AFOSR 68-2154

AD 677202

1. This document has been approved for public release and sale; its distribution is unlimited.

THE SHUFORD-MASSENGILL CORPORATION

A LOGICAL ANALYSIS OF GUESSING

H. EDWARD MASSENGILL and EMIR H. SHUFORD, Jr.

THE FIRST SEMIANNUAL TECHNICAL REPORT (WHICH COVERS THE PERIOD MAY 1966 THROUGH OCTOBER 1966) OF WORK PERFORMED UNDER CONTRACT NUMBER AF 49 (638) - 1744, ARPA ORDER NUMBER 833, BY THE SHUFORD - MASSENGILL CORPORATION, P.O BOX 26, LEXINGTON, MASSACHUSETTS, 02173

DECISION-THEORETIC PSYCHOMETRICS: AN INTERIM REPORT, NOVEMBER 1966

Emir H. Shuford, Jr. and H. Edward Massengill

ABSTRACT

In Section A, A Logical Analysis of Guessing, appropriate test-taking strategies are derived for six major test-scoring procedures. Three commonly used definitions of guessing are interpreted as corresponding degree-of-confidence distributions. The ability of the testing procedures to separate these distributions from those representing higher degrees of knowledge is considered with the major result that only admissible probability measurement performs satisfactorily.

In Section B, The Effect of Guessing on the Quality of Personnel and Counseling Decisions, the fundamental probability distributions for total test scores are derived by assuming that each person knows the answers to some items and guesses on the remaining items. Analysis of a 10-item test shows that guessing levels encountered in practice (a) seriously degrade the value of selection, placement, and counseling decisions, (b) significantly impair test reliability and validity, and (c) magnify the influence of testwiseness.

In Section C, The Worth of Individualizing Instruction, equations are developed for expressing the cost and gain for applying an instructional sequence. The expected return from assigning instruction on the basis of (1) admissible probability measurement, (2) admissible choice testing, (3) conventional choice testing, (4) prior information only, and (5) matching the average student is computed for each of seven distributions of state of knowledge. The performance of (1) is outstanding; that of (2), (3), and (4) is disappointing, while (5) does surprisingly well.

DECISION-THEORETIC PSYCHOMETRICS: AN INTERIM REPORT, NOVEMBER 1966

Emir H. Shuford, Jr. and H. Edward Massengill

Probably the most significant development in applied mathematics occuring in this century is the conjoining of probability theory and utility theory to yield what is now commonly referred to as decision theory. The basic foundations for this area of mathematics have been provided by Ramsey (1926), de Finetti (1937), and Savage (1954). The major quantitative techniques have been integrated and extended by Raiffa and Schlaifer (1961). Decision theory, like all applied mathematics, is a tool, the use of which guarantees one vital property, consistency in thought and action. The domain of application of decision theory is quite broad since, in principle, it applies to all behavior. Given the decision maker's view of the decision problem, his information, and his values, decision theory aids the decision maker by placing certain constraints on his behavior. These constraints are those implied by the necessity for mathematically consistent and coherent behavior. (See de Finetti, 1937.) Thus, it should be clear that decision theory is not a moral system for dictating the choices of people, but rather is an aid for understanding the logical and mathematical implications of a decision problem (Toda & Shuford, 1965).

The first major application of decision theory to psychometrics was reported by Cronbach and Gleser (1965). They used decision theory to study factors affecting the quality of institutional decisions made on the basis of testing information. These include the typical personnel decisions such as selection, classification, and placement. Such decisions are called institutional because they are made on behalf of an institution, say, one of the military departments, a company, or a school. Confining themselves to institutional decisions, Cronbach and Gleser had to deal only with situations in which the utilities could, in principal, be defined in monetary terms and the probabilities could be interpreted in terms of relative frequencies of occurrence in large populations of individuals. Within this context, Cronbach and Gleser were able to develop many fresh and interesting insights into the psychometrics of conventional testing.

At about the same time, a number of widely scattered investigators were using decision theory to develop procedures for measuring an individual's subjective probabilities; Masanao Toda (1963), in Japan; van Naerssen (1961) and de Finetti (1962), in Europe; and Roby (1965), in the United States, independently developed measurement procedures having the property that an individual could maximize his expected utility if, and only if, he honestly expressed his subjective probabilities. Shuford, Albert, and Massengill (1966) integrated and extended this work under the rubric of admissible probability measurement procedures. This conceptual

development and the consequent realization of practicable methods for use in educational and personnel testing appear to have profound implications for psychometric theory and practice (See Massengill & Shuford, 1965; Shuford, 1965; Shuford & Massengill, 1965). In essence, an individual's subjective probability, degree of confidence, or degree of belief in the correctness of answers to objective and semi-objective test items can now, for the first time, be measured in a valid and defensible manner. Admissible probability measurement procedures can be substituted for conventional choice methods in all power tests.

But what is the point of doing this? Why should the conventional procedures which have served so well in the past be replaced? The key to the most general answer that can be given lies in the notion of information. Testing should be used to provide information to someone. How much information can the test provide? This depends, in part, on the method of testing used. In conventional choice testing, each item can provide at most a few bits of information, since only several discrete responses are available to the taker of the test. In admissible probability testing, the taker of the test can respond to each item with a nearly continuous probability distribution. Thus, an order of magnitude increase in information is possible. So, in general, admissible probability testing provides a great deal more information than does conventional testing.

So, on the one hand, we have the application of decision theory to the analysis of institutional decisions providing techniques for arriving at the value of testing information and, on the other hand, we have the application of decision theory to individual decisions creating new testing methods which yield vastly more test information. This looks like the beginning of a revolution in psychometrics. This revolution should be informed by knowledge--knowledge as to how valuable this additional test information will prove to be in practice. What gains can be expected from incorporation of admissible probability measurement procedures into existing education and personnel practices? What totally new and highly effective practices can now be developed to exploit this additional information? Psychometric theory judiciously interpreted and applied can serve to guide these developments but, in order to have an integrated theory which is consistent throughout from the level of a person responding to a test item up to the level of setting personnel policies on a national scale, decision theory must be used.

So this is what decision-theoretic psychometrics is all about. In this report we begin an attack on three different problem areas: (1) A Logical Analysis of Guessing, (2) The Effect of Guessing on the Quality of Personnel and

Counseling Decisions, and (3) The Worth of Individualizing Instruction. The first two studies are mainly concerned with the benefits accruing from substituting admissible probability procedures in current educational and personnel practices while the third study begins to consider the probable benefits of adopting new educational practices.

The first study is concerned with the logic of guessing, both from the point of view of the person taking the test and from the point of view of the person interpreting test data. There is really quite a bit of confusion in the literature as to just what guessing is. Here we are able to use decision theory to explicitly define guessing and, hopefully, to eliminate the confusion. A rather surprising result of this analysis is that constructed-response or fill-in-the-blank tests can be affected just as much by guessing as multiple-choice and true and false tests. This is a dramatic contradiction of the generally held opinion that constructed-response tests are unaffected by guessing. Another surprising result is that none of the techniques devised and advocated over the years as a means of eliminating guessing actually work. They do not penalize guessing. And finally, intuitive explanations are offered for the remarkable increase in reliability observed as a result of changing to an admissible procedure.

The second study develops an explicit, and not too unrealistic, model for standard achievement and ability tests. Numerical methods are then used to compute the degree to which guessing degrades the value of test information for several classes of decisions based upon the results of testing. The most surprising result of this study has to do with the area of selection and classification testing. It is generally thought that the nature of these personnel decisions, for example, where utility is linear in the actual achievement or ability level of an individual assigned to a group, is such that guessing either has no effect whatsoever or can be compensated for by a simple correction for guessing. This widely held opinion is contradicted by the results of this study which indicate that the quality of selection and classification decisions can be seriously degraded by the effects of guessing. A second, possibly less surprising but more dramatic result is in the area of counseling decisions where the test results are used to estimate a person's ability or achievement level. Here, the results of this study show, not only that guessing seriously degrades the value of these estimates, but that under a wide range of conditions, an individual would be better advised if he were not given any test and just assigned an average ability level than if he were sent through a testing program and the procedures recommended in test manuals and text books were used to estimate his ability or achievement level. In other words, in these

situations the value of testing is not just low, it is negative and can represent a serious injustice to a person. Next, since variances and test reliabilities are important in research studies and factor analyses, a model for test-retest reliability is defined. Numerical computations show the loss in test reliability due to guessing is quite dramatic. Finally, some consideration is given to the effect of the individual difference of test-wiseness from the point of view of the individual and of the institution.

The third study develops an explicit formulation for a class of decision processes necessary to individualized instruction. Numerical methods are used to compute the value of (a) precisely tailoring the instruction to the ability level of each person, (b) choosing to treat all persons as being either completely misinformed, maximally uncertain, or completely informed, (c) using conventional choice testing to decide which of the three ways to treat each person, (d) using an admissible choice procedure to decide which way to treat each person, and (e) matching instruction to the average person. The relative effectiveness of these various instructional strategies is investigated for different distributions of initial knowledge levels among persons. One of the more surprising results is that individualizing instruction sometimes yields quite trivial or no improvement over more rigid procedures. Of some interest is the finding that choice testing, either conventional or admissible, is of value over a rather limited range of conditions. When choice testing is of value, admissible choice testing evidences a slight superiority over conventional choice testing. In these situations, admissible probability measurement, of course, yields quite major gains in the value of the instructional strategy.

