DOCUMENT RESUME

ED 104 901 TM 004 346

AUTHOh Sibley, William L.
TITLE An Experimental Implementation of Computer Assisted Admissible Probability Testing.
INSTITUTION Rand Corp., Santa Monica, Calif.
REPORT NO P-5174
PUB DATE Feb 74
NOTE 28p.; Paper presented at the Military Association Conference (San Antonio, Texas, October 28-November 2, 1973)

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS *Computer Oriented Programs; Computers; Interaction; Military Training; *Multiple Choice Tests; *Probability; *Response Style (Tests); Scoring Formulas; Statistical Analysis; Student Evaluation; Test Construction; *Testing; Test Validity
IDENTIFIERS *Admissible Probability Testing

ABSTRACT

The use of computers in areas of testing, selection, and placement processes for those in military services' training programs are viewed in this paper. Also discussed is a review of the motivational and theoretical foundation of admissible probability testing, the role of the computer in admissible probability testing, and the authors' experience with computer-based systems. (Author/DEP)

AN EXPERIMENTAL IMPLEMENTATION OF COMPUTER ASSISTED ADMISSIBLE PROBABILITY TESTING

William L. Sibley

February 1974

P-5174

The Rand Paper Series

Papers are issued by The Rand Corporation as a service to its professional staff. Their purpose is to facilitate the exchange of ideas among those who share the author's research interests; Papers are not reports prepared in fulfillment of Rand's contracts or grants. Views expressed in a Paper are the author's own, and are not necessarily shared by Rand or its research sponsors.

The Rand Corporation
Santa Monica, California 90406

AN EXPERIMENTAL IMPLEMENTATION OF COMPUTER ASSISTED ADMISSIBLE PROBABILITY TESTING*

INTRODUCTION

The work reported in this paper was conducted under the Computer Technologies in Training Project of the ARPA-sponsored Manpower and Training Management Program of the Rand Corporation. As indicated by the project title, we are interested in ways computer technology can be used to enhance the military services' training program as well as to help the selection and placement processes. In particular, we have recently been investigating the use of computers in testing.

As reported in 1969 by W. C. Gardner, Jr. and E. H. Shuford, Jr., there is considerable experimental evidence that testing systems that incorporate admissible probability measurement have wide application in important areas of military testing and evaluation [1, 2]. The nature of Admissible Probability Testing and its extensions invites the use of computers in their application as I shall try to make clear in what follows.

The remainder of the discussion is divided into three parts:

- A review of the motivations and theoretical foundations of Admissible Probability Testing (APT).
- A description of the role of the computer in Admissible Probability Testing.
- A brief discussion of our experience to date with the computer-based system.

*Presented at the 1973 Military Association Conference, San Antonio, Texas, 28 October-2 November 1973, Air Force Human Resources Lab.

MOTIVATIONS AND FOUNDATIONS

The prime motivation for Admissible Probability Testing is simple and straightforward: we need to measure better the "state of knowledge" of trainees in a large variety of subject areas. An improved measuring technique would be of considerable value not only in testing a trainee but also in evaluating and planning curricula. As indicated by Gardner, planning is especially critical for those courses of instruction that are individually tailored to the trainee [1].

A second, and no less important, motivation is that the estimation of probabilities is a critical skill in some endeavors; for example, weather prediction, intelligence analysis, etc. I believe that it will become apparent that Admissible Probability Testing is especially applicable in the training of students to use effectively all of their facts and reasons in the estimation of probabilities.

A corollary of the need for training specialists in probability estimation is the idea that, in fact, we are all engaged in probability estimation. That is, there is a need for us all to convey accurately and effectively our opinions of the possibilities of various outcomes. If that is the case, we must learn to speak the language of probabilities precisely and with a common understanding.

So what is Admissible Probability Testing and how is it applied? I shall talk about it in terms of our specific application even though it has greater generality than may be apparent from my discussion.

