Confidence Testing

A New Tool for Measurement

THE SHUFORD-MASSENGILL CORPORATION
LEXINGTON
MASS. 02173

CONFIDENCE TESTING: A NEW TOOL FOR MEASUREMENT*

Emir H. Shuford, Jr.

Many of you must be aware that off and on during the last forty or fifty years people have been experimenting with confidence-marking in educational testing. More frequently, people have been concerned with the related "correction-for-guessing" problem. Andrew Ahlgren of Harvard's Project Physics has written an excellent review of the relevant literature (Ahlgren, 1967) so I will not repeat this history. Suffice it to say, however, that many people have felt the limitation of choice testing and have desired a mode of responding to a test question which better reveals the student's state of knowledge. In spite of this, these experimental studies and appeals have had a negligible impact on the theory and practice of educational testing.

So, what's new about confidence testing? There is a great deal that is new and most of it stems from the recent extension of logic and mathematics into the domain of decision making as expounded by Frank P. Ramsey (1926), Bruno de Finetti (1937), L. J. Savage (1954) and by now many others too numerous to mention. Although mathematical decision theory was applied first to the problems of business and the military, one can look back now and know that it was just a matter of time until decision theorists would begin to take a hard look at the problems of education and of educational testing in particular. This process has

---

*Paper read at the 11th Annual Conference of the Military Testing Association, 15-19 September, 1969, Statler Hilton Hotel, hosted by the U.S. Coast Guard Training Center, Governors Island, New York. This research was supported by the Advanced Research Projects Agency of the Department of Defense, and was monitored by the Air Force Office of Scientific Research under Contract No. F44620-69-C-0068.

now begun.

From a background in decision theory, it is quite natural to look at a student taking a test as a decision maker who, in most instances, is trying to act so as to maximize his expected test score. Each item in the test poses a decision problem to the student. He tries to recall and to evaluate all relevant information and he takes account of the scoring system in arriving at his answer. Because the scoring system is pretty much the same for every item in a test, the interest of the test-giver must be focused on the information that the student brings to bear in answering the questions. Now this student-specific information is reflected in the probabilities that the student holds that each answer may be counted as correct. Getting at these student probabilities is, therefore, the fundamental measurement problem in educational testing.

During the past decade, decision theorists scattered around the world have had this insight and have gone on to devise methods for better revealing these student probabilities and have advocated their use in educational testing. I doubt that I have a complete list, but these people should certainly be included: Bruno de Finetti (1965) in Italy, Shuford, Albert & Massengill (1966) in the U.S., Masanao Toda (1968) in Japan, and R. F. van Naerssen (1961) in the Netherlands.

Of the people working in this area, I feel that I have been most fortunate because since May of 1966 our research efforts have been supported by the Advanced Research Projects Agency (ARPA) of the Department of Defense. While undertaking new theoretical studies for ARPA we continued to devote internal corporate resources to the development of materials designed to facilitate the gathering of confidence data on a large scale. These materials first became available in the fall of 1967 at which time ARPA authorized the collection of confidence testing data at a number of military installations including, most notably, the Academic Instructor Course of the Air University. This organization has done a considerable amount of experimenting with and evaluation of confidence testing and some of the results will be reported by Capt. Gardner next on the program. Early this summer, revised and improved materials for confidence testing became available and are being used not only in the Academic Instructor Course but also in pilot programs at the Air Force's Chanute Technical Training

Center, the U.S. Army Signal Center and School at Fort Monmouth, the Naval Service Schools Commands at Great Lakes and at San Diego, and the Naval Air Basic Training Command at Pensacola.

I feel that I can best prepare the way for Capt. Gardner's paper by going into the logic of the problem faced by a student attempting to answer a test question and then I will go on to contrast confidence testing with other forms of testing.

Consider first the simple true-false type of test item. Choice testing allows the student to choose either one of two answers - "True" or "False." The situation confronting the student is, thus, a decision problem with two possible courses of action and four possible outcomes which are:

1. The student answers "True" and "True" turns out to be the right answer,
2. The student answers "True" and "True" turns out to be the wrong answer,
3. The student answers "False" and "False" turns out to be the right answer,
4. The student answers "False" and "False" turns out to be the wrong answer.