A. A Logical Analysis of Guessing

When an individual sits down to take an objective test there are two things which determine his score on the test. First, there is his knowledge about the items on the test. Second, there is the strategy which he uses in answering the items. Once the student is in the testing situation there is little he can do about increasing his knowledge but he can guarantee himself of making the best expected score on the test given the amount of knowledge he has by using an appropriate strategy.

Suppose a student were to go to a mathematician for advice about what test-taking strategy he should use. In order to give an individual this advice, a mathematician would need to know the particular scoring system which was going to be used in grading the test. Given this information the mathematician could determine the test-taking strategy which would allow the student to make his highest expected score given the knowledge he has at the time which he takes the test.

In an analogous fashion, suppose an individual is planning to give a test and is interested in having the test yield the maximum amount of information about the knowledge of each person who takes the test. There are two determinants governing how much information can be obtained about the knowledge of a person taking the test. One has to do with the particular test items which are used on the test and the other with the scoring system which is used to grade the test. Thus, an individual with a set of test items could also consider going to the mathematician to obtain information concerning which of many possible scoring systems would give him the most information about each person taking the test in question.

This section deals first with the type of advice that could be given to an individual taking a test and then with the type which could be given to an individual giving a test. For the person taking the test, the problem is how to achieve his highest expected score given the knowledge he has at the time that he takes the test. For the individual giving a test the problem is how to get the most information concerning each person taking the test. In this section we will examine the various proposed scoring systems and illustrate the strategies which the mathematician would recommend to a person wanting to maximize his expected test score. Then we will examine these scoring systems in the light of the information they provide an individual giving a test.

KNOWLEDGE

We will define a person's knowledge about a given test question as his degree of confidence in the correctness of each of the possible answers to the question

(Shuford & Massengill, 1965). Since there are many possible degree-of-confidence distributions for an item with m alternatives and since the individual does not always know which distribution he will have for a test item, he needs to obtain information from the mathematician which will indicate the best strategy for any possible degree-of-confidence distribution.

For two and three alternative test items the possible degree-of-confidence distributions can easily be represented graphically. In our discussions of the various scoring systems we will use this graphic method of presentation. Though we will talk only in terms of two and three alternative items it should be realized that the results can be generalized to items with any number of alternatives.

Figure 1 shows the representation of all the possible degree-of-confidence distributions for a three alternative question. Each point within the graph represents a set of three degree-of-confidence values: c_1=the degree of confidence in A_1(Alternative 1), c_2= the degree of confidence in A_2, and c_3= the degree of confidence in A_3(c_3= 1-c_1-c_2). The arrow in Figure 1 shows the point in the graph for which c_1=.2, c_2=.1, c_3=.7.

Notice that the scale for c_1 moves from the left hand side of the triangle to the lower right hand corner going from zero to one. The scale for c_2 moves from the right hand side of the triangle down to the left hand corner going from zero to one. The scale for c_3 moves from the bottom of the triangle to the top going from zero to one. This triangle also has the property that the base line, i.e., the line going from A_2 to A_1 represents (but on an expanded scale) all of the possible degree-of-confidence values for a two-alternative item. This means that we can use the representation in Figure 1 to talk about both two and three alternative items.

Figure 2 illustrates some special points within the trianguiar coordinate representation introduced in Figure 1. One point of interest is that at which c_1=c_2=c_3=1/3. This is the point in the very center of the triangle. The other points of interest are actually continua of points. For example,the line running from the center of the triangle out to the right hand side for which c_1=c_3>c_2. The other two cases of interest are analogous. With this fundamental information concerning the representation of an individual's knowledge for a test item, we are prepared to examine the various possible scoring systems.

THE CONVENTIONAL CHOICE SYSTEM

The conventional choice system is familiar to all who have taken objective tests. This is the scoring system which gives an individual one point if he chooses the correct answer and zero points if he chooses an incorrect answer or skips the question.

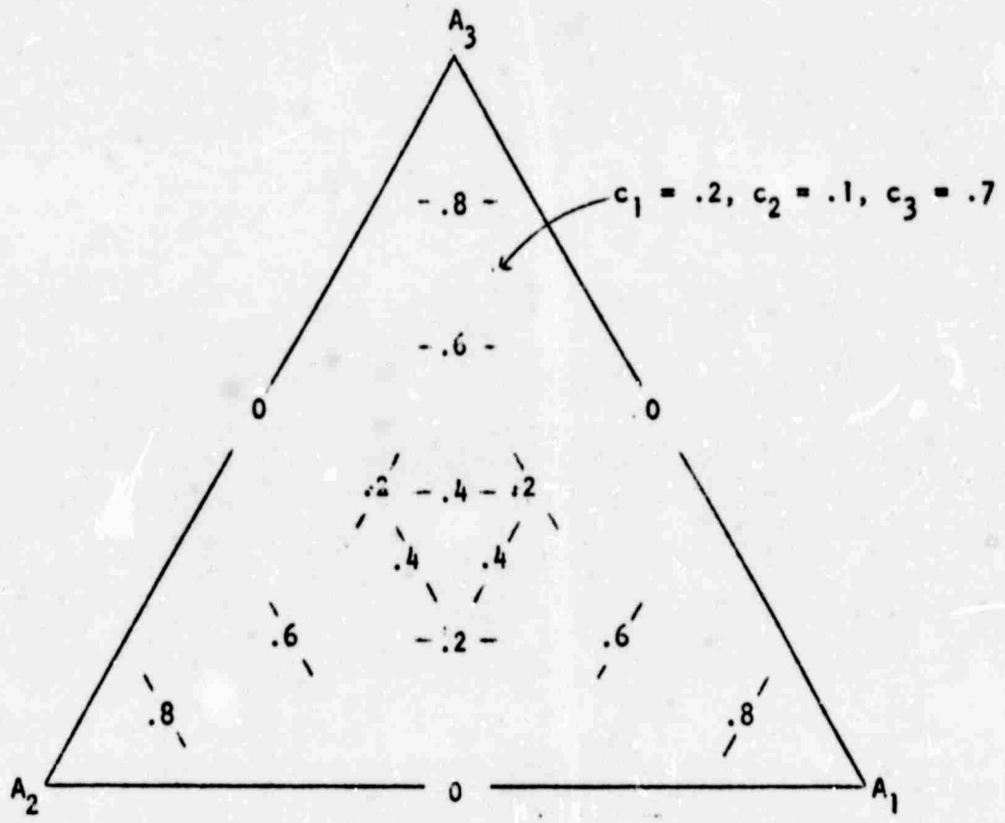


Figure 1. A representation of all possible degrees of confidence for a three alternative question.

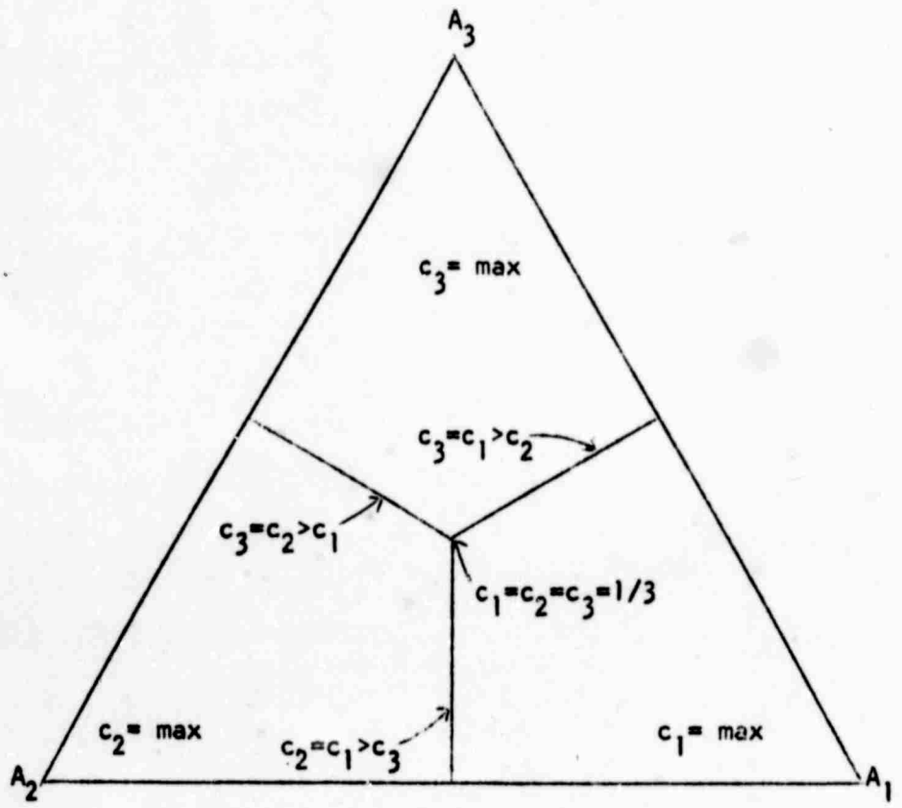


Figure 2. Illustration of special points and areas within the triangle representing the knowledge states for a three-alternative question.

