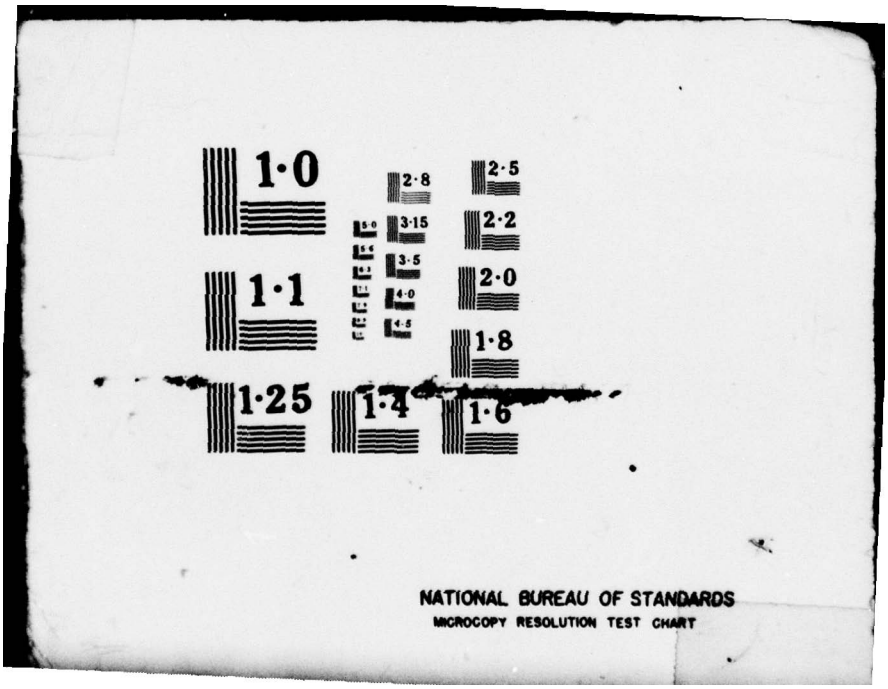
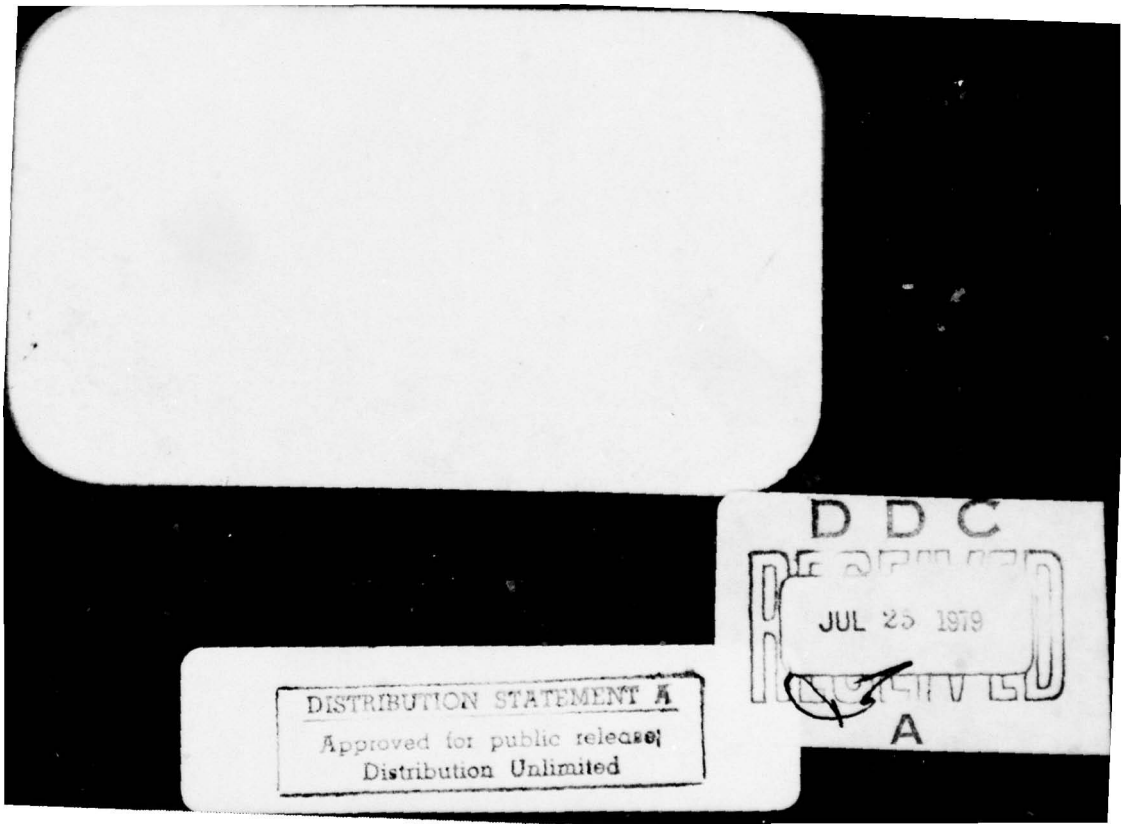
AD-A071 613 GEORGE WASHINGTON UNIV ALEXANDRIA VA HUMAN RESOURCES--ETC F/G 5/9
SCALES AND STANDARDS FOR MILITARY TRAINING RESEARCH.(U)
OCT 59 R G SMITH

