to carry a rating of 2 than 1, 4 than 5, and 5 than 6.

Form 5

Method. In Form 5, subjects were required to assign an estimated numerical probability to the likelihood of an event reported under each of the 25 joint evaluations generated by combining reliability ratings A through E with accuracy ratings 1 through 5 in all possible ways. The instructions to subjects gave the following example and qualifications:

> Suppose a report has come through that your camp will be attacked tomorrow and the report has been rated A1. What would you then say is the probability, or number of chances in 100, that your camp will indeed be attacked tomorrow? For each evaluation you must assign one and only one number. If you should feel that the probability under certain evaluations should assume a range, then assign a probability in the middle of that range; in short, the number you assign should be your one best estimate for the probability.
>
> ...Reliability rating F and accuracy rating 6 have been omitted since these ratings indicate that the reliability/accuracy cannot be judged--therefore no numerical value could possibly be assigned for them.

Table 1

NORMALIZED MEAN CONDITIONAL PROBABILITIES (FORM 4)

Normalized Mean Conditional Probability for Accuracy Rating Given Reliability Rating

		1	2	3	4	5	6
Given Reliability Rating	A	.28	.25	.20	.12	.08	.07
	B	.23	.25	.20	.13	.10	.09
	C	.19	.21	.23	.14	.12	.11
	D	.14	.14	.18	.21	.19	.14
	E	.10	.12	.14	.21	.24	.19
	F	.09	.11	.13	.19	.21	.27

Normalized Mean Conditional Probability for Reliability Rating Given Accuracy Rating

		A	B	C	D	E	F
Given Accuracy Rating	1	.26	.25	.20	.13	.08	.08
	2	.22	.24	.21	.13	.11	.09
	3	.16	.20	.24	.16	.13	.11
	4	.12	.13	.17	.21	.20	.17
	5	.08	.10	.15	.21	.25	.21
	6	.08	.10	.14	.18	.20	.30

Note. Modal probability within each row is circled.

On a single response sheet, the combined ratings were arranged systematically in a 5-row X 5-column matrix. Rows represented the 1-5 accuracy categories and columns showed the A-E reliability categories. The verbal descriptors for the reliability and accuracy ratings were listed on the bottom of the response sheet for reference.

Results. A Reliability (5) X Accuracy (5) X Session (2) X Subjects (37) analysis of variance was performed on the probability assignment data. As indicated in the summary of results (Table 2), the significant sources of variance (at the $p < .001$ level) were: reliability, accuracy, and Reliability X Accuracy. An estimated-variance component analysis[7] showed that 4.5 times more response variance could be attributed to accuracy than to reliability, and that about 37 times more variance could be attributed to reliability than to the Reliability X Accuracy interaction.

Probability assignments were pooled over sessions and subjects, and the mean (based on 74 values) for each joint rating is given in Table 3. Across ratings, the average range in assignments was .45; the range was narrowest for A1 (.88 to 1.00) and widest for E1 (.35 to 1.00). Inspection of the means reveals a pattern characterized by the greater weight attached to the accuracy rating. Holding reliability (i) constant, the decline in mean assigned probability from R_iA_1 to R_iA_5 averaged across reliability ratings is .61 (or .15 per level in rating). Holding accuracy (i) constant, the decline in mean assigned probability from R_AA_i to R_EA_i averaged across accuracy ratings is .28 (or .07 per level in rating). In terms of the probabilistic impact of a report, therefore, a decrease in two levels of reliability rating had about the same effect as a decrease in one level of accuracy rating.

Rank order of probability assignments to the 25 joint accuracy and reliability ratings was used to evaluate consistency between subjects. Correlations for rank order were computed between all 666 possible pairings of subjects using their mean assignments across the two sessions. The mean correlation coefficient was .89 (SD = .10) with a range of .46 to 1.00, indicating a very high degree of intersubject agreement for ranking.