Consider the multiple choice test item in Figure 1. We are generally required to respond to such a question with what amounts to a probability distribution as is illustrated in Figure 2. However, it would be my personal tendency to respond with the probability distribution illustrated in Figure 3. In fact, if confronted with the question in Figure

QUESTION: The 11th president of the United States was:

1. George Washington
2. James Polk
3. Franklin Pierce

Figure 1: Multiple Choice Test Item

QUESTION: The 11th president of the United States was:

$$\begin{bmatrix}0\\1\\0\end{bmatrix}\quad\begin{array}{ll}1. & \text{George Washington}\\2. & \text{James Polk}\\3. & \text{Franklin Pierce}\end{array}$$

Probability
Distribution

Figure 2: A Standard Response

QUESTION: The 11th president of the United States was: . .

$$\begin{array}{cl} \begin{bmatrix} 0 \\ 1/2 \\ 1/2 \end{bmatrix} & \begin{array}{l} \text{1. George Washington} \\ \text{2. James Polk} \\ \text{3. Franklin Pierce} \end{array} \end{array}$$

Revised
Probability
Distribution

Figure 3: A Revised Response

QUESTION: The 11th president of the United States was:

$$\begin{array}{cl} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} & \begin{array}{l} \text{1. John Tyler} \\ \text{2. James Polk} \\ \text{3. Franklin Pierce} \end{array} \end{array}$$

Figure 4: An "unable to distinguish" Response

4, my response might very well be, "I can't distinguish among the alternatives." That is, as far as I can tell from my facts and reasons, each alternative is equally likely. I think it is evident that requiring a response of zeroes and a single one can mask a large variety of states of knowledge as well as introduce guessing on the part of the student.

If such freedom of response is allowed, there remains the problem of how to score such a response. That is precisely the function of the "admissible," or "proper," or "reproducing" scoring system [3]. Such a scoring system encourages the strategy of responding with exactly what you perceive the probability distribution to be based on your own facts and reasons.

The admissible scoring system can be motivated in terms of a gambling example (a more complete discussion can be found in [3]). For the sake of simplicity, suppose that a student were faced with a true-false question. Then he may be viewed as a gambler faced with determining the probability of occurrence of the event "the correct response is 'true.'" Suppose also that there were a gambling house in which there were $\phi(u)du$ wagers available at odds $(1-u)/u$ (the "correct odds" for an event of probability u). If the student believes that the probability of the event taking place were p, he would take all wagers at odds better than $(1-p)/p$. Similarly, he would take all wagers on the event not taking place at odds better than those appropriate for probability 1-p.

Figure 5 illustrates the mathematics of admissible scoring systems. Line (1) corresponds to our true-or-false discussion. Line (2) extends those notions to a n-foil multiple choice question. Line (3) adjusts the equations to give a score of zero for the "unable to distinguish" response of $p_i=\frac{1}{n}$ for i=1,n. What these integrals represent is the net settlement made between the gambling house and the student, after the correct alternative has been revealed. It is the sum of settlements on an infinite number of

(1) True or False

payoff (if "true" occurs) $= \int_{o}^{p} \phi(u)\frac{1-u}{u}du - \int_{o}^{1-p} \phi(u)du$

(2) Multiple Choice (n foils)

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

(3) Adjusted Multiple Choice

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

Figure 5: Admissible Scoring Systems

infinitesimal wagers, each of which appeared to the student to be favorable to him.

Figure 6 illustrates the logarithmic admissible scoring system which has the property of depending on only the probability ascribed to each foil independent of the probabilities ascribed to the remaining foils. Line (2) of Figure 6 illustrates the particular version of the logarithmic system we use in Admissible Probability Testing. No probability less than .01 may be assigned to avoid infinitely large penalties.

The gambling illustration shows that it is in the student's best interest to gauge accurately his probabilities and thus to make no unnecessary bets. That is, the scoring system does encourage him to reproduce his own probabilities accurately.

However, how does the student learn to "play the game" and improve the way he brings his facts and reasons to bear on the determination of the probability he ascribes to each alternative? As Gardner points out, "It seems likely that the more exposure the students have to [probability] tests, the more realistically they will evaluate test items and their own knowledge, thus providing a more valid indication as to their actual level of achievement" [1]. We believe the answer to the question lies in the computer system I will describe next.

(1) The Logarithmic Scoring System $\phi(u)=1$:

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

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

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

(2) An adjusted system for 3 foils and a score range of about 100 points:

