AD 42123
ASTIA FILE COPY

BUREAU OF RESEARCH AND SERVICE
College of Education
University of Illinois
Urbana, Illinois

A CONSIDERATION OF INFORMATION THEORY AND UTILITY THEORY AS TOOLS FOR PSYCHOMETRIC PROBLEMS

Lee J. Cronbach

Technical Report Number 1
under Contract N6Ori-07146
with the Office of Naval Research

College of Education
University of Illinois
Urbana, Illinois

November, 1953

A CONSIDERATION OF INFORMATION THEORY AND UTILITY THEORY AS TOOLS FOR PSYCHOMETRIC PROBLEMS

Lee J. Cronbach

Technical Report Number 1
under Contract N6Ori-07146
with the Office of Naval Research

College of Education
University of Illinois
Urbana, Illinois

November, 1953

CONTENTS

Introduction 1

I. Measurement as a Communication Process 5

Shannon's model for the communication system 5

Tests viewed as communication systems 8

Assignment of persons to categories as the aim of testing 9
General meaning assigned to the term "test"
Use of categories in reporting test results
Summary of the test as a communication system

Ways of using communication theory 11

Implications of the communication analogy 12

II. Measures of Uncertainty and Information in Terms of Message Length 15

A derivation of Shannon's measure of uncertainty 15
Degree of confidence 15
The standard item
Standard items required to reach certainty 18
Interpretation as sequential analysis
Interpretation in terms of coding
Interpretation as log confidence
Cautions in employing H to measure uncertainty

Information as the reduction of uncertainty 24
Measure of residual uncertainty after testing 24
The confidence continuum and the measure R 25

Exhaustiveness and dependability 27
The index of exhaustiveness 27
The index of dependability 27
Significance of the formulas 28
Non-symmetric validity relations
Relations involving a fallible criterion 33

Application of the formulas to ordered scales 36
The rectangular distribution 36
The normal distribution 37
Runder assumptions of normality 38

Summary of major implications 41

III. Measures of Uncertainty and Information in Terms of Correct Decisions 43

A measure of average uncertainty 43
Previously published formulas for measuring discriminating power 44
Geometric interpretation as a dispersion measure 44
Residual uncertainty after testing 45

Gain in average confidence (information) 46
Information yielded by a series of independent items 48
Exhaustiveness and dependability 49

Significance of confidence formulas 50

IV. An Introductory Statement of Utility Theory, Giving its Implications regarding the Information Formulas 51

Utility theory 51
Basic data required 51
Transition matrix
Evaluation matrix
Interpretation matrix
Calculation of utility 52

Review of the average confidence formulas 54

Review of the average log confidence formulas 54

V. Analysis of a Psychiatric Screening Test by Utility Theory 56

Correlational validity 57

Utility Analysis 58
Introduction of evaluations 58
Determining entries for the evaluation matrix 59
Likelihood ratios as a basis for cutting scores 59a
Choice of cutting score 60
Optimal strategy at high and low E.R. 61
Comparison with conclusions of correlational analysis 61

Utility curves for various strategies 61
Decisions with the test compared to a priori decisions 61
Strategies to be compared 61a
Results
Limited strategies 61e
Reduction of costs 62

Conclusions 63

References

INTRODUCTION

In any field of investigation, the adoption of a new mathematical model often permits one to investigate questions which had been overlooked or had been incapable of treatment under the model formerly employed. Many investigators have found the communication model introduced by Shannon (26) a stimulating conceptualization and point of departure.

It has seemed to several writers that this approach would be particularly useful for test theory. Both Hick (19) and Miller (22) have suggested this, but question whether communication theory will yield fundamentally new results, or will contribute chiefly by presenting old results in a fresh perspective. Either of these contributions might be significant.

The present writer began to examine testing problems in terms of Shannon's information theory in 1949. It became clear that this new model had several values. A particularly striking feature is the generality which permits one to formulate statements about tests which will apply to devices which categorize, devices which measure along a single numerical scale, devices which yield scores along several dimensions, and even those instruments which lead to verbal descriptions of the person tested. This general approach suggests that we inquire how much useful information a testing procedure tells us. The older approaches to test evaluation ask how accurate the test is in making any single measurement and prediction, and make no provision for evaluating as a whole the information yield of a test which provides many measures or predictions.

A preliminary report in 1952 (11) presented a statement of testing problems in the language of information theory. In this report, no attempt was made to reexamine the theory developed by Shannon; rather, his formulas were taken over bodily and translated into implications. Criticisms of the preliminary report were solicited, and these drew attention to the fact that the model itself required careful reconsideration. While the informational approach led to new insight regarding tests, there was reason to think that alternative mathematical treatments involving some of the same conceptions deserved consideration, and might be even more appropriate. Under support from the Office of Naval Research, we undertook to study the basis of Shannon's formulas, to examine other possible formulations of the testing problem, and to identify implications for test analysis.

The comparison of various approaches has made clear that while the Shannon measure has interesting properties, it does not correspond perfectly to the requirements of psychometrics. Assumptions are embodied in the Shannon model which are not appropriate for test analysis. At the same time, we have identified even more places than before where the examination of a test in terms of the communication model suggests concepts or questions which are worthy of serious thought.

An alternative approach which we shall discuss in this report is an information analysis based on average probabilities in contrast to Shannon's use of an average of log probabilities. This analysis includes and extends the formulas for discriminating power of a test which have been proposed several times in recent years. This series of formulas has approximately the same implications as those based on the Shannon measure, even though the mathematical formulas differ. We find this second method of analysis superior in some respects to that of Shannon --- but unfortunately it too has serious limitations for test analysis.

The first product of our work, then, is that we have found out what not to use and why not. It is valuable to disclose the limitations of the Shannon formulas or the discrimination formulas for psychometrics, since the findings warn investigators in this field against using these schemes of analysis indiscriminately.

