DOCUMENT RESUME

ED 093 995 TM 00? 851

AUTHOR Shuford, Emir H., Jr.
TITLE The Student as an Assessor of Uncertainty: Some Statistical Measures Useful for Feedback to the Student.
INSTITUTION Rand Corp., Santa Monica, Calif.
PUB DATE [Apr 74]
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974); Some charts may have marginal legibility

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS Computer Programs; *Confidence Testing; *Feedback; Prediction; *Probability; *Response Style (Tests)
IDENTIFIERS *Decision Theoretic Testing

ABSTRACT

A dicussion is provided of some statistical measures and graphical information that, when used as feedback to the student, facilitates his ability to assess his own uncertainty. These measures and graphs, which result from the application of least squares analysis and information theory to decision-theoretic testing, provide the student with the capability to compare perceived information with actual information. The possibility of improving his ability to communicate uncertainty using the language of probability is discussed. (Author/RC)

THE STUDENT AS AN ASSESSOR OF UNCERTAINTY
SOME STATISTICAL MEASURES USEFUL FOR FEEDBACK TO THE STUDENT

Emir H. Shuford, Jr.
The Rand Corporation

Paper presented at the 1974
Annual Meeting of the
American Educational Research Association
April, 1974

THE STUDENT AS AN ASSESSOR OF UNCERTAINTY
SOME STATISTICAL MEASURES USEFUL FOR FEEDBACK TO THE STUDENT

For the past year and a half, I have been using computers to administer decision-theoretic tests. By using a graphics terminal and by exploiting the nearly instantaneous analytical capabilities of the computer, a student can be provided with an environment for understanding the nature of decision-theoretic testing and, possibly, for improving his ability to communicate uncertainty using the language of probability.

Figure 1 shows a three-alternative multiple-choice test item as it appeared on the screen of the graphics terminal. The subject responds by touching a light pen anywhere within or on the edges of the triangle. For any response as indicated by the "X", the computer displays the possible item scores based on a truncated logarithmic scoring system (Shuford, Albert & Massengill; 1966).

Each point on the triangle corresponds to a probability distribution over the three answers as illustrated by Figure 2. The subject can change his response any number of times and when he is satisfied with the set of possible scores he can see the correct answer to the question as shown in Figure 3. Before moving on to the next question, the subject sees a cumulative graph of his test score up to now.

Upon completing a test of from 15 to 20 items, the subject sees an analysis of his test performance as illustrated by Figure 4. Much of this analysis is based upon an evaluation of the external predictive

validity of the subject's responses. A subject is, in effect, making probabilistic predictions as to which answer will be judged correct. If a subject had responded to a very large number of test items, we could do an analysis such as that shown in Figure 5. For this subject, the differences between the observed relative frequency and the ideal proportion indicated by the dashed line can be attributed to sampling fluctuations so we can conclude that he is unbiased in his use of probabilities. [Strictly speaking, the probabilities should be treated as triplets as in Shuford & Brown (1974).]

By assuming that the relation between relative frequency and probability as used by a subject can be approximated by a linear function and by using a least-squares estimation procedure (Brown & Shuford; 1973: Sibley; 1974: Shuford & Brown; 1974) it is possible to make inferences about a subject's bias from much less data than that used in the above. Figure 6 illustrates two linear fits -- one for a subject (I) who undervalues his information, the other for a subject (II) who overvalues his information. These functions are used to eliminate the bias from a subject's responses by deriving a new set of revised probabilities, e.g., whenever subject II stated that the probability of an answer being correct was one, the revised probability would be changed to match the relative frequency of .85.

These revised probabilities are used to compute a new test score which, if the subject is biased, will be larger than his original test score. The difference between these two scores is the basis for the

statement -- "YOU CAN IMPROVE YOUR SCORE BY 37 POINTS BY MORE REALISTIC USE OF YOUR KNOWLEDGE." -- shown in Figure 4. The difference between this new test score and a perfect score is the basis for the statement -- "YOU CAN IMPROVE YOU SCORE BY 224 POINTS BY MORE STUDY."

These revised probabilities are use also to estimate the actual amount of information (Shannon & Weaver; 1949) the subject possesses with respect to the test. This absolute measure is rescaled and displayed as "ACTUAL KNOWLEDGE" as shown in Figure 4. "PERCEIVED KNOWLEDGE" is, of course, computed using the original probabilities given by the subject. The subject in Figure 4 undervalues his knowledge because his test performance indicates he actually possesses more information than he thinks he does.

This analysis in terms of actual vs. perceived information, as shown in Figure 7, is closely related to the old Arabian proverb --

He who knows, and knows that he knows,
He is wise, follow him.
He who knows, and knows not that he knows,
He is asleep, awaken him.
He who knows not, and knows not that he knows not,
He is a fool, shun him.
He who knows not, and knows that he knows not,
He is a child, teach him.

What will actually happen when people are allowed to express their knowledge in terms of probabilities? Hopefully, we will find wise

men and children, possibly a few sleepers, but certainly no fools. I wish I could give a definitive answer to this question. All I have are sometentative but suggestive results reinforced, fortunately, by some of the findings that Dave McMullen will report later in this symposium.

At Rand we have demonstrated, and tried out, computer-administered decision-theoretic testing to many different people using as sample tests Reader's Digest vocabulary tests; Humanities, Natural Sciences, and Social Sciences items from a workbook for the College Level Examination Program tests; and a mid-term post-graduate level test in Econometrics. About half way through these demonstrations we decided to begin keeping a permanent record of what people were doing at the terminal.

Figure 3 compares the two information measures for the first test taken by each of 66 individuals. Most of the data points fall below the diagonal, indicating that most of the "subjects" at least initially overvalue their knowledge of these subject matter areas. A few people fall close to the diagonal, suggesting that there may exist some people who can discriminate what they know well from what they know less well with a high degree of accuracy.

What happens when people take more tests and, thus, gain more experience with decision-theoretic testing? We find that many of these subjects can reduce their score loss due to lack of realism (Sibley, 1974). I think that this improvement comes as they begin

to experience the consequences of the admissible scoring system (Shuford, Albert & Massengill; 1966) and learn to reduce their risk-taking tendencies by making their utilities more nearly linear in points earned or lost. There does, however, appear to be a limit to this improvement.

A number of people were encouraged or challenged to take more tests and to try to be as realistic and to score as well as they possibly could. It should be remembered that there is no conflict between these goals when an admissible scoring system is used (Shuford & Brown; 1974). So I now have 11 subjects who have taken an appreciable number of tests -- enough so I could discard the early ones taken while they were learning the procedures and the consequences of the admissible scoring system.

Figure 9 shows the apparently stable state behavior of the most biased of the 11 subjects. The line designated $\bar{I}_A$ is located at the mean of the actual information measures while the line designated $\bar{I}_P$ is located at the mean of the perceived information measures. The intersectio of the two lines gives a gross indication of actual vs. perceived information for those tests the subject decided to attempt. By taking the ratio of $\bar{I}_P$ to $\bar{I}_A$ we can obtain a rough measure of the extent and direction of bias. The ratio for this subject is 2.44 indicating that she thought that she had almost two and one-half times as much information as she actually had.

Figure 10 tables some personal characteristics for the 11 subjects listed in decreasing order of bias which goes down almost to the unbiased value of 1.00. Notice that no subject yielded an overall ratio less than one which would indicate a person who typically undervalued his information. Figure 11 compares the information measures for subject B. Although apparently striving to reduce bias and to improve his score, this subject was also unable to do so. Figures 12 through 18 show subjects with decreasing amounts of bias who were more and more often successful in producing a realistic assessment of their uncertainty. Figures 19 and 20 are for the two most accurate subjects who were remarkedly consistent in demonstrating their ability to accurately assess their uncertainties.

In conclusion, the introduction of decision-theoretic testing makes it possible to define and to measure for the first time a human ability, call it realism, which may prove to be a very important determinant of individual and team performance. For example, to what extent and in what manner is an unrealistic student handicapped in his attempts to learn and to study effectively? For another example, does a team of realistic individuals tend to outperform a team of overvaluing individuals and, if so, for what types of tasks? Answers to these and many other questions must await further research.

I have shown here that some people can be very realistic over a wide range of subject matter while others characteristically overvalue their information. We do not yet know what deficits in this ability exist within different subgroups of the population nor do we know to

what extent or what it takes to educate people to become more realistic. The results summarized in Figure 10 certainly prove that level of education does not insure realism in assessing and communicating uncertainty.

The decision theorist, L. J. Savage, in his posthumously published article on the "Elicitation of Personal Probabilities and Expectations" (Savage; 1971) correctly conjectured that people would be found who tended to overvalue their information. I suspect that the remainder of his statement will also prove to be prophetic and a useful guide for future research and applications of decision-theoretic testing. For this reason, I repeat it here.

> "Though requiring more student time per item, these [decision-theoretic testing] methods should result in more discrimination per item than ordinary multiple-choice tests, with a possible net gain. Also, they seem to open a wealth of opportunities for the educational experimenter.
>
> Above all, the educational advantage of training people -- possibly beginning in early childhood -- to assay the strengths of their own opinions and to meet risk with judgment seems inestimable. The usual tests and language habits of our culture tend to promote confusion between certainty and belief. They encourage both the vice of acting and speaking as though we were certain when we are only fairly sure and that of acting and speaking as though the opinions we do have were worthless when they are not very strong."

REFERENCES

Brown, T. A. and E. H. Shuford, Jr. (1973) Quantifying Uncertainty into Numerical Probabilities for the Reporting of Intelligence, The Rand Corporation, R-1185-ARPA.

Savage, L. J. (1971) "Elicitation of Personal Probabilities and Expectations," Journal of the American Statistical Association, Vol. 66, pp. 783-801.