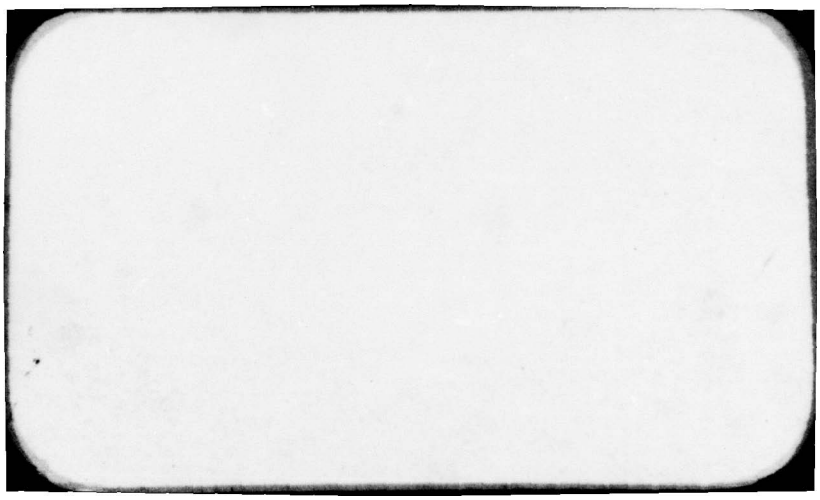
UNCLASSIFIED NL

1 OF 1
AD
A071613

END
DATE
FILMED
8-79
DDC


1·0
1·1
1·25
1·4
1·6
1·8
2·0
2·2
2·5
2·8
3·15
3·5
4·0
4·5
NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART


DDC
JUL 25 1979
A
DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited



(9) 5823

RESEARCH MEMORANDUM

(6) SCALES AND STANDARDS FOR MILITARY TRAINING RESEARCH,

by

(10) Robert G. Smith, Jr.

(11) Oct 59 (12) 32p.

Approved:

ROBERT G. SMITH, JR.
Director of Research

U. S. Army Air Defense
Human Research Unit
Fort Bliss, Texas

173 200

OCT 1959

TABLE OF CONTENTS

INTRODUCTION

This report describes new concepts for two related problems:

1. Scaling of proficiency measures
2. Setting proficiency standards for training

It is believed that when the methodological problems for applying these concepts are solved, military training researchers will possess more powerful tools for evaluating training programs and generalizing research findings from various specific studies.

THE PROBLEM

The procedures used to develop proficiency tests for military training research results in scores which have definite limitations. A review of the procedures used in developing proficiency tests will clarify the nature of these limitations:

1. The tests are commonly preceded by a job analysis and represent a sample of tasks required by a particular job. This means that scores are specific to a given job.

2. Scoring procedures for a given task are based upon various considerations, such as judgments of the seriousness of errors, or ease of observation of behavior. When scores for the tasks are combined, the resulting total score is in terms of units which are

an unknown quantity with respect to such major classifications of measures as rank-order, equal-interval or ratio scales (10).

3. Norms for the scores are based upon a specific sample of subjects. Furthermore, these norms are generally expressed in standard scores. This means that the score represents a crude approximation to an individual's rank order in a given sample.

The limitations described above lead to serious shortcomings, both of a practical and a research nature:

1. The use of these scores in training research renders practical recommendations difficult to make in certain situations. Where a less expensive training program yields measured proficiency equal to or greater than that developed by a more expensive training program, then there is little difficulty in making appropriate recommendations. However, when a more expensive training program also produces a higher level of proficiency, there is usually little basis upon which to make a decision.