Our work has also indicated what will be a more suitable line of attack. The problem of the tester is to employ whatever testing time is available in whatever way will most improve the decisions made. The value of the tests is to be judged by the improvement in decisions. Therefore a "utility theory" is called for. The two systems of information analysis we have considered are in essence special cases of utility analysis which invoke somewhat limiting assumptions. We expect the utility theory to provide a proper formal demonstration of the conclusions suggested by the information analogy, and the utility approach, being more general, should also lead to conclusions not covered by the information treatments. Moreover, the explicit utility theory should have substantial value in organizing and clarifying test theory.

At this point, we have not studied the utility approach thoroughly, and are not ready to present it in final form. In order to introduce it, we provide a brief sketch of the essential concepts. Ultimately, we expect to replace the present report with a discussion organized around utility theory.

While exploring utility theory, we worked through various examples. One of these worked examples is presented here, because its conclusions are interesting in themselves. The presentation provides a simple illustration of utility analysis and the types of result that can be expected from it.

The five papers which follow are a record, therefore, of the thinking developed during the past year. We may comment briefly on the nature and possible significance of each.

Section I describes the communication model and shows that the test can be viewed as a communication system. This is a purely verbal presentation. It sets forth a series of analogies which are useful in thinking intuitively about tests, regardless of the mathematical system adopted. It is still the basis for applying mathematical information models. This section is essentially a restatement of materials given in the 1952 report.

Section II presents the formulas for analyzing information according to Shannon's assumptions, and clarifies the meaning of those formulas. While we do not recommend that tests be treated by these formulas, many of them suggest important ways of looking at tests. We expect ultimately to develop comparible but more adequate formulas within the utility model. Of particular importance in Section II are the examination of testing as a problem in sequential analysis (following Evans (14)) and the consideration of the concepts of exhaustiveness and dependability.

Section III presents the formulas for analyzing information by arithmetic averages. While these formulas are also to be replaced by a more adequate system eventually, they are somewhat closer in conception to utility theory than Shannon's. Persons interested in previously proposed evaluations of tests in terms of their "discriminating power" will find that Section III clarifies the significance of those formulas.

Section IV discusses briefly the essential concepts of utility theory, together with the reasons for preferring this schema to the formulations presented in II and III. This section indicates the line of attack we intend to follow in further studies, and provides a foundation for understanding Section V.

Section V is a study by means of utility theory of one particular test, intended for psychiatric screening of recruits. This provides a demonstration of the type of conclusions utility theory can yield, which would be overlooked in conventional validity analysis. The procedures of this section may be applied to other screening tests.

These reports are primarily designed to share our present thinking with others interested in these problems, rather than to provide a final statement of conclusions. Criticisms of the concepts developed in this report will be welcomed.

Our present thinking has depended to a very large extent upon the stimulation provided by comments on the 1952 report from the following persons: Yehoshua Bar-Hillel, Massachusetts Institute of Technology; Raymond H. Burros, Training Research Laboratory, University of Illinois; A. S. C. Ehrenberg, Institute of Psychiatry, University of London; George A. Ferguson, McGill University; W. R. Garner, Institute for Cooperative Research, Johns Hopkins University; W. E. Hick, Applied Psychology Research Unit, Psychological Laboratory, Cambridge, Massachusetts; F. M. Lord, Educational Testing Service; William McGill, Massachusetts Institute of Technology; Brockway McMillan, Bell Telephone Laboratories; George Miller, Massachusetts Institute of Technology; Frederick Mosteller, Department of Social Relations, Harvard University; Henry Quastler, Control Systems Laboratory, University of Illinois; B. O. Smith, College of Education, University of Illinois.

Particular credit for the work presented here should go to Dr. Eugene Burdock, now with Carnegie Corporation. He served as Research Associate with this project in 1952-1953, and rendered valuable assistance. He was author

or co-author of many file memoranda which represented way stations toward the present paper. Dr. Goldine Gleser of Washington University Medical School, Department of Neuropsychiatry, has helped substantially in the preparation of the present report.

I. MEASUREMENT AS A COMMUNICATION PROCESS

Tests have customarily been described in language derived from the measurement problems of the physical sciences. The very use of the term measurement suggests that the function of a psychological or educational test is analogous to that of the meter stick. Employing physical measurement as a model, the psychologist has gone on to employ such associated concepts as scale, error of measurement, and reliability as principal constructs for the evaluation of tests. Recently there has been increased awareness that many instruments for measurement in the behavioral sciences lack the properties of physical measuring scales. Coombs, among others, has suggested that tests can be more soundly interpreted if other models are used as a basis for measurement theory (7).

An alternative model also offered by the physical sciences suggests new ways of looking at tests. The communication engineer, dealing with problems in electronics particularly, has made extensive use of a schematic model of the communication system and of a mathematical theory of communication. It is possible to describe tests also as communication systems. The communication analogy throws new light on the function of tests, and leads to new ways of evaluating the power of any test.

Shannon's Model for the Communication System

Although an excellent summary of communication theory is already provided in a recent article by George Miller (22), a brief account of the communication model is offered here.