The score table given below represents the situation for the conventional scoring system.

		CORRECT ANSWER		
		A_1	A_2	A_3
CHOICES	a_1: Choose A_1	1	0	0
	a_2: Choose A_2	0	1	0
	a_3: Choose A_3	0	0	1
	$\tilde{a}$: Omit item	0	0	0

The rows of the table represent the possible choices that a person has while the columns represent the possible answers to the question. The numbers within the table represent the score he will receive if he chooses a particular answer or chooses to skip the question and a given answer is correct. Thus, for example, if a person chooses A_1 and A_1 is correct he will receive one point while if he chooses A_1 and A_2 is correct he will receive zero points. For the conventional scoring system, the table for a two-alternative item may be obtained by omitting column A_3 and row a_3, i.e., the scores do not depend upon the number of alternatives.

Once we have the possible score values for this scoring system we can determine the conditional expected score of an individual for each possible choice. This expected score is conditional upon the individual's degree-of-confidence distribution. Thus for each point in the triangle in Figure 1 there will be an expected score for each possible choice. See Figure 3. Our main interest is in determining which choice has the maximum expected score for each point. Thus, there are some points for which a_1 gives the maximum expected score, some for which a_3 does, some for which a_2 does, and none for which $\tilde{a}$ does. Notice that the expected scores along the A_2A_1 line are the expected scores for a two-alternative question.

Having determined the maximum expected score for each point in the triangle we can show the particular regions of the triangle for which the individual should choose A_1, A_2, A_3, etc. Figure 4 shows these decision regions. By comparing Figure 4 with Figure 2 we can see that the two figures are exactly alike except for labeling. Thus Figure 4 can be interpreted as recommending that when c_1 is maximum A_1 should be chosen; when c_2 is maximum A_2 should be chosen; etc. When $c_3=c_1>c_2$, the line running from the center of the triangle to the right hand edge, then either A_1 or A_3 may be chosen. Analogous remark may be made for the other two lines running from the center of the triangle to the left hand side and to the

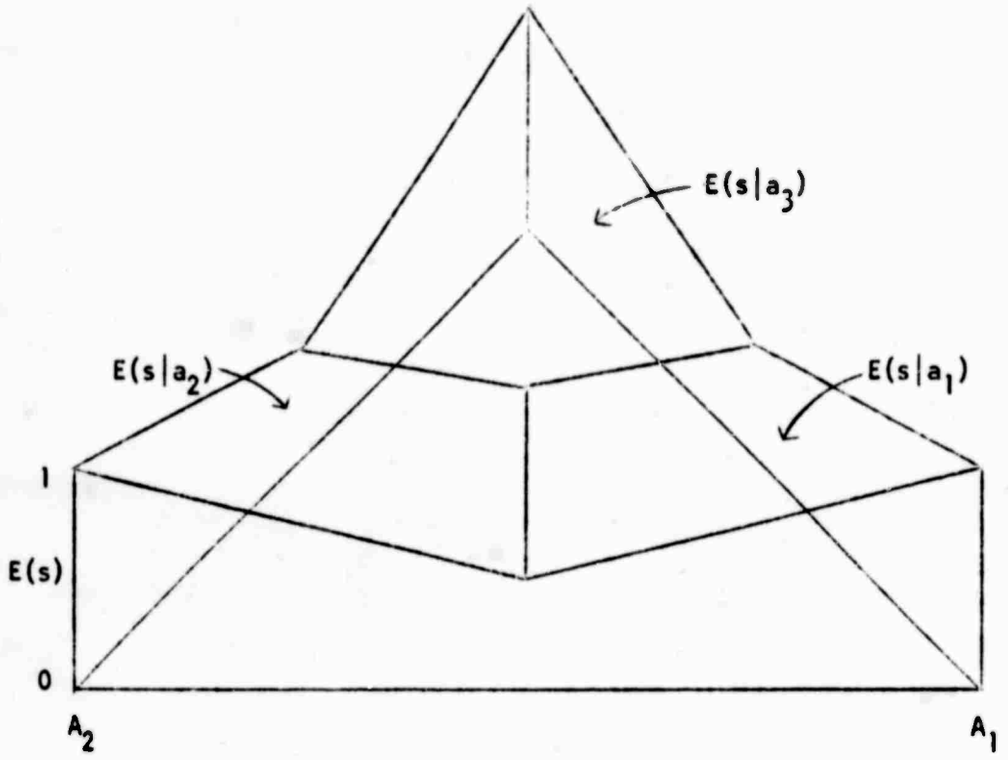


Figure 3. The maximum expected scores for a three-alternative question given the conventional scoring system.

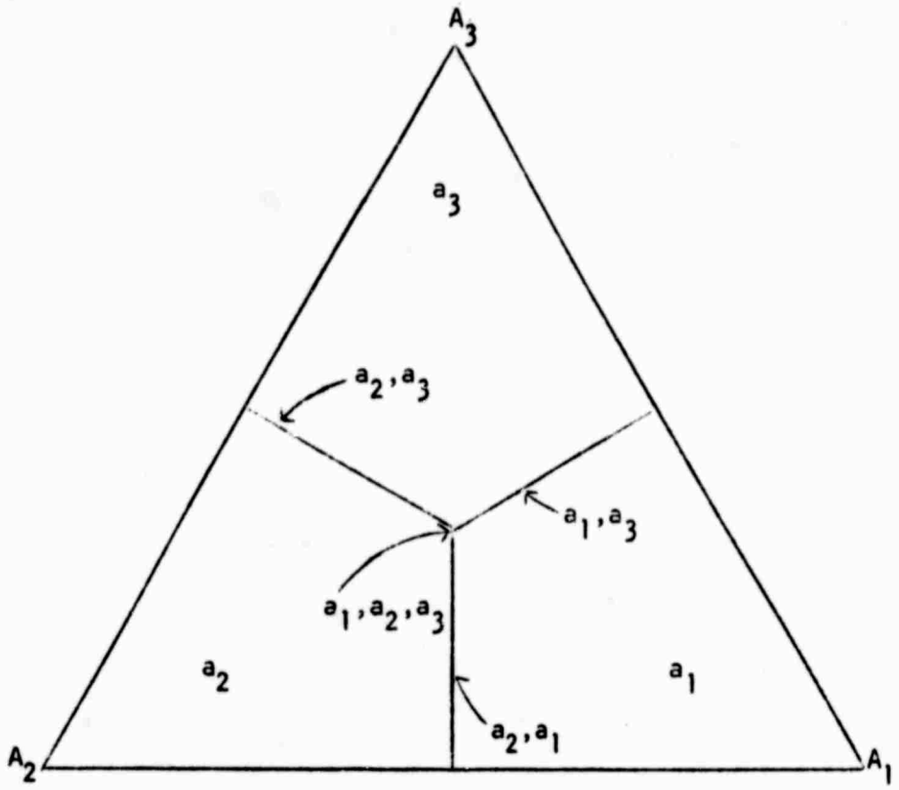


Figure 4. The decision regions for a three-alternative question given the conventional scoring system.

base line. If all three of the degrees of confidence are equal, i.e., the middle point in the triangle, either a_1, a_2 or a_3 may be chosen. Notice that the individual should never skip a question.

We can interpret the decision rule for a two-alternative question by looking at the base line of the triangle in Figure 4 and we see that if c_1 is maximum, A_1 should be chosen; if c_2 is maximum, A_2 should be chosen; whereas if $c_1=c_2$, either A_1 or A_2 may be chosen. Here again, the individual should never skip a question.

Now we can summarize the recommended strategy for an individual who wishes to maximize his expected score under the conventional scoring system. That strategy is: never skip a question, give that answer for which you have the highest degree of confidence, and if two or more possible choices have the maximum degree-of-confidence, then choose either one of them.

THE CORRECTION SYSTEM

For this scoring system, an individual receives one point if his answer is correct, $-1/(m-1)$ if his answer is incorrect, and zero if he skips the question. This scoring system is derived from the correction for guessing formula: $R-W/(m-1)$. With this scoring system as well as all others which we discuss in this section, we will rescale the points into the same units as those used for the conventional scoring system. This will not change the recommended strategies. See the score table below.

	A_1	A_2	A_3
a_1	1	0	0
a_2	0	1	0
a_3	0	0	1
$\tilde{a}$	1/m	1/m	1/m

Now we can obtain the expected score for each of the possible choices. Figure 5 shows the maximum scores for the correction system. Notice that for both the two-alternative and three-alternative item there is one point in each expected score graph for which the expected score for skipping is equal to the maximum expected score. For the case of the two-alternative question, this situation arises when $c_1=c_2=.5$. For the three-alternative question the situation arises when the individual has equal degrees of confidence on all of the possible answers. Notice also, however, that in each of the two situations the other choices also

*The expected scores of all possible acts (including $\tilde{a}$) are equal at $c_1=c_2=c_3=1/3$.

