DOCUMENT RESUME

ED 093 947 TM 003 765

AUTHOR Brown, T. A.
TITLE A Theory of How External Incentives Affect, and Are Affected by, Computer-aided Admissible Probability Testing.
PUB DATE [Apr 74]
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th , Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Computer Assisted Instruction; *Confidence Testing; *Motivation; *Multiple Choice Tests; Response Style (Tests); Scoring Formulas
IDENTIFIERS *Admissible Probability Testing

ABSTRACT

Admissible probability testing is a way of administering multiple choice tests in which a student states his subjective probability that each alternative answer is correct. His response is then scored by an admissible scoring system designed so that the student will perceive that is is in his interest to report his true subjective probability. With regard to admissible probability tests, two issues are treated surrounding the relation between external incentives and optimal student behavior. It is shown that excessibe competition or the use of a strict "pass-fail" system can lead to responses which misrepresent the student's true state of knowledge, and that the use of admissible scoring systems should influence students to study fewer topics to a higher degree of mastery than do other objective scoring systems. Issues treated here are theoretical. Controlled field experiments will discover whether the advantages and dangers theoretically inherent in computer-aided admissible probability testing will show up in real life. (Author/RC)

A THEORY OF HOW EXTERNAL INCENTIVES AFFECT, AND ARE AFFECTED BY, COMPUTER-AIDED ADMISSIBLE PROBABILITY TESTING

By T. A. Brown

April 1974

Session 13.22

A THEORY OF HOW EXTERNAL INCENTIVES AFFECT, AND ARE AFFECTED BY, COMPUTER-AIDED ADMISSIBLE PROBABILITY TESTING

Let me begin by reviewing for you some of the basic facts about admissible probability testing. It is a relatively new way of administering multiple choice tests (see [1], [3], [4]). Instead of asking a student to choose just one of the alternative answers to a question, you ask him to state his "subjective probability" that each alternative answer is correct. His response is then scored by an "admissible scoring system." Such a scoring system is designed so that the student will perceive that it is in his interest to report his true subjective probability. What do we mean by true subjective probability? L. J. Savage defined subjective probability in terms of betting behavior: he would say that if an individual was willing to bet $2.00 to $1.00 that a given event would take place, but unwilling to bet $2.01 to $1.00 on the event, then that individual's subjective probability that the event will take place is two-thirds. An admissible scoring system may be viewed as a system in which we take what a student asserts his subjective probability to be, and make some bets in his name which he would consider good bets if he truly believes in the subjective probability he has asserted to us. Obviously the student is generally only hurting himself in such a system if he exaggerates or understates his subjective probability. There are, of course,

a great many distinct "admissible scoring systems," but in this talk we will generally limit our attention to the logarithmic scoring system. For definiteness and for convenience in graphical display we will consider only two-alternative multiple-choice tests (such as true-false tests). If we normalize the scoring system so that the student gets zero if he expresses complete ignorance (i.e., if he specifies 1/2 as his subjective probability for each of the two alternative answers) and one if he expresses complete and accurate certainty (i.e., if he assigns probability 1 to the correct response and probability 0 to the incorrect response), then our scoring system is simply

$$\log_2 (2p)$$

where p is the probability the student ascribes to the correct alternative.

The purpose of my talk today will be to examine the following two points:

- How is a student's study behavior affected by the knowledge that he will face an admissible probability test rather than a conventional true-false test?
- How will the student's response to particular questions on the test be affected by his knowledge of how the total score achieved will affect him?

First we consider how the use of admissible scoring systems affects, in theory at least, a student's study behavior. To approach this question, we begin by observing that if a student feels that, on some specific true-false question, there is probability p that "True" is the correct answer and probability 1 - p that "False" is the correct answer then his subjective expected score will be

$$p \log_2 (2p) + (1 - p) \log_2 \left(2(1 - p)\right)$$

Figure 1 shows this function in graphical form (the solid line). The dashed line in Figure 1 shows the expected score if the student is facing a true-false test taken and scored in the conventional way, with +1 given for choosing the correct alternative and -1 given for choosing the incorrect alternative. If the student thinks there is a .8 chance that "True" is the correct answer, he will mark "True" and feel he has probability .8 of winning a point and probability .2 of losing a point, for a net expected gain of .6 points.

Notice that as the student acquires information to move his probability away from the state of being uninformed (p = .5), the optimal expected score from the simple procedure increases in proportion to the distance moved along the probability scale while that from the logarithmic scoring procedure increases only slightly at first and then more and more as higher levels of mastery are achieved. Thus, the logarithmic procedure requires a higher level of mastery to yield any given optimal expected score than does the simple choice