$$= 50 \log(3p_i) \qquad .01 \leq p_i \leq .98$$

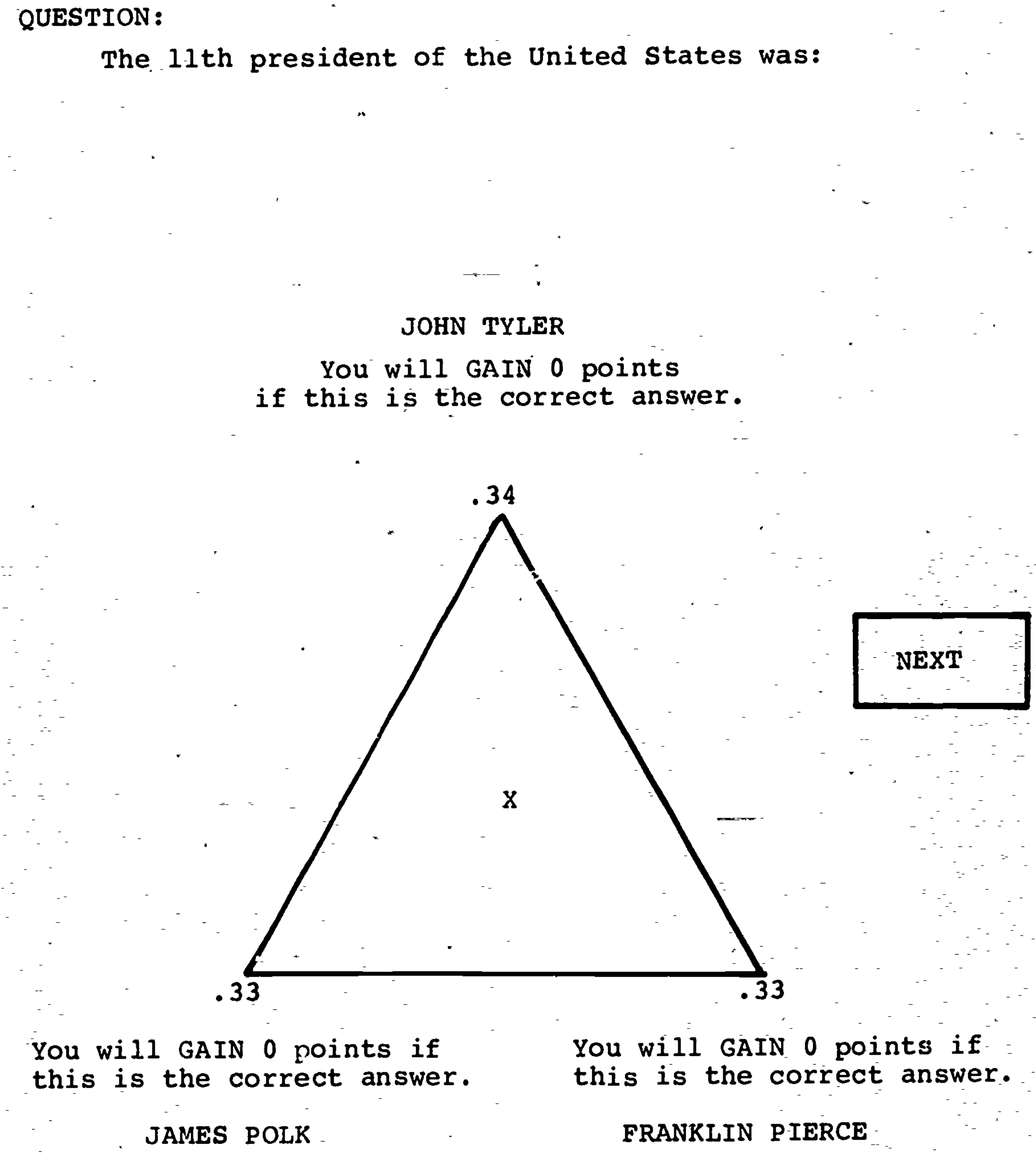
Figure 6: The Logarithmic Scoring System

THE ROLE OF THE COMPUTER

As you will see in the figures that follow, the computer-based system we have developed requires a computer terminal with graphics capabilities and some means of allowing the student to interact with it in elapsed times of at most a few seconds. The original system was developed on a highly interactive console connected to a powerful IBM 370/155 computer. A newer system exists on a "smart terminal," the IMLAC PDS-ID, which supports the system independently of any larger computer. My discussion will relate to the latter system.

Figure 7 illustrates the display the student sees while answering a question. He assigns a probability to each alternative by choosing a point in the interior or on the boundary of the equilateral triangle. In such a triangle, the sum of the distances of the point from each side of the triangle is a constant and can be scaled to 1. Thus each point determines a probability distribution for the three alternatives. The probabilities appear at the vertices of the triangle and the appropriate logarithmic score appears close by. Figure 7 shows the results for the "unable to distinguish" case. The probabilities are constrained to sum to one. Also, the probabilities must be greater than .01 because of the behavior of the logarithm near zero.