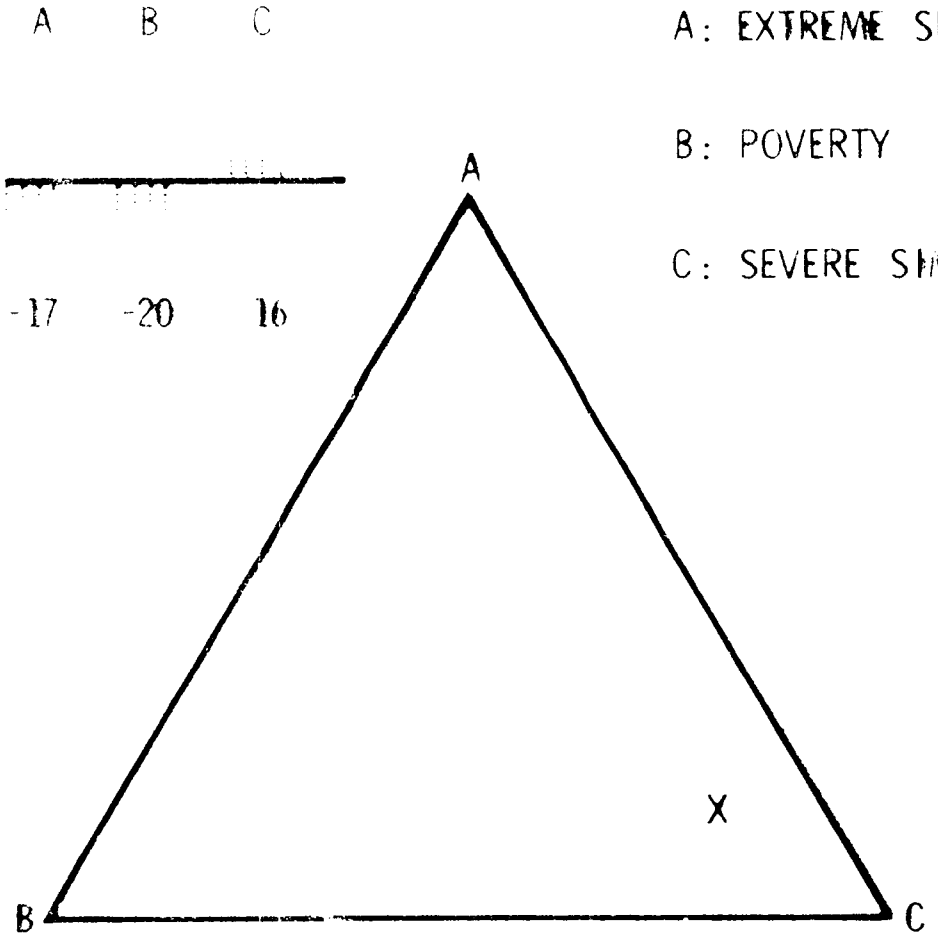
Shannon, C. E. and W. Weaver (1949) The Mathematical Theory of Communication. Urbana: The University of Illinois Press.

Sibley, W. L. (1974) An Experimental Implementation of Computer-Assisted Admissible Probability Testing, The Rand Corporation, P-5174.

Shuford, E. H., Jr., A. Albert, and H. E. Massengill (1966) "Admissible Probability Measurement Procedures," Psychometrika, Vol. 31, pp. 125-145.

Shuford, E. H., Jr. and T. A. Brown (1974) A Rationale and Specifications of Computer-Administered Admissible Probability Measurement, The Rand Corporation, R-1371-ARPA.

1. AUSTERITY



A: EXTREME SUFFERING

B: POVERTY

C: SEVERE SIMPLICITY

TEST SELECT

QUESTION ANSWER

PREVIOUS QUESTION

NEXT QUESTION

Fig. 1

THE EQUILATERAL RESPONSE TRIANGLE

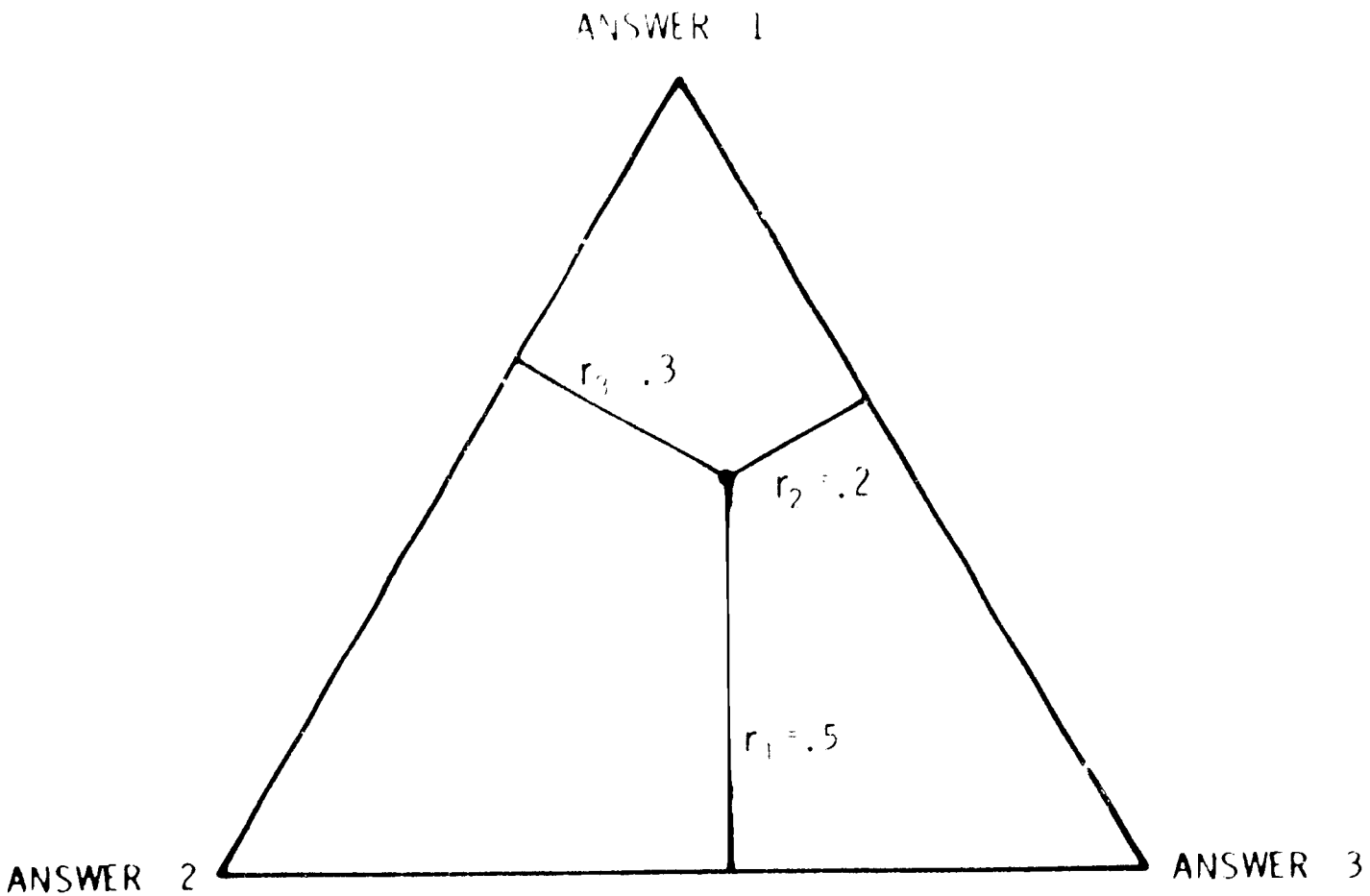
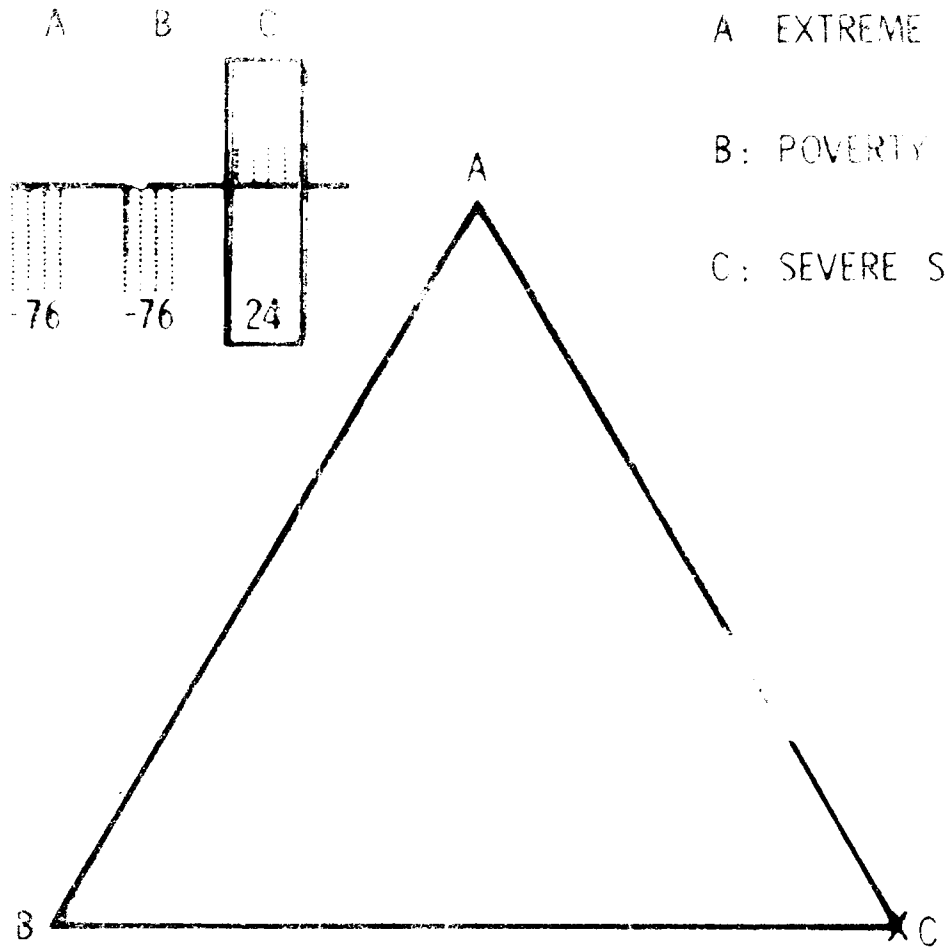


