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

THE SHUFORD-MASSENGILL CORPORATION
LEXINGTON
MASS. 02173

13

DECISION-THEORETIC PSYCHOMETRICS

Final Technical Report

Emir H. Shuford, Jr.

Sponsored By

Advanced Research Projects Agency

ARPA Order No. 833

ARPA Order Number 833

Program Code Number 9D20

Name of Contractor
The Shuford-Massengill Corp.
4 Lincoln Terrace
Lexington, Mass. 02173

Effective Date of Contract–68 Nov. 01
Contract Expiration Date–70 June 30
Amount of Contract–$159,456

Contract Number F44620-69-C-0068

Principal Investigator
Dr. Emir H. Shuford, Jr.
Area Code *617* 861-8700

Project Scientist
Dr. Glen Finch
Area Code *202* OX4-5045

Short Title of Work
Decision-Theoretic Psychometrics

This final report reviews the major items of work performed under this contract during the 20 month period 1 November 1968 through 30 June 1970.

Work Statement A. *Prepare item analyses of military classification tests in order to provide methods for shortening the tests and to improve the reliability.*

Technical Report No. AFOSR-69-0408-TR "ITEM ANALYSIS BASED ON CONFIDENCE RESPONSES" by E.H. Shuford and H.E. Massengill.

ABSTRACT

In examining the behavior of a group of subjects responding to a test question, a distribution of student likelihoods is obtained for each answer to the item. While the two-point distribution from choice testing can be completely characterized by one number based on the proportion of subjects choosing the answer, the results from confidence testing can be fully characterized only by specifying the complete distribution of student likelihoods. Such distributions are analyzed for the responses of 98 students to 16 four-alternative items. The distributions are found to be quite complex in shape and clearly cannot be characterized by using only one parameter.

Techniques are derived for computing both a difficulty index and a validity index from confidence data. Although yielding essentially the same information as that available from choice testing, these confidence-based indexes have much smaller sampling variabilities as indicated by relative efficiencies on the order of 1 1/2 times that of the choice testing index. Two graphical procedures are devised and applied to the 16 items to indicate the ability of these items to discriminate between better and poorer students. One procedure compares confidence distributions of the upper and lower subjects for each of the four answers, while the other compares the frequency of various states of knowledge in the upper and lower groups.

Item analysis based on confidence test data yields all the same type of information available from choice testing but it does it with greater efficiency. The use of confidence testing to obtain item analysis data also provides qualitatively different information which, in principle, cannot be obtained from choice testing.

Work Statement B. *Relate confidence measures to other test performance measures.*

Technical Report No. AFOSR-69-1329-TR "A NEW METHOD FOR PREDICTING PERFORMANCE" by E.H. Shuford and D.L. Gibson.

ABSTRACT

A small-scale pilot study had three subjects give likelihood estimates for successfully performing target shooting tasks. Data analysis indicates that the likelihood estimate is a better predictor of future performance than is test performance itself. Further, the likelihood estimate is a better predictor of current test performance than is a success-failure prediction.

Although these results are by no means definitive and do not have the realism and relevance of field studies with military personnel, they do suggest the possibility of considerable gain from the introduction of APMP into performance testing programs. If similar results are obtained in the field, it would be possible to greatly increase the predictive power, reliability, and validity of military performance testing programs, and, by querying students about all the job relevant tasks but actually administering only a small random sample of tasks, to vastly increase the scope of military performance testing. It appears well worthwhile to conduct further studies aimed at evaluating this new application of Admissible Probability Measurement.

Work Statement C. *Determine the optimal use that can be made of admissible probability measurement to improve the efficiency and effectiveness of military personnel selection, classification, training, and education.*

Proceedings of the 11th Annual Conference Military Testing Association, Sept. 1969, pp. 234-250. "Confidence Testing: A New Tool for Measurement" by E.H. Shuford, Jr. Also appears as AFOSR-69-2348TR.