The student may choose different points in the triangle until he is satisfied with the indicated probability or score distribution. The nearer he approaches a vertex, the nearer that probability approaches .93 and the other two approach .01. When he is satisfied with the distribution he selects the NEXT option and the terminal displays his cumulative score as illustrated by Figure 8.



QUESTION:

The 11th president of the United States was:

JOHN TYLER

You will GAIN 0 points
if this is the correct answer.

You will GAIN 0 points if
this is the correct answer.

JAMES POLK

You will GAIN 0 points if
this is the correct answer.

FRANKLIN PIERCE

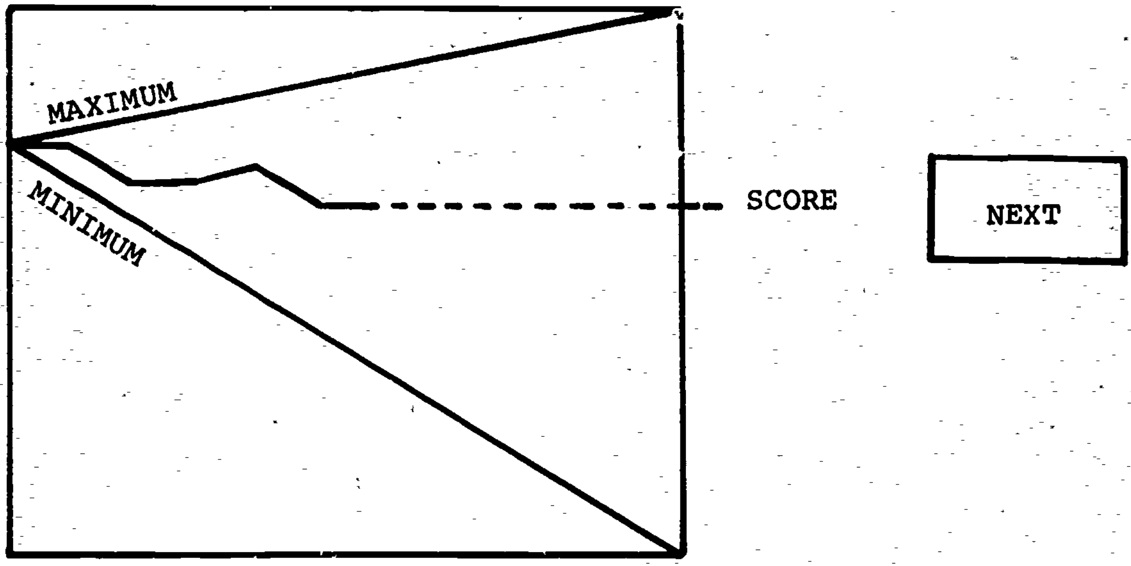
Figure 7: Probability Assignment

QUESTION:

The 11th president of the United States was:

JOHN TYLER

You will GAIN 0 points if this is the correct answer.



You will GAIN 0 points if this is the correct answer.

JAMES POLK

You will GAIN 0 points if this is the correct answer.

FRANKLIN PIERCE

Figure 8: Cumulative Score

The cumulative score is presented in the form of the maximum (or minimum) score attainable question by question. In the case illustrated in Figure 8, nothing has been gained or lost by the score assignments in Figure 7 (the final segment in the broken line in Figure 8). However, a probability of .98 assigned to a correct answer would result in a gain of 23 points while .98 assigned to an incorrect alternative would result in a loss of 76 points. Thus, the student has question by question feedback about the relationship of his probability assignments to his score and may adjust his actions accordingly. That is, he may develop a "feel" for the scoring system. Under usual test administration procedures, the delay from testing to reporting the scores is so long as to prevent the development of a "feel" for the scoring system. Moreover, the scores are usually not reported on an item by item basis. This portion of the computer-based system gives the student a maximum amount of experience with the scoring system.

Even though the student may learn to express his own probabilities accurately the question remains as to how those probabilities relate to reality. That question can be answered in part by his External Validity Graph [3].