Fig. 2

The Rand Corporation
Santa Monica, CA

1. AUSTERITY



A EXTREME SUFFERING

B: POVERTY

C: SEVERE SIMPLICITY

TEST SELECT

QUESTION ANSWER

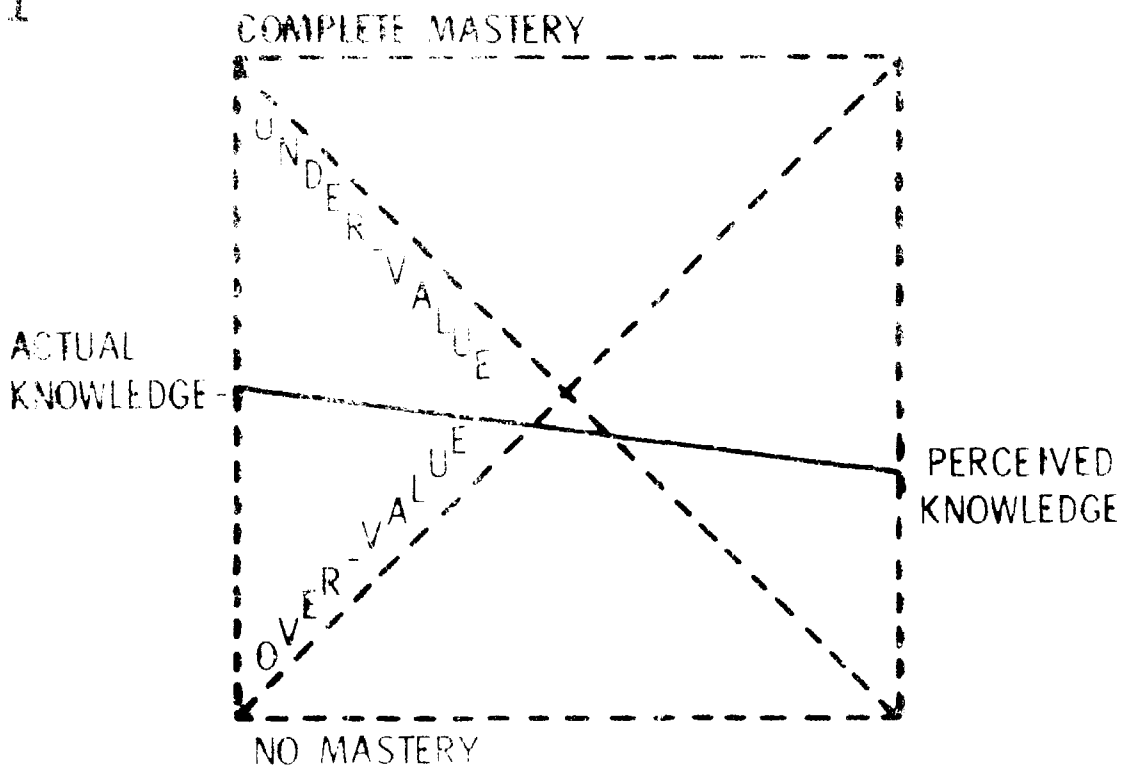
PREVIOUS QUESTION

NEXT QUESTION

Fig. 3

DIGEST 1

03/09/73 13.48.40 W. L. SIBLEY



YOU TEND TO UNDER VALUE YOUR KNOWLEDGE.

BET MORE HEAVILY ON ANSWERS YOU FEEL ARE CORRECT.

YOU CAN IMPROVE YOUR SCORE BY 37 POINTS BY MORE REALISTIC USE OF YOUR KNOWLEDGE.

YOU CAN IMPROVE YOUR SCORE BY 224 POINTS BY MORE STUDY.

YOU CAN IMPROVE YOUR SCORE BY 261 POINTS OVERALL.

NEW TEST

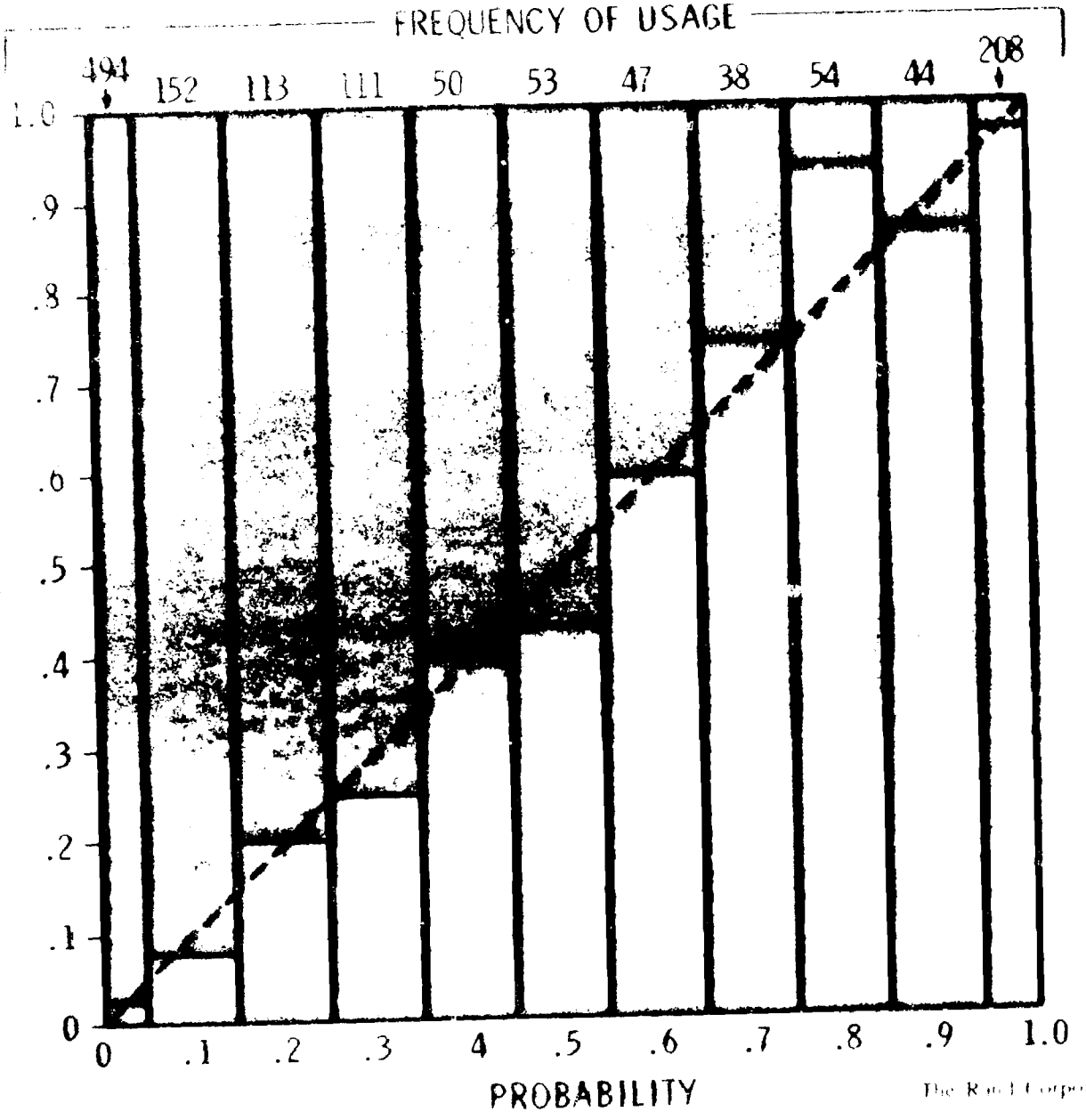
RECORD

PLOT

OLD TEST

Fig. 4

AN EXTERNAL VALIDITY GRAPH



Proportion assigned to correct answers

Proportion assigned to incorrect answers

B 15 · 5"