2. The relative nature of the norms used in current proficiency scores provide little basis for defining satisfactory performance. One of the important uses of proficiency tests is as quality control measures for the graduates of training programs, both formal and on-the-job. An individual's score on a proficiency test usually provides no guidance as to whether he is satisfactorily trained.

3. One of the most important uses of measures is to provide indices whereby direct comparisons may be made of objects and situations of widely varying characteristics. Examples of such indices would be: Amount of learning per student-week; amount of proficiency per instructor; or the amount of proficiency per dollar cost. Such indices would provide important tools for training managers in evaluating the efficiency of training programs. These indices are typically formed by the algebraic process of division, although other processes may be used as well. The process of division is legitimately performed only upon ratio scales. The uncertainty with regard to the basic nature of the scales used in current proficiency tests means that such indices cannot be formed. Thus a powerful means for comparing widely different training situations, and thus increasing the generality of research, is lost.

4. Since the dimensions and units used in the typical proficiency test are specific to the particular research study, it is not possible to make direct comparisons of the effects of different experimenters and relate them to a common basis. It frequently occurs that different researchers are taking common approaches to common training problems, although with variations in procedure. Because each of these researchers will be using as evaluative criteria proficiency tests developed for particular jobs,

and yielding scores which are specific to the particular samples used, there can be no common basis of comparison.

The preceding comments point out the need for proficiency measures with the following characteristics:

1. The proficiency measures should be ratio scales. More mathematical operations can be performed on ratio scales than upon other kinds of scales. With ratio scales it is possible to develop new and useful indices involving various ratios for comparison of degrees of proficiency.

2. Proficiency measures should be expressed in terms which are sufficiently general to permit comparisons of the results of widely different researchers. In other words, they should be capable of measuring Proficiency in general, rather than Proficiency-as-a-NIKE-AJAX-Platoon-Leader, for example.

3. In situations in which the need for practical recommendations is paramount, proficiency measures should permit the making of a broader range of recommendations concerning levels of proficiency in relation to other criteria, particularly criteria which are related to the cost of training.

The purpose of this report is to propose new scales for training research which will have the characteristics described above, and to discuss problems associated with setting proficiency standards.

RATIONALE

The purpose of this section is to propose a model for the determination of satisfactory performance based generally upon decision theory. The method here proposed will be called Consequence Analysis because it assumes that the effect of an error or a lack of proficiency can be determined only through an analysis of the consequences of making the error.

The general rationale underlying Consequence Analysis is as follows: The making of an error has a consequence. These consequences may be different depending upon the situation in which the error is made. The cost of each consequence can be estimated or determined. Finally, the expected cost of an error can be determined by multiplying the cost of each consequence by the probability of the occurrence of the consequence and summing over consequences. The end result of this analysis will be the expected cost of the error.

PROCEDURES

The initial step in consequence analysis is to identify all possible errors that can be made on the proficiency test. In a multiple-choice question the selection of each mislead on an item

may have different consequences. It has been frequently recognized that some wrong answers are "wronger" than others. In a performance test, it is quite likely that the making of different types of errors may have different consequences.

When all the possible errors that can be made on a proficiency test have been identified, it is necessary to identify the consequences of the errors. At this stage of the analysis the services of a group of qualified job incumbents would seem to be a necessity. It is important to keep in mind that a given error may have different consequences under different conditions and that the same consequences may have a different cost under different conditions. Accordingly it is important to catalog not only the consequances of making the error, but the conditions under which these consequences may occur. A cost estimate should be assigned to each combination of consequence and situation. In many instances these cost estimates can be made quite accurately if we will make the effort to determine them. In other instances it may be necessary to make less accurate estimates.

Each combination of consequence and situation has in addition to a cost figure, a probability of occurrence. Again these probabilities are to be estimated as accurately as is feasible. The final step in consequence analysis is to multiply the cost of each consequence-situation combination by its associated probability.

When these problems are summed the result is an estimate of the expected cost of the error. Figure 1 shows a format which can be used in Consequence Analysis.

Figure 1

FORMAT FOR CONSEQUENCE ANALYSIS

Error: _____

	Cost	Probability	Expected Cost (Cp)
Consequence Situation			
Situation			
Consequence Situation			
Situation			

Expected cost of error = Σ Cp

It is well recognized that in practical application the model just proposed will yield results only as accurate as the estimates which go into it. It seems quite reasonable to expect that the ingenuity of researchers will yield improvements in methodology which

will make for more accurate estimates of the values which enter into the determination of the expected costs of an error.