Form 6

Method. In Form 6, subjects were required to assign an estimated numerical probability to the likelihood of an event reported under each reliability rating (A-E) appearing by itself and an event under each accuracy rating (1-5) appearing by itself. The reliability ratings were presented in Form 6a, and subjects were given the following example in the instructions.

[7] Vaughan, G. M., and M. C. Corballis. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, 1969, *72*, 204-213.

Table 2

SUMMARY OF ANALYSIS OF VARIANCE ON PROBABILITIES ASSIGNED TO JOINT RATINGS (FORM 5)

Source of Variance	Degrees of Freedom	Mean Square	F-ratio
Between subjects	36		
Subjects within groups (S)	36	.2525	
Within subjects	1813		
Reliability (R)	4	4.7515	123.17*
R x S	144	.0386	
Accuracy (A)	4	21.8247	311.32*
A x S	144	.0701	
Session (E)	1	.0157	.08
E x S	36	.1868	
R x A	16	.0382	5.86*
R x A x S	576	.0065	
R x E	4	.0130	1.02
R x E x S	144	.0127	
A x E	4	.0163	.51
A x E x S	144	.0321	
R x A x E	16	.0041	.75
R x A x E x S	576	.0054	
Total	1849		

*p < .001.

Table 3

MEAN PROBABILITIES ASSIGNED TO JOINT RATINGS (FORM 5)

		Reliability of Source (i)				
		A	B	C	D	E
Accuracy of Report (j)	1	.96	.92	.87	.81	.75
	2	.86	.81	.74	.64	.56
	3	.74	.67	.60	.48	.40
	4	.55	.48	.42	.32	.24
	5	.38	.31	.25	.19	.14

> Suppose a report has come through that your camp will be attacked tomorrow and the reliability of the source was rated A but no rating was made for the accuracy of the information. (Note that the fact that accuracy was not rated at all does not mean that the accuracy rating should be 6 -- it means only that the report was submitted without any accuracy rating.) What would you then say is the probability, or number of chances in 100, that the camp will be attacked tomorrow?

The accuracy ratings were presented in Form 6b with similar instructions.

Ratings in each form were arranged in descending order (A-E/1-5) on a single response sheet, and the associated verbal descriptor was given alongside each rating (e.g., A - Completely reliable).

<u>Results.</u> A Reliability (5) x Session (2) x Subject (37) analysis of variance performed on Form 6a data resulted in a highly statistically significant effect for reliability rating (F(4,144) = 483.38, p < .001); and an Accuracy (5) x Session (2) x Subject (37) analysis of variance on Form 6b data resulted in a highly statistically significant effect for accuracy rating (F(4,144) = 465.91, p < .001). In neither analysis was session significant.

Probability assignments were pooled over sessions and subjects; the mean, range, and standard deviation for each rating is given for reliability and for accuracy in Table 4. The mean probabilities assigned to the scale values described a linear trend that was statistically significant for both reliability ($F(1,144) = 25.75$, $p < .001$) and accuracy ($F(1,144) = 25.02$, $p < .001$). With the exception of the highest rating, the accuracy ratings were assigned a mean probability that closely approximates successive multiples of .20. To a lesser extent, the means for the reliability ratings show a similar pattern.

A wide amount of disparity in intersubject numerical interpretations of reliability and accuracy scales was shown by the range and standard deviation for the probabilities assigned to each rating. In addition, as the degree of reliability and accuracy decline, the changes in standard deviation generally indicate that subjects' numerical interpretations of the ratings become more divergent.

Intrasubject Comparison (Form 5 vs Form 6)

Multiple linear regression analyses were performed to determine the degree to which the probabilities assigned to the joint ratings (Form 5) could be predicted from a linear combination of the probabilities assigned to the individual reliability and accuracy ratings (Form 6). An analysis was performed for each subject using his mean data pooled across the two sessions. For the 37 subjects, the individual ratings accounted for an average of 93.1% of the variance of the joint ratings, 71.3% being attributed to accuracy and 21.8% to reliability.