REALISM FUNCTIONS

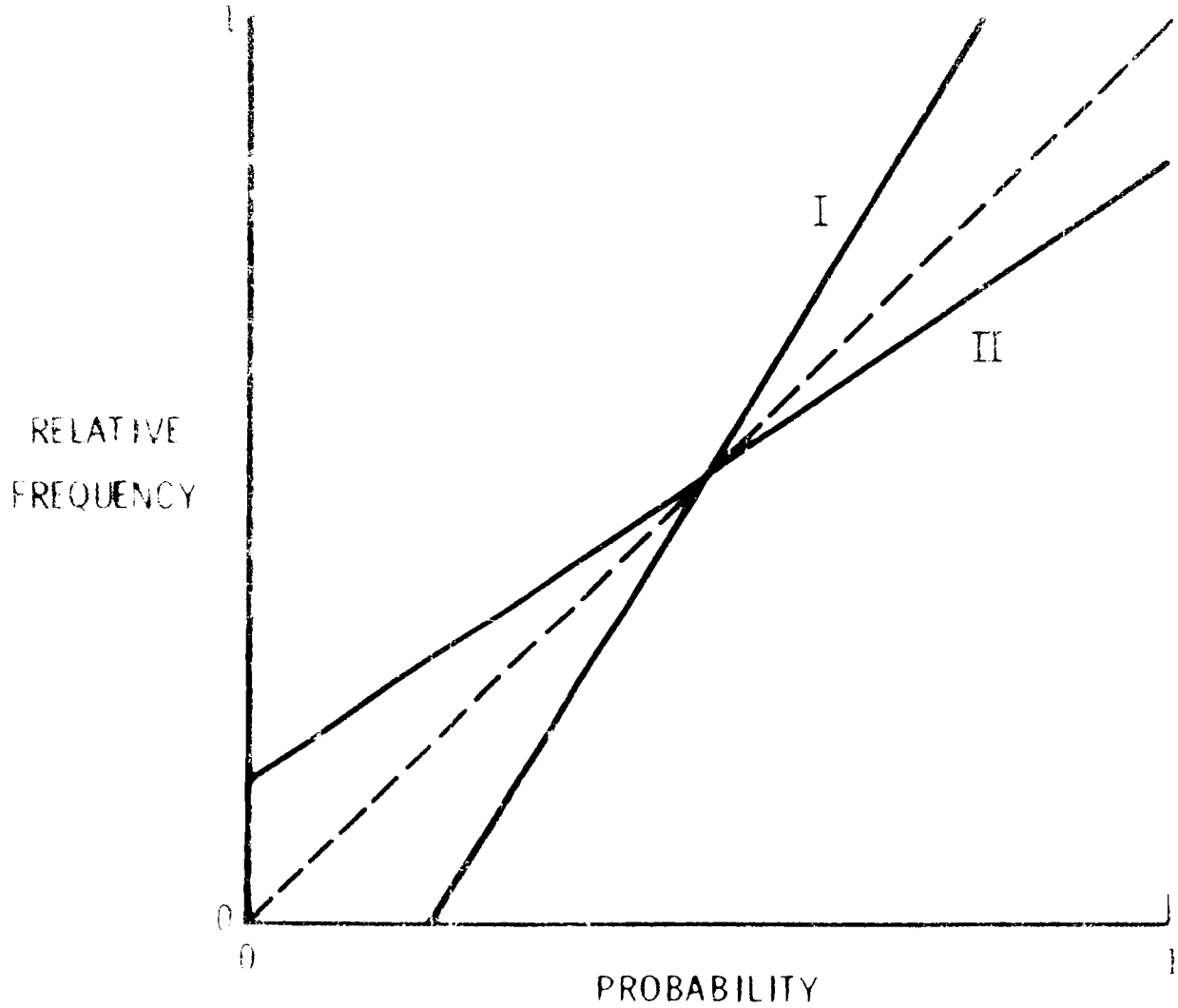


Fig. 6

THE INFORMATION SQUARE

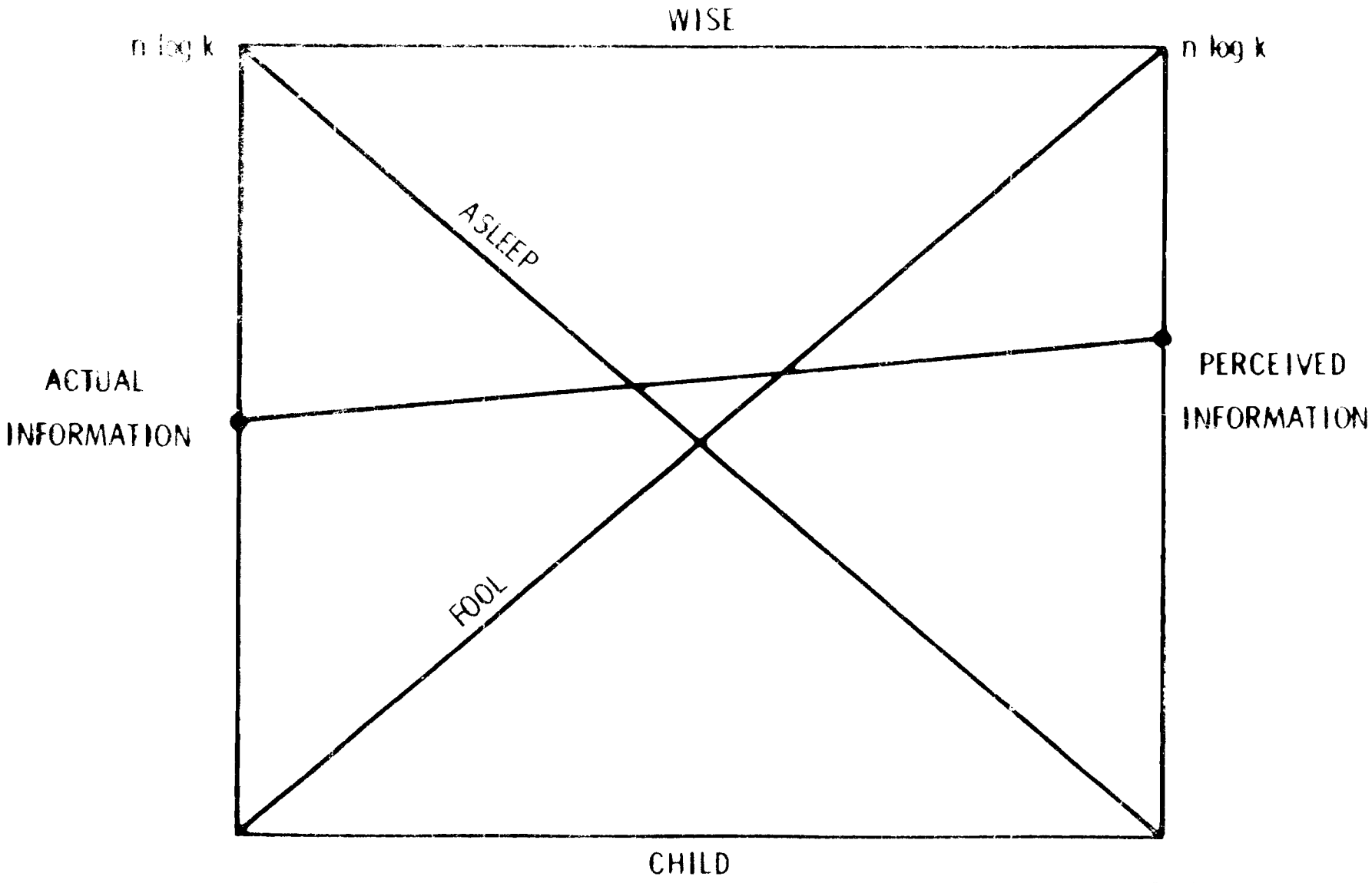


Fig. 7

INFORMATION COMPARISONS

(First Test)