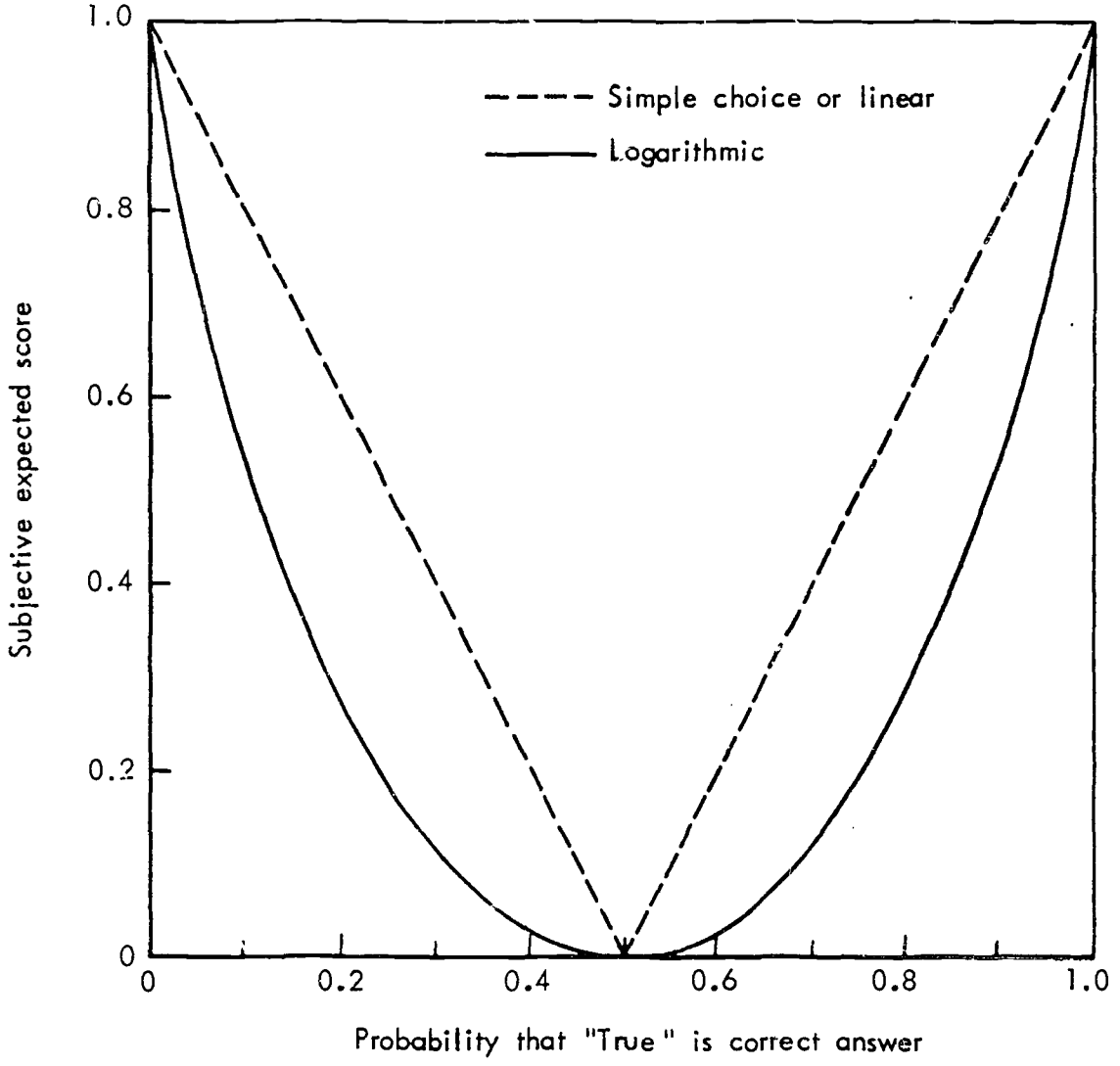

Fig. 1 — Optimal expected score as a function of probability in the case of two answers

procedure and, in this sense, serves as a more stringent incentive system for learning.

Because lower levels of mastery often require less effort to achieve than do the higher levels, admissible scoring systems may prove to be reward systems which can motivate students to achieve higher levels of mastery of a subject matter than they do under a conventional system. To investigate this quantitatively, assume that the student has, for each question, an exponential "learning curve" of the form

$$p = 1 - \frac{1}{2} \exp(-2\lambda c)$$

where c represents the cost to the student in time and energy, say, of the effort he puts into studying the question; λ is a parameter which reflects the "easiness" or rate of learning of the subject matter of the question; and p is the student's subjective probability associated with the correct answer (see Figure 2). Thus, if the student puts no study at all into the question (i.e., c = 0), his probability for the correct answer is .5. As he invests effort in studying the subject matter his probability increases asymptotically toward 1.0.