Within this decision-theoretic framework the student, in effect, associates a utility or worth to each of these four different outcomes. If the test of which this true-false item is a part is to be scored in the usual manner by giving one point for each right answer but no points for a wrong answer, the student should prefer giving a right answer to giving a wrong answer and he should be indifferent as to whether the right and wrong answers are arrived at through answering either "True" or "False." In other words, there are just two different utilities for the student, not four, and these utilities depend only upon whether or not his answer is right.

Now this student wants to pick the correct answer in order to receive a good grade. If the situation were a deterministic one where the student knows beyond the shadow of a doubt which answer is correct, there would be no problem. But this is generally not the case in educational testing. If it were, we might as well give the answer key to the students before administering the test. The information available to the student concerning a test question sometimes is of such quality as to leave no doubt as to which answer will be counted as correct.

At other times, however, the student may lack complete assurance as to which answer will be counted correct.

The quality and quantity of the student's information can be subsumed in his probability that "True" is the right answer and, the complement of this, his probability that "False" is the right answer. From decision theory we have the formulation that the "value" of responding "True" should be computed by multiplying the probability that "True" is correct times the utility, say one point, of giving the right answer and adding to this the product of the probability that "True" is incorrect times the utility, say zero point, of giving the wrong answer. This function is shown for all possible probability values in the right-hand portion of Fig. 1. The student's expected score for responding "False" is computed in a like manner and is shown in the left-hand portion of Fig. 1.

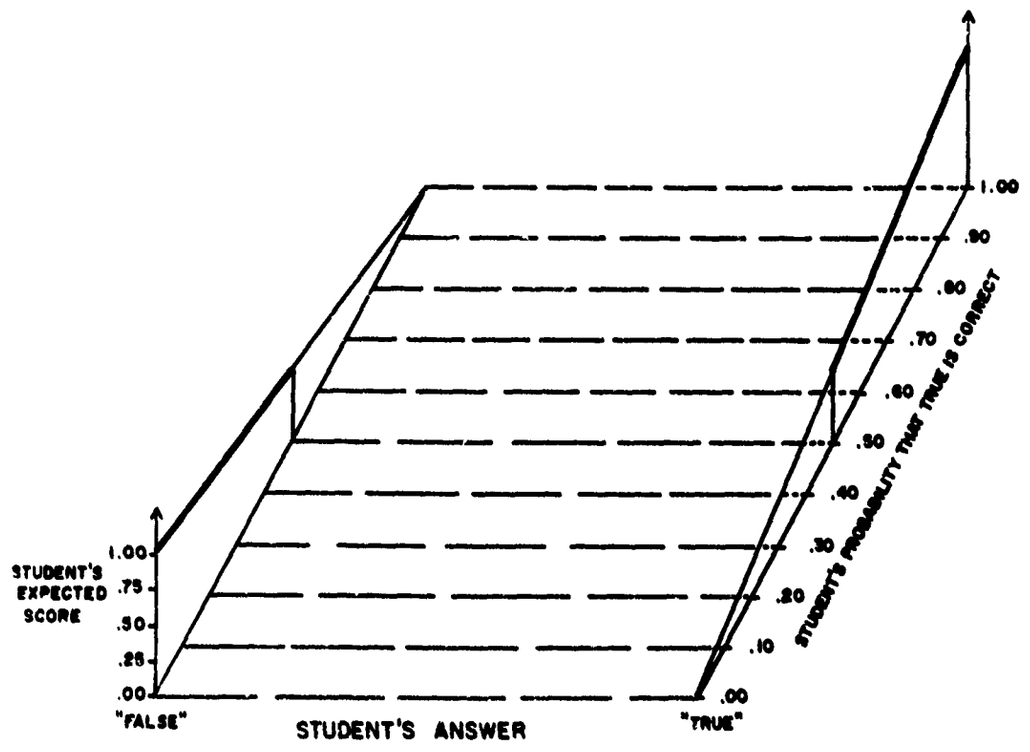

Figure 1

In a testing situation, the student can reflect, retrieve, and evaluate relevant information which sooner or later results in a probability between zero

and one. Where the probability falls in Fig. 1 depends upon the particular question asked the student and his state of knowledge with respect to that question. Wherever it falls, decision theory implies that the student can do best by choosing that answer which yields the highest expected item score. From Fig. 1 it is easy to see that if the student's probability that "True" is correct is less than .50, he would be well advised to answer "False" while if his probability is greater than .50, he would be well advised to answer "True." If his probability is exactly .50, he is in a guessing situation (Massengill & Shuford, 1966, 1967; Shuford & Massengill, 1966) and may as well choose either answer.