The External Validity Graph can be used to determine how well the student does as an estimator of probability. It can be thought of as his "track record" as an estimator. We may collect all those events which he claims to have a probability of, say, .80 of occurring, and determine what percentage of them did, in fact, occur. If he were a perfect estimator, we would expect roughly 80% of them to occur. Figure 9 illustrates such a graph. The diagonal AB represents a perfect estimator. If we collect, say for 20 questions, all the probabilities used by a student and compute the relative frequency with which each particular probability is associated with a correct foil, we may construct an external validity graph for that student and that

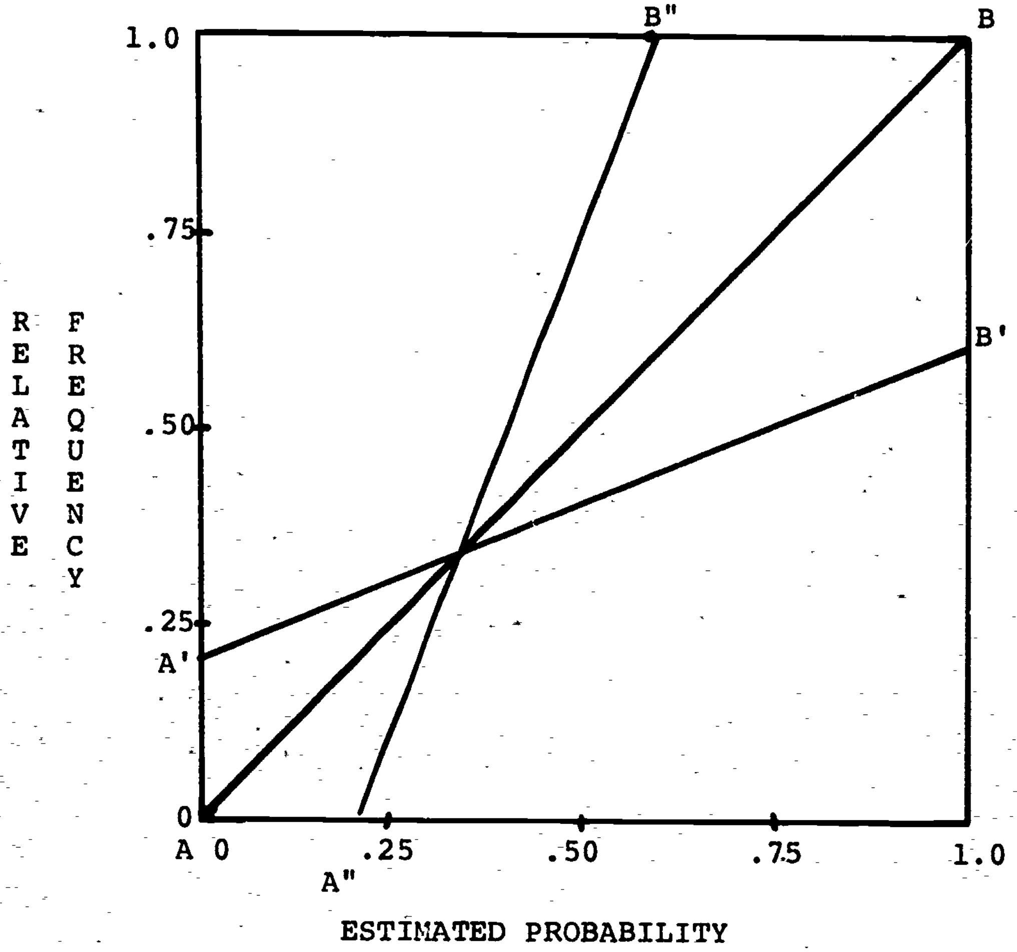
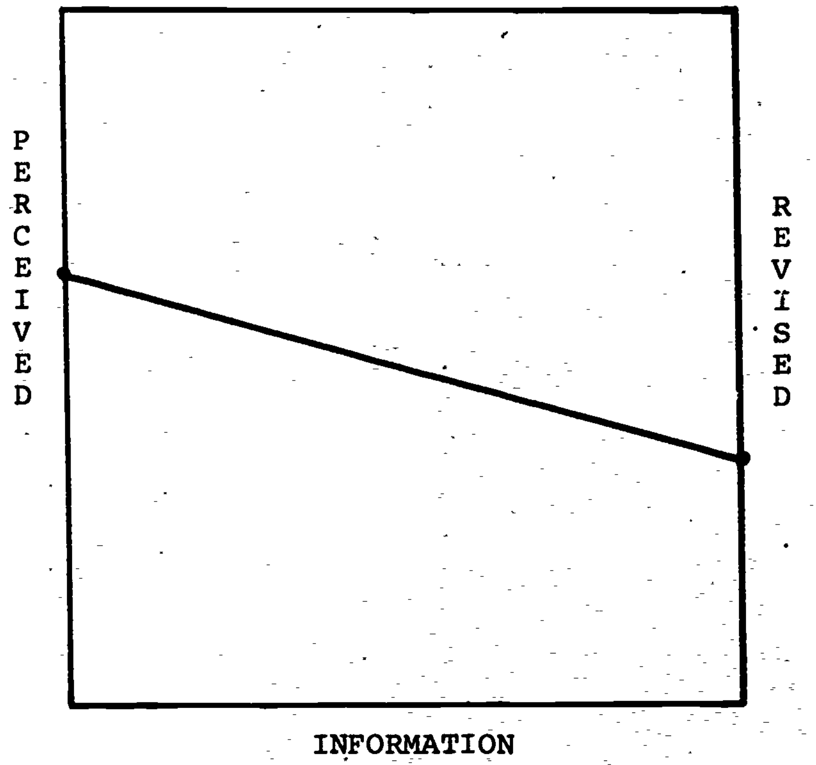