Actually, the individual impact of the accuracy rating was stronger than that of the reliability rating in determining how 33 of the 37 subjects interpreted the joint ratings. Reliability accounted for more explained variance than did accuracy for only two subjects; reliability and accuracy accounted for the same percentage of variance (50%) for two other subjects.

Performance Feedback

Method. After all subjects had completed Form 6 during the second experimental session, they were presented with summary feedback about the overall group performance on Forms 4, 5, and 6, based upon data collected in the first session from 60 subjects. The feedback data were presented in tables similar in both content and structure to Tables 1, 3, and 4.

Each data table was displayed before the entire group on a view-graph screen and also was given to each subject on a mimeographed sheet. The tables were all thoroughly explained and questions from subjects were answered. Subjects were informed that the data from those who responded to Form 4 as if ratings of reliability and accuracy were totally independent were not used in the summary analysis. The feedback presentation lasted about 15 minutes.

Table 4

MEAN, RANGE, AND STANDARD DEVIATION FOR PROBABILITIES ASSIGNED TO INDIVIDUAL RATINGS (FORM 6)

Rating	Verbal Descriptor	Mean	Range	Standard Deviation
Reliability				
A	Completely reliable	.86	.65-.99	.11
B	Usually reliable	.73	.55-.90	.12
C	Fairly reliable	.57	.40-.80	.13
D	Not usually reliable	.36	.15-.70	.15
E	Unreliable	.18	.05-.53	.15
Accuracy				
1	Confirmed by other sources	.93	.70-1.00	.09
2	Probably true	.79	.53-.90	.10
3	Possibly true	.61	.40-.80	.14
4	Doubtfully true	.38	.15-.65	.16
5	Improbable	.21	.03-.53	.16

Immediately after receiving the feedback, subjects were asked to respond again to the attitude question used in Form 1. This was done to obtain a measure of their change in attitude toward the currently used scales as a function of participating in the experimental tasks as well as being exposed to the feedback. They were not specifically told that the form was being repeated, nor were they directed to respond in accordance with any impressions obtained from the feedback.

Results. The mean response score for all subjects was .46 on the second administration of Form 1 compared with 1.18 at the beginning of the experiment. Although the overall attitude change toward the scales was significant (Wilcoxin[8] test: $Z = 2.40$, $p < .02$), it moved only from just above to just below "slightly adequate."

Questionnaire

At the conclusion of the second session, after all experimental data had been collected, subjects were asked to complete a brief questionnaire. The questions, together with the number of subjects who answered each response alternative, are listed below.

1. Do you feel that any problems that might accompany the use of the A-F reliability scale and 1-6 accuracy scale are mainly due to?

6 inadequacy of the scales themselves

31 inability of intelligence officers to correctly assess and interpret the ratings

2. Which of the following is usually easier to rate?

18 source reliability

19 information accuracy

3. If you were given an intelligence report and could choose between knowing either the reliability of the source or the accuracy of the information, which would you choose (in other words, which is more important)?

2 source reliability

35 information accuracy

4. When making a source-reliability rating and an information-accuracy rating for a single spot report, do you feel that you can truly make the two judgments independently?

13 yes 24 no

5. Do you feel that the currently used double-dimension rating scale (reliability and accuracy) should be replaced by a single-dimension scale (for example, likelihood that the report is correct)?

21 yes 16 no

[8] Siegel, 1956, op. cit.

6. If a single-dimension scale is designed, do you feel that it should employ ratings in terms of:

__8__ verbal descriptors only?

__11__ numbers (like probabilities) only?

__18__ both verbal descriptors and numbers?

Although the subjects who participated in the questionnaire had equally divided opinions about whether reliability or accuracy was easier to rate, nearly all considered the availability of the accuracy rating to be the more important. Two-thirds of the subjects felt that they could not truly make the judgment of reliability and accuracy independently, and a majority of the subjects favored the proposal of developing a new single-dimension scale.