IMPLICATIONS

With further effort being devoted to improving the accuracy of the various estimates used in Consequence Analysis and in increasing the efficiency of its application, Consequence Analysis may be expected to provide a powerful tool for determining the answers to a number of important practical questions which training researchers frequently face.

The principal usefulness of Consequence Analysis is that it provides a metric for lack of proficiency which can be balanced against the training costs required to overcome this lack.

Psychologists have frequently been unable to justify to research consumers or themselves the adoption of training methods which increase proficiency but at the same time cost more money. Consequence Analysis, by providing a monetary yardstick, may be very useful in converting improved proficiency into a saving which can be set against training costs.

The problem of optimum length of training programs also finds an evaluative instrument in Consequence Analysis. It is conceivable that in some instances, reducing the length of a course is

an action that one cannot afford because it costs too much in the consequences of errors.

It is common practice to graduate an individual from training provided he performs correctly on a test sampling the content of the training program. The use of Consequence Analysis in weighting test items is likely to result in graduates who have learned those skills and knowledges whose cost, if left unlearned, is of major importance.

Along similar lines, Consequence Analysis may result in important gains by using it to determine the cost of promotion from one sub-unit of training to the next. It might be more profitable to have an individual repeat one sub-unit of training than to promote him to the next one.

It should be recognized that Consequence Analysis is likely to find its widest application in those jobs in which the tasks involve well-defined procedures. Many of the technical tasks performed by military personnel are of this nature. It is from consideration of training problems for these individuals that Consequence Analysis was conceived.

At the same time, it should be possible to take a more positive approach. If exceptionally meritorious behavior were identified by means of approaches like the Critical Incidents technique, Consequence Analysis would be applied to these behaviors. Instead of costs, savings would be entered into the analysis tables.

DETERMINING THE LEVEL OF PROFICIENCY DESIRED OF HUMAN COMPONENTS OF MISSILE SYSTEMS

INTRODUCTION

The concept of a weapon system includes not only the equipment involved in the system but the human components as well. Both the human and the equipment components of a weapon system must operate at a high degree of reliability in order for the weapon system to be effective.

Those concerned with the reliability of equipment components, such as Lusser (3), have developed a set of concepts and procedures for setting reliability standards. Similar concepts and procedures for determining proficiency standards of the human component, however are presently lacking.

The purpose of this section is to consider concepts and procedures that are related to equipment reliability and examine, by analogy, their implications for human proficiency. It is felt that the application to the human component of requirements similar to those of the equipment component of a weapon system will shed new light on the adequacy of our present notions about setting proficiency standards for humans. These concepts and procedures have been adapted from Lusser (3).

THE MEASUREMENT OF RELIABILITY

The reliability of equipment components is defined as the probability of successful functioning under operating conditions. The reliability of the over-all system consists of the product of the reliabilities of all of the components of the system.

$$P_{total} = P_1P_2P_3 \ldots\ldots\ldots P_n$$

When human operators and maintenance personnel are included as components of the over-all system, it is quite clear that there is a serious need for a high degree of reliability in terms of probability of correct performance, for these personnel.