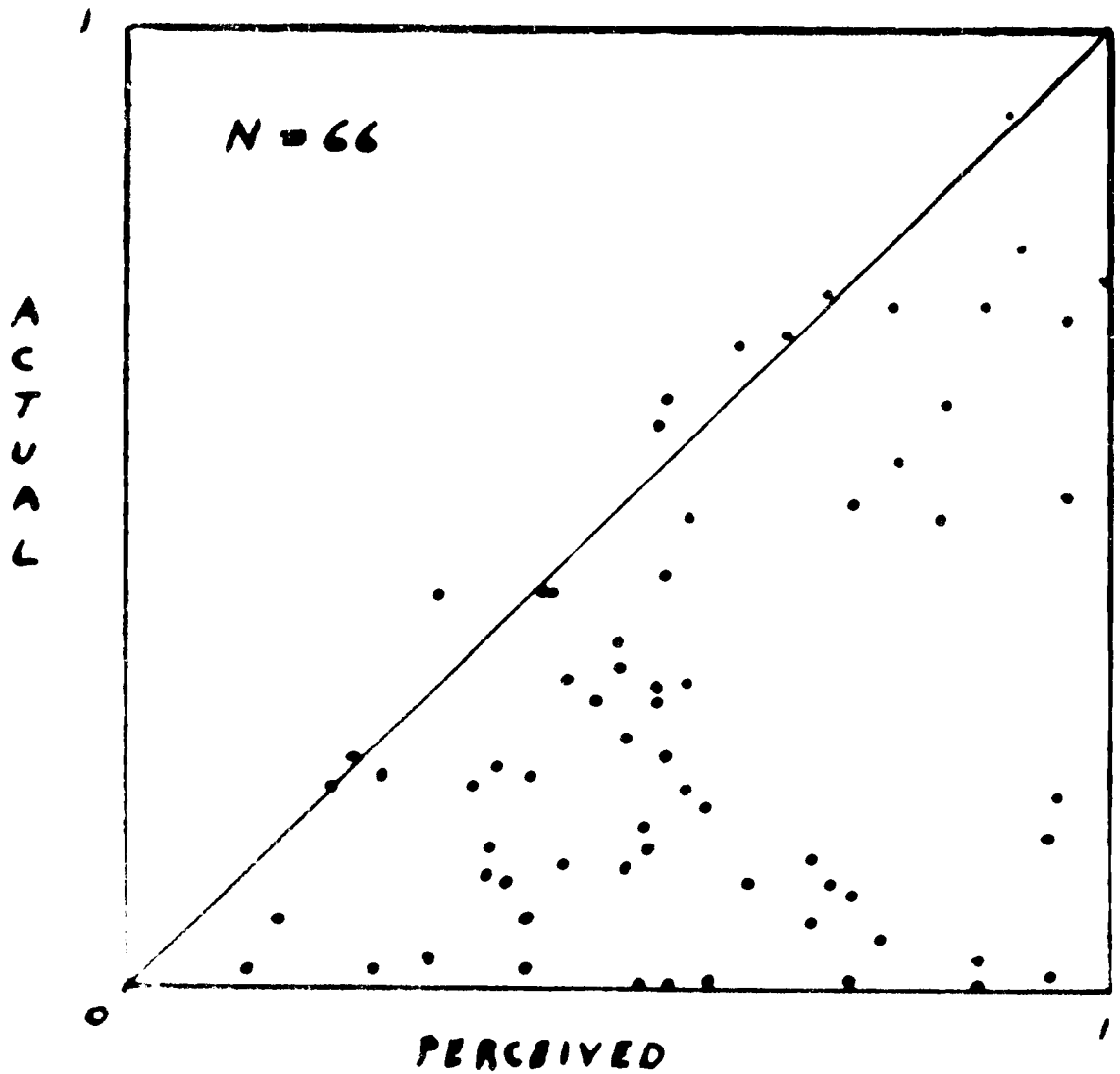


Fig. 8

INFORMATION COMPARISONS

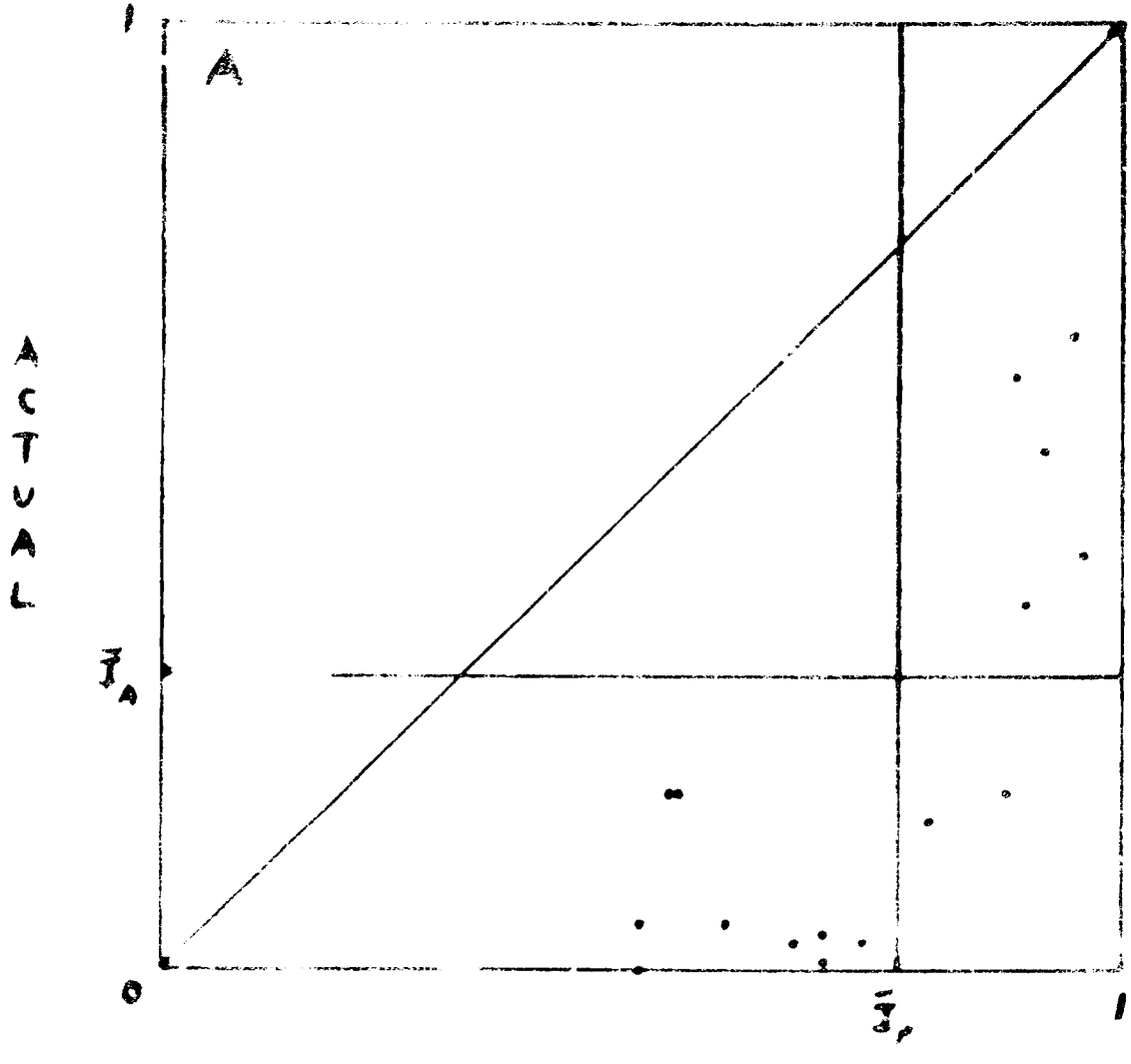


Fig. 9

SUBJECT CHARACTERISTICS

SUBJECT	I_p/I_A	I_A	TESTS	SEX	AGE	EDUCATION
A	2.44	.31	18	WOMAN	20-30	MASTER'S +
B	2.42	.17	13	MAN	30-40	DOCTORATE
C	2.26	.28	7	MAN	50-60	DOCTORATE
D	2.11	.32	27	WOMAN	20-30	BACHELOR'S
E	1.81	.18	20	WOMAN	20-30	SOME COLLEGE
F	1.67	.40	12	WOMAN	50-60	DOCTORATE
G	1.62	.30	20	WOMAN	30-40	BACHELOR'S
H	1.33	.35	9	GIRL	9	THIRD GRADE
I	1.32	.38	21	GIRL	12	FIFTH GRADE
J	1.02	.71	34	MAN	40-50	DOCTORATE
K	1.004	.85	8	MAN	40-50	DOCTORATE

Fig. 1a

INFORMATION COMPARISONS

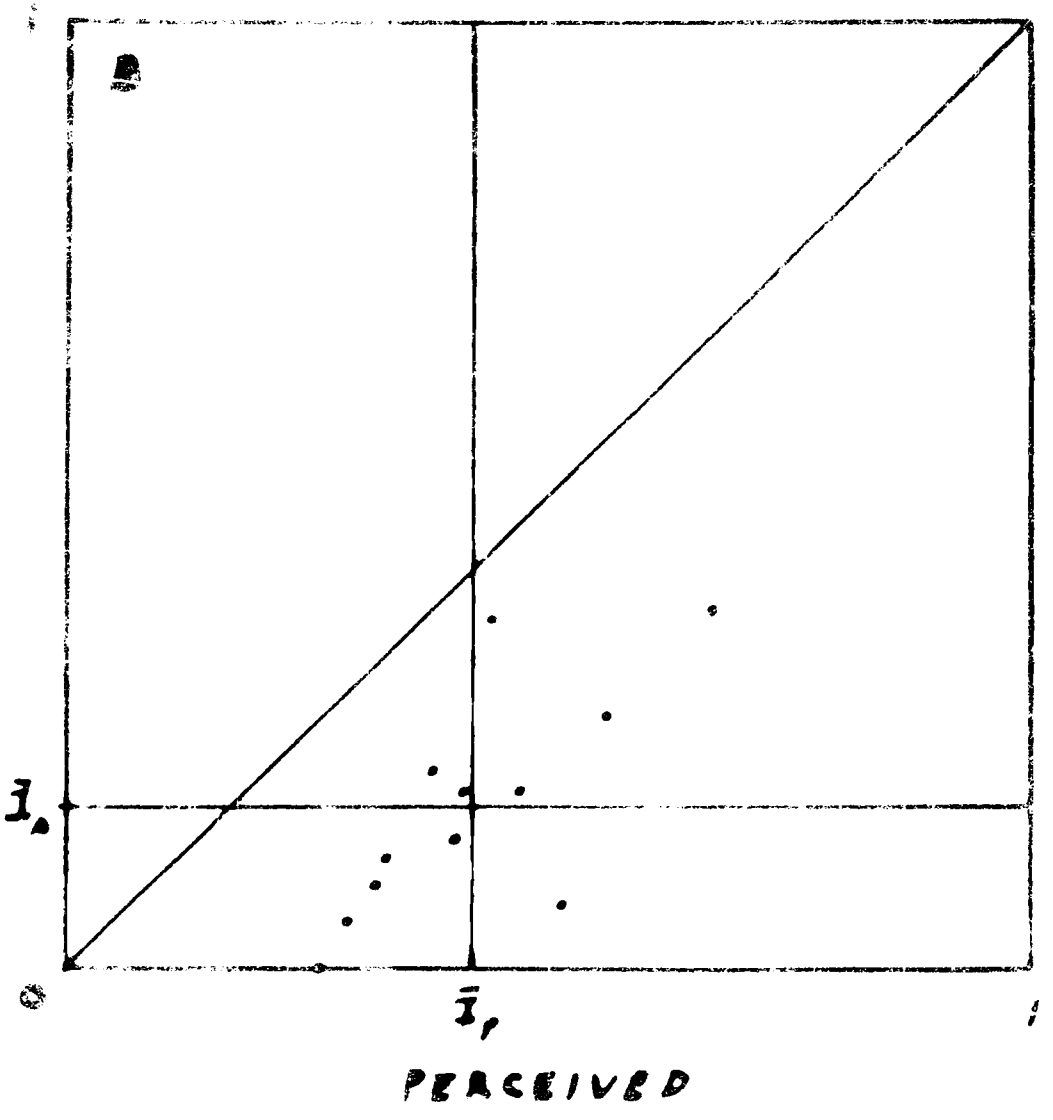
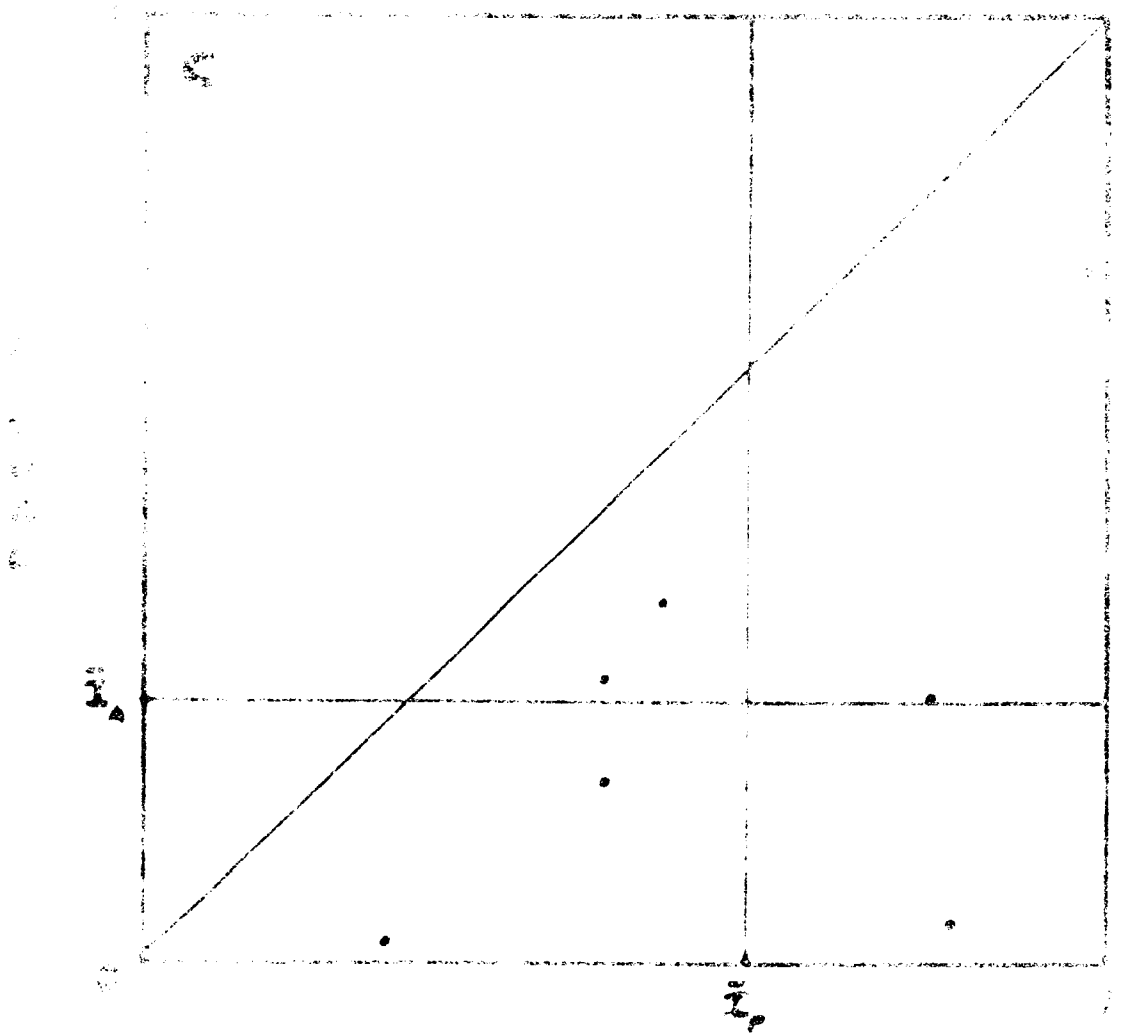