These expected score functions for choice testing are shown from a different point of view in Fig. 2. This might show even more clearly what the

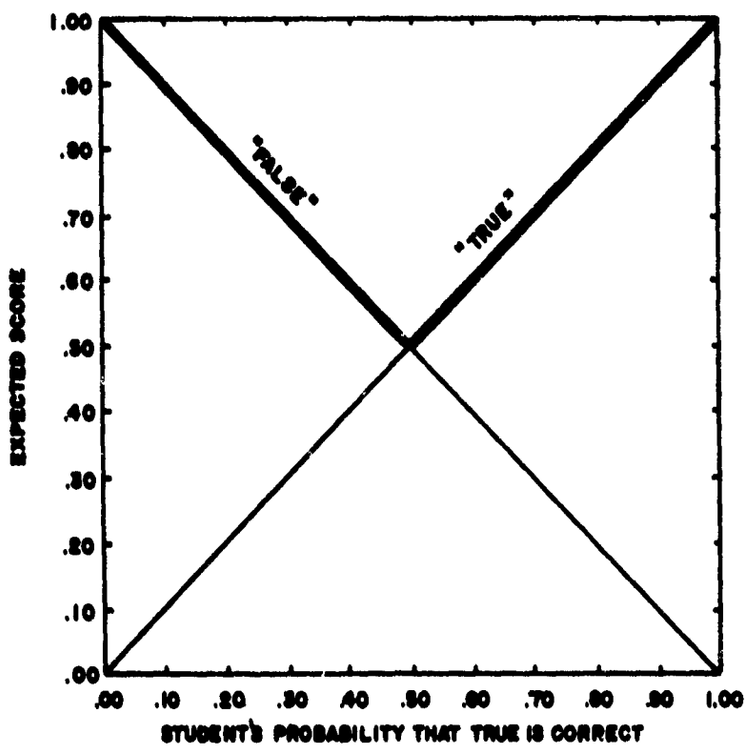

Figure 2

student's choice should be for the different probabilities. Fig. 2 also implies something about the student's motivation in preparing for a test. It shows that on an expected basis the student will be rewarded for studying to develop good

information that will justify probabilities as close as possible to zero or one. The more extreme the probability, the higher is the expected item score.

Fig. 3 shows the student's best choice or decision rule as a function of

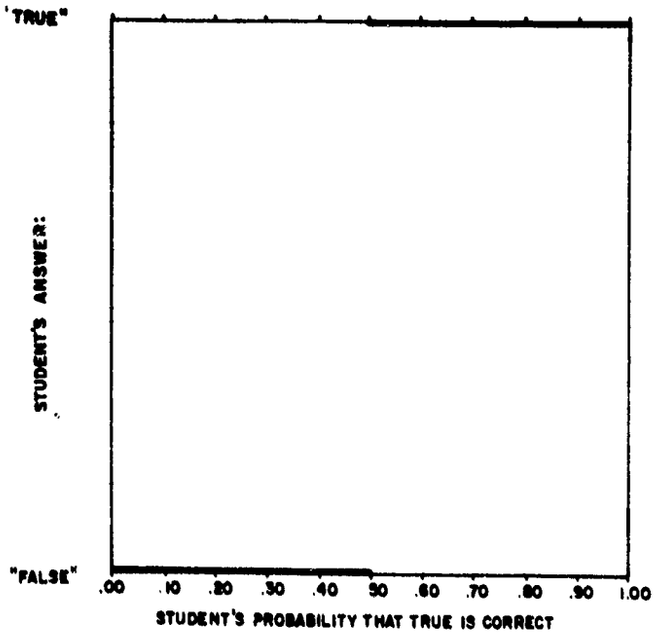

Figure 3

his probabilities for a true-false type of test item. It expresses what we have discovered above. The student should answer "False" if his probability that "True" is correct is less than .50 while he should answer "True" if his probability is greater than .50. Fig. 3 contains another implication though. It shows that the item score actually received by the student will be the same whether his information is of poor quality and leads to a probability of .51 or is of high quality and leads to a probability close to one. In this sense then, choice testing does not differentially reward higher levels of student achievement. Another implication of Fig. 3 is that choice testing does not tell very much about the student's state of knowledge. If the student answers "True," all we know is that he considers "True" more like to be the correct answer than is "False." If the student answers "False," all we can infer is that he considers "True" less likely to be the correct answer than is "False." Choice testing forces all sorts

of different states of knowledge to be lumped together. The method of responding itself sets up a barrier which filters out all sorts of information about student achievement. And by filtering information at the source, it makes it impossible to retrieve the information lost from the data.