The mathematical theory of communication is most closely associated with Claude Shannon. He developed the system while analyzing problems of coding in cryptanalysis, and later applied it to problems of transmission in electrical and electronic communication systems (25,26). It has been employed independently by Wiener in the study of servo-mechanisms, and has had a substantial number of applications to all types of dynamic phenomena. Indeed, its protean adaptability is both the greatest advantage and greatest disadvantage of the model. The information concept can be used to describe almost any process, but has so many different possible interpretations that it is hard to focus on the exact meaning of such a concept as "amount of information transmitted." Quastler, who has edited a symposium on applications of the theory to biological processes, makes this comment: (24, p.11)

> "Information Theory" is a name remarkably apt to be misunderstood. The theory deals, in a quantitative way, with something called "information" which, however, has nothing to do with meaning. On the other hand, the "information" of the theory is related to such diverse activities as arranging, constraining, designing, determining, differentiating, messaging, ordering, organizing, planning, restricting, selecting, specializing, specifying, and systematizing; it can be used in connection with all operations which aim at descreasing such quantities as disorder, entropy, generality, ignorance, indistinctness, noise, randomness, uncertainty, variability, and at increasing the amount or degree of certainty, design, differentiation, distinctiveness, individualization, information, lawfulness, orderliness,

particularity, regularity, specificity, uniqueness. All these quantities refer to some difference between general and specific; in this sense, they can be measured with a common yardstick. Furthermore, measures which are appropriate exist, due to the developments of Information Theory.

A communication system involves a transmitter, a receiver, and a channel connecting them (Figure 1). The communication is sent because the receiver is uncertain, and an appropriate communication should reduce his uncertainty. If we think of the teletype, as used to transmit stock quotations, we can see the essential features of the system.

(1) Ensemble of possible messages. The transmitter can choose among a great variety of possible messages. Some of these are more likely to occur than others, but until the message is selected, there is some uncertainty as to what it will actually be. The receiver can expect that today's quotation on U. S. Steel will be within a few points of yesterday's, but he cannot be certain of this, nor can he predict the precise quotation that will be transmitted.

(2) True message, or input. When the market closes, the true quotation (one of the possible messages) is available at the transmitter but the receiver is uncertain what it is. This message put into the system is called an input. For any input, some output is received at the other end of the system.

(3) Channel. The channel includes all the equipment used in conveying the message. Actually, the channel can be divided into a chain of smaller communication systems (operator's visual system, operator's motor system, teletypewriter, cable, etc.), for each segment of the channel has its own input and output. But we lose nothing by regarding a sequence of these miniscule systems as a single communication channel.

The channel has its own response properties. For any given input, there is some distribution of outputs which the channel may yield. The channel filters out some parts of the input and fails to transmit them; it may introduce errors of various sorts due to mechanical failures, friction, and interference.

(4) Channel capacity. Each communication channel has a limit or capacity. Only some specified number of transmissions per unit of time can take place, either because of an agreed rate of transmission or because the physical properties of the system prevent more rapid transmission.

(5) Noise. When a particular message is sent several times, there may be variation in the output. Thus, when the letter L is typed at the transmitter, interference (static, or cross-talk between circuits) may cause K

A priori

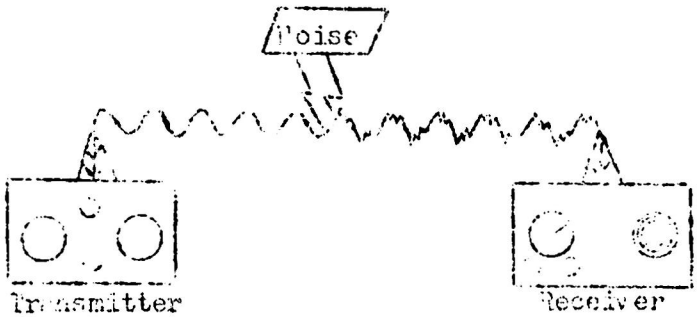


True Message (Input)

No Message Obtained

No Knowledge of State at Transmitter

A posteriori

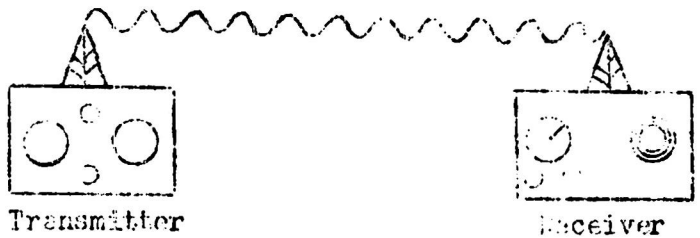


True Message (Input)

Obtained Message (Output)

Imperfect Knowledge of State at Transmitter

Ideal



True Message (Input)

Output Message same as True Message

Perfect Knowledge of State at Transmitter

Figure 1. The Communication Model.

sometimes to be printed at the receiver. Such variation introduced in the channel is called noise.

(6) Receiver. The receiver may be regarded as a person who wishes to make some decisions (i.e., accept or reject some hypotheses) and who desires to reduce his uncertainty in order to make better decisions. Information regarding today's stock quotation is needed as a basis for tomorrow's decisions.

Before the message is transmitted, we consider the receiver as having some initial or a priori uncertainty. After the transmission, the receiver has a message containing some degree of error, great or negligible. There is no possible harm in such a misprint as "U S STEEK 47½" which can be instantly corrected. It would be very important that the numerals be transmitted without error. So long as error is possible, the receiver has some a posteriori uncertainty after receiving the message.

Tests Viewed as Communication Systems

The person who gives a psychological or educational test is in the position of the receiver. He has some degree of uncertainty, and wishes to become more certain before making decisions. The employer judging an applicant, the clinician diagnosing a patient, the teacher evaluating a pupil's proficiency — each assumes the existence of some true classification or description for the subject which he wishes to determine as certainly as he can. The testing procedure is a device for making a more certain decision as to the probable true message or description of the individual. While, with the teletype, the true message is actually known to someone when the transmission is sent, the true message desired in testing is ordinarily unknown and not directly observable.

Consider the employer who has ten vacancies to fill, and fifty applicants. A priori, without further information, he could only select men by chance. He is uncertain whom he should employ. To gain information he gives each man a ten-day tryout. This is his information-getting device. At the end of this time, he may have observed sufficient differences to warrant a definite decision as to which ten men to hire. Or he might be certain about eight men, and have another group of twelve who seem equally qualified for the two remaining vacancies. The tryout has decreased his uncertainty. A test might have been used in place of the tryout, and it would similarly have reduced his uncertainty to some degree.