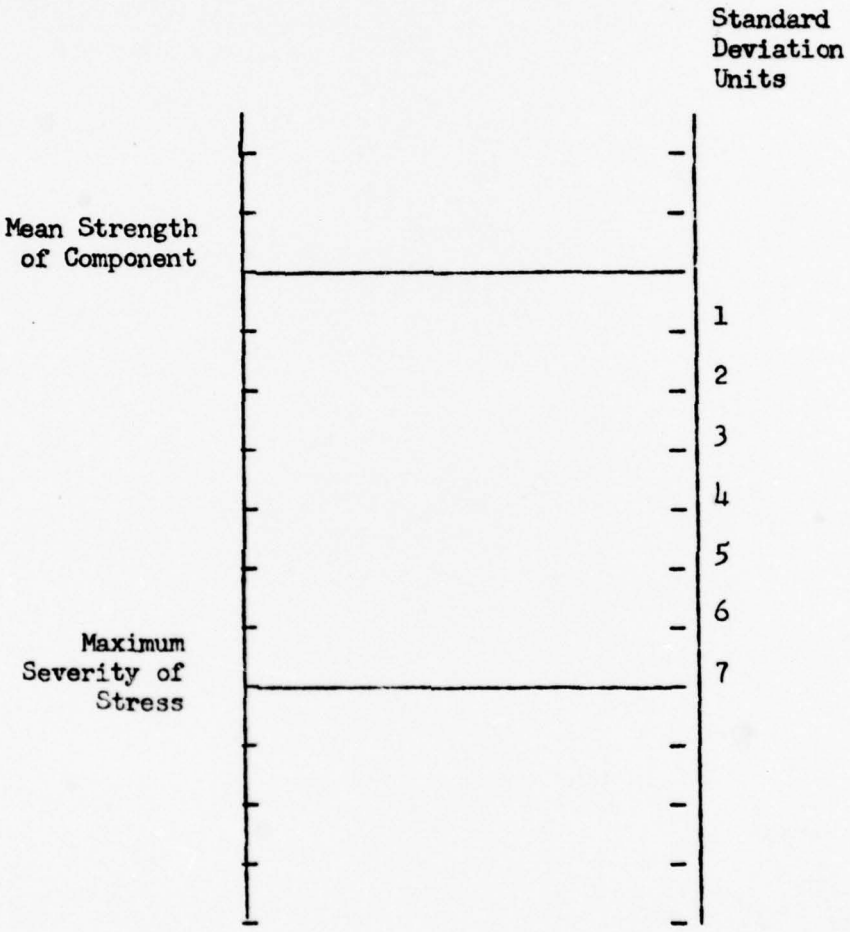
THE SAFETY MARGIN

Safety Margins Applied to Components

Lusser proposes that the average strength of a component be separated from the maximum severity of stress to which that component will be exposed by means of a safety margin which is measured in standard deviation units. The maximum stress is called the reliability boundary, and the safety margin is then the difference between the reliability boundary and the mean strength of the component, measured in standard deviation units which are based upon measures of the strength of the component. (Figure 2)

Figure 2

SAFETY MARGIN FOR EQUIPMENT COMPONENTS



Difference Between Human and Machine Components

In order to apply the model developed by Lusser (3) to the determination of reliability standards for human components of weapon systems, it is necessary to clearly describe the differences between human and machine components. 1) The strength of machine components is measured in continuous measures, based upon their resistance to a given force. On the other hand, the human equivalent of strength is proficiency, which is usually measured by noncontinuous variables based upon the presence or absence of error. 2) Although components may vary among each other, variability from one time period to another for the same individual must be considered for the human as well as differences between humans. 3) For machine components, maximum stress can be specified on the same scale and with the same units as the strength of the component. For humans, the equivalent of maximum stress cannot be so quantitively determined.

THE HUMAN ANALOGY

In order to carry out the analogy between determination of reliability standards for machine components and a similar determination for human components of weapon systems, the following are needed: 1) A definition for the human of resistance to stress and the reliability boundary. 2) A continuous scale for measuring resistance to stress. 3) A procedure for at least ranking stresses

so that conditions of maximum stress can be determined.

Resistance to Stress

In a machine component, strength is measured by subjecting it to various forces. For the human component, the equivalent of strength would be correct task performance. Any environmental change which increases errors for a given individual or group of individuals, can be considered stressful. Therefore, resistance to errors can be used as a measure of stress.

Naturally, errors are not equal in importance. Some errors have minor consequences. Others have major consequences. The notion of Consequence Analysis - of determining the cost of the consequences of errors should be considered here.

The Reliability Boundary

For machine components, as stated above, the reliability boundary is the maximum severity of stress to which a component will be subjected. However, for machine components, the stress and strength of the component are both measured in the same unit. This is not the case for human components of a system. In the previous section, the strength of a human component has been defined in terms of lack of errors in performance. Similarly, stress has been defined as an environmental change which increases errors. In order to avoid a circularity of definitions, a different basis must be used for determining the reliability boundary for the human components of a weapon system.

There are several possibilities for defining the reliability boundary in cost terms.

Finan (1), for example, makes the point that in training we must be certain that the proficiency of our troops exceeds that of our potential enemies. While this is undoubtedly the ideal, there are many problems involved in securing accurate data.

Another possibility for defining the reliability boundary is in terms of the cost of training. If the cost of training a person is established, then the cost of not training him could be established by Consequence Analysis.

Still another possibility is to define the reliability boundary as the cost of the equipment which the person maintains or operates. Or, in some instances, the cost of failure to accomplish the unit mission might be appropriate.

Further work should explore the suitability of these various bases for defining the reliability boundary. Such problems as the relative stringency of the various boundaries should be studied.

A Continuous Scale for Resistance to Stress

In order to determine the safety margin, strength or its equivalent for humans, proficiency must be expressed in continuous terms. However, an error is a single point occurrence. There is then a need for a procedure for converting errors into a continuous scale.