Figure 9: External Validity Graph and Realism Lines

test material. We then assume that the relationship between them is roughly linear and apply a simple linear regression. We call the resulting line the student's realism line. The lines A'B' and A"B" in Figure 9 illustrate possible results. The line A'B' has the following interpretation: about 60% of the events given a probability of 1.0 of occurring did in fact occur. We interpret this to mean that the student tends to over-value his facts and reasons. Similarly, the line A"B" can be interpreted as: the events given a probability of .60 (or greater) of occurring did in fact occur 100% of the time. The student tends to under-value his facts and reasons. We may further interpret the realism line as: "when he estimated x he should have estimated Ax+B" where Ax+B is the equation of his realism line. We may then transform the student's probabilities using his realism line and recalculate his score on the basis of the new probability assignments. This "more realistic" score can then be used to decompose his original score into the portion attributable to lack of information. The lower portion of Figure 10 illustrates such a score decomposition.

The scores are actuall, reported as the difference between the student's scores and the maximum attainable for a given length test. The over-all gain is that maximum less the un-revised score. The gain attainable from more information is that maximum less the revised score. The gain from more "realistic" use of probabilities is simply the revised score less the original score.

The upper portion of Figure 10 brings us to another interesting aspect of the logarithmic scoring system. We call the square in the upper part of Figure 10 the student's information square. The term "information" arises from a very interesting relation between the logarithmic scoring system and the concept of "entropy" in information theory. That relation is developed below.



YOU TEND TO OVER-VALUE YOUR INFORMATION.

YOU CAN IMPROVE YOUR SCORE BY 261 POINTS OVERALL.

YOU CAN IMPROVE YOUR SCORE BY 236 POINTS BY GAINING MORE INFORMATION ABOUT THE SUBJECT.

YOU CAN IMPROVE YOUR SCORE BY 25 POINTS BY MORE REALISTIC USE OF YOUR INFORMATION.

Figure 10: Information Square and Score Decomposition

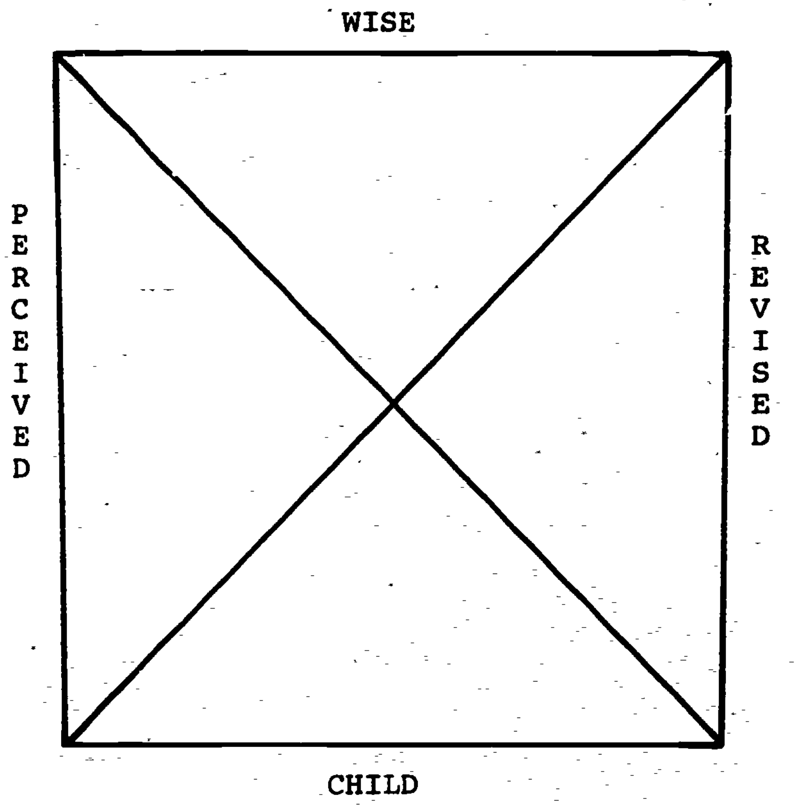
Figure 11 illustrates the expected score for a given question in our three foil system. The expression in the parentheses is called in information theory the "entropy" of the partition p_i. It represents the expected amount of information which will be conveyed by the event itself; or, in other words, the expected surprise content of the event. Equation (2) of Figure 11 is the expected score normalized to the range 0 to 1. Equation (2) can be interpreted as follows: if there is no surprise content in the event then there is perfect information [eq(2) = 1]; if there is maximum surprise content in the event then there is no information [eq(2) = 0]. The upper part of Figure 10 plots on the PERCEIVED axis the point corresponding to the relative information computed from the student's original probability distributions. The relative information derived from the probabilities as transformed by the realism line (the regression line in the external validity graph) is plotted on the REVISED axis and the two points are connected by a line. This information square has an interesting interpretation in terms of an Arabian proverb as depicted in Figure 12.