DISCUSSION

Subjects viewed the source reliability and information accuracy scales (Form 1) as "slightly adequate" for rating information to meet Army intelligence processing requirements. However, more than 80% of the group favored the view that problems accompanying the scales are due to the inability of intelligence officers to correctly assess and interpret the ratings. This view is typical of many experienced intelligence officers who became indoctrinated under the present system. Even subjects who had experience with the scales showed only 71% familiarity with them according to a quiz (Form 2). Yet, most of the data from the present experiment and other related research [9, 10, 11] point to inadequacies of the scales themselves. The ambiguity and insensitivity of the reliability and accuracy scales, as well as the interdependence of the two, are largely functions of their intrinsic qualitative nature and structure.

A majority of subjects (as well as many analysts and researchers) do not perceive the reliability and accuracy ratings as independent of each other. The expectations of subjects about how frequently each accuracy rating would be a concomitant of each reliability rating (Form 4) are consistent with empirical data [12] showing how ratings actually were

[9] Baker et al., 1968, op. cit.

[10] Kelly, C. W., and C. R. Peterson. Probability estimates and probabilistic procedures in current-intelligence analysis. Gaithersburg, Maryland: International Business Machines Corporation. Report 71-5047 January 1971.

[11] Baker and Mace, in preparation, op. cit.

[12] Baker et al., 1968, op. cit.

juxtaposed in a field exercise. It is quite reasonable that most subjects perceive the level of accuracy to be highly correlated with the level of reliability because the expected accuracy of a given report is equivalent to the reliability of its source when no other information is available. In addition, source reliability is typically only a means to the goal of ascertaining accuracy. In the present study, the results are no surprise when the measured influence and judged importance of accuracy ratings exceeded that of reliability ratings in every comparison (Form 3; Form 5; Form 5 vs Form 6). In summary, the assumption that reliability and accuracy ratings are generally independent is not supported on logical or empirical grounds.

Another problem with the scales is the wide difference of opinion with respect to the absolute level of probability suggested by each rating, as reflected by the sizable ranges and standard deviations for responses. For example, assigned probabilities (Form 6) for both Fairly reliable and Possibly true ranged from .40 to .80. Large ranges and standard deviations have also been obtained in other studies of encoding verbal expressions into probabilities.[13,14,15,16] In contrast, the rank ordering of ratings in Form 5 indicated much intersubject agreement according to a correlational analysis. Further, Meeland and Rhyne[17] found very high intercorrelations for relative confidence assignments between different groups of subjects including collectors, analysts, and users of intelligence information. Although intersubject agreement on the meaning of one rating relative to another is encouraging, the large variability in the numerical interpretation of each rating raises doubts about the effectiveness of the qualitative rating scales to communicate specific levels of judgment.

[13] Lichtenstein, S., and R. J. Newman. Empirical scaling of common verbal phrases associated with numerical probabilities. Psychonomic Science, 1967, 9, 563-564.

[14] Levine, J. M., and D. Eldredge. The effects of ancillary information upon photointerpreter performance. ARI Technical Paper 255. September 1974. (AD 785 706)

[15] Kelly and Peterson, 1971, op. cit.

[16] Johnson, E. M. Numerical encoding of qualitative expressions of uncertainty. ARI Technical Paper 250. December 1973. (AD 780 814)

[17] Meeland and Rhyne, 1967, op. cit.

In addition to other problems, the rating scales lack sensitivity in grading the degree of reliability and accuracy. The results for Form 6 indicate that subjects were able to make a clear quantitative distinction between the five meaningful levels of each scale. However, the average size of the difference between mean probabilities assigned to adjacent levels (.175) suggests that there is room for finer discriminations, which are not permitted by only five categories. Research has shown, for example, that more information can be transmitted by a rating scale that uses nine categories rather than five.[18] The sensitivity of the scales is further reduced by the availability of the ratings of F and 6 which allow the analyst an easy way out of an especially difficult evaluation.