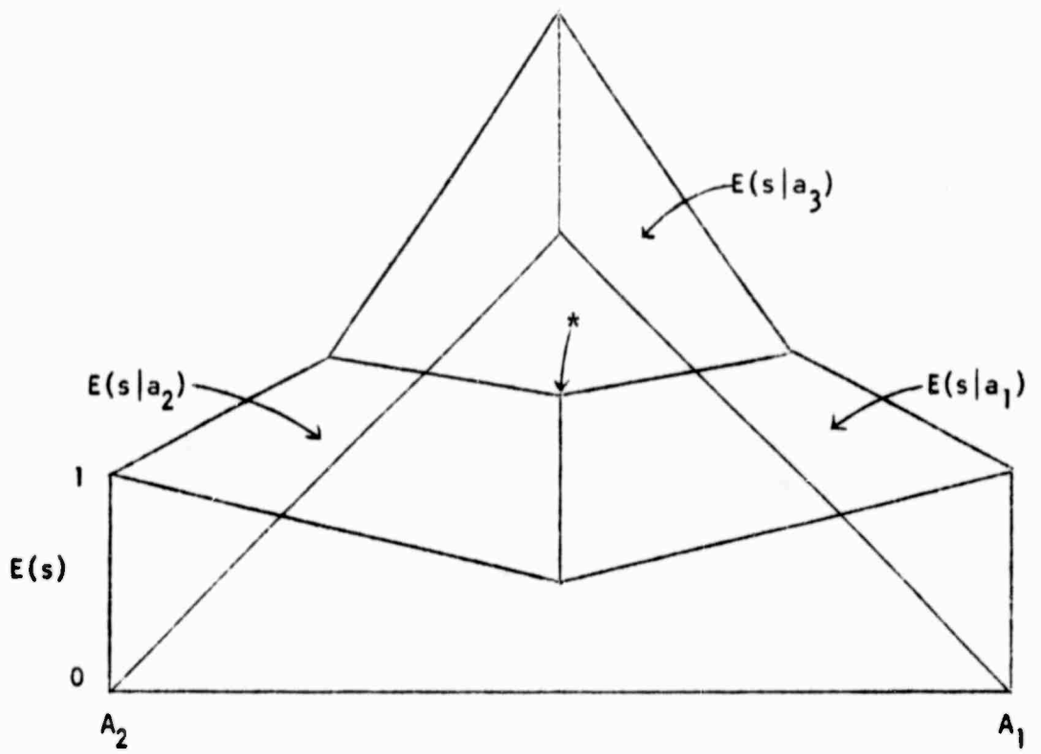


Figure 5. The maximum expected scores for a three-alternative question given the correction scoring system.

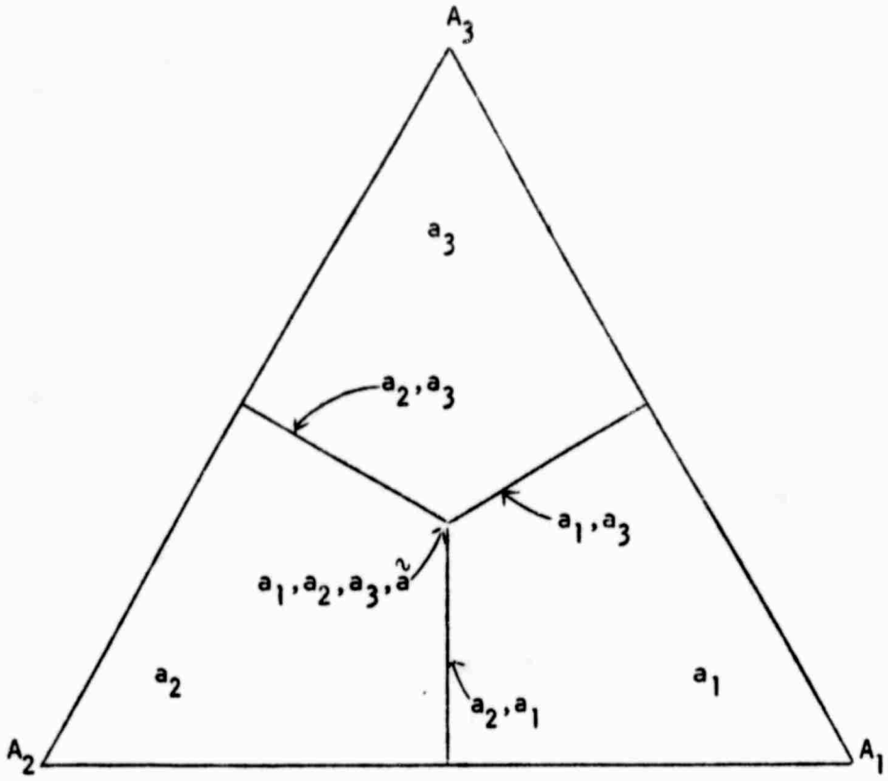


Figure 6. The decision regions for a three-alternative question using the correction scoring system.

yield the maximum expected score, i.e., all of the expected scores are equal.

The decision rule for the correction system is the same as that for the conventional choice system except at the point for which the individual has equal confidence in each of the possible answers. In this situation, the person can either choose from among the possible answers or skip the question. Whereas in the conventional choice system he should never skip a question. Figure 6 shows the decision regions both for the two-alternative and three-alternative question.

We can summarize the advice to a person taking a test under the correction system as: behave exactly as you would for the conventional choice system except when you are equally uncertain between all of the alternatives. In this case you have the additional option of skipping the question.

THE ADMISSIBLE CHOICE SYSTEM

The admissible choice scoring system comes from the same family of scoring systems as the conventional choice scoring system and the correction system. The table below shows the scoring system.

	A_1	A_2	A_3
a_1	1	0	0
a_2	0	1	0
a_3	0	0	1
$\tilde{a}$	q	q	q

Notice that the individual receives q points if he skips a question, where q must be greater than .5 in order to qualify as an admissible choice system. Figure 7 shows the maximum expected scores for q = .75. Figure 8 shows the resulting decision regions. The value of q determines a cutoff point Z, such that if a person has a degree of confidence greater than Z for an answer, he should choose this answer. If his largest degree of confidence is less than Z, he should skip the question. If he has a degree of confidence exactly equal to Z, he may either choose the answer for which he has this degree of confidence or skip the question.

Thus far we have discussed three members of one family of scoring systems: that family for which the student receives one point for a correct answer, zero points for an incorrect answer and q points for skipping the question. For the conventional scoring system, q = 0. For the correction system, q = 1/m. For the admissible choice system, q > .5.

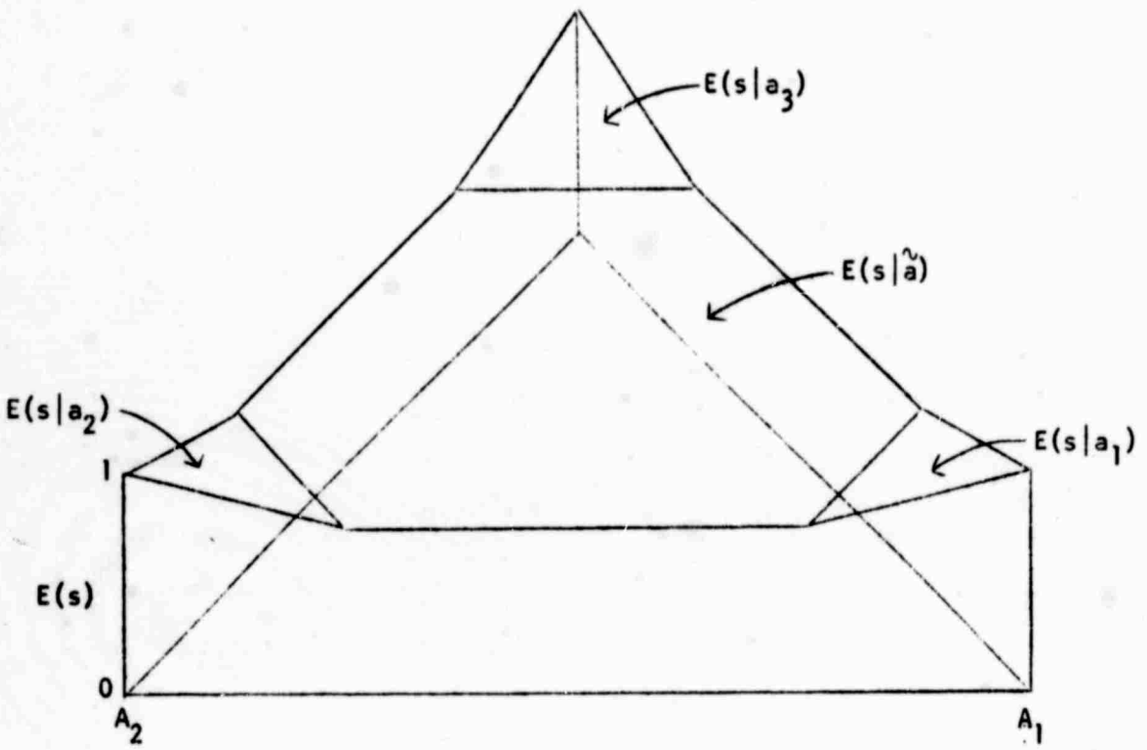


Figure 7. The maximum expected scores for a three-alternative question using the admissible choice scoring system for q=3/4.