(1) $$\text{Expected Score} = \sum_{i=1}^{3} p_i (50 \log 3 p_i)$$

$$= 50 \left[\log 3 - \left(-\sum_{i=1}^{3} p_i \log p_i\right) \right]$$

(2) $$\text{Relative Information} = \frac{50 \left[\log 3 - \left(-\sum_{i=1}^{3} p_i \log p_i\right) \right]}{50 \log 3}$$

Figure 11: Information Theory Analogues



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.

Figure 12: An Arabian Proverb

EXPERIENCE TO DATE

The few observations I would like to make here are based largely on anecdotal evidence. We have used every opportunity to test our approach before committing to formal experimentation. We have gleaned our test students from among the Rand professional and secretarial staffs (and their children), high school and college students, casual visitors and attendees at various Rand-sponsored meetings. It is gratifying to report that not one of our subjects has had difficulty with the mechanics of using the system.

Figure 13 illustrates the relation between the actual and perceived information measures for the first exposure of 66 subjects to the test system. This figure supports what we have observed: subjects tend to over-value their facts and reasons at the outset. Our observation has been that most subjects tend to choose the vertices of the triangle rather than try to differentiate more finely among their probability assignments. This seems to be a result of previous training in answering multiple choice items.

Figure 14 illustrates the amount of improvement in realism experienced by our subjects. Among the fifteen individuals who have taken a test on the computer two or more times, the average loss in score due to lack of realism on the first test was 108 points; the average loss in score due to lack of realism on the last test they took was 15 points.