The following beneficial uses of admissible probability measurement (APM) are described in this report:

1. *Selection and classification testing as:*
   - a. *an improved procedure for item and test development.*
   - b. *a method of test administration which increases test validity and reliability.*
   - c. *a measure of the ability to realistically assess information.*
2. *Instruction and learning through:*
   - a. *better feedback from testing and assessment programs.*
   - 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.*
4. *Test and evaluation of new weapons systems by providing a measure of human performance which is not only inexpensive but is more sensitive in detecting tasks subject to operational degradation.*
5. *Internal reporting procedures and the resultant organizational decisions by:*
   - a. *incorporating confidence as a new concise dimension for reporting.*
   - b. *orienting personnel toward the realistic assessment of information.*

As of the date of this final report, APM is in operational use at several schools providing better feedback from the testing and assessment programs, see Item 2(a) above. APM is also in operational use improving internal reporting procedures and organizational decisions, see Item 5 above.

As for the other uses of APM, they appear to require varying amounts of further development. In selection and classification testing, item and test development using APM data is fairly straightforward and certain benefits would result whether or not APM were used in the administration of the resulting selection and classification test. The same procedures would be highly effective also in evaluating the relative effectiveness of existing tests in order to decide between competitive testing programs.

A large number of studies have demonstrated that test reliability and validity can be improved by changing the method of administration from the forced-choice method over to some method of responding with weights and more studies continue to appear. For example, Armstrong and Mooney (1969) report appreciable gains in test-retest reliability while Hambleton, Roberts and Traub (1970) report an eight-fold increase in test validity. Shuford and Gibson (1969) developed a fundamental method for measuring the predictive validity of *any* testing method and used it to evaluate the effectiveness of APM when used with performance tests. They found that APM yields a startling reduction in error variance when used to predict future performance confirming a finding by Ahlgren (1967) that the superiority of confidence scores in predicting retention and future grades increased with the time between prediction and confirmation. Ahlgren's data also indicated that confidence scores were less biased by personality factors than were the choice scores of conventional testing.

In spite of the large amount of evidence that APM can yield significant benefits when used for the administration of selection and classification tests, to our knowledge it is correct to say not only that no tests are actually being administered with APM but also that no research is being conducted leading to the application of APM in this area.

Two negative factors probably help to account for this state of affairs. First, the economics of commercial test publishing are such that a company is not rewarded for investing heavily in a new test which would undermine its market for an existing test. Second, the data from APM call for new kinds of test statistics and new ways of analyzing test results. The old ways of test analysis can no longer be routinely applied to the data.

APM can yield a measure of a person's ability to realistically assess information. Except for extreme cases, this measure is independent of the difficulty of the test(s) used for the analysis. There are strong logical reasons for suspecting that this ability may prove quite useful for predicting success in training and on-the-job. Further research is required to determine how to fit this ability measure into a prediction equation and to assess its contribution to the classification process.

APM comes directly from decision theory. It can be viewed as a way of helping a person use his information to make probabilistic predictions and, as such, it becomes the natural foundation upon which to base a course on the logic of effective decision making. A rather straightforward curriculum development effort could very well produce an effective course of instruction in this increasingly important area. A different, but related, curriculum development effort could orient the decision making toward strategies for effective study and learning.

Job knowledge tests are now used for promotion within the enlisted grades in all of the military services. At present, these tests are typically written examinations. Substitution of APM for the choice method of administration in current use would result in different test scores yielding a changed rank ordering of individuals. The changes in rank order are caused by two major factors. First, guessing is reduced if not eliminated by APM, thus improving the fairness and validity of the promotion process. Second, APM values job knowledge in a fundamentally different way. Choice testing assumes that each item of knowledge is independent of all other knowledge. APM assumes that in practice different items of knowledge may be combined to allow the correct performance of a job. Whenever this latter description is more appropriate, APM will yield scores which are more objectively related to job performance. Finally, APM has been used to reduce the costs and increase the predictive validity of performance testing to the extent that it may have become economically feasible to use performance testing for promotion decisions. Further research is needed to estimate both the benefits and costs of using APM in this area.

The test and evaluation of new man-machine systems has relied heavily on expert judgment. The need for more "objective" measures has led to greater use of instrumentation and analysis during the test and evaluation phase. In the same way that organizations are using APM to make executive judgments and forecasts more objective and quantitative, APM could be used to quantify the judgment and forecasts of test pilots and other experts involved in man-machine system development. This appears to be a rather straightforward and beneficial application of APM.