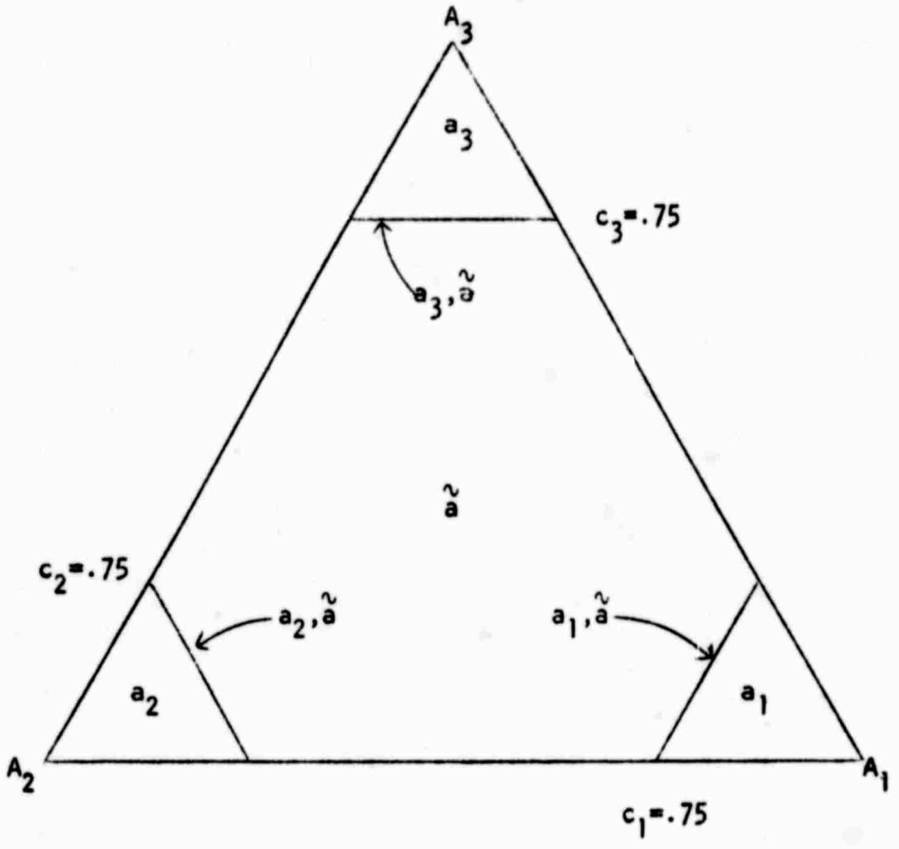


Figure 8. The decision regions for a three-alternative question using the admissible choice scoring system with q=3/4.

THE CONFIDENCE WEIGHTING SYSTEM

Ebel (1965, pp. 130-135) discusses what he calls confidence weighting of responses to true-false test items. In the confidence weighting system the person is to choose between five responses to a question. He can say that the first alternative is probably true, or possibly true, that the second alternative is probably true or possibly true, or he can skip the question.

Ebel discusses two versions of this scoring system. The table below shows the scoring system for one of these versions.

		A_1	A_2
a_1':	A_1 "Probably True"	1	0
a_1 :	A_1 "Possibly True"	3/4	1/2
a_2'':	A_2 "Probably True"	0	1
a_2 :	A_2 "Possibly True"	1/2	3/4
$\tilde{a}$:	Omit Item	5/8	5/8

Figure 9 shows the maximum expected scores yielded by this scoring system, while the decision regions are given in Figure 10. The second version of the scoring system is similar except that the cutoff point is equal to 3/4 rather than 2/3.

THE ADMISSIBLE CATEGORY SYSTEM

The confidence weighting system above is one of a family of scoring systems for which the individual can choose from more than one response for each alternative. The table below shows such a scoring system.

	A_1	A_2	A_3
a_1'	1	0	0
a_1	u	v	v
a_2'	0	1	0
a_2	v	u	v
a_3'	0	0	1
a_3	v	v	u
$\tilde{a}$	q	q	q

Figure 11 shows the maximum expected scores for u=7/8, v=4/8, and q=3/4 while

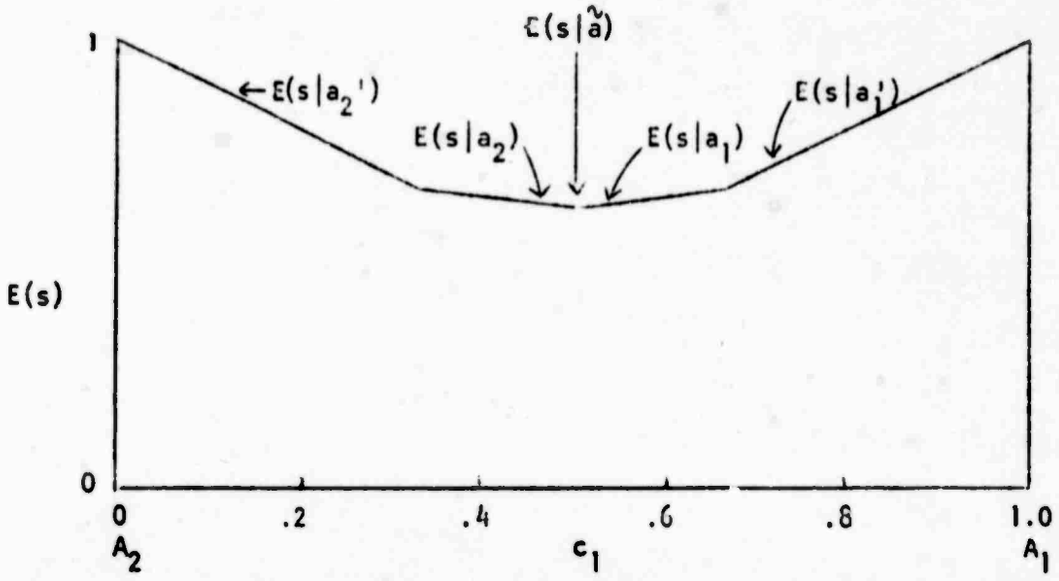


Figure 9. The maximum expected scores for the confidence weighting scoring system with q=5/8.

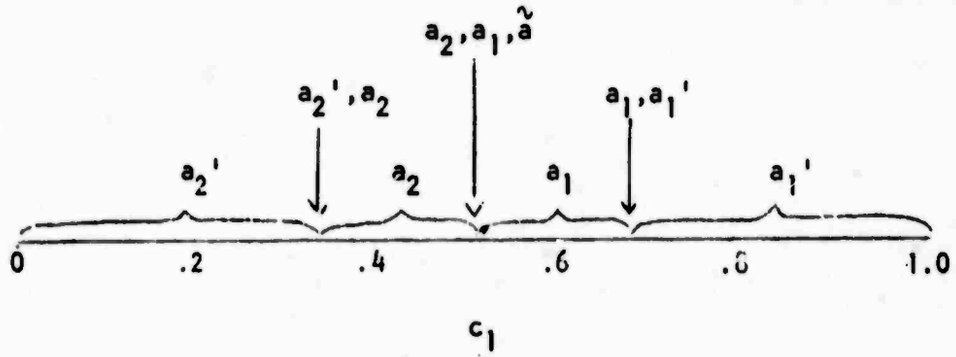


Figure 10. The decision regions for the confidence weighting scoring system with q=5/8.