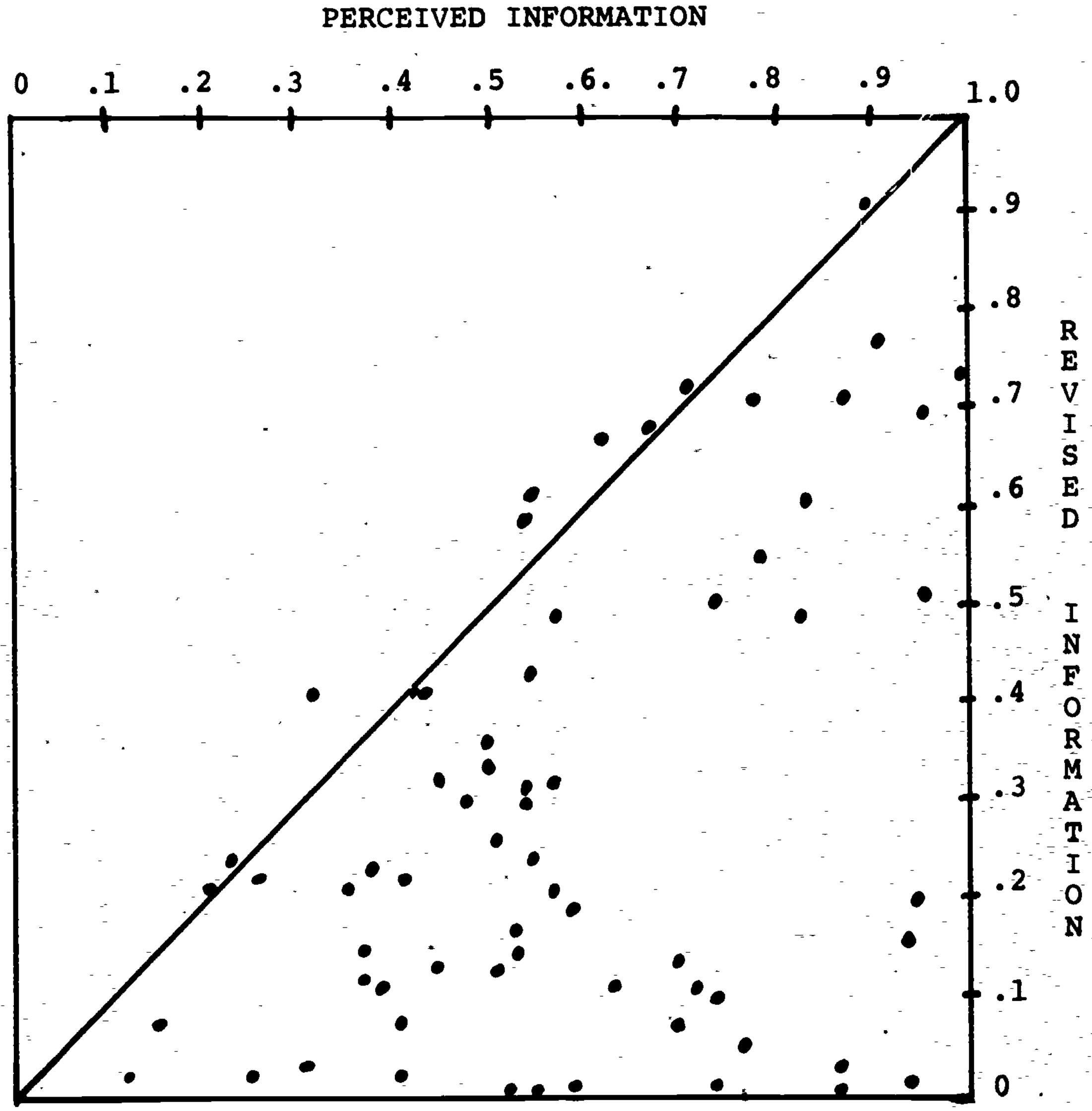


Figure 13: Relative Information on First Exposure (66 Subjects)

Tests	No. of Cases	No. Showing Improvement
1	8	N. A.
2	7	6
3	4	4
4	2	2
5	1	1
⋮		
11	1	1

Av. loss due to lack of realism on first test: 108
" " " " " " " " last "' : 15

Figure 14: Effect of Experience on Realism Score

THE FUTURE

The near term future of this work as we see it consists of at least three parts:

- An on-going improvement of our test and interpretation ideas.
- Several small-scale formal experiments to verify our design.
- An implementation of our ideas on the University of Illinois' PLATO IV system.

This final point indicates our desire to make Admissible Probability Measurement available to a large community of potential users. Within that community we hope to encourage experimentation on a larger scale than we can support with our present facilities.

CONCLUDING REMARKS

At the beginning of the paper, I made reference to the possible importance of Admissible Probability Testing for selection and training in the military services. As a result of his experiment, Gardner [1] found the following hypotheses to be supported:

- The scores attained on a [probability] test will be a more accurate assessment of student knowledge than will choice scores.
- [Probability] test results will be more reliable than choice test results since there is significant reduction in guessing.
- [Probability] tests yield more information that can be used in policy-making for a multi-tracked curriculum than do choice tests.

We have taken these results as an encouragement to continue to develop and extend Computer Assisted Admissible Probability Testing.

ACKNOWLEDGEMENTS

I would like to thank Dr. Emir H. Shuford, Jr., and Dr. Thomas A. Brown, both of the Rand Corporation, for their obviously very important contributions to this work.

REFERENCES

1. Gardner, Willie C. (1969), *The Use of Confidence Testing in the Academic Instructor Course.* Proceedings of the 1969 Annual Meeting of the Military Testing Association, September 15-19, New York, N.Y.

2. Shuford, Emir H. (1969), *Confidence Testing: A New Tool for Measurement.* Proceedings of the 1969 Annual Meeting of the Military Testing Association, September 15-19, New York, N.Y.

3. Brown, Thomas A., *Probabilistic Forecasts and Reproducing Scoring Systems*, RM-6299-ARPA, June 1970, The Rand Corporation, Santa Monica, California.