All elements of the communication system are to be observed in testing:

(1) Ensemble of possible messages = Criterion distribution. The "possible messages" are the true descriptions or scores which might arise in the population being studied; the receiver wants to know which of these applies to the person under study.

(2) Input=Criterion classification. Each person tested has a true or criterion classification.

(3) Channel. The entire operation of testing, which may include observation, scoring, and interpretation of performance, may be regarded as part of the communication system. Each of these is an opportunity for information to be lost (discarded or distorted), and it may sometimes be profitable to study one specific link in the chain separately.

(4) Channel capacity. While the engineer ordinarily thinks of capacity in terms of information transmitted per unit time, the tester may often be more interested in the information yield per test administered. Each test has a certain capacity, determined by its discriminating power.

(5) Noise=Error in testing. Typical errors which cause a single input (true classification) to yield different outputs (obtained classification) are sampling errors, fluctuations in the subject over time, chance responses due to guessing, and error in scoring.

(6) Receiver=Test user. The receiver arrives at some final obtained score or obtained classification on which a decision is based.

Assignment of Persons to Categories as the Aim of Testing

The preceding section has been quite general, making no assumptions regarding the form of the messages being transmitted. These might be descriptions such as are derived from a projective test protocol, categorical designations such as "literate" - "illiterate", or numerical scores such as an IQ. The mathematical theory of communication, however, is stated in terms of qualitative unordered categories. To examine tests in the light of Shannon's mathematical formulation, we shall think of our criterion "messages" as categorical classifications. It can be readily shown that this formulation can be extended to continuous (numerical) scales and to complex descriptions.

General Meaning Assigned to the Term "Test"

The usage of such terms as "test" and "measurement" varies among different writers. Some would restrict the term "test" to devices yielding quantitative statements about individual differences . Others would restrict the word "measurement" to processes which locate individuals on a scale having demonstrably equal intervals, or might even require that the scale have an absolute zero. Since our interest is to develop a highly general theory, applicable to the widest variety of studies of individual differences among persons, we will not apply such restrictions.

We shall speak of any procedure employed to compare two or more persons as a "test". This means that our discussion will be relevant to many devices

which are not essentially quantitative, and which would not be called "measurements" in the ordinary senses of that term. An interview, for example, or a clinical counseling session, permits an employer or a counselor to make judgments about the individual he talks with. This judgment may be expressed in a complex verbal statement which is formally quite unlike the numerical result a test would yield. The purposes of interview and test may be much the same, however; namely, to permit a more confident decision about one or more questions. It is therefore reasonable to inquire which of the two procedures is preferable, in any given problem. It may be necessary to develop our method of analysis somewhat differently for various types of test report, but this does not make less valuable the overall conception within which these detailed models fit.

It is to be noted that a test is defined as a process for making comparisons. Perhaps we could equally well regard a test as a process for describing any one individual. The statement that "John is male" is significant, however, only because it is possible for John to be non-male. There is no use, from the point of view of making decisions about John, in reporting a truism. The purpose of any test is to find out how one individual differs from others; thus all testing is an attempt to make discriminations between individuals. Sometimes, it is true, our result is expressed in a statement about John alone. But the statement "John is six feet tall" describes John on a scale which has meaning only because other persons or objects have different heights. The user of a quantitative statement brings to bear a frame of reference for interpreting the scale, and this frame of reference arises out of experience with other individuals. Hence any result from a test is an implicit comparison or discrimination.

Use of Categories in Reporting Test Results

A test is used to aid in making decisions or statements of the form "S belongs in category i". For the psychologist, S will ordinarily be a person or animal, but the same model holds when a sociologist reports a characteristic of a neighborhood, or when an educator, let us say, judges a school building. In behavioral testing, typical statements naming a category explicitly are the following:

"S is paranoid schizophrenic"
"S is a poor risk for medical school"
"S is a superior officer"
"S has a speech defect"
"S is in a state of high anxiety"
"S needs further work on the addition combinations"

That these are categorical statements is instantly seen by noting that the word "not" or "no" inserted in the statement causes it to describe a different category of persons from whom S is being differentiated. Sometimes the category named is one of a whole set of possible categories (e.g., the various categories used to classify patients); sometimes there are only two categories in a set (good risk, poor risk, for example).

Categories are likewise used when a result is expressed in numerical terms, even though we think of numerical scales as continuous rather than discrete. "John is six years old" places John in a category of "persons having CA between 5-6 and 6-6", under the usual definition of this class interval. If the measurement is more precise, the category becomes smaller. "John's four-year grade average was 3.1523" locates John in a category "3.15225 - 3.15235".

When we have a set of categories, it may be important to recognize whether the categories are unordered, ordered, or partially ordered; whether the category system is unidimensional (e.g., red, black) or multidimensional (red-odd, red-even, ...); and whether, if ordered, any statement is to be made about the magnitude of the scale interval represented in the category (8). These differences may influence the way in which the communication or utility model is to be applied. Shannon states the problem, however, in terms of system of **k** categories, without restricting the category system to any of the above types.

Summary of the Test as a Communication System

We have stated that the tester desires to make decisions about an individual. In order to do so as wisely as possible, he would like to know what categories this individual falls into. The categories that concern the tester will of course be those relevant to his problem. The tester has some a priori expectation about the subject's classification. If this a priori knowledge is too uncertain to be a basis for action, the tester seeks to reduce uncertainty by means of the testing device. If the test is useful, it permits the tester to assign the person to a category and to make decisions about him with greater confidence than he had before testing. It is consistent both with common usage and with communication theory to say that the test has contributed "information" about the individual. The person's proper classification is an unknown "message" to be discovered. After administering the test, we have an "obtained message" which may be an accurate basis for classifying him, or which may be to some degree untrustworthy.