Fig. 11



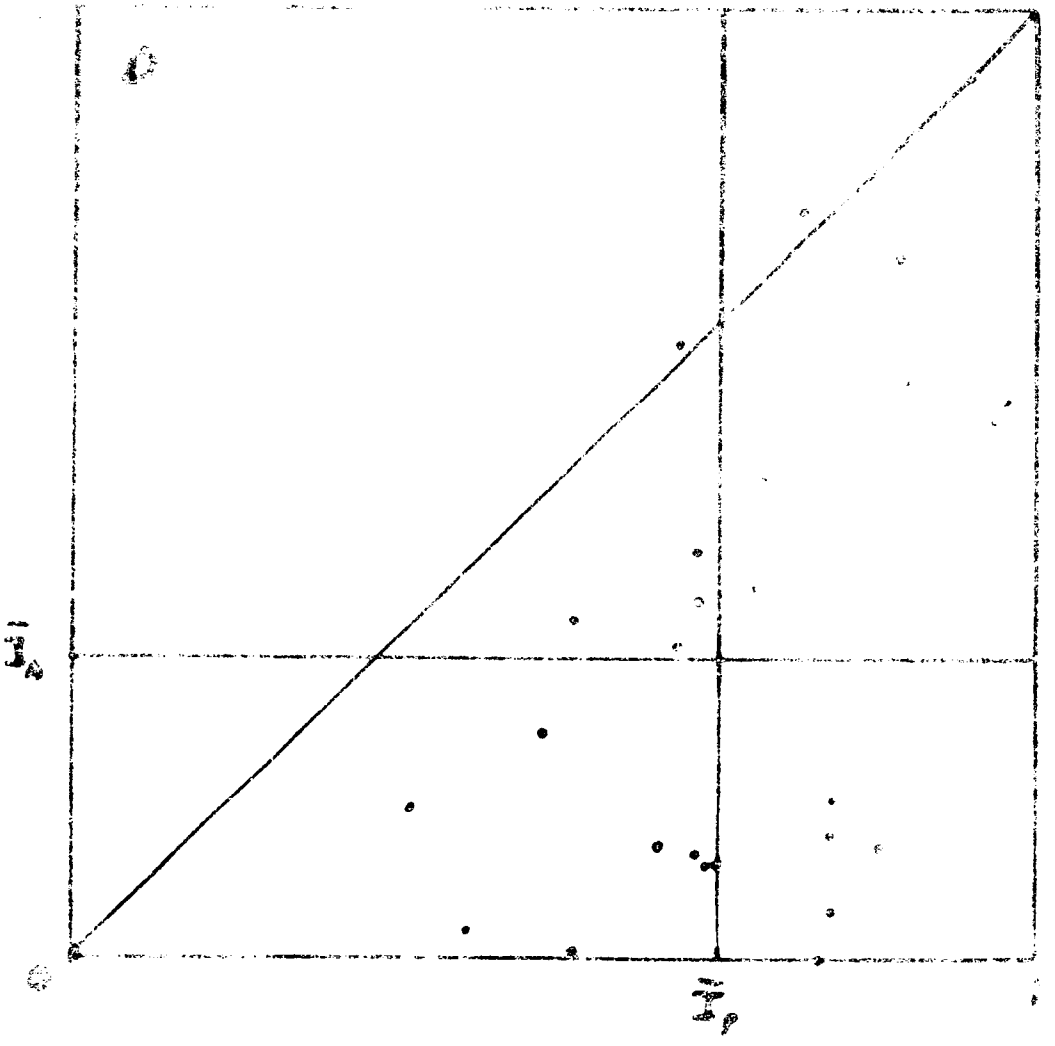


Fig. 13

INFORMATION COMPARISONS

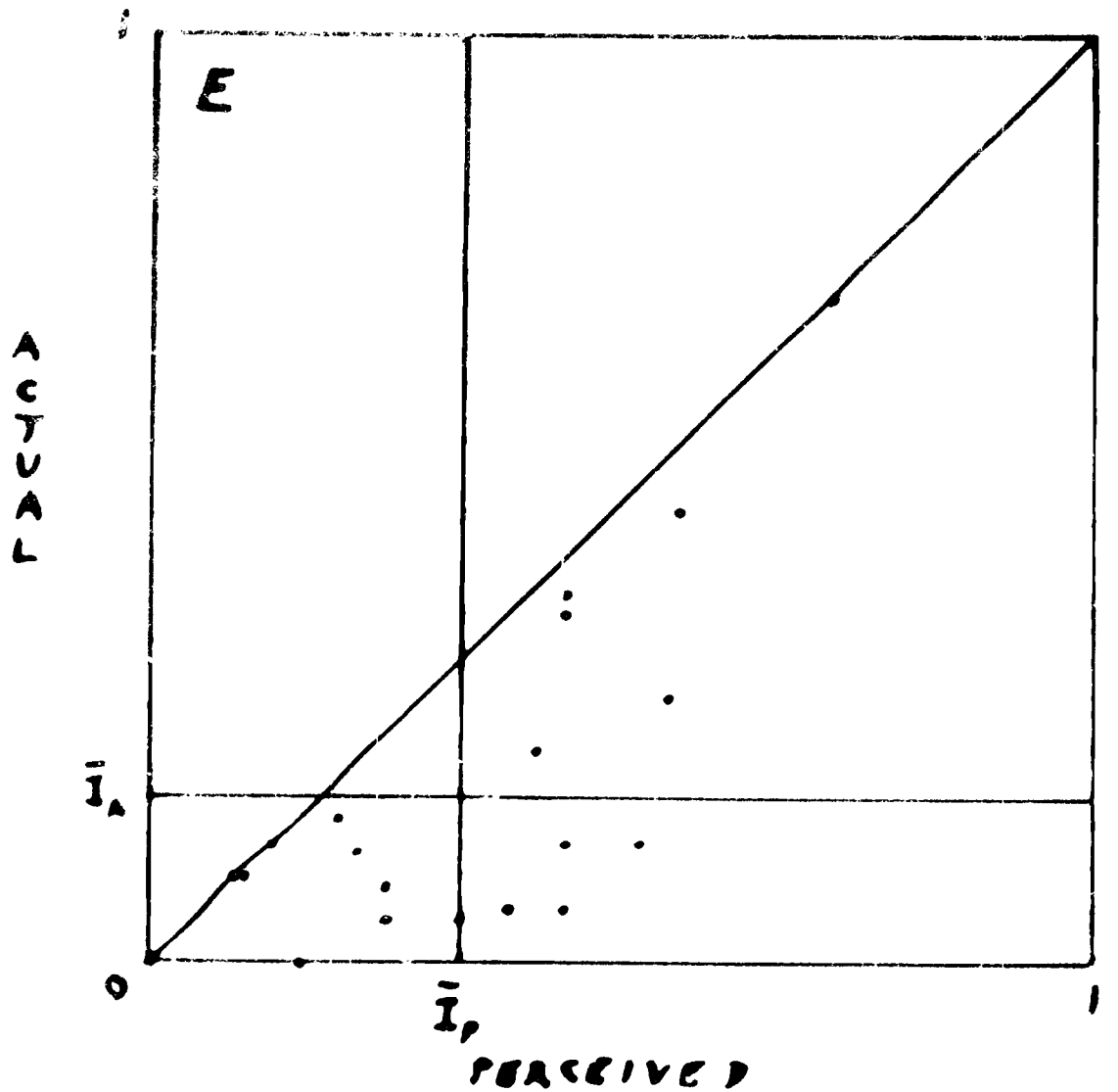
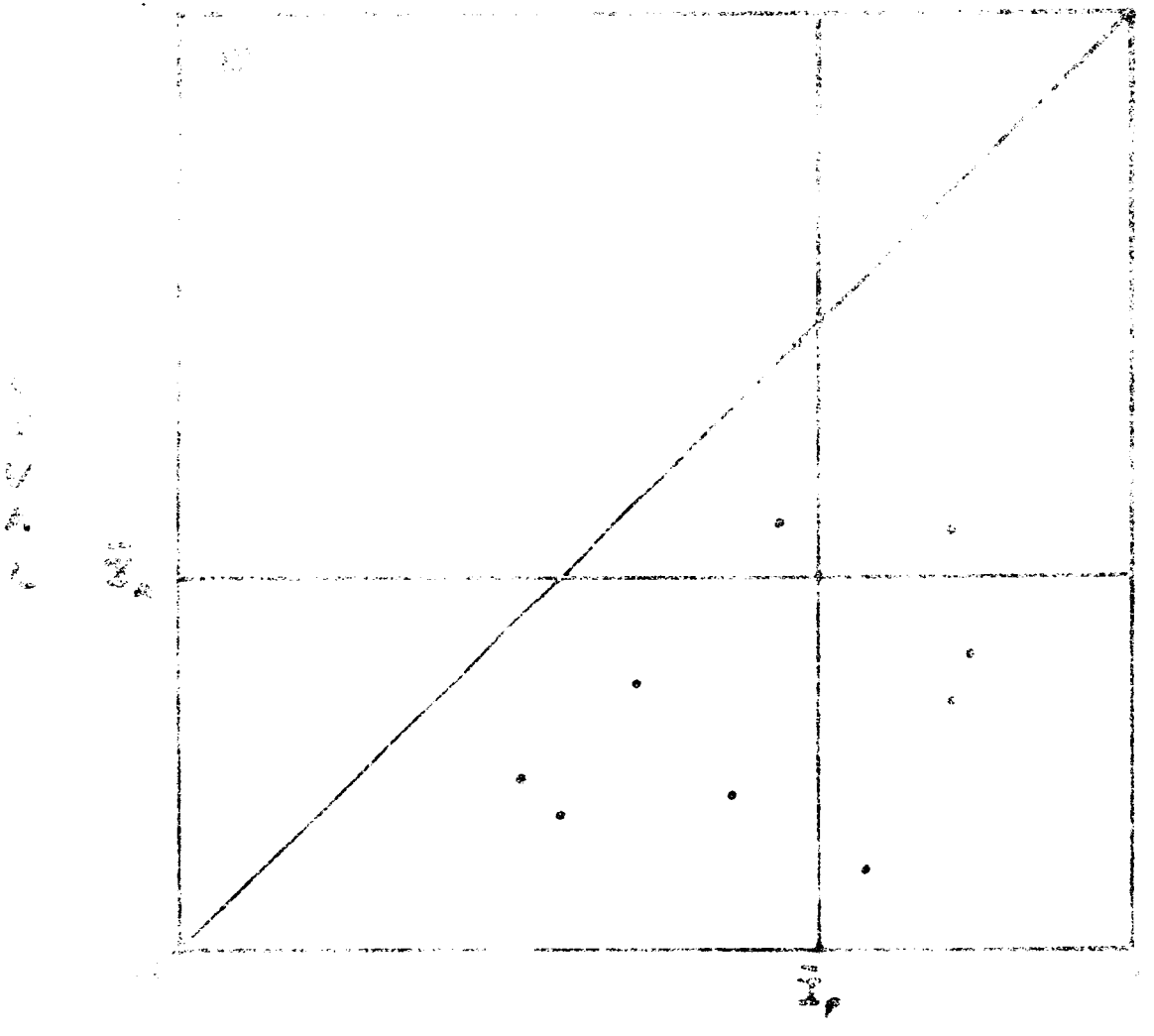