There are two good ways of modeling the way a student will choose to spend his study time and effort. You may either assume that he has a fixed amount of time available and seeks to allocate it across the questions he expects in such a way as to maximize his optimal expected score; or

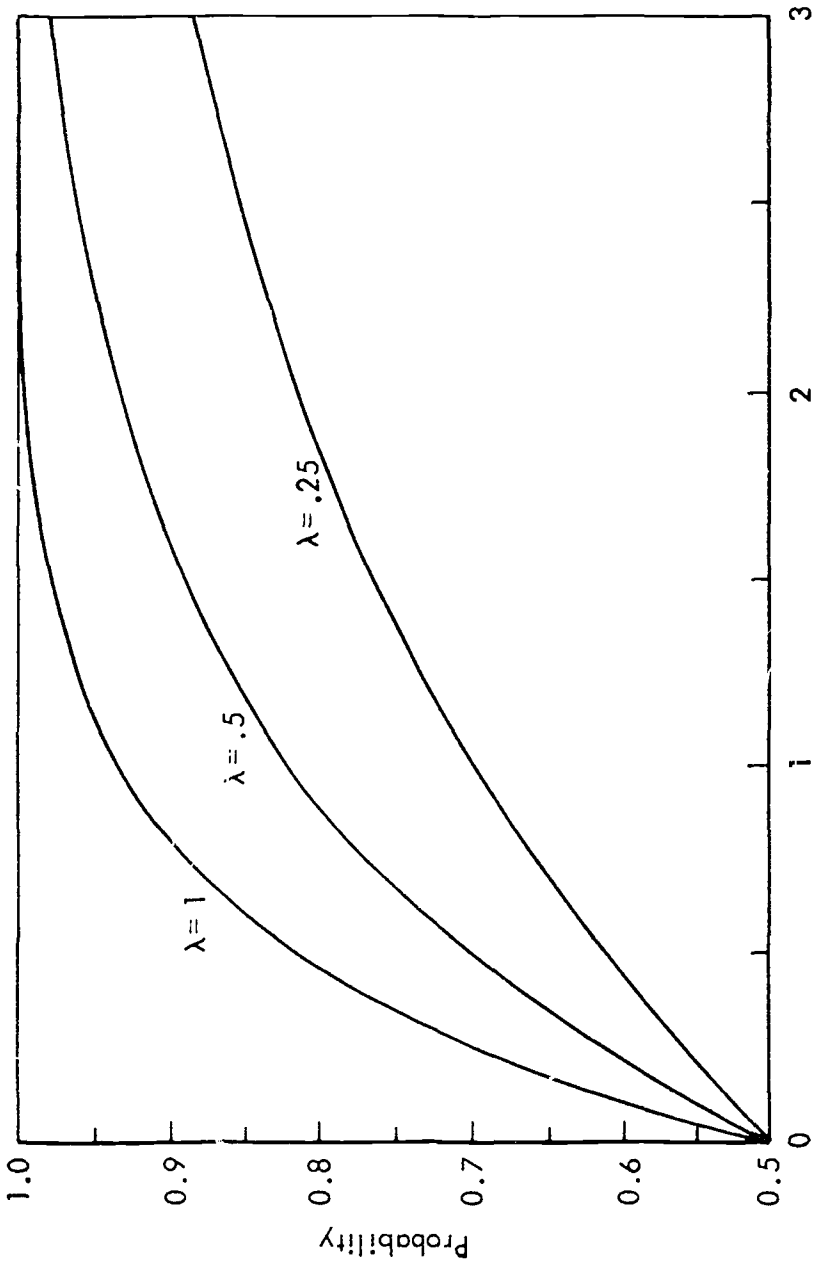

Fig. 2 — Probability as a function of effort, c, where $p = 1 - 1/2 \exp(-2\lambda c)$

you may assume that there is some "exchange rate" between study time and score (e.g., one point of score is worth three minutes of time) and that he will "spend" his time on each question in such a way as to maximize his "profit," i.e., the difference between his expected score on a question and the value of the time he expends on studying it. These approaches will be discussed separately, but it will become apparent that their solutions are closely related.

First, suppose that the student has a fixed and limited amount of study time available and wishes to allocate it over the questions likely to be asked in such a way that he will maximize his expected score. Figure 1 expresses expected score as a function of subjective probability, and Figure 2 expresses subjective probability as a function of effort. We can combine these curves to get expected score (E) as a function of cost in study time (c). This new function is shown in Figure 3 for both the conventional scoring system and the logarithmic scoring system (assuming λ, the parameter reflecting easiness, is one-half). The maximum return (in terms of expected score) per unit of effort may be found graphically by measuring the slope of the steepest line through the origin which is tangent to the optimal expected score function. Analytically, it can be determined by finding the point where the derivative of $\left(\frac{E}{c}\right)$ with respect to c is zero. Now in fact