The essential problems of test theory are (1) to find appropriate measures for the increase in confidence or goodness of decisions permitted by a test, and (2) to determine how tests can be designed to be most effective.

Ways of Using Communication Theory

Communication theory (information theory) can be used in two ways. First, it offers an analogy which we can use to describe tests and testing problems. Second, it offers a mathematical system, containing definite postulates, formulas, and theorems. These two uses make different demands upon the theory, and it is well to distinguish them.

Analogies in thinking are helpful, indeed indispensable. The Lewinian school of thought, for instance, communicates effectively with such terms as "force", "barrier", and "distance". This permits a reader to visualize a

conclusion which would otherwise be unduly abstract. The concept of homeostasis or dynamic equilibrium, borrowed from chemistry and physiology, is similarly used to emphasize certain aspects of adjustive action.

Analogy is primarily a contributor to creative thinking - to perceiving problems in a new light, or to imagining new solutions. Analogy is much less useful for criticizing hypotheses or for establishing precise results. Analogies are rarely perfect, and any failure of the event under study to correspond to the model means some degree of error in the conclusions.

When a model is used as a basis for rigorous deduction or for mathematical treatment of a problem, careful scrutiny of its details is necessary. Either we must be certain that the model fits the problem in all details, or we must locate the precise points of non-correspondence and ascertain just how much effect these "errors" have upon the result we are deriving.

Information theory has been used in both these ways, i.e., formally and informally. In the informal use, the investigator tries to state his problem in informational terms. This may clarify his problem and suggest variables whose effect he should measure in his experiment. A formal use of the theory is represented when the investigator adopts the Shannon formulas and computes his results in terms of them. Thus Hick (19) suggests, as an example, that the optimum item difficulty may be determined by maximizing the rate of transmission of information, which is, under certain conditions, a function of the item difficulty. Since this function is rather flat, any failure of the formulas to fit the case under study may lead to a sizeable error in determining the optimum difficulty. Before we can accept conclusions from any such formal use of the Shannon formula, we need to know whether it does fit the testing problem perfectly, or how serious would be the effects of its failure to fit.

In the ensuing sections we present the Shannon formulas in such a way that some of their characteristics are apparent. We also develop a set of reasonable alternative formulas. The formula may be applied formally only in situations which fit the underlying model in all details; this is rarely the case in problems of test analysis.

It is not necessary to be so cautious in using the communication model informally. Used simply as an analogy, it is remarkably fruitful in suggesting ideas. We shall point to such inferences as we go along.

Implications of the Communication Analogy

Two simple but important alterations in concepts regarding tests result from the general model as presented in this chapter.

(1) The value of a test should be judged by its ability to reduce the a priori uncertainty. (Cf. 10, p. 65). This seems obvious if we view a test as a communication.

This is not, however, the concept most often employed in judging tests. The test has almost always been judged by a correlation (or the like) stating the validity of the test. A correlation is a measure of improvement over chance decisions. If the potential user of the test is already able to make inferences with better than chance accuracy, using whatever information about the individual he has prior to testing, the validity coefficient overestimates the possible contribution of the test. An example shows the significance of this distinction.

Without giving a mental test, the school can probably make fairly good estimates of pupils' mental ages. From age, grade in school, and the teacher's opinion, the score on a group test could probably be predicted a priori, with only moderate error in most cases. The contribution of the test should be judged, not by ability to report scholastic ability to a receiver who has no knowledge, but by its ability to give information not already at hand. The mental test does give some essential new information. A conventional coefficient does not report the validity over and beyond a priori knowledge; it estimates validity over all pupils in a broad class, crediting the test for information overlapping what the teacher knows before the test is given. In contrast, the TAT and the like are regarded as inaccurate by usual standards. But since the school probably is completely ignorant about pupils' fantasies and the needs they signify, whatever valid information TAT reports is new information. Hence test analysis which discounts a priori information will tend to encourage the use of tests which measure things not now known, and to discourage use of tests, however accurate, which duplicate information that can be gleaned from records already at hand.

In quite a different context, we may point to the many current studies of the effectiveness of the clinician in making diagnoses or discriptions. A common plan is to determine how many correct judgments the clinician can make upon the basis of some set of data, compared to the number he might make by chance. One might provide the judge with data about the individual and ask him to predict certain aspects of the individual's behavior. Instead of comparing his success with chance, we might compare it with the judges' success in predicting solely from knowledge that the individual belongs to a certain group (e.g., patients). There is evidence that the clinician sometimes does better when making the latter type of prediction, from very little information, than when he predicts from more complete data (17). That is to say, the individualized procedure -- whether better than chance or not -- is poorer than the prediction based on a sterotype of S's group. The evaluation of the clinical procedure should be based on improvement over the best a priori procedure, rather than on improvement over chance. Dailey has shown one method of estimating the goodness of a clinical procedure by comparing it with the best a priori estimate (13).

(2) Multidimensional measures can be compared with unidimensional measures. In previous psychometric analysis, it has been customary to judge a test against a single criterion. A multidimensional test may be treated by multiple correlation, to determine if it predicts this criterion better than does some unidimensional measure. The question is rarely raised, however, whether the multidimensional test which can be interpreted so as to predict many criteria gives more information than some unidimensional test aimed toward just one criterion. Some mathematical methods, such as canonical correlation, are available for this problem, but they have been given little

application in psychometrics.

In communication analysis, the question just raised is a very reasonable one. Given a certain amount of transmitter time for our message, we might either seek a highly dependable report on some one thing, or we might prefer to have less complete or less accurate reports on many things. To take a simple example: Late on Saturday afternoon, one radio listener might prefer a fifteen-minute sports report in which dozens of football scores were read off --- very little information about each of many separate questions or uncertainties. Another listener would much prefer a fifteen-minute report describing a particular game play-by-play --- much information on one closely related set of questions; no information on other questions.