Work Statement D. *Experimentally test military personnel with confidence testing techniques in actual military personnel and training operations to determine the practical utility of these procedures.*

Proceedings of the 11th Annual Conference Military Testing Association, Sept. 1969, p. 252-306. "The Use of Confidence Testing in the Academic Instructor Course " by Major W.C. Gardner, Jr. Also appears as AFOSR-70-0143-TR.

Air Force ROTC Education Bulletin, Oct. 1969, p. 5-7. "Confidence Testing" by Major W.C. Gardner, Jr.

USAF Instructors Journal, Winter 1969-70, p. 4-10. "Confidence Testing" by Major W.C. Gardner, Jr.

Air Force ROTC Education Bulletin, April 1970, p. 4-6. "How to Reward Achievement" by E.H. Shuford, Jr.

At the onset of this contract it was known that at least some populations of military personnel given adequate instruction in the techniques of APM could use it to take tests and that the data so obtained yielded information over and above that available from choice testing. Thanks to the cooperation of many individuals in the Air Force, Army, Navy, by the end of this contract APM had been used experimentally in fourteen different military training programs. In all cases, APM was used with test questions in current use at the school. No special test questions were required. The subjects in the experimental groups pretty well covered the range of military and civilian personnel in the Department of Defense. There were new recruits undergoing basic training and there were senior non-comissioned officers in advanced technical training. There were cadets in training to become officers and there were senior colonels taking a course in academic instruction. And so on.

In all experiments, the subjects proved to be able to learn the techniques of APM and provided test data which yielded additional information when compared with choice testing. As these experimental tryouts were designed to do, they pointed up many areas that needed improvement and, of more importance, discussions with staff and students at some of the schools revealed the surprising finding that a major reorientation in thinking about APM would greatly increase the practical utility of these procedures. These findings are summarized in the following recommendations.

Recommendation I.

The notion and ideas of confidence testing should not be used in conjunction with APM because they interfere with the proper working of this new procedure.

Throughout almost all of our research with APM we have identified it as *one* type of confidence testing, a generic term in use for many years in the area of educational measurement. As a consequence, we introduced APM to staff and students as a way of measuring *confidence.* We did not foresee the negative influence this would have on the operation of APM and how this would offset much of the practical utility of the method.

Many of the instructors who favored APM valued it for the type of logical thinking and judgmental processes it encourages in their students. Many of the students who favored APM also valued it for this reason, but many went on to point out that thinking in terms of feelings of confidence blocked further thought about the question being considered and that it was no longer enough just to recall an isolated fact to answer a question. They felt (correctly according to the mathematical theorems of APM) that they could score best on the test by coming up with information, reasoning about it to develop arguments for or against each answer, and by accurately assessing the validity of this reasoning. These students reported that they found it difficult to do this type of reasoning, but they universally considered it an important skill to be mastered.

Toward the very end of the contract period, we abandoned all reference to confidence in introducing APM to instructors and students. The results are rather dramatic. By focusing attention on wisely placing score on the possible answers and by not having to deal with the confidence, people can learn the technique of APM in less than one-fourth the time previously required by the "confidence approach". This savings in instructional time is somewhat offset by the students devoting more thought (and time) to answering questions during a test, but this is just what many instructors want–a setting which encourages logical thinking and judgment in the students. They want it for one or both of two reasons–in the expectation that practice may improve the reasoning ability of the students and that the subject matter will become more meaningful and, thus, better remembered by the students.

This reorientation toward answering questions shows up in the data too. The students are more discriminating and the information gained from APM is even greater than before. This indicates, of course, that the gains found in earlier studies tended to underestimate the potential of APM.

Recommendation II.

The truncated logarithmic scoring system, because of its unique properties, should be adopted as the standard measure of achievement and performance.