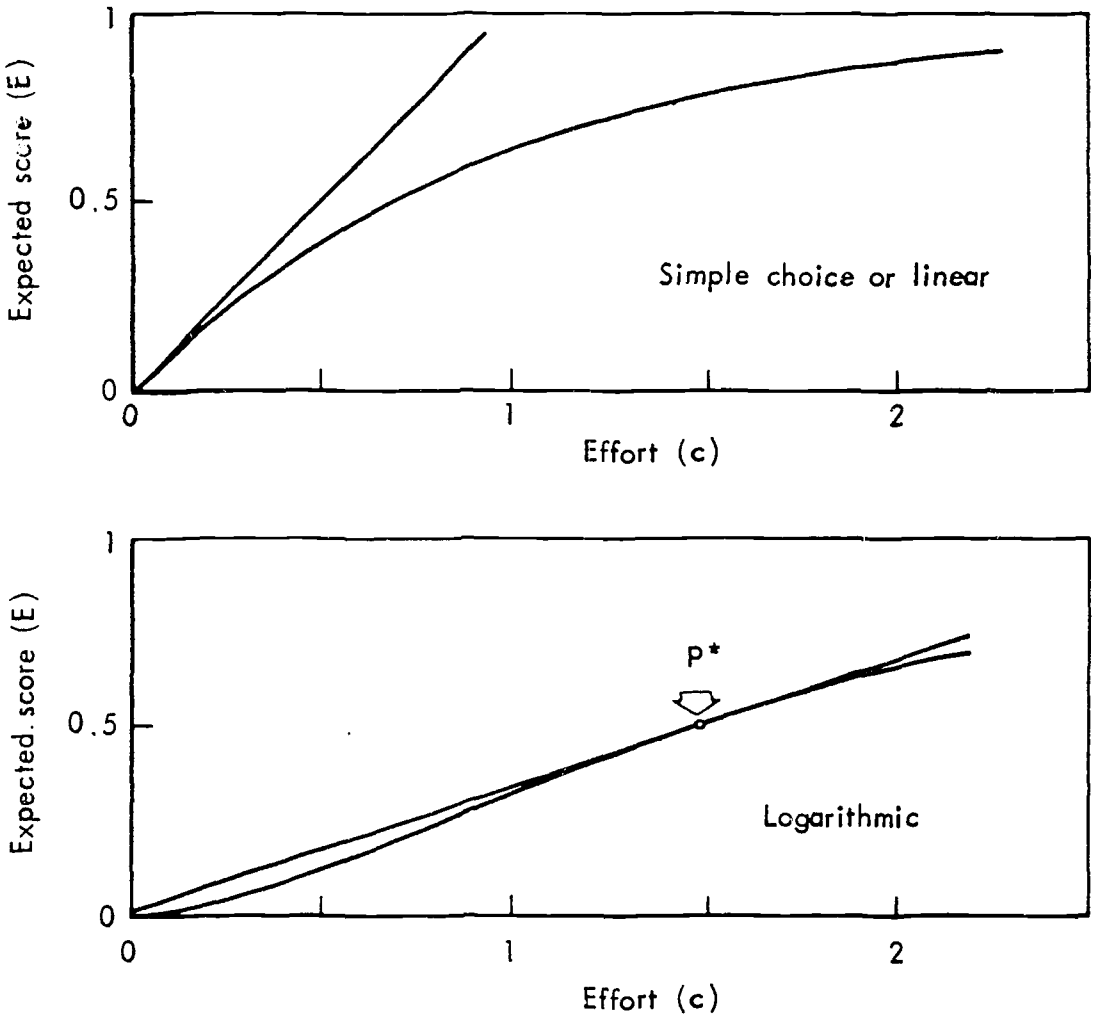

Fig.3 — Optimal expected score as a function of effort (c) when $\lambda = 0.5$

$$\frac{d}{dc}\left(\frac{E}{c}\right) = \frac{1}{c}\cdot\frac{dE}{dp}\cdot\frac{dp}{dc} - \frac{E}{c^2} =$$

$$\frac{-(1 - p)\ \log\ [2(1 - p)]\frac{dE}{dp} - E}{c^2}$$

Because of the particular form chosen p(c), the numerator of this expression depends on p alone, not on c or λ. Thus, there exists a "critical value" of p, say p^*, for any given scoring rule such that on any question, regardless of what λ may be, the student will get maximum reward per unit effort to bring his probability for the correct answer up to p^*.

It is easy to calculate p^* for any given scoring rule. To be specific:

Scoring Rule	Critical Probability
Simple Choice or Linear	.5
Logarithmic	.891....

An allocation procedure which yields an approximately optimal solution to the overall problem (and an exactly optimal solution in most cases) is as follows. Arrange the questions in order of increasing difficulty (so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$). The student should work on the first question until he has expended enough effort so that $p \geq p^*$ and the ratio of marginal return to marginal cost (that is, $\frac{dE}{dc}$) is just equal to the maximal achievable gain per unit effort on the second question. Then he should work on the second question until $p \geq p^*$ and then work on the first and second questions (keeping marginal return ratios equal)

until the marginal return ratios equal the maximal achievable gain per unit effort on the third question. The process is continued until the student has expended all the effort he has available.

This allocation procedure will yield the true optimum for the scoring rules considered above if the student "runs out of gas" at a point where every question which has been worked on at all has been worked on to a point where $p \geq p^*$. In more complicated, non-reproducing scoring procedures that do not have steadily diminishing marginal returns for $p \geq p^*$, the allocation procedure described above will not work so well.