Sometimes noise or error in transmission makes it very important to send just one single message with maximum precision (as when transmitting an S-O-S, together with the ship's position). The operator sends his limited message over and over, even though he could use the same time to convey additional (but less important) information. These examples make it clear that it is meaningful to ask whether, in a given situation, we are wiser to emphasize breadth of coverage at the risk of thoroughness, or vice versa. In communication parlance, we "can trade bandwidth for fidelity"if we wish.

One of the major problems posed for test evaluation is to find a proper formula for comparing the procedure that gives moderately dependable predictions on many dimensions, with the procedure which is more dependable but less comprehensive. It may well be that the interview, capable of touching on dozens of dimensions in a half-hour, is frequently a better personnel classification procedure than the test which answers just one question with great precision, in the same half-hour. While our analogical use of information theory poses this question, to answer it accurately will require that we develop a rigorous treatment using utility theory.

II. MEASURES OF UNCERTAINTY AND INFORMATION IN TERMS OF MESSAGE LENGTH

A Derivation of Shannon's Measure of Uncertainty

Following the general conception of a communication system presented in Section I, we consider a transmitter, channel, and receiver. The data available at the receiver are used to infer the state of affairs at the transmitter. We shall adopt the following terminology and notation:

General	Significance in Testing	Notation
Possible states at the transmitter (alternatives) allowed)	Criterion classifications	$x = a,b,c...k$
Transmissions	Persons	$S = A,B,C,...N$
Possible states at the receiver (alternatives allowed)	Categories in which results are reported	$y = \alpha,\beta,\delta...\kappa$

Any person S has his proper or criterion classification x_s. The receiver at any time has certain data y_s about S, and desires to infer x_s. y_s includes whatever responses or scores describing S are available; when the tester first approaches his problem he may have no differentiating information, in which event y is the same for all subjects.

Using whatever data y_s are available, a judgment as to the proper x_s will ordinarily be made with some degree of uncertainty. As more data are obtained about S, the receiver hopes to increase the confidence with which he can infer x_s. We may evaluate a data-gathering procedure by determining how much it increases confidence, or reduces uncertainty. We might speak of measuring changes in "degree of certainty", but this conflicts with the dictionary usage which regards certainty as all-or-none. In common speech, one does not refer to "degrees" of certainty. Therefore we shall use the terms "degree of confidence" (or "confidence") and "degree of uncertainty" (or "uncertainty") as antonyms. The precise relation between these two will depend on the formulas used to define them, but in general we may think of degree of uncertainty as "1.00 minus degree of confidence".

Degree of Confidence

When the receiver infers that $x_s = a$, he is accepting a hypothesis. The probability that this hypothesis is true we may call his confidence regarding that hypothesis. We shall let Conf (a, y_s) denote the confidence that $x_s = a$, for a person having a given y_s.

$$\text{Conf}\ (a, y_s) = \Pr\{x_s = a \mid y = y_s\} = p_{a/y_s} \qquad (1)$$

Table 1. Specimen Transition Matrix

		Outputs (Responses) (y)					
		α	β	γ		κ	
Inputs or Criterion Classifications (x)	a	$p_{\alpha/a}$	$p_{\beta/a}$	$p_{\gamma/a}$	...	$p_{\kappa/a}$	1.00
	b	$p_{\alpha/b}$	$p_{\beta/b}$	$p_{\gamma/b}$	...	$p_{\kappa/b}$	1.00
	c	$p_{\alpha/c}$	$p_{\beta/c}$	$p_{\gamma/c}$	...	$p_{\kappa/c}$	1.00
	.						
	.						
	.						
	k	$p_{\alpha/k}$	$p_{\beta/k}$	$p_{\gamma/k}$		$p_{\kappa/k}$	1.00
	Pooled	p_{α}	p_{β}	p_{γ}	...	p_{κ}	1.00

We may regard $p_{y/a}$ as the probability that y will be received when a is transmitted. (If there is variable error in transmission, $p_{y/a}$ will not be 1.00 for any y.) Similarly, $p_{a/y}$ is the probability that this y belongs to a population of responses received when a is transmitted many times.

The response characteristics of the communication channel define a set of probabilities $p_{y/x}$. Table 1 shows a "transition matrix" giving these values. These constitute a statement of the frequency and character of errors associated with the test performance of persons falling in any criterion class. Further, if we are dealing with a particular population of persons, this population is described by a distribution of p_x over categories. When p_x is known, we can determine the probabilities p_y of various responses:

$$p_y = \sum_x p_{xy} = \sum_x p_x p_{y/x} \tag{2}$$

Further,

$$p_{x/y} = \frac{p_{xy}}{p_y} \tag{3}$$

These are the usual equations stating relations between joint and conditional probabilities.

It is well to note at this point that relations to be developed in this Section are reversible. While we speak of a communication channel as having a direction (transmitter-to-receiver), we do no mathematical violence if we reverse the direction. This makes little sense in physical communication problems, where we visualize a message passing along the channel as time progresses, but there are occasions when it is of interest to view a process in reverse. In testing, we shall at times be interested in inferring (predicting) an obtained score or a response from a true score. A suitable substitution of x for y in our formulas makes them appropriate for this reversed inference.

Now, applying (2) and (3) to (1), we obtain the useful computing formula

$$\text{Conf}\ (a, y_s) = \frac{p_a\, p_{y_s/a}}{p_{y_s}} \qquad (4)$$

Shannon points to the desirability of measuring the confidence (or uncertainty) of a receiver averaged over an entire set of inferences.