A useful way of doing this would be to use Consequence Analyses and convert the errors to the cost of their consequences. The continuous scale required for determining safety margins would then be the cost of the consequences of making errors.

Procedures for Scaling Stresses

If we accept the number of errors an individual makes as an inverse measure of his resistance to stress, then any environmental condition which increases errors is a stress. The number of errors made on a given task has been a matter of concern to test and measurement researchers for some time. One of the standard items of information one obtains on a proficiency test is the proportion of errors. This concern with errors has led to a considerable amount of information concerning task and environmental characteristics which make for increase in errors. Included among these factors are the following: Degradation of stimulus cues, increased time requirements, fatigue, the performance of concurrent tasks, and negative transfer, to mention a few.

The importance of methods for scaling stresses becomes more critical at the stage of quality control through proficiency testing than it does at the point of determining the reliability standards for human components of weapon systems. It is especially important that proficiency measures be devised which will test the limits of human performance under the most extreme conditions under which the weapon system will be employed.

Determining the Safety Margin

Having established the reliability boundary as the cost of a missile, how many standard deviations above this point should be the performance of the human components of the weapon system, when that performance is measured in terms of the consequences of errors? Lusser points out that there is no fixed procedure for determining the safety margin. How many standard deviation units must be included in the safety margin will depend upon the presence of various contingencies, each with its own particular contribution to the over-all safety margin. The following contingencies are adapted from Lusser's discussion, but are not direct translations of his list of contingencies. The particular margins contributed by each contingency are again judged in their relative weight by the frame of reference presented by Lusser's set of contingencies. The contingencies and their weights are listed below:

1.	Uncertainty in determining service conditions.	1
2.	Uncertainty in methods of evaluation of personnel	1
3.	Uncertainty in estimating reliability of supervision	2
4.	Uncertainty in estimating consequences of errors	2
5.	Employment in low-risk equipment, which can simply be repaired and set right again.	0
6.	Employment in high-risk equipment, in which human error can make for complete loss.	5

7. Employment in ultra high-risk equipment, in which human life or national prestige may be affected. 10

8. Less than complete sampling of tasks in proficiency tests. 2

9. Scatter in proficiency test scores. 1 - 3

10. Deviation of proficiency test conditions from those of maximum stress. 1 - 3

The total safety margin is determined by taking the square root of the sum of the squares of each of the contingency margins, for example:

$$\text{Safety Margin} = \sqrt{1^2+1^2+2^2+2^2+5^2+2^2+3^2+2^2}$$

$$= \sqrt{52} \quad = 7.3$$

This result is presented graphically in Figure 3.