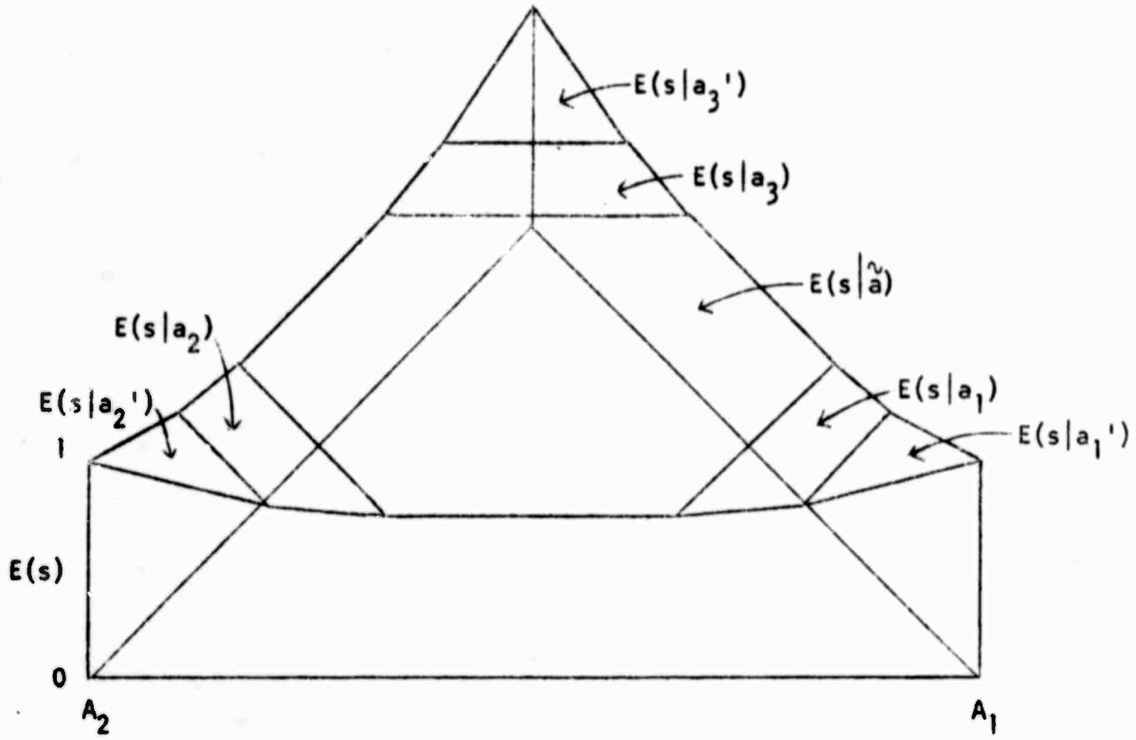


Figure 11. The maximum expected scores for the admissible category scoring system with two categories and u=7/8, v=4/8, q=3/4.

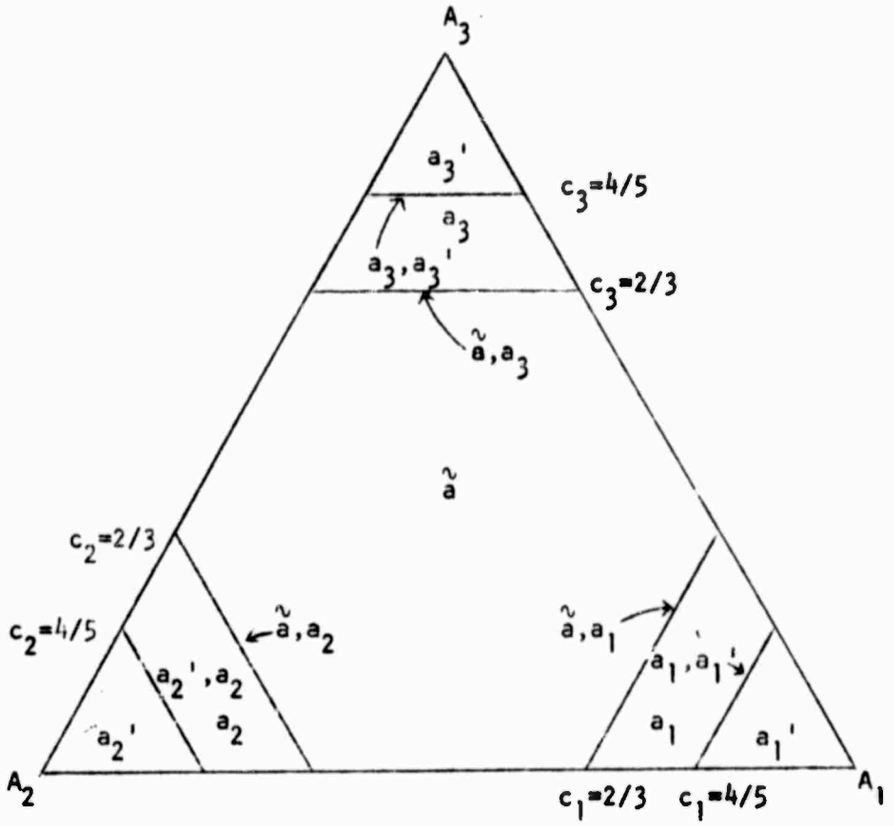


Figure 12. The decision regions for the admissible category scoring system with u=7/8, v=4/8, and q=3/4.

Figure 12 shows the decision regions.

The process suggested in the above table can be extended to include as many categories per alternative as desired. But if the resulting scoring system is to be admissible, it must always be optimal for the individual to skip the question when his maximum degree-of-confidence is less than or equal to .5.

COOMBS-MILHOLLAND-WOMER SYSTEM

The table below shows the scoring system proposed by Coombs, Milholland, and Womer (1955).

		A_1	A_2	A_3
a_{123}:	Choose A_1, A_2 and A_3	1/2	1/2	1/2
a_{12} :	Choose A_1 and A_2	1/4	1/4	1
a_{13} :	Choose A_1 and A_3	1/4	1	1/4
a_{23} :	Choose A_2 and A_3	1	1/4	1/4
a_1 :	Choose A_1	0	3/4	3/4
a_2 :	Choose A_2	3/4	0	3/4
a_3 :	Choose A_3	3/4	3/4	0
$\tilde{a}$:	Omit item	1/2	1/2	1/2

This system differs from those we have discussed in that there are situations in which the individual may respond with more than one of the possible answers since the individual deletes answers which he believes are incorrect.

Figure 13 shows the maximum expected scores, while Figure 14 shows the decision regions for a three-alternative question. The decision regions for a two-alternative question are exactly like those of the correction system except the individual has the option of responding with a_{12} at c=.5.

ADMISSIBLE CONFIDENCE

The final scoring we will consider is the admissible confidence procedure. This scoring system has the property that an individual maximizes his expected test score if and only if he responds to each possible answer of an item with his degree of confidence in the correctness of that answer. For further explanation of this system see Shuford, Albert, and Massengill, (1966) and Shuford and Massengill, (1965).