Let us see how we can have the student respond so as to retain this information. In essence, we want the student to respond with his probabilities. We do not want to have the student incorporate his probabilities into a trivial decision problem which masks all sorts of differences in his states of knowledge. In other words, I am saying that we want to measure the student's confidence in the answers. Further, we must insist that the student's response be scored in such a way that it is in his best interest to reveal his probabilities (Shuford, Massengill & Organist, 1965). This means that one must use an admissible scoring system, that is, one that has the property that any student with any state of knowledge can maximize his expected score if and only if he responds with his actual probabilities (Shuford, Albert & Massengill, 1966).

Fig. 4 illustrates this situation for a true-false item. It is similar to Fig. 1 but with some important differences. Note first that the center of the graph is now filled in. This is so because the student is no longer constrained to respond with either of the two extreme values but can now also use the probability values lying between zero and one.

The curve at the top of the graph shows the item score that the student will receive given that "True" is the correct answer while the curve at the bottom of the graph shows the item score in the event that "False" is the correct answer. Notice that item score is a monotonic increasing function of the probability assigned to the correct answer. Here we have another significant contrast with choice testing - the more good information the student has, the higher will be his item score and this is true every time, not just on an average basis.

The curve in the middle of the graph shows the student's expected item score given that he has a .50 probability that "True" is correct and it shows this for all the probability values that can be assigned by the student. This curve has its maximum at .50 so to the extent that the student's response deviates from .50 he is throwing away score points. The two other curves showing expected item scores for student probabilities of .25 and .75 also reach a maximum

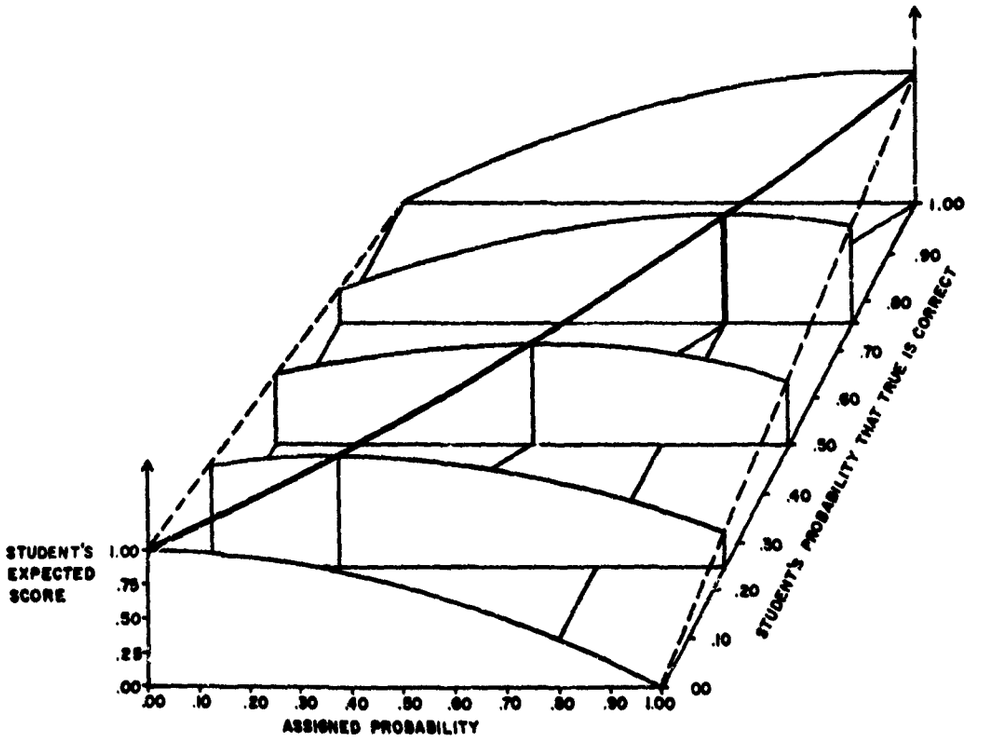

Figure 4

where the probability assigned by the student equals his actual probability.