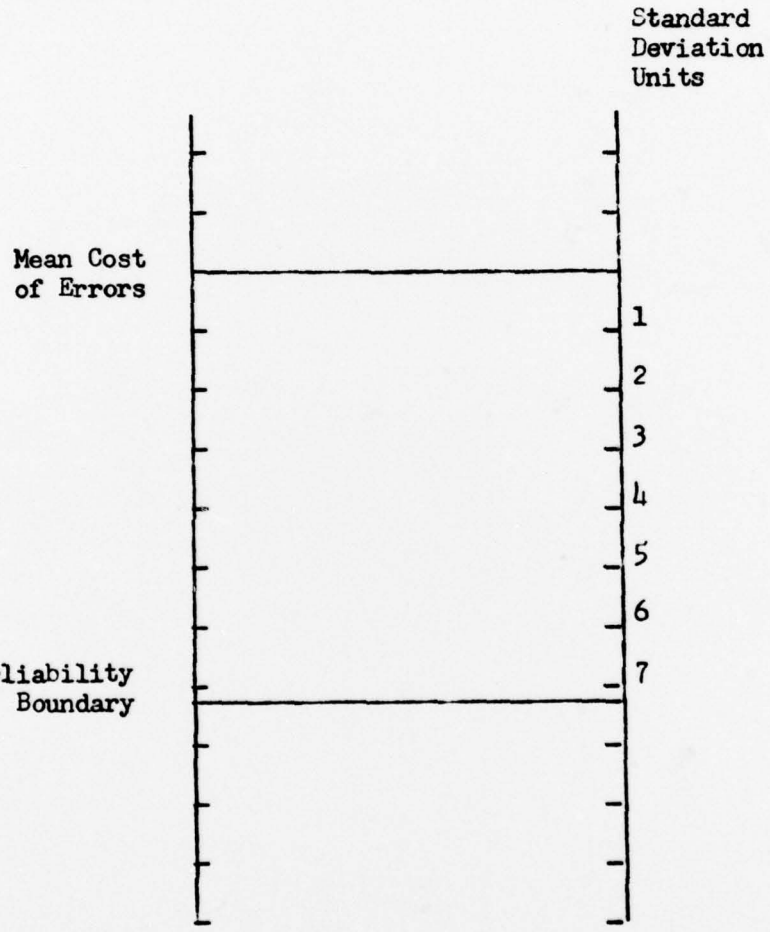
IMPLICATIONS FOR TRAINING RESEARCH

The reliability of a missile system is the product of the reliabilities of the individual components. The reliability of the human components should be equal to that of the equipment components if missile systems or weapon systems in general are to be reliable.

This analysis of the problem of insuring reliability of human

Figure 3

SAFETY MARGIN FOR HUMAN COMPONENTS



(Safety Margin = 7.3 Standard Deviations)

components has indicated a number of instances in which our present procedures and expectations regarding proficiency testing are highly inadequate. These instances will be described below.

Proficiency test scores are generally relative to the group from which they are obtained. They are either made relative by means of standardization procedures such as percentile ranking or standard scoring, or the difficulty of the items is adjusted to this group. For the human components of missile systems, this relativity is inadequate. A meaningful ratio scale is required. It is proposed that scaling errors in terms of the cost of the consequences of the errors would make for such a scale.

At the present time there is no absolute standard against which to measure the adequacy of training. The adequacy of training must be measured by comparison of one training program with another. The use of the Safety Margin for the evaluation of training would permit the direct measurement of the adequacy of training.

Proficiency testing and achievement testing make much use of written tests because they are relatively inexpensive. By putting both proficiency and school achievement testing in a context of quality control of components, the conclusion is reached that:

1. Testing must occur in realistic situations, covering actual tasks to be performed under a wide range of conditions.

2. Attention must be given to testing the limits of human performance, especially under the most stressful conditions expected to occur in the actual employment of the missile system.

This analysis has indicated the need for new standards of rigor in developing and applying proficiency tests. Since present standards of training adequacy are based on existing concepts of proficiency measurement, the new standards may be expected to have considerable impact upon conceptions of what constitutes adequate training. It is very likely that present standards of training adequacy must be revised upward to a considerable extent.

INFORMATION MODELS

Consequence Analysis as a method of scaling proficiency test scores appears to have its greatest potential value for those situations in which it is desired to develop a basis for practical recommendations concerning training. In many researches the matter of practical recommendations is not as important. Another possibility for scaling proficiency tests which possesses both the requirements of a ratio scale and independence on particular units of measurement is given by information theory.

In the following discussion of the application of information theory models to proficiency measurement, technical discussions of formulae will be avoided. The interested reader is referred to the following references (2, 4, 5, 6, 7, 8, 9).

What is Information?

Information is equivalent to uncertainty or variance (5, 8). If a situation is highly uncertain, with many possible alternatives that might occur, we obtain more information by observing what actually occurs than we obtain in a situation which was more certain and with fewer possible alternatives. The concept of variance is similarly related to the amount of information. A large amount of variance means that there is uncertainty about what will actually occur. Then a particular observation will yield a large amount of information. On the other hand, if the variance is small, making a particular observation does not yield as much information, since there are fewer possibilities of various occurrences.

The unit of information used in studies in the information theory framework is the bit, which stands for binary digit. A bit is that amount of information required to reduce the number of alternatives by one-half. The bit is thus independent of the particular units and dimensions used to measure variance or uncertainty, and thus will permit the comparison of results obtained in widely different experimental situations.

Several different models based on information theory and measurement have been used in psychology. Two of these appear to be of particular value for training research. These are the redundancy model and the transmission model.

The Redundancy Model

The redundancy model has been applied primarily in studies of language (6, 7). The maximum amount of information is contained in situations where all alternatives are equally likely to occur. Thus, since the English language contains primarily 26 letters and a space, the maximum amount of information would be indicated by English if the occurrence of letters and spaces were all equally likely. Of course, it is obvious that English does not operate this way. The letter "q", for instance, never occurs except just prior to the letter "u". There are also other constraints placed upon the usage of the symbols of the English alphabet by our language habits. These constraints, then, mean that less than maximum information is transmitted using the English alphabet. Accordingly, the alphabet when used to express language is redundant.

One way of looking at training is to consider it a process for bringing responses under the control of appropriate stimuli. Thus, the range of possible responses to a given stimulus is reduced, and we may consider that the relative redundancy of the responses to these stimuli has increased. In terms of information theory, then, the purpose of training is to increase redundancy.