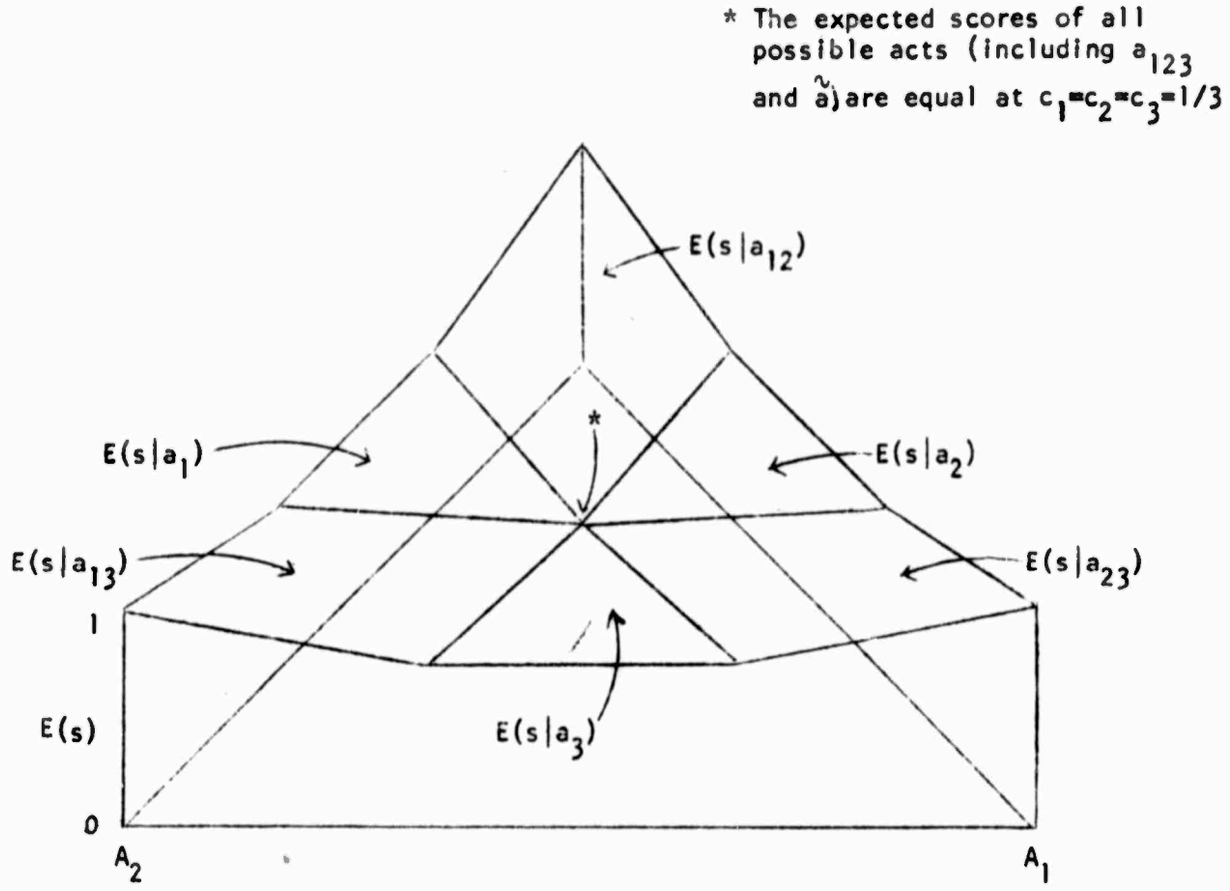
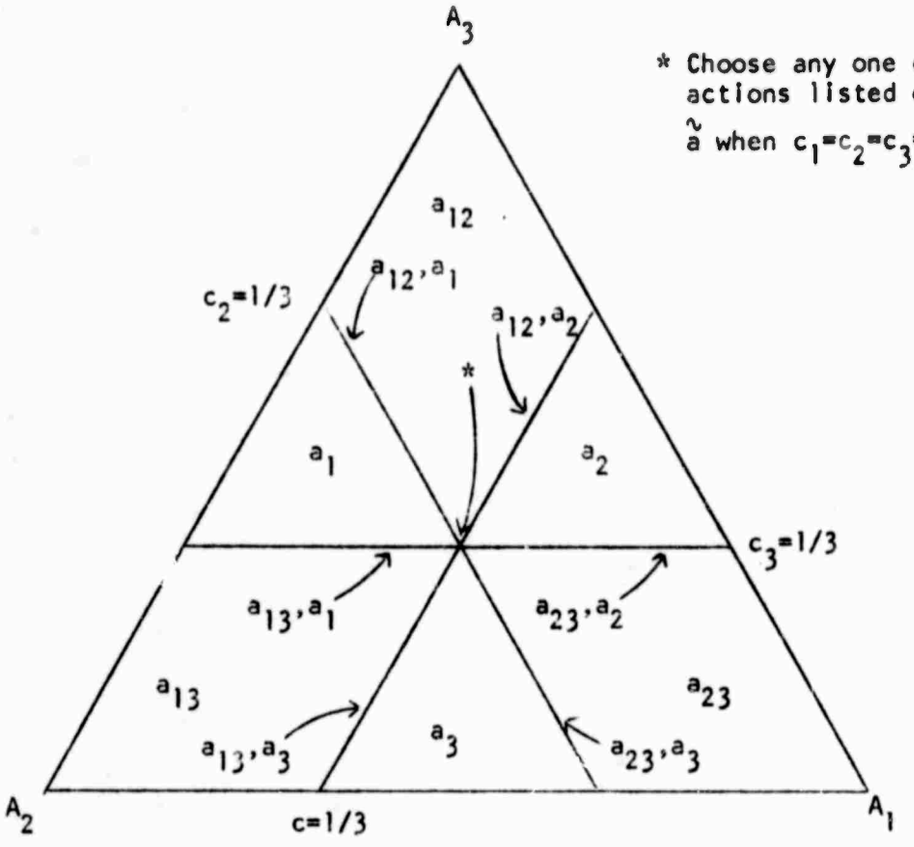


Figure 13. The maximum expected scores for the Coombs-Milholland-Womer scoring system.



* Choose any one of the actions listed or a_{123} or $\tilde{a}$ when $c_1=c_2=c_3=1/3$.

Figure 14. The decision regions for The Coombs-Milholland-Womer scoring system.

GUESSING

It is generally recognized that guessing presents a problem in the interpretation of objective test results. This problem has to do with the fact that a person can get the correct answer to a test question even when he doesn't "know" the answer. However, there seems to be some confusion as to exactly what is meant by the term. A review of references to guessing in books on testing and a look at definitions of guessing in various dictionaries seem to indicate at least three different ideas concerning the meaning of guessing.

1. Guessing is answering a question when not completely sure which answer is the correct answer. This seems to be the equivalent of the dictionary definition "to conclude from merely probable grounds". Ebel (1965, p. 230) talks about "rational" guessing as acting on the basis of insufficient evidence. It is not clear from these ideas where the cutoff point dividing sure and not sure is meant to be. If "sure" means "completely certain" then all of the points within Figure 15 except the end points: A_1, A_2, and A_3 would represent guessing. If it means "fairly certain", fewer of the points would represent guessing.

2. Guessing is answering a question when all of the possible answers are considered to be equally likely. This is equivalent to the dictionary definition of "making a conclusion without evidence". This is the type of guessing which Coombs, Milholland, and Womer (1955, p. 22) refer to in their treatment of the correction for guessing. It is also the type of guessing which Ebel (1965, p. 229) refers to as "blind" guessing (as opposed to rational guessing).

   The second definition specifies only one point for an item with m alternatives. From Figure 15 we see that the guessing point for a three alternative item is at O and for a two-alternative item is N.

3. Guessing is answering a question when the answer chosen is regarded as being equally likely with some, but not necessarily all of the possible answers. From Figure 15, we see that for a two-alternative question definitions 2 and 3 are equivalent but for a three-alternative question guessing includes the lines OL, OM and ON.

Now consider which of the scoring systems are able to distinguish guessing situations from other situations. Examination of the decision regions of the conventional scoring system indicates that it doesn't distinguish guessing under any of the definitions. If a person skips an item under the correction scoring procedure, then we can be sure that he has encountered a definition 2 situation. If the person does not skip the item, however, we cannot infer the absence of a definition 2 situation.

If q were set sufficiently high in an admissible choice scoring system,

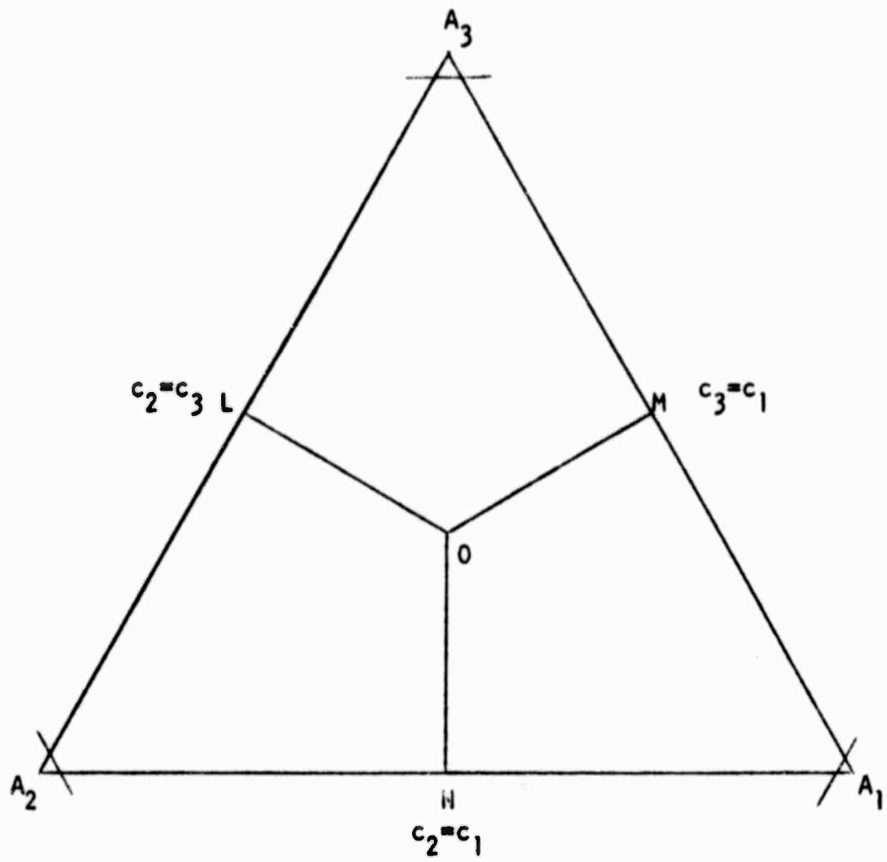


Figure 15. Figure for explicitly defining the three definitions of guessing.

then a person skipping an item indicates the existence of a definition 1 situation. If q is set sufficiently close to 1/2, then skipping an item represents the presence of definition 2 guessing but only for a two alternative item.

For the Ebel's confidence weighting system, if the cutoff point is set sufficiently high, definition 1 situations can be distinguished as in the case of admissible choice, while if a student skips an item a definition 2 situation is implied as in the case of the correction system.

For the Coombs-Milholland-Womer system, skipping an item implies a definition 2 guessing situation. As before, if the person does not skip the item, however, we cannot infer the absence of a definition 2 situation.

As we have seen above none of the discrete choice systems identify definition 1 and 2 guessing situations very well and are totally incapable of detecting the existence of definition 3 guessing situations. On the other hand, admissible confidence systems can distinguish all three types of guessing situations. This is so because when the response scale of an admissible confidence system is sufficiently fine-grained, any distribution of confidence can be effectively determined.