IMPLICATIONS

The results, in general, show that several difficulties exist in subjective interpretation of the currently used rating scales for source reliability and information accuracy. A suggested effort is to design and validate a new, more effective system to communicate evaluations of intelligence data. Findings of the present study indicate that the two-dimensional evaluation should be replaced because: 1) the accuracy rating dominates the interpretation of a joing accuracy and reliability rating and 2) there is frequently an undeniable correlation between the two scales.

A single evaluation of an intelligence report could be rated in terms of its likelihood of being borne out by truth or reality. A specific rating would 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; temporal and spatial factors; etc. This likelihood rating would be associated with the report and used in subsequent data communication and processing. This scheme does not decrease, by any means, the benefit to be derived from formally maintaining an empirically determined estimate of a source's reliability and continuously revising it according to accumulated evidence. This reliability rating would be available to serve in data-collection management. Furthermore, in the absence of information other than the identity of the source, the analyst/system could rely upon that source's latest reliability index as a best estimate for the probable truth for a given report. It might also be feasible and desirable to maintain, and selectively utilize, separate reliability ratings for different categories of information furnished by the same frequently used source.

The use of a well-defined numerical scale to express probable truth should substantially reduce ambiguity in communicating intelligence evaluations, and offer other practical advantages. Given a basic understanding

[18] Bendig, A. W. Transmitted information and the length of rating scales. Journal of Experimental Psychology, 1954, 47, 303-308.

of probability, individual differences have little plac , for example, in an interpretation of the statement: "the probability that camp X will be attacked tomorrow is .70." When numbers rather then words are used in probability statements, differences in interpretation due to context[19] might be less likely to arise. NATO military analysts have stated additional cogent reasons for switching to a quantitatively oriented rating system:[20] it would be applicable to the current manual procedures, compatible with future automated procedures, and equally comprehensible across language barriers. Compatibility with automated procedures would also enable quantified likelihoods to be directly input into computerized models designed to process and make inferences from probabilistic information; a scale sensitive enough to allow quantification of extremely low or high likelihoods would be particularly useful in such applications. Preliminary research has shown that a quantitative scale can be successfully used to rate photointerpreter reports[21] and can be accepted by sophisticated intelligence analysts.[22]

An initial step in designing such a rating system would be to determine in which form the likelihoods might best be expressed. Two candidates that show promise are a probability scale (e.g., 0.00 to 1.00 or 0% to 100%) and some kind of odds scale. Another possibility worth investigating is whether a careful assignment of verbal annotations to certain numerical values on the scale can be beneficial. Extensive laboratory and field evaluations of any new rating system devised, including comparisons with the old system, would be desirable before it could be accepted and implemented.

[19] Johnson, 1973, op. cit.

[20] Letter, from NATO Assistant Chief of Staff for Intelligence to Military Agency for Standardization OTAN/NATO, Autoronte Brussels/Zaventem B-1110, Brussels 39, Belgium, [MAS (Army) (69) 559], dated 20 February 1970, subject: Proposed Agenda Item for Next Meeting of the Intelligence Procedures Interservice Working Party (NU).

[21] Samet, M. G. Checker confidence statements as affected by performance of initial image interpreter. ARI Technical Research Note 214. September 1969. (AD 700 127)

[22] Kelly and Peterson, 1971, op. cit.

However, 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 percent probability, x percent should turn out to be true). In confronting these difficulties, full advantage should be taken of potentially effective training procedures, including the provision of performance feedback, that rely upon interactive computer aids.[23] For example, the computer might prove helpful in guiding the analyst to acquire the information necessary to make an appropriate likelihood judgment, and in facilitating the encoding of that judgment into a numerically stated likelihood devoid of response bias.