Although only five curves are shown in Fig. 4, there exists a different curve for each different student probability. With this scoring system, each of these curves reaches its maximum at the point where the student's response corresponds exactly with his probability that "True" is correct. The maxima of this series of curves when connected would trace out the heavy diagonal line shown in Fig. 4. Thus we have that for any probability the student may have, his expected item score will be decreased whenever he responds with any value other than his probability that "True" is correct.

Fig. 5 shows a different view of the expected item score that the student can achieve as a function of his probability that "True" is correct. This graph corresponds to that shown in Fig. 2 for choice testing and as in that case implies that the student should be motivated to study in preparing for a test. Whereas choice testing tends to reward the student in equal increments as he develops

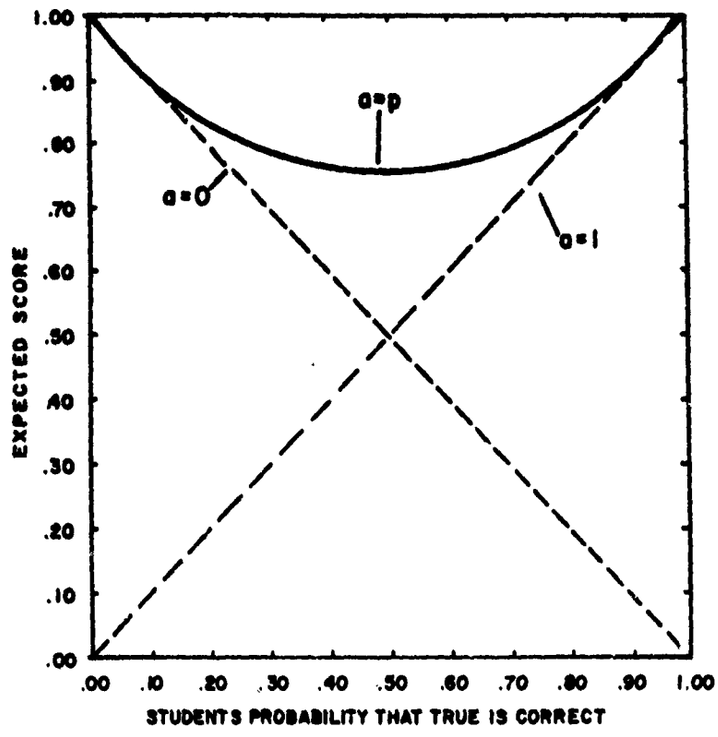

Figure 5

information to move his probability toward either zero or one, confidence testing tends to reward the student in greater and greater increments.

The student can achieve these maximum expected item scores only if he follows the optimal strategy illustrated in Fig. 6. This is a graphical way of saying that the probability assigned by the student should correspond exactly with his probability that "True" is correct. It also says something else. It says that by observing the response of such a student we can know his actual probability. There is a one-to-one relation between response and probability. In fact it is even better than that - it is the identity relation. There is no distortion, no filtering, no loss of information built into this method of testing.

Although I have just described only the true-false type of test item, the same logic applies rather directly to test items with more than two possible answers so confidence testing is well adapted to the multiple-choice format. It can also be used with completion or fill-in-the-blank items, essay items, and