This new approach to APM requires a different interpretation of the underlying mathematics, but one which tends to even stronger results than before. In our original article (Shuford, Albert, & Massengill; 1966), we identified the quantity, p with confidence and used the theorems to prove that the subject could maximize his expected test score *if and only if* he honestly revealed his confidence in the answers. Now, by identifying the quantity, p with the probabilistic prediction justified by the information and reasoning available to the student as defined and evaluated by the personal realism graph, the theorems now prove that APM automatically rewards each student according to the quality of his knowledge and his skill at applying this knowledge. This means the APM can be an almost unbelievably powerful system for shaping behavior to deal effectively with reality.

Of all the admissible scoring systems, the truncated logarithmic has some unique properties which are quite compelling in many applications. For example, it is the only one which yields a total test score which measures (in the information-theoretic sense) the amount of useful information demonstrated by the student. It is also the measure to use when combining information from different sources and for evaluating these sources.

Finally, the total score yielded by the truncated logarithmic scoring system values knowledge in a way fundamentally different from choice testing. The counting of right answers to obtain the test score which is almost universally used in choice testing treats the structure of knowledge as being composed of nothing but independent and unrelated segments of information. From this point of view, education and training is like pouring water in a bucket. The more "water in the bucket", the better. But the implications go beyond this when we ask that the test score reflect the student's ability to perform outside of the testing situation. In this event, the logic of choice testing says that a student is able to perform in *exact proportion* to the amount of water in the bucket.

This is not always true. Consider just one task to be performed by some students, for example, the task of driving a car and suppose the criterion is passing a driving test to obtain a license. These are some of the things the student has to know if he is going to succeed at the task. He has to know what the steering wheel, brake and accelerator are used for in controlling the car. These segments of information are *necessary* conditions for driving a car. But they are not *sufficient* conditons. Even more to the point, misinformation can be a *sufficient* condition which guarantees that the student cannot drive, for example, suppose that the student had the *misinformation* that the accelerator was for stopping the car. Letting a "1" represent complete and correct information about a segment and a "0" represent complete misinformation, we can consider the meaning of two scoring rules–one yielding a total score by summing the "0's" and "1's", the other, multiplying the "0's" and "1's". The summation rule characterizes choice testing but only the multiplication rule reflects performance.

The total score obtained from the truncated logarithmic scoring system is most like the multiplication rule but with some refinements. Let p_i be the probability underlying the correct answer to the *i*th question on a test. Then the multiplication rule would score the student with the quantity, $p_1 \times p_2 \times p_3 \ldots . \times p_n$. This is the same thing as the antilog of the *sum* of the log probabilities. This is basically what the total test score is when using the truncated logarithmic scoring system expect for the truncation at $p \leqslant .01$. One interesting interpretation of the truncation is that *not all* information segments in the test are necessarily related to the performance of *any* relevant task. It is a compromise and in many cases a closer approximation to reality because it does not assume that one instance of complete misinformation, $p = 0$, is sufficient to prevent the student from properly performing all tasks requiring any of the information segments in the test.

Recommendation III.

Operational use of APM should be implemented at school only after thorough indoctrination and training of instructors.

It is almost a truism that the success of any educational innovation depends upon the training and attitudes of the teachers. This may be somewhat less true of those educational technologies that, in effect, take instruction out of the hands of the teacher, e.g., programmed instruction and some types of computer-assisted instruction. The philosophy behind the experimental field tests in this project was to determine the minimal conditions or threshold at which APM could have practical utility in military training operations. Putting this philosophy into practice meant, of course, that in searching for the threshold, the conditions had to be reduced enough to produce a number of gross failures of APM. The Up-and-Down method of threshold determination proved effective for this purpose. This philosophy contrasts with the more usual strategy of completely rewriting course materials and using exceptional or extremely well-trained instructors. The latter strategy makes the research look good but may lead to one of two deficiencies: (1) It may grossly overestimate the practical benefits of innovations which actually require such preparation and support but do not get them in other applications. (2) It may mean that other innovations which do not require such support are actually oversupported in such applications with a consequent waste of some resources.