The standard item. Shannon introduces what we may call the "standard message" as a device for expressing degree of uncertainty. The unit or standard for measuring information is the so-called bit, and a standard message is one which conveys "one bit of information". In thinking of testing problems, we might think of the standard message as a "standard item". While the standard binary message is ordinarily described as a message which divides persons into two equal categories with no error, we shall employ the following, slightly more general, definition. A standard item is one for which $\frac{p_{x/y}}{p_x}$ is 2 or 0, for each x and y. That is to say, if a report from a standard item is added to whatever prior information we have, we can reject some hypotheses as having zero probability; and the probability of any other hypothesis being true is doubled over what it was when inference was made without the information given by the item. From (3), it follows that $\frac{p_{y/x}}{p_y}$ also is 2 or 0, for a standard item.

Table 2. Illustrative Transition Matrix for a Standard Item

(Cell entries show $p_{y/x}$)

Inputs (x)	Outputs (y)			
	1	2	3	
a	$2p_\alpha$	0	0	
b	$2p_\beta$	0	0	
•	•	•	•	
g	$2p_\gamma$	$2p_\gamma$	0	
h	$2p_\delta$	$2p_\delta$	0	
i	0	$2p_\varepsilon$	0	
•	•	•	•	
m	0	$2p_\zeta$	$2p_\zeta$	
n	0	$2p_\eta$	$2p_\eta$	
•	•	•	•	
q	0	0	$2p_\theta$	
•	•	•	•	
•	•	•	•	
k	0	0	0 ...	
	p_α	p_β	p_δ	1.00

In Table 2, a transition matrix for a standard item is presented, showing the probability that any y will arise, for a given x.

Standard Items Required to Reach Certainty

Shannon's measure of uncertainty can be shown to be the number of standard items per individual required to provide complete certainty regarding all individuals. In the general communication case, it is the average number of standard (binary) symbols that must be received to provide the receiver complete certainty as to the transmitted message.

Suppose a series of standard items is administered to a population of individuals, each item having the same type of transition matrix and each independent of the others. Independence must be defined by the following conditions (where 1 and 2 designate any two items). For any y and x,

$$\left.\begin{aligned} p_{y_1 y_2/x} &= p_{y_1/x} p_{y_2/x} \\ \text{and} \qquad p_{y_1 y_2} &= p_{y_1} p_{y_2} \end{aligned}\right\} \tag{5}$$

This is to say, the items are uncorrelated, and uncorrelated with x held constant. Each item measures a separate portion or aspect of the criterion, as in that sort of test where we seek to make correlation between item and criterion positive, and correlation between items zero.

For any person, the series of items generates a series of responses which we may call y_{t_s}. By definition (1), the confidence in classifying S as an x after t items ("t messages received") is

$$\text{Conf}\ (x, y_t) = p_{x/y_t} \tag{6}$$

$$p_{x/y_t} = \frac{p_x p_{y_t/x}}{p_{y_t}} \tag{7}$$

From (5),

$$\text{Conf}\ (x, y_t) = \frac{p_x p_{y_1/x} p_{y_2/x} \cdots}{p_{y_1} p_{y_2} \cdots} \tag{8}$$

The ratios $\frac{p_{y_1/x}}{p_{y_1}}$, etc. can be 2 or 0. Each response must have non-zero probability in order to have arisen from the given x. Therefore

$$\text{Conf}\ (x_s, y_{t_s}) = p_{x_s} (2)^t \tag{9}$$

This follows from the definition of a standard item.

Now our question is, what value of t is required, to make the receiver completely certain? When x = a, we may denote by t_a the number of standard items required to make Conf $(a, y_t) = 1$.

$$1 = p_a \cdot 2^{t_a} \tag{10}$$

$$0 = \log p_a + t_a \log 2 \tag{11}$$

As Shannon points out (26,p.4), the choice of base for the logarithm here is completely arbitrary. It does not affect conclusions from the theory, and may be regarded as a convention. However, if we choose base 2 for the logarithm, our equation simplifies to

$$t_a = - \log_2 p_a \tag{12}$$

If information is transmitted about one person at a time, we of course have no way of transmitting (say) 2.3 standard items. If $- \log p_a = 2.3$, we would have to transmit 3 items to classify a person as a with certainty. In general, the number of items required to classify an individual is the integer next larger than $- \log p_a$ (unless $- \log p_a$ is an integer). That is, when information is transmitted for just one person in any item,

$$- \log p_a \leq t_a < 1 - \log p_a.$$

In a sample where w_x is the proportion of persons in any x category, we may let $\bar{t}$ signify the average number of items required per person.

$$- \sum_x w_x \log p_x \leq \bar{t} \leq 1 - \sum_x w_x \log p_x \tag{13}$$

As N becomes very large the number of persons for whom x is the true classification approaches Np_x. Therefore,

$$- \Sigma p_x \log p_x \leq \lim_{N \to \infty} \bar{t} \leq 1 - \Sigma p_x \log p_x \tag{14}$$

It is of interest to note that, for any value of a, there is a limited number of items that can simultaneously satisfy the independence conditions (5). Not more than $- \log p_a$ such items can be found.

Interpretation as sequential analysis. Obtaining certainty by means of a test may be thought of as a problem in sequential analysis. Suppose items are administered to a person, one at a time, until an accurate inference can be made regarding his true category. We would continue testing any person until some desired level of confidence is reached, at which point we would

make a decision regarding him and proceed to the next person (14). He would probably have to administer more items to some persons than others. In quality control a testing procedure is evaluated by determining the average sample number; it is the average number of objects which must be tested in order to arrive at decisions about the population from which the objects are sampled. In the communication problem, $\bar{t}$ is the average sample number; i.e., the average number of standard messages required to reach a desired degree of confidence for members of a given ensemble of messages.