It seems appropriate here to attempt to correct the widely-held misconception that guessing cannot occur in a fill-in-the blank or constructed-response test. In responding to an item of this type, the student is either (a) unable to think of <u>any</u> answer, or (b) he is able to think of one or more potential answers to the question. If (a), he <u>must</u> skip the item. If (b), he is, in effect, faced with a multiple-choice item where the possible answers (assumed to be mutually exclusive) have been provided by the student's own efforts. For example, if the student is able to think of only one potential answer, his state of knowledge can be represented by a distribution for which c_1 is his degree of confidence that his potential answer is correct while c_2 is his degree of confidence that his potential answer is not correct, i.e., that some other, unthought of, answer is correct. If the student thinks of two potential answers, he is in the three-alternative situation represented in Figure 15 and it should now be clear that the different definitions of guessing can be applied.

CORRECTION FOR GUESSING

An individual who scores tests with the conventional choice scoring system may have heard that guessing by persons taking a conventional choice test causes ambiguity in the interpretation of test results. If the individual knows about the correction-for-guessing formula, R-W/(m-1), he may wonder if the application of this formula can solve the guessing problem associated with the use of the conventional scoring system. If he were to ask a mathematician, the mathematician would have to tell him that it is very unlikely that it can. We will see why in the following discussion.

Let us assume that the alternatives of the questions on a test are arranged in a random manner. For the conventional-choice scoring system, the person must, if he wants to maximize expected test score, choose an answer for each question. For those instances in which the person is in a guessing situation, we can determine the probability that he will choose the correct answer given the above assumption and given that the person behaves optimally.

There are two primary strategies which a person might use in order to choose an answer when he is in a guessing situation. 1. He might choose his answer according to its position in the set of eligible answers, eg., he might choose the answer in the first position. 2. He might pick an answer randomly from the set of eligible answers. It can be shown that regardless of whether the person chooses according to position or chooses randomly or mixes these two strategies, his probability of getting the correct answer is 1/n, where n is the number of eligible answers in the guessing situation under consideration.

Remember that for a three-alternative item, there are two types of guessing situations: one with three possible answers (n=3) and one with two (n=2). In general, for an item with m alternatives, there are m-1 types of guessing situations.

We can write the general equation for a person's test score on a conventional choice test as

$$(1) \qquad E(R) = K_1 + \frac{1}{2}K_2 + \frac{1}{3}K_3 + \ldots + \frac{1}{n}K_n + \ldots + \frac{1}{m}K_m + 0\left(N - \sum_{i=1}^{m} K_i\right)$$

E(R) is the person's expected test score, i.e., his expected number of correct answers on the average. K_1 is the number of situations in which the person's degree of confidence in the correct answer is larger than his degree of confidence in any one of the incorrect answers. If he behaves optimally, he will make one point for each such situation. K_n for n = 2,3, ,..., m is the number of guessing situations with n candidates for choice and 1/n is the probability that he will make one point in such a situation. N is the total number of questions on

on the test. And $N-\sum_{i=1}^{m} K_i$ is the number of situations in which the person is misinformed, i.e., has a larger degree of confidence in an incorrect answer than in the correct answer.

The correction-for-guessing formula is derived from the equation

$$R = K_1 + \frac{1}{m}(N-K_1), \tag{2}$$

where R is the person's observed test score, $N-K_1$ is the number of n=m guessing situations, and $N = R + W$, i.e., the number of right answers plus the number of wrong answers. If we solve for K_1,

$$K_1 = \frac{mR - N}{m - 1}$$

$$= \frac{mR - R - W}{m - 1}, \qquad N = R + W$$

$$= R - \frac{W}{m-1}.$$

This, of course, is the correction-for-guessing formula.

Equation (2) is actually an "average" score, i.e., the score the person could expect to receive on the average. Thus, the use of this formula is based on the assumption that the person's observed score is equal to his average score. But we have seen that Equation 1 is the general equation for a person's expected test score. Now let us see under what conditions the two equations are equivalent.

First, it is assumed in Equation 2 that the only guessing situations involved are definition 2 situations, i.e., all of the possible answers to a question are candidates for choice. For this assumption Equation 1 becomes

$$E(R) = K_1 + \frac{1}{m}K_m + 0[N-(K_1+K_m)]. \tag{1'}$$

Second, it is assumed in Equation 2 that $N-(K_1+K_m) = 0$, i.e., that there are no questions for which the person is misinformed. Thus if we set $K_m = N-K_1$ in Equation 1', we obtain Equation 2.

This means that if the correction-for-guessing formula is going to work for a given individual,

1. Any guessing situations involved must be definition 2 situations.
2. He cannot be misinformed on any items.
3. His observed test score must be equal to his expected test score.

Situations of this type are very rare. How rare will become evident as admissible confidence procedures are more widely used.

REFERENCES

Coombs, C. H., Milholland, J. E. & Womer, F. B. (1955) The assessment of partial knowledge. J. ed. psychol. Meas., 16, 13-37.

Cronbach, L. J. & Gleser, G. C. (1965) Psychological tests and personnel decisions. Urbana: University of Illinois Press.

Ebel, R. L. (1965) Measuring educational achievement. Englewood Cliffs: Prentice-Hall.

de Finetti, B. (1937) La prévision: ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincaré, 7. [Translated and reprinted as "Foresight: its logical laws, its subjective sources" in H. E. Kyburg, Jr. & H. E. Smokler (Eds.) Studies in subjective probabilities. New York: Wiley, 1964]

de Finetti, B. (1962) Does it make sense to speak of good probability appraisers? In I. J. Good (Gen. Ed.) The scientist speculates. New York: Basic Books, 357-364.

Nassengill, H. E. & Shuford, E. H., Jr. (1965) Direct vs. indirect assessment of simple knowledge structures. ESD-TR-65-542, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Mass.

van Naerssen, R. F. (1961) A scale for the measurement of subjective probability. Acta Psychologica, 159-166.

Raiffa, H. & Schlaifer, R. (1961) Applied statistical decision theory. Boston: Division of Research, Harvard Business School.

Ramsey, F. P. (1926) The foundation of mathematics and other logical essays. New York: Humanities Press.

Roby, T. B. (1965) Belief states: a preliminary empirical study. ESD-TDR-64-238, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Mass.

Savage, L. J. (1954) The foundations of statistics. New York: Wiley.

Shuford, E. H., Jr. (1964) Some Bayesian learning processes. In Shelly, M. W. & Bryan, G. L. (Eds.) Human judgments and optimality. New York: Wiley, pp. 127-152.

Shuford, E. H., Jr. (1965) Cybernetic testing. ESD-TR-65-467, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Mass.

Shuford, E. H., Jr., Albert, A., & Massengill, H. E. (1966) Admissible probability measurement procedures. Psychometrika, 31, 125-145.

Shuford, E. H., Jr. & Massengill, H. E. (1965) On communication and control in the educational process. ESD-TR-65-563, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Mass.

Toda, M. (1963) Measurement of subjective probability distributions. ESD-TDR-63-407, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Mass.

Toda, M. & Shuford, E. H., Jr. (1965) Logic of systems: introduction to the formal theory of structure. In General systems theory yearbook: 1965.

Toda, M. & Shuford, E. H., Jr. (1965) Utility, induced utilites, and small worlds. Behavioral Science, 10, 238-254.

Watanabe, S. M. (1960) Information theoretical aspects of inductive and deductive inference. IBM J. Research and Develop., 4.

Watanabe, S. M. (1966) A quantative study of certain formal aspects of knowing and guessing. New York: Wiley, (in press).

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
The Shuford-Massengill Corporation
Box 26
Lexington, Massachusetts 02173

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE
DECISION-THEORETIC PSYCHOMETRICS: A LOGICAL ANALYSIS OF GUESSING

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Scientific Interim

5. AUTHOR(S) (First name, middle initial, last name)
H. Edward Massengill
Emir H. Shuford, Jr.

6. REPORT DATE
November 1966

7a. TOTAL NO. OF PAGES
32

7b. NO. OF REFS
20

8a. CONTRACT OR GRANT NO.
AF 49(638)1744

b. PROJECT NO
920F-9719

c. 6154501R

d. 681313

9a. ORIGINATOR'S REPORT NUMBER(S)
SMC R-4

9b. OTHER REPORT NO(S): (Any other numbers that may be assigned this report)
AFOSR-68-2154

10. DISTRIBUTION STATEMENT
1. This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES
TECH, OTHER

12. SPONSORING MILITARY ACTIVITY
Air Force Office of Scientific Research
1400 Wilson Boulevard (SRLB)
Arlington, Virginia 22209

13. ABSTRACT
In Section A, A Logical Analysis of Guessing, appropriate test-taking strategies are derived for six major test-scoring procedures. Three commonly used definitions of guessing are interpreted as corresponding degree-of-confidence distributions. The ability of the testing procedures to separate these distributions from those representing higher degrees of knowledge is considered with the major result that only admissible probability measurement performs satisfactorily.

DD FORM 1 NOV 65 1473

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Test-taking strategies Guessing Counseling Decisions						