Fig. 14



PERCEIVED

INFORMATION COMPARISONS

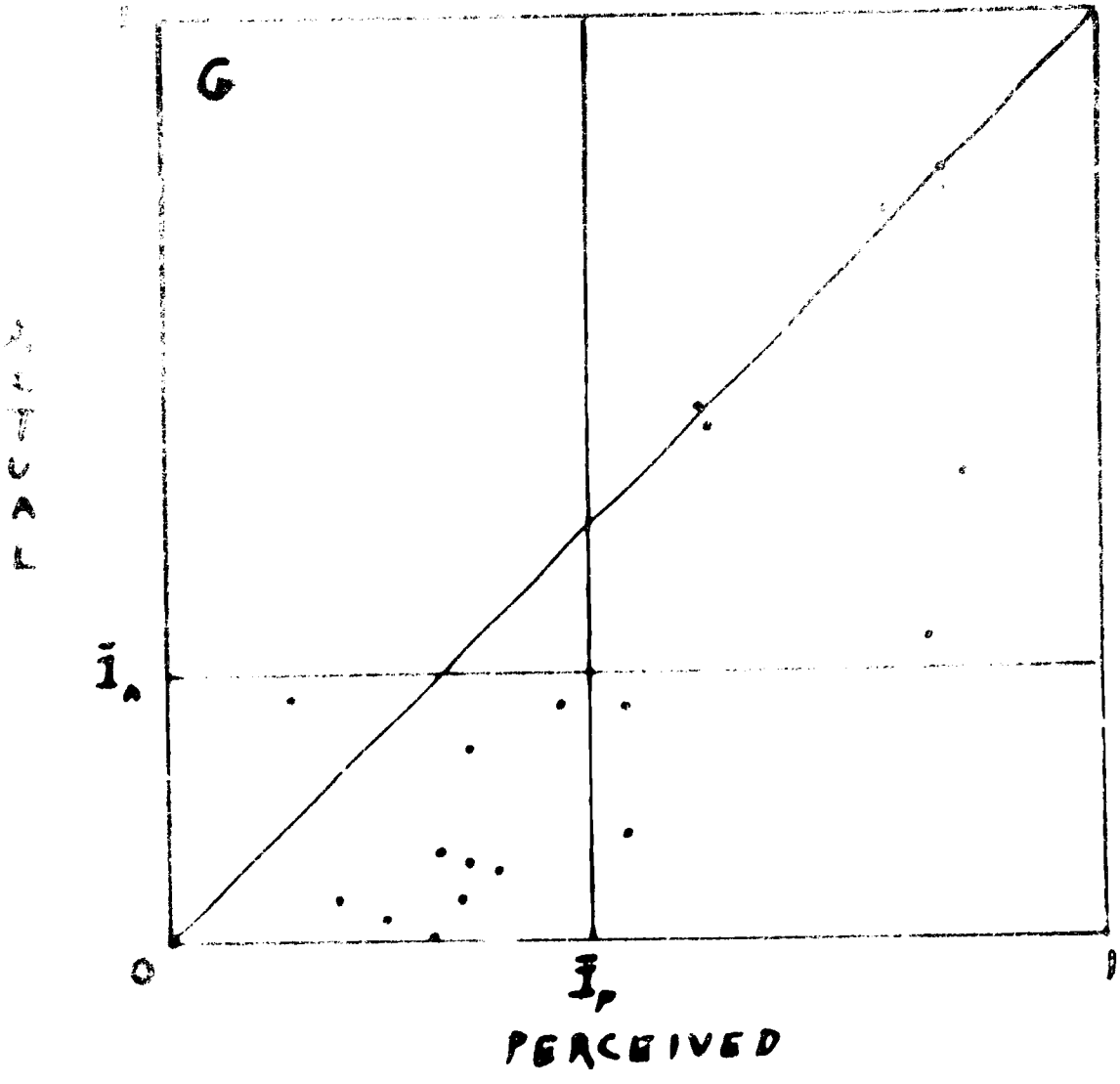


Fig. 16

INFORMATION COMPARISONS

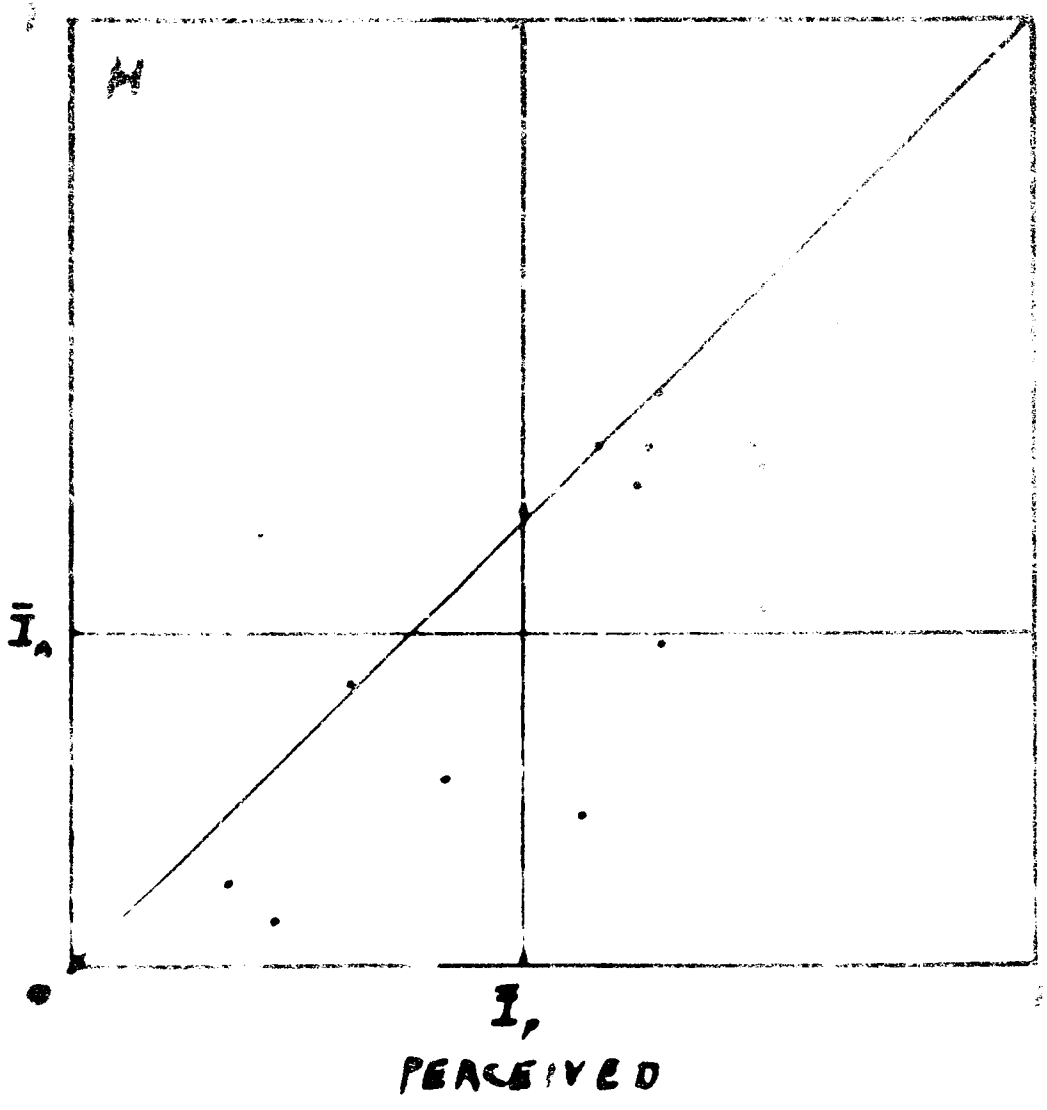


Fig. 17

INFORMATION COMPARISONS