One of the major advantages of the redundancy model is that there are already available certain important baselines. Estimates of the amount of information in single letters and words in connected

English have already been developed (6). Thus, since proficiency tests are samples of English text, the techniques of computing the amount of information in a proficiency test can be applied and the results related to the additional estimates of redundancy in English text. The number of different responses given to the same item of a proficiency test can be expected to be less for a trained group of subjects than for an untrained group of subjects. Thus these results when measured in information theory terms can be used as means for computing the relative amount of redundancy developed by training.

Another possibility for the application of the redundancy model lies in the currently active area of automated instruction. One of the presumably desirable characteristics of certain types of automatic teaching procedures is that the content should be programmed in such a way that the student never makes a mistake. Stated another way, this requirement means that responses to stimuli should be completely redundant. The techniques of information measurement can be applied then to determining the degree of redundancy attained in a given program or the effect of different procedures in approaching this high level of redundancy.

Another possible use of the redundancy model is in research on the effectiveness of various types of job aids. Since the job aid can be interpreted as a means of reducing the variability of on-the-job behavior, the redundancy model would apply here also.

The Transmission Model

The transmission model considers the human to be a channel for transmitting information. There is input in the form of stimuli. There is output in the form of responses. Information is transmitted through the human to the extent that responses are highly correlated with stimuli. Thus, whereas information is equivalent to variance, transmitted information is equivalent to covariance or correlation.

As the amount of information in the input is increased, there is normally an increase in the amount of information in the output. There is generally a limit to the amount of information transmitted through the channel, however, and eventually a point is reached at which additional amounts of information in the input does not result in additional information being transmitted through the channel. The maximum amount of information which can be transmitted through the communication channel is called the channel capacity.

Another possible way of looking at training is to consider it a process for increasing the channel capacity of the individual. Thus, an individual with greater training would be expected to be able to transmit more information than an individual with little training. In such an individual there would be a high correlation between the stimulus inputs and the response outputs.

The technique of data analysis for the transmission model are

different from those in the redundancy model. In the transmission model the analysis techniques are more complicated than in the redundancy model (2, 4).

However, the transmission model has one major advantage over the redundancy model. This is that various stimulus or input components can be analyzed in a method similar to the way the effect of different variables can be isolated in an analysis of variance. Then the amount of transmitted information attributable to each component of the stimulus can be identified (4).

Most of the kinds of analysis which can be performed using the redundancy model can also be performed with the transmission model. The choice must be based upon the complexity of the analysis desired.

The use of the transmission model in prior research on memory suggests that one way of increasing the channel capacity of the human is by recoding the material submitted to him (8). Thus, the channel capacity, or the maximum amount of learning, can be increased by recoding information into a set of symbols, each symbol of which carries more information with it.

SUMMARY

There is a definite need for proficiency measures in military training which have the characteristics of ratio scales with widely general dimensions. For studies with practical implications these measures also need to be criterion-related.

Models for proficiency measures based on decision theory and information theory are described and possible uses discussed. Consideration is given to the problem of specifying proficiency standards.

REFERENCES

1. Finan, J. L. Focus on man. Army, 1959, Vol. 9, No. 12, 28-35.

2. Garner, W. R. and Hake, H. W. The amount of information in absolute judgments. Psychol. Rev., 1951, 58, 446-459.

3. Lusser, R. Reliability through Safety Margins. Redstone Arsenal, Alabama, 1958.

4. McGill, W. J. Multivariate Transmission of Information and its Relation to Analysis of Variance. Human Factors Operations Research Laboratories, Air Research and Development Command; HFORL Report No. 32.

5. Miller, G. A. What is information measurement? Amer. Psychologist, 1953, 8, 3-11.

6. Miller, G. A. Information theory and the study of speech. In Current trends in information theory, Pittsburgh: Univ. of Pittsburgh Press, 1953.

7. Miller, G. A. Communication. In Stone, C. . (Ed.) Annual Review of Psychology, Vol. 5. Stanford, Calif. Annual Reviews, Inc. 1954.

8. Miller, G. A. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev., 1956, 63, 81-97.

9. Newman, E. B., and Gerstman, L. J. A new method for analyzing printed English. J. exp. Psychol., 1952, 44, 114-125.

10. Stevens, S. S. Mathematics, measurement, and psychophysics. In Stevens, S. S., (Ed.), Handbook of experimental psychology, New York: John Wiley and Sons, 1951.