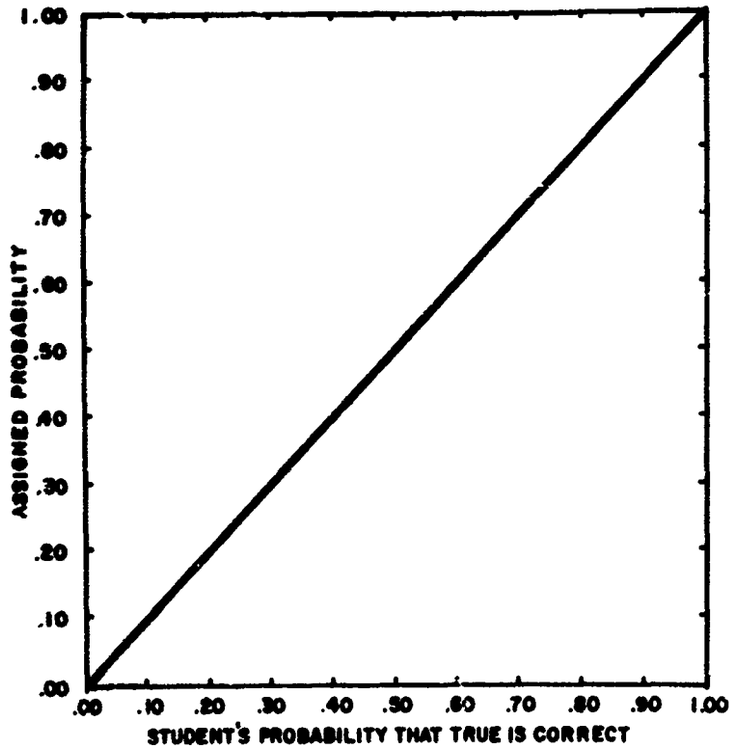

Figure 6

in conjunction with performance tests, but the scoring procedure has to be modified slightly. In an objective examination the possible answers are specified and are not under the control of the student, whereas in these other types of examinations the student acts so as to produce an answer for later evaluation. Here, we must motivate the student by giving him extra score for producing a satisfactory answer or performance (Shuford, Albert & Massengill, 1966).

In the case of a completion item, the student produces his answer and then reveals his probability that his answer is correct. He receives his confidence score according to whether or not his answer is correct and in addition one or more bonus points for producing a correct answer. In the case of a performance item, the student reveals his probability that he will successfully perform the specified task and then attempts the task. He receives his confidence score according to whether or not his performance is satisfactory along with one or more bonus points for a satisfactory performance.

The situation is illustrated by Fig. 7. Notice that it is quite similar to Fig. 4 except that now the maximum score ranges up to two points achieved by

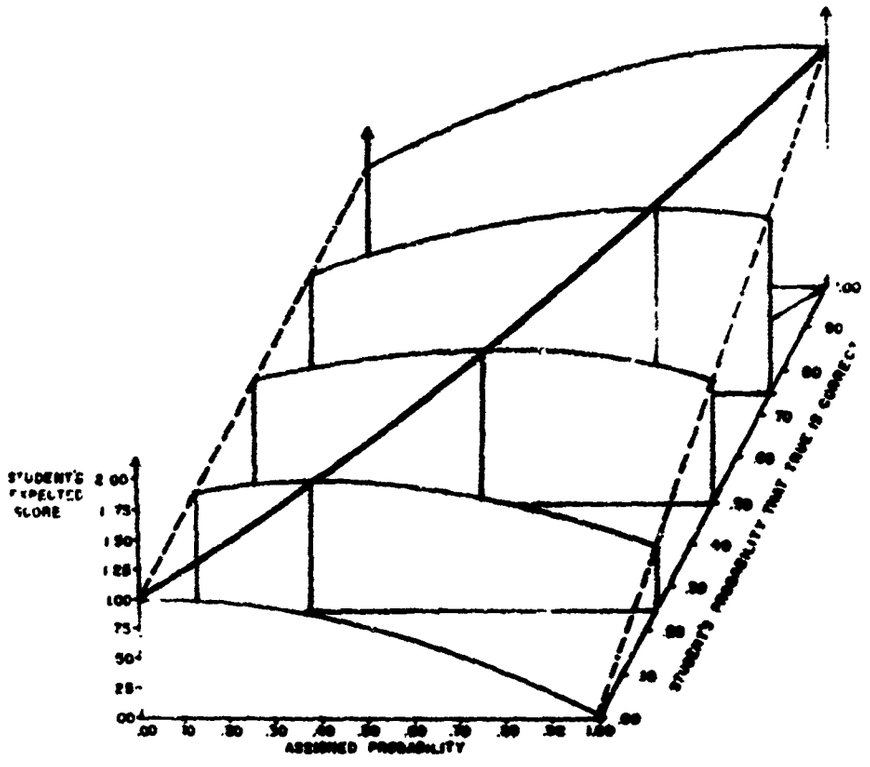

Figure 7

the student who is certain of his answer which is in fact correct. If the student is certain that his answer is incorrect and it is in fact wrong, he will receive an item score of one point. The student can maximize his expected score <u>if and only if</u> he tries his best to produce a correct answer or successful performance <u>and</u> honestly states his probability of doing so. This is also illustrated in Fig. 8 which shows the maximum expected item score that can be achieved by the student. Notice that the reward is such as to motivate the student toward high achievement.