Now obviously a "real-life" student will not go through a careful quantitative analysis of how to allocate his study efforts, but the quantitative model (which may come to represent the behavior of experienced, test-wise students fairly well) does catch one aspect of study behavior which is worth remarking: the use of a logarithmic scoring rule encourages the student to study fewer questions to a higher degree of mastery, while the conventional simple-choice procedure encourages the study of more questions to a lower degree of mastery. Which incentive system is to be preferred depends upon the particular learning situation at hand.

Neither incentive system offers a panacea when study time is strictly limited. On the one hand, use of the conventional simple-choice procedure may mean that none of the subject matter will be remembered more than a few hours or days beyond the time of taking the test. On the other

hand, use of the logarithmic procedure may mean that while some of the subject matter will be remembered, the student will not know enough of the subject matter for it to be of any use to him.

An alternative way of modeling the student's study incentives is to assume that his study time is not strictly limited and that his time has a value to him which is commensurable to the value of the test score he may earn. If the total amount of time which he may spend on study is flexible, he would perhaps attempt to maximize his "profit" on each test question. That is to say, he would choose an expenditure of time c^* on each question which maximizes $E(c) - sc$, where s is the value, in units of test score, of a single unit of time (or study effort). We will assume that the units of time (or study effort) have been chosen in such a way that $s = 1$.

Within the context of the quantitative model it is an easy task to calculate, as a function of λ, the optimal investment strategy and maximal point under both the simple choice and the logarithmic scoring rules. The results of these calculations are graphed in Figure 4. For a given λ the simple choice procedure allows the larger profit and, in this sense, is a more lenient reward system than is the logarithmic. Under the simple choice procedure it does not pay to ever work on a question where $\lambda < .5$ while under the logarithmic the student cannot make a profit if $\lambda < 1.5$. If $\lambda \geq 1.5$, the student will expend considerably more effort under the logarithmic scoring rule.