[23] Samet, M. G. Computer-controlled differential review-time payoff as a training aid. In Proceedings 16th Annual Meeting Human Factors Society, October 1972, pp. 374-376.

REFERENCES

Baker, J. D., and D. J. Mace. Certitude judgments revisited. ARI Technical Paper. In preparation.

Baker, J. D., J. M. McKendry, and D. J. Mace. Certitude judgments in an operational environment. ARI Technical Research Note 200. November 1968. (AD 681 232)

Bendig, A. W. Transmitted information and the length of rating scales. Journal of Experimental Psychology, 1954, 47, 303-308.

Department of the Army Field Manual No. 30-5, Combat Intelligence. February 1971.

Johnson, E. M. Numerical encoding of qualitative expressions of uncertainty. ARI Technical Paper 250. December 1973. (AD 780 814)

Kelly, C. W., and C. R. Peterson. Probability estimates and probabilistic procedures in current-intelligence analysis. Gaithersburg, Maryland: International Business Machines Corporation Report 71-5047. January 1971.

Letter, from NATO Assistant Chief of Staff for Intelligence to Military Agency for Standardization OTAN/NATO, Autoronte Brussels/Zaventem B-1110, Brussels 39, Belgium, [MAS (Army) (69) 559], dated 20 February 1970, subject: Proposed Agenda Item for Next Meeting of the Intelligence Procedures Interservice Working Party (NU).

Levine, J. M., and D. Eldredge. The effects of ancillary information upon photointerpreter performance. ARI Technical Paper 255. September 1974. (AD 785 706)

Lichtenstein, S., and R. J. Newman. Empirical scaling of common verbal phrases associated with numerical probabilities. Psychonomic Science, 1967, 9, 563-564.

Meeland, T., and R. F. Rhyne. A confidence scale for intelligence reports: An application of magnitude estimation scaling. Menlo Park, Calif.: Stanford Research Institute Technical Note 4923-31. June 1967.

Samet, M. G. Checker confidence statements as affected by performance of initial image interpreter. ARI Technical Research Note 214. September 1969. (AD 700 127)

Samet, M. G. Computer-controlled differential review-time payoff as a training aid. In Proceedings 16th Annual Meeting Human Factors Society, October 1972, pp. 374-376.

Siegel, S. *Nonparametric statistics*. New York: McGraw-Hill, 1956.

Vaughan, G. M., and M. C. Corballis. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, 1969, *72*, 204-213.

INTRODUCTORY BRIEFING TO SUBJECTS

INTRODUCTORY BRIEFING TO SUBJECTS

The advent of the automated Tactical Operations Systems (TOS) and the Integrated Battlefield Control System (IBCS) concept will enhance capability to process and utilize ratings for the reliability of a given source and for the accuracy of the information that it provides. Such ratings when validly made materially facilitate subsequent processing and contribute to the adequacy of the overall intelligence estimate; they can be useless or even degrade processing when poorly made.

The method used by the Army to make data evaluations, as given in FM 30-5, is to rate source reliability on a scale from A through F and to rate information accuracy on a scale from 1 through 6. The purpose of this study is to investigate in detail, the nature and properties of these two scales and to determine how they are understood and interpreted by intelligence officers. The results of the study promise to benefit personnel who might make or use spot reports; G2 staffs would benefit directly as would the users of their product indirectly. At the completion of our experimental sessions, we will be glad to give you more details and to answer any questions that you might have.

To make the results of this study meaningful, it is important that you make every effort to respond as honestly as you can. The data from this study will be used strictly for research; your individual performance will never be associated with your name nor be used to bear personal consequences

Although all of you will complete the same tasks, please note that all of you will not always be working on the same task at the same time. Also, within each task the order of presentation of problems will often be different for each of you.

We will occasionally pace you by announcing the approximate amount of time you will have to complete a task.

Please remember to fill in your identification on each response form.

Preceding page blank