We found that in order for APM to have practical utility at any level, instructors had to understand and support the logic of APM. The frequent use of the notions of confidence testing undoubtedly made the task of instructor training more difficult and raised all sorts of side issues which become irrelevant under the new approach. Even so, the one or two hours up to a half-day of instruction time available at most experimental locations would not be enough for some instructors. The logic and meaning of APM is best appreciated by putting oneself in the role of a student to answer many types of questions and to evaluate the quality of your logical thought processes. A two-day workshop should suffice to train almost any instructor to use APM very proficiently with his classes.

At the most basic level of application, APM can be substituted for the choice method and using the same existing test questions will produce the practical benefits of improving the fairness, reliability, and validity of testing and substantially reduce errors in pass-fail decisions taken at the school, and improve retention of the subject matter. It also allows each student to evaluate the quality of his thinking. This requires very little extra time from the students or instructors. Even at this level, APM begins to focus attention away from going through the motions of teaching and testing and on to what's happening to student achievement and understanding and in these terms how good is instruction (presentation and materials) and how good are the test questions. At some schools, the instructors are not prepared to cope with this shift in emphasis. While this characteristic of APM may be viewed as having considerable practical utility, in fairness to the instructors they should be given whatever additional training and support that may be required to prepare them for this shift in emphasis.

There are schools where good formal systems and administrative procedures for quality control of curriculum, instruction, and testing have been developed and put into operation. The instructors in these schools proved to be in a position to make effective use of APM and to appreciate the power of the additional and unambiguous information about student understanding of the subject matter. Only a minimal amount of instructor indoctrination was required for this type of situation.

Recommendation IV.

In every application of APM in an instructional setting, each student should continually assess the quality of his thinking processes by keeping a running record showing if he tends either to overvalue or undervalue the validity of his reasoning as reflected in his score settings.

Many students show much improvement when following this type of procedure. Students express great interest in improving this skill. The value that students place on this skill in most instances is far greater than the desire just to make better test scores when using APM. (As mentioned earlier, a fundamental characteristic of APM is that it rewards valid reasoning.) Instructors and school administration also value this skill, especially where they perceive logical thinking and decision making as important to the job for which they are training students. To cite some examples, trouble-shooting in the repair and maintenance of equipment, officer training, and basic training where the students go on for many months of additional study and training.

Another reason for this recommendation is that it leads to more discriminating use of the possible score allocations thus increasing the power of the test data from APM.

References

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

Armstrong, Robert J. and Mooney, Robert F. (1969) *A Study Concerning the Reliability of Probabilistic or Confidence Testing.* A report to the 10th Annual Conference of the Educational Research Association of New York State.

Hambleton, Ronald K., Roberts, Dennis M. and Traub, Ross E. (1970) A comparison of the reliability and validity of two methods for assessing partial knowledge in a multiple-choice test. *Journal of Educational Measurement,* 7, 75-82.

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

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
Four Lincoln Terrace
Lexington, Massachusetts 02173

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

DECISION-THEORETIC PSYCHOMETRICS: FINAL TECHNICAL REPORT

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

Final covering 20 month period 1 Nov 1968 through 30 Jun 1970

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

Emir H. Shuford, Jr.

6. REPORT DATE

July 1970

7a. TOTAL NO. OF PAGES

9

7b. NO. OF REFS

4

8a. CONTRACT OR GRANT NO. F44620-69-C-0068 (ARPA)

b. PROJECT NO. 9719

c. 61101D

d. 681313

9a. ORIGINATOR'S REPORT NUMBER(S)

SMC R-23

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

AFOSR 70-2291TR

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

The major items of work performed under this contract are reviewed and four recommendations are stated and justified.

Recommendation I. The notion and ideas of confidence testing should not be used in conjunction with admissible probability measurement because they interfere with the proper working of this new procedure.

Recommendation II. The truncated logarithmic scoring system, because of its unique properties, should be adopted as the standard measure of achievement and performance.

Recommendation III. Operational use of admissible probability measurement should be implemented at a school only after thorough indoctrination and training of instructors.

Recommendation IV. In every application of admissible probability measurement in an instructional setting, each student should continually assess the quality of his thinking process by keeping a running record showing if he tends either to overvalue or undervalue the validity of his reasoning as reflected in his score settings.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
decision making						
reasoning						
thinking						
instruction						
personnel testing						
confidence						
training						
human performance						