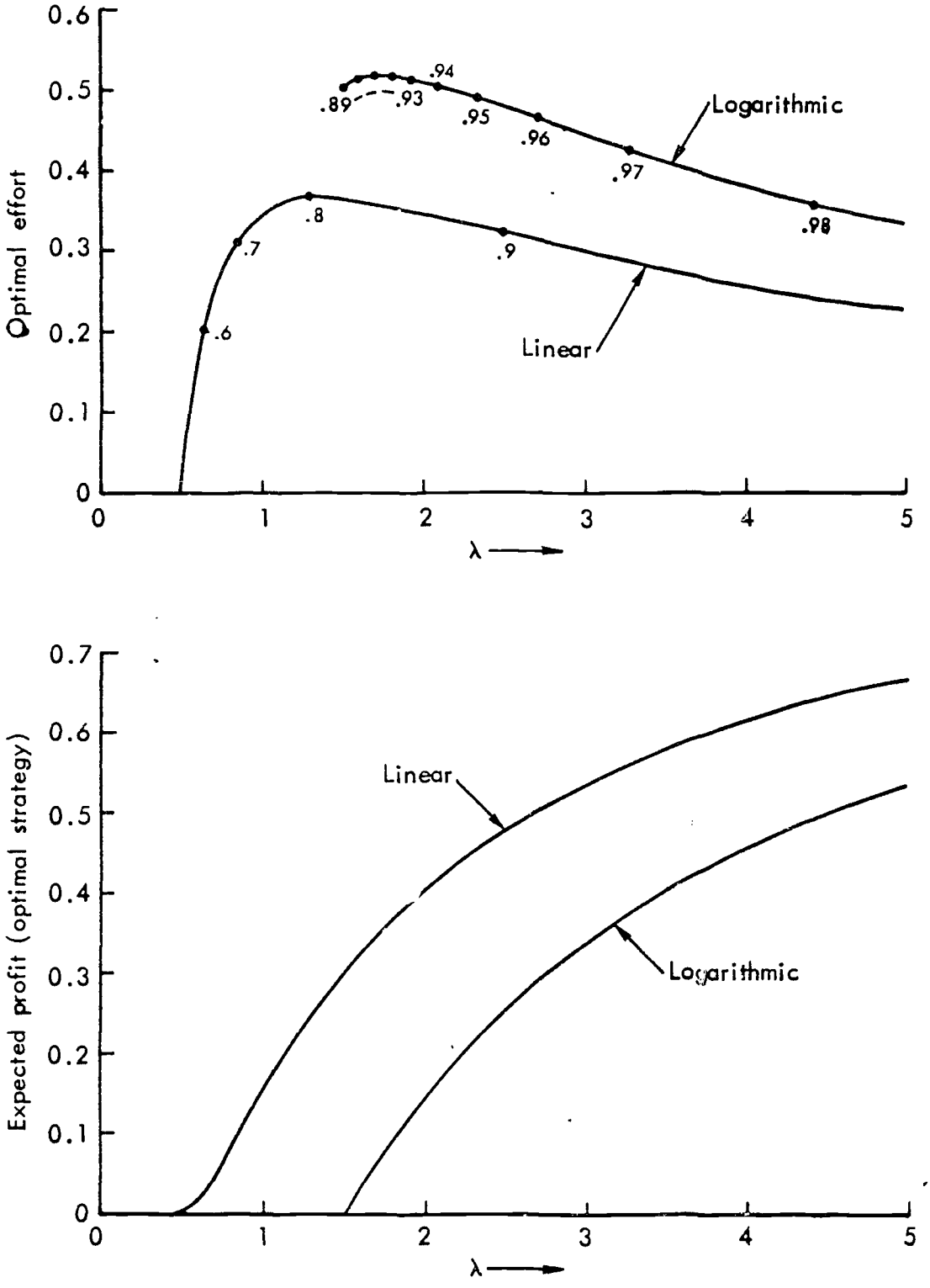

Fig. 4 — Optimal investments and profits as a function of rate of learning

Note, by the way, that if the student studies a question at all under the "maximum profit" approach, he studies it at least up to the level where his probability exceeds p^*, the critical probability of the "optimal allocation" approach.

Thus, the same basic pattern appears under the "maximum profit" approach as under the "optimal allocation" approach. Specifically, the student is theoretically motivated to study fewer questions (through avoidance of the harder ones with $\lambda < 1.5$) but to a higher degree of mastery under the logarithmic scoring rule than under the conventional simple choice procedure. However, the student may be induced to study all of the questions by increasing the reward for learning or by increasing the rate of learning (λ) either through improving learning efficiency or through reorganization of the subject matter.

Rational economic models of the kind we have been discussing here only catch one side of the problem of motivating desirable study behavior by students. General morale is just as important as the realization that they'll get a better score by following better study habits. It seems to me that a student's morale is higher if he regards his instructor as a friend and guide rather than as an antagonist to be outwitted. I believe the conventional multiple-choice test encourages the latter point of view, for the student is forced to express complete confidence in the foil he selects, even if he selects it for very flimsy reasons indeed. He is forced into a pattern

of deception on at least some of the questions, implicitly claiming to be certain of his response when in fact he is not certain at all. An admissible scoring system, on the other hand, enables and encourages the student to give honest answers to all the questions, freely and frankly identifying where there are gaps in his knowledge. The examination thus becomes more a communication device from student to instructor than a duel of wits with a substantial chance element. So I believe you can make a case that admissible probability testing, if it is once understood and accepted by the students, will be a superior motivational influence in terms of morale and attitude as well as in terms of hard considerations of effort versus score trade-offs.

Whether these effects will be observable in real students in real-life situations will be an interesting matter to investigate empirically.

Now let us turn to the question of how a student's behavior on an admissible probability test may be affected by his knowledge of how the total test score is to be used. The simple-choice procedure is relatively insensitive to the reward structure within which it is embedded. As a consequence of this property of the widely used simple-choice scoring procedure, test givers have probably gotten in the habit of ignoring external reward structures. An admissible scoring system makes a test a more sensitive instrument, and this sensitivity opens the door for certain distortions in behavior if the final score is not going to be used properly.

Let me explain by means of an example. Suppose a student is facing a 20 question test, and on each question assesses probability .8 that one alternative is correct and .2 that the other is correct. If he answers each question in this way, he will perceive himself as having a certain probability of achieving any given score. This probability may be closely approximated by a normal distribution (Fig. 5). If, instead of reporting probability .8 for the more likely alternative and .2 for the less likely one, the student reports probability .75 and .25 respectively, what will happen to his perceived distribution of score? The mean of the distribution will slip from 5.56 down to 5.35, but the standard deviation will go from 3.57 to 2.84. At the cost of a small loss in expected value the student is able to get a substantial reduction in the variance of his distribution. On the other hand, if he reports .85 and .15 on each question, the mean declines to 5.30 while the standard deviation goes up to 4.48. When you are close to the optimal response, you can buy a big change in standard deviation by a small sacrifice in expected score. Students will undoubtedly sense this, and under certain circumstances it may introduce a systematic bias into their responses.

For example, suppose that some special prize is to be given to whichever student gets the best score on a given test. This will tend to make students overstate their probabilities (or, to put it another way, to appear to overvalue their information), because the chance of getting