In summary, mathematical decision theory has provided a logical framework which allows us not only to better understand choice testing but also to derive a scoring system and procedures which make possible the direct measurement of the confidence that a student holds in the occurrence of an event. When used

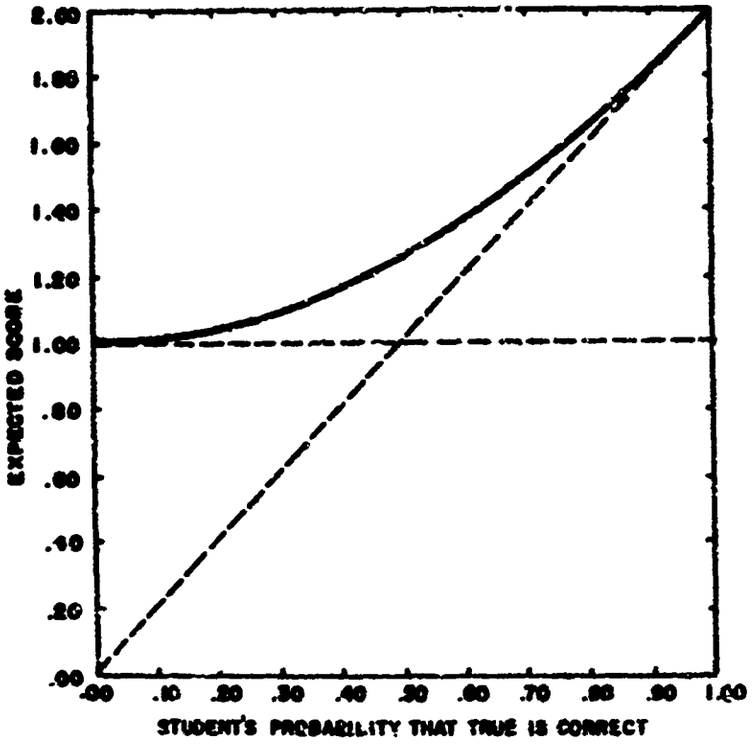

Figure 8

with written examinations and performance tests, confidence testing is capable of yielding a precise measure of student knowledge and skill.

Technical problems concerned with how to give a student a concrete representation of the notions and logic of confidence testing have now been solved

1. To insure the conservation of probability,
2. To display the possible scores for each allocation of confidence over the alternatives, and
3. To allow the student to express his confidence with any degree of precision he desires.

These advances have been incorporated into a methodology and a device called the SCoRule™ which can be used for the administration of almost any written or performance test.

That confidence exists and is meaningful is reflected by the following findings:

1. The more confidence a person holds in a possible event, the more likely the event will be confirmed (Massengill & Shuford, 1968, 1969; Shuford & Gibson, 1969; Shuford & Massengill, 1968),
2. Confidence is better able to predict the retention of information and future performance than is the simple correct-incorrect measure of traditional testing (Ahlgren, 1967; Shuford & Gibson, 1969),
3. Of two students correctly performing a task, the one with the higher confidence in his performance is more likely to correctly perform the same or a related task in the future than is the other student who has a lower degree of confidence (Shuford & Gibson, 1969),
4. Students differ considerably with respect to realism in assessing their knowledge and skill and this realism can be improved through training (Gardner, 1969; Shuford, 1969).

The scope of application of confidence testing is quite wide. It can be used to yield benefits in at least the following areas:

1. Selection and classification testing as:
   (a) an improved tool for item and test development (Shuford & Massengill, 1969a),
   (b) a method of test administration which increases test validity and reliability (Ahlgren, 1969; Armstrong & Mooney, 1969; Gardner, 1969; Shuford & Massengill, 1966a, 1967b, 1968),
   (c) a measure of the ability to realistically assess information.
2. Instruction and learning through:
   (a) better feedback from testing and assessment programs (Gardner, 1969; Massengill & Shuford, 1969; Shuford & Massengill, 1966b, 1967a, 1969b),
   (b) the development of curriculum to teach effective decision making through increased realism.
3. Assignment, retention, and promotion decisions by providing fairer, cheaper, and more objective measurement of job knowledge and performance (Shuford & Gibson, 1969).
4. Test and evaluation of new weapon systems by providing a measure of human performance which is not only inexpensive but is more sensitive

in detecting tasks subject to operational degradation (Shuford & Gibson, 1969).

5. Internal reporting procedures and the resultant organizational decisions by:
   (a) incorporating confidence as a new and concise dimension for reporting statements,
   (b) orienting personnel toward the realistic assessment of information.

References

Ahlgren, Andrew (1967), Confidence on Achievement Tests and the Prediction of Retention, Ph.D. Dissertation, Harvard Graduate School of Education, Cambridge, Mass.