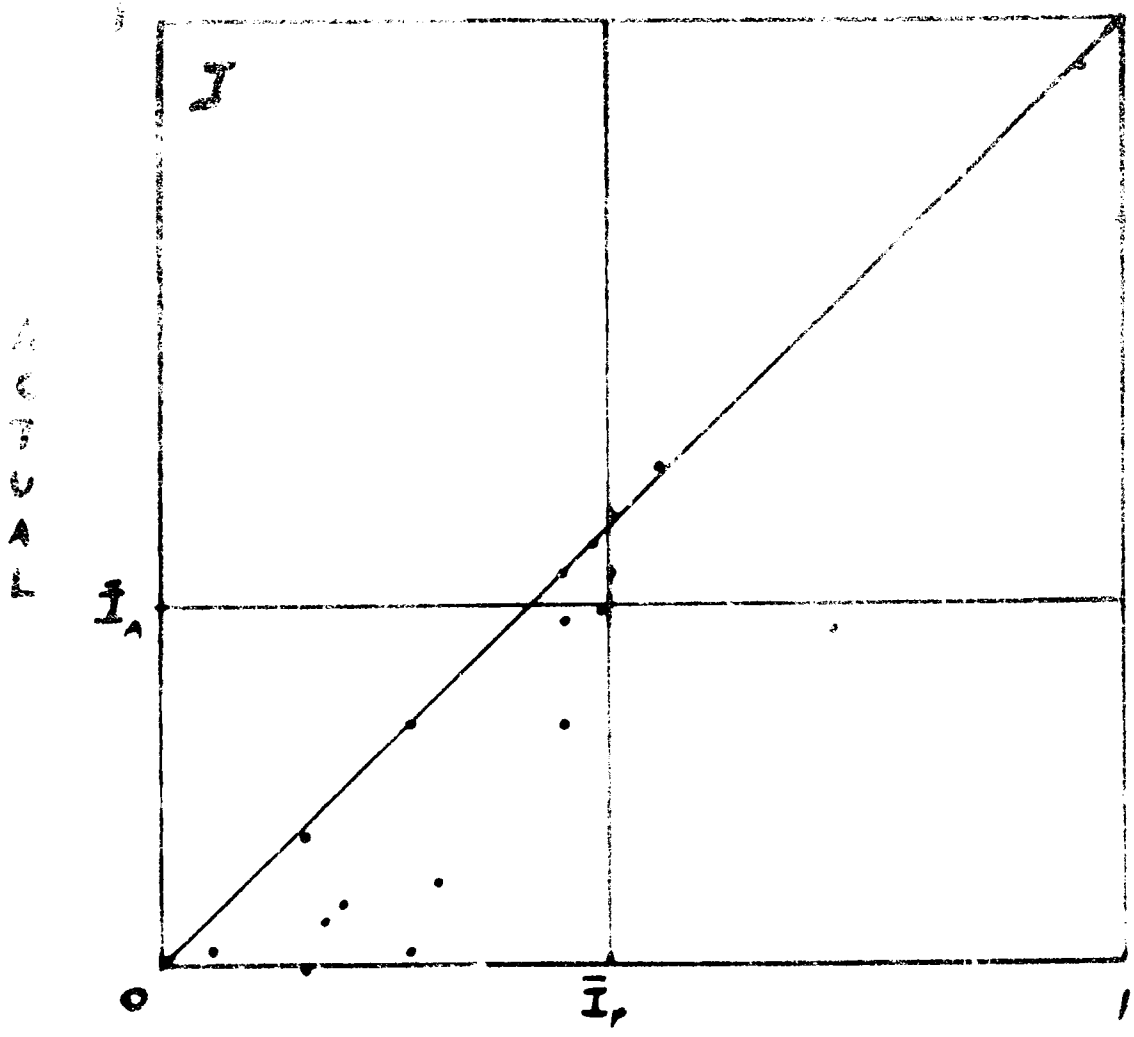


Fig. 18

INFORMATION COMPARISONS

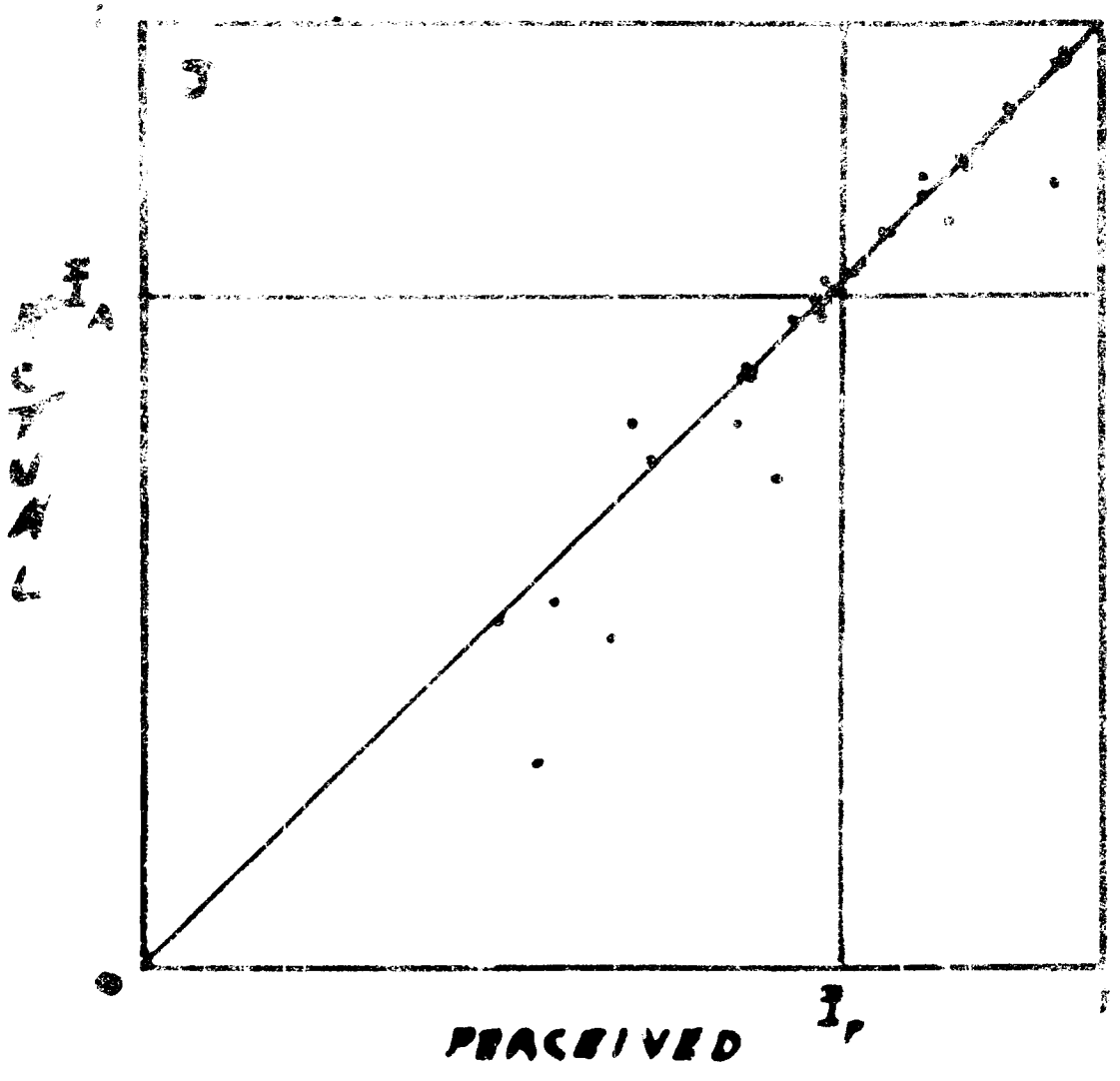


Fig. 19

INFORMATION COMPARISONS

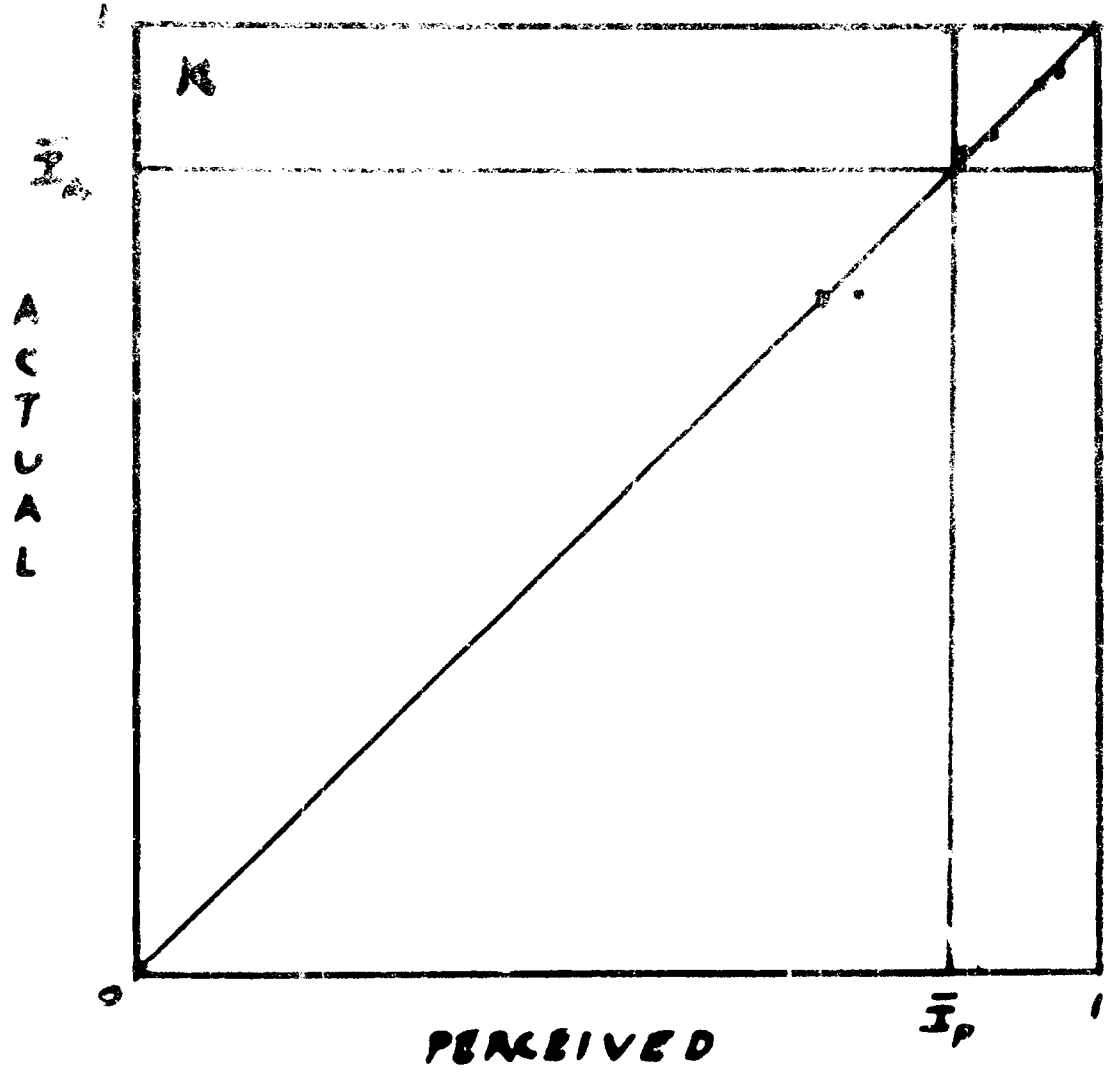


Fig. 20