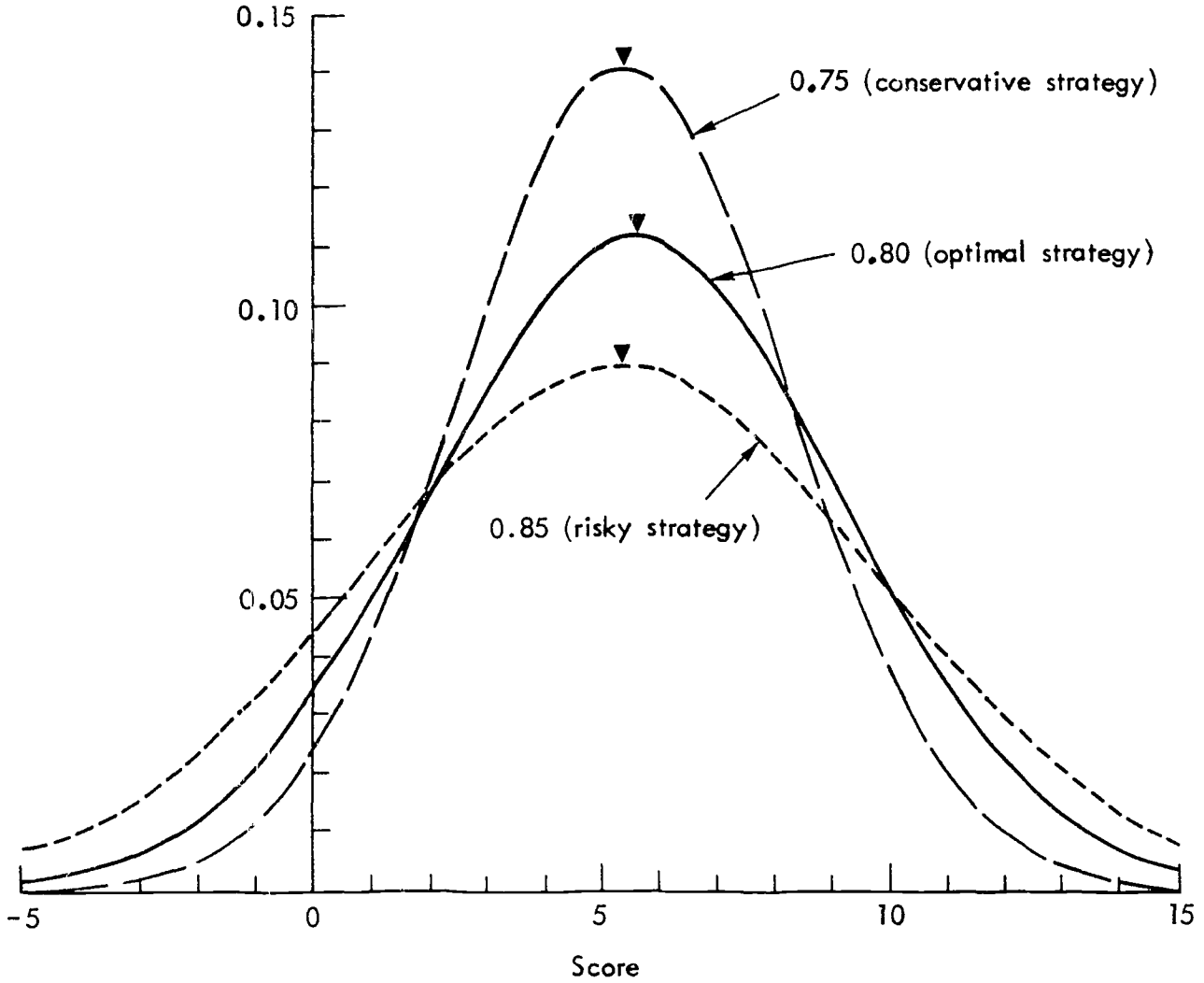

Fig.5 — Distribution of score under three alternative strategies

a really high score will be worth more than the risk of getting an unusually low score (which will be no worse for the student than a mediocre score). The precise quantitative measurement of this effect is very difficult in general, because it involves a multi-person game which is affected not only by each player's perception of the difficulty of the questions but also by his perception of the ability of the other players. Even the case of just two students competing for a prize on a test consisting of one three-alternative question is surprisingly complex (see [2], pp. 12-13).

The special case in which a prize is awarded only in the event that the student makes a perfect score is very easy to understand. With this reward structure, the student should always express absolute certainty no matter how great his uncertainty is in fact. If he fails to do so, he will foreclose any possibility of making a perfect score.

Another context in which a student will be motivated to give responses other than his true personal probabilities is a "pass-fail" system, where he passes the course if he achieves a certain test score or better, and fails the course otherwise. The general problem of determining an optimal response strategy under these circumstances is mathematically very complex, and no general solution is known. The following simplified example, however, can be solved. It illustrates very clearly how the imposition of a "pass-fail" reward structure on top of a reproducing scoring system may

undermine the incentive for responses which accurately reflect uncertainty.

Suppose that a student is facing an exam consisting of n two-alternative items. Suppose these questions all "look alike" to the student in the sense that on each question he has a fixed probability distribution, p and 1 - p, with $p \geq 1/2$. Suppose that he requires a total score T on the test in order to pass. He wants to choose a fixed response r to assign to the preferred answer to each question. What value of r should he choose in order to maximize his probability of passing the test? It is not hard to show that he will have the maximal probability of passing if he chooses r in such a way that his expected score on each question is $\frac{T}{n}$. Note that this r does not depend on p at all! So the student's optimal test-taking strategy depends only on what score he must make in order to pass, and not on his level of knowledge with respect to each test item. In short, this reward structure utterly destroys the reproducing character of the scoring rule. Figure 6 illustrates the student's probability of passing as a function of his response strategy in the particular case where $p = .8$, $n = 20$, and $T = .58$. Note that the student will be about nine times as likely to fail the test if he pursues the "maximum expected value" strategy as he will be if he follows the "maximum probability of passing" strategy.