Ahlgren, Andrew (1969), Reliability, Predictive Validity, and Personality Bias of Confidence-Weighted Scores. Remarks delivered in the symposium, "Confidence on Achievement Tests - Theory, Applications," at the 1969 meeting of the AERA and NCME.

Armstrong, Robert J. & Robert F. Mooney (1969), Confidence Testing: Is It Reliable? Paper presented at National Council on Measurement in Education, 1969 Annual Meeting, Los Angeles, California.

de Finetti, Bruno (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 Henry E. Kyburg, Jr., & Howard E. Smokler, eds., Studies in Subjective Probabilities, Wiley, New York, 1964].

de Finetti, Bruno (1965), Methods for discriminating levels of partial knowledge concerning a test item, The British Journal of Mathematical and Statistical Psychology, 18, 87-123.

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.

Massengill, H. Edward & Emir H. Shuford, Jr. (1966), Decision-Theoretic Psychometrics: a Logical Analysis of Guessing, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Massengill, H. Edward & Emir H. Shuford, Jr. (1967), What Pupils and Teachers Should Know about Guessing, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Massengill, H. Edward & Emir H. Shuford, Jr. (1968), A Report on the Effect of Degree of Confidence in Student Testing, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Massengill, H. Edward & Emir H. Shuford, Jr. (1969), Confidence Testing at the Academic Instructor Course of the Air University: August and September, 1968, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Ramsey, Frank P. (1926), The Foundations of Mathematics and Other Logical Essays, The Humanities Press, New York.

Savage, L. J. (1954), The Foundations of Statistics, Wiley, New York.

Shuford, Emir H., Jr. (1969), Systems of Confidence Weighting: Theory and Practice, Lexington, Massachusetts: The Shuford-Massengill Corporation.

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

Shuford, Emir H., Jr. & H. Edward Massengill (1966a), Decision-Theoretic Psychometrics: the Effect of Guessing on the Quality of Personnel and Counseling Decisions, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, Emir H., Jr. & H. Edward Massengill (1966b), Decision-Theoretic Psychometrics: the Worth of Individualizing Instruction, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, Emir H., Jr. & H. Edward Massengill (1967a), The Relative Effectiveness of Five Instructional Strategies, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, Emir H., Jr. & H. Edward Massengill (1967b), How to Shorten a Test and Increase Its Reliability and Validity, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, Emir H., Jr. & H. Edward Massengill (1968), Airman Qualifying Examination - 66 Administered as a Confidence Test, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, Emir H., Jr. & H. Edward Massengill (1969a), Confidence Testing at the Officer Training School, Lackland Air Force Base, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, Emir H., Jr. & H. Edward Massengill (1969b), Item Analysis Based on Confidence Responses, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, Emir H., Jr., H. Edward Massengill & Walter E. Organist (1965), Communication and Control in the Educational Process, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, Emir H., Jr. & Duncan L. Gibson (1969), A New Method for Predicting Performance, Lexington, Massachusetts: The Shuford-Massengill Corporation.

Toda, Masanao (1968), Algebraic models in dynamic decision theory, in C. A. J. Vlek, ed., Algebraic Models in Psychology. Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof," NATO, The Hague, August 5-17, 1968.

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

UNCLASSIFIED

Security Classification

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)	2a. REPORT SECURITY CLASSIFICATION
The Shuford-Massengill Corporation One Wallis Court Lexington, Massachusetts 02173	UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

CONFIDENCE TESTING: A NEW TOOL FOR MEASUREMENT

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

Scientific Interim

5. AUTHOR(S) (First name, middle initial, last name)

Emir H. Shuford, Jr.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 1969	17	25

8a. CONTRACT OR GRANT NO. F44620-69-C-0068 (ARPA)	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO. 9719	SMC R-19
c. 61101D	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AFOSR 69-2348TR
d.	

10. DISTRIBUTION STATEMENT

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

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

13. ABSTRACT Mathematical decision theory has provided a logical framework whcih allows us not only to better understand choice testing but also to derive a scoring system and procedures which make possible the direct measurement of the confidence that a student holds in the occurrence of an event. When used with written examinations and performance tests, confidence testing is capable of yielding a precise measurement of student knowledge and skill. Graphical procedures are used to contrast choice and confidence testing.

Four types of experimental findings which confirm that confidence exists and is meaningful are described.

Five major areas of application of confidence testing and the relevant benefits are outlined.

DD FORM 1 NOV 68 1473

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
decision theory educational measurement probability selection and classification testing training human performance						