In an actual situation the reproducing character of the scoring rule would not be completely washed out,

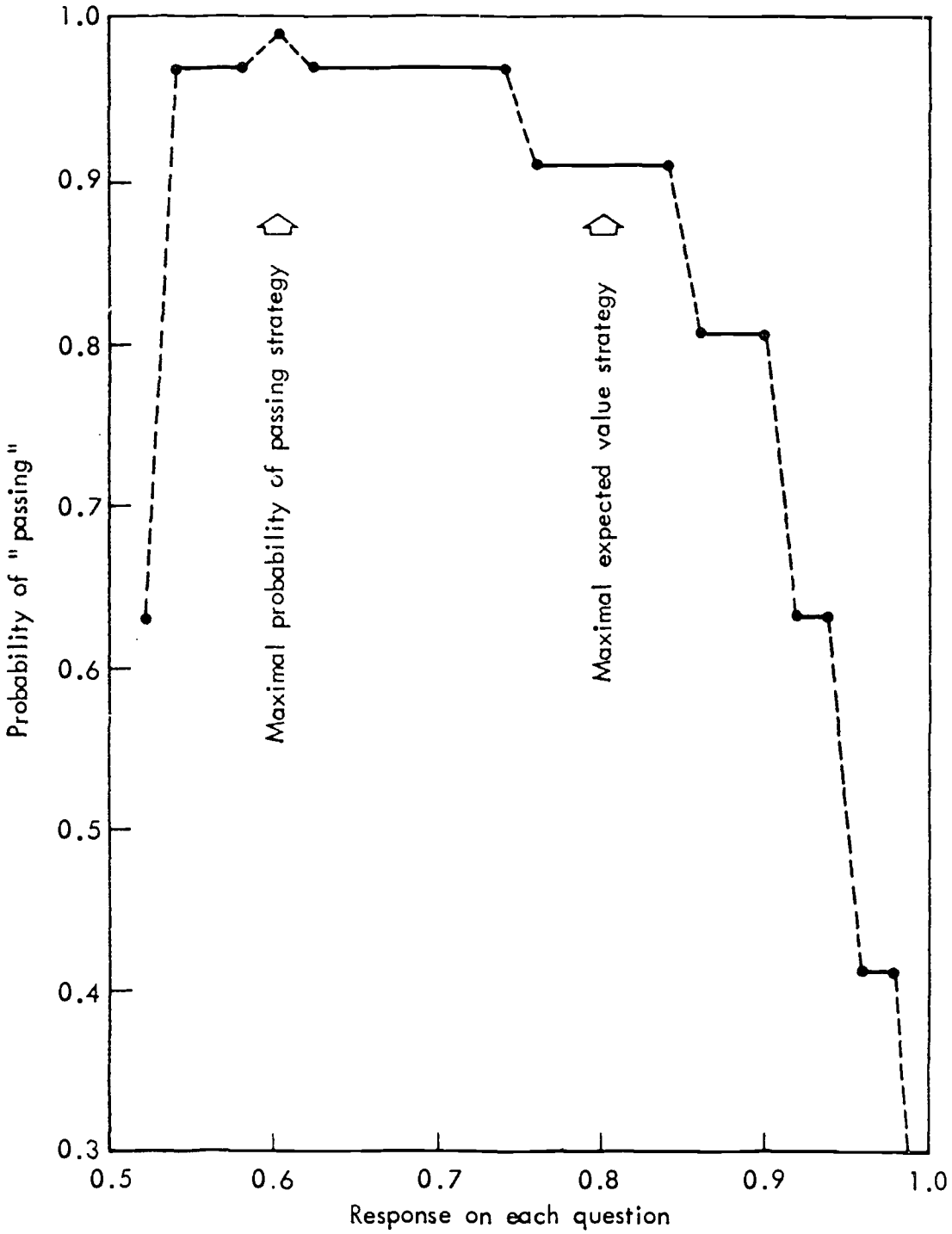

Fig.6 — Probability of "passing" as a function of response to each question in a 20 question test; Logarithmic scoring system; probability 0.8 on each question; Score required to pass T = 0.58

however, because the student would not have precisely the same probability distribution for each item. It seems intuitively evident (although a rigorous proof has not yet been discovered) that his best strategy would be to let his responses vary with his subjective probabilities, but hedge all of them either up or down.

The best remedy is probably to avoid creating reward structures which put a highly non-linear value on points earned under an allegedly reproducing scoring rule. Another (partial) remedy is to avoid letting the student know how many questions there are on a test, or how difficult they are, before he begins to take it.

Everything I've said today is, of course, theoretical. We don't know if actual students will exhibit the behavior which our theory predicts or not. We're anxiously looking forward to controlled field experiments such as are now going on at Air University, to discover whether the advantages and dangers theoretically inherent in computer-aided admissible probability testing show up in real life.

REFERENCES

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

2. Brown, T. A., *Probabilistic Forecasts and Reproducing Scoring Systems*, The Rand Corporation, RM-6299-ARPA, June 1970.

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

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