If we are willing to discontinue testing at some confidence level less than 1.00 (say C'), then the required number of items t'_a is a simple function of t_a.

$$C' = p_a \cdot 2^{t'_a} \tag{15}$$

$$t'_a = t_a + \log C' \tag{16}$$

(t', by this formula, may not be an integer). If C' is the same for all categories,

$$\bar{t}' = \bar{t} + \log C' \tag{17}$$

This demonstrates that an additive correction makes $\bar{t}$ a statement of the number of standard items required to classify persons at any desired confidence level. We might require different confidence levels for different categories, larger C' attending the more important categories where we want to minimize risk of error. If each x has some C'_x,

$$\bar{t}' = \bar{t} + \Sigma p_x \log C'_x \tag{18}$$

Interpretation in terms of coding. Shannon uses information theory to develop theorems regarding coding, and it is in that context that his formulas have particular relevance. Any message to be transmitted is encoded to fit a particular channel. Only a one-to-one encoding is involved in the typewriter, where the "channel" has one symbol for each letter. In other channels, such as telegraph, telephone, or teletypewriter, the transmitted symbol is encoded into dots-and-dashes, vibrations, or other new symbols. In order to measure the transmission capacity and rate of systems using various symbol systems, Shannon evaluates each one by determining how each one compares with a noise-free binary transmission system.

A binary system has two states: "on-off", "push-pull", etc. The two states may be symbolized by 1 and 0. By shifting from one state to another, the binary system transmits messages of the form 101, or 00011, or 1101. The electronic computer uses a binary system of this type. If, a priori, 1 and 0 are equally likely at a given instant, the transmission of either symbol without error ("noise-free"case) doubles the confidence of the receiver. A standard message (i.e., one bit) has been transmitted.

Any complex message can be encoded in binary form. We may agree, for instance, that "red-even" will be transmitted as "11," or that the letter "g" will be represented by 11001. Any set of sixteen equally likely alternatives can be encoded perfectly into four binary symbols (e.g., 1101). In general, 2^t equally probable messages can be encoded into messages of t binary digits.

If the transmissions are not equally likely, it is economical to use shorter codes for the most common messages. Thus, consider how we might encode four alternatives, knowing their probabilities of occurrence.

Alternative	p_x	Code	No. of digits	Weighted digits
A	.50	0	1	.50
B	.25	10	2	.50
C	.125	110	3	.375
D	.125	111	3	.375
			Weighted avg. digits per letter	1.75

The message ACB would be encoded 011010. Each possible code has a unique interpretation. The transmission 1100100101110 is equivalent to CABABDA. Thus, in thirteen standard messages we convey seven letters, using an average of 1.86 bits per letter. If we had coded each alternative in a two-digit pattern, our average would be 2.0. By using fewer digits for A than for C and D, we have a relatively efficient code. Over a longer series of transmissions in which each letter appeared in the specified proportions p_A, etc., the average number of bits required per letter would drop to exactly $-\Sigma\, p_x \log p_x$ (1.75).

Coding of one letter at a time will not be so efficient as this if the p_x are not integral powers of 0.5 as in the foregoing example. It is possible, however, to encode patterns of letters, and this permits economy of message space. For example, compare single-letter and two-letter-pattern coding, when the p_x for two alternatives are .64 and .36.

Alternatives	p	Best single letter code	Digits	Weighted Digits
A	.64	1	1	.64
B	.36	0	1	.36
		Weighted avg. digits per letter		1.00

Suppose A and B are independent, so that the sequence AB has probability $p_A p_B$, etc. Then we may consider sequences of <u>two</u> letters.

Alternative sequence	p	Best code	Digits	Weighted digits
AA	.41	0	1	.41
AB	.23	10	2	.46
[illegible]	.23	110	3	.69
BB	.13	111	3	.39
			Sum	1.95
		Digits per pattern transmitted		1.95
		Digits per letter transmitted		.975

Coding of two letter sequences permits a saving of 2.5% in average message length. There is still some inefficiency, however. Reporting by the first digit that the sequence is AA or not AA conveys less than one bit of information because $p_{AA} \neq .50$. By building longer sequences, we can arrive at one which is equal to a power of 0.5, to any desired degree of approximation.

Shannon demonstrates that if indefinitely long sequences may be encoded, it is possible to transmit messages in exactly - $\Sigma p \log p$ standard items per symbol, on the average. This minimum number of items is his measure of uncertainty, in terms of message space required to produce certainty. He calls this H_x.

$$H_x = - \Sigma \; p_x \log p_x \tag{19}$$

For the probabilities given above, H_x is .943 bits per letter.

The testing problem is not truly comparable to Shannon's encoding problem for which he uses (19) and derivative formulas. Given a person who belongs in a category having probability .25, it is theoretically possible to find two independent items, each dividing the group in half, so that their combined information identifies the category the person belongs to. We have essentially "encoded" the information in terms of responses to two items. Actual test items contain error, but this can be taken into account.

What is not reasonable in testing is to think of encoding two or more persons, i.e., a sequence of persons; it is not possible to "send two or more persons through the channel" at once. Since such encoding of patterns is not possible, formula (19) is only approximately a statement of the number of items required to classify a person on the average. Formula (14) holds precisely, but is not manageable. We shall employ (19) as a basis for subsequent formulas, therefore, with the warning that these results are interpretable only approximately. Since $H + 1 > \bar{t} > H$, results from subsequent formulas may be in error by as much as one bit. No exact statement may be made regarding the error introduced into our ratio formulas by this approximation. The failure of Shannon's formula to have an exact meaning for the testing problem means that his theorems regarding capacity and encoding cannot be regarded as necessarily true or meaningful in psychometrics.

Interpretation as log confidence. Another interpretation of Shannon's measure results from considering a long sequence of transmitted symbols. Any series of transmissions may be regarded as one single message. This sequence has a particular frequency of occurrence. If the elements in the message are independent as defined in (5), we can compute this probability easily.

Let the sequence v contain N_i elements of type i.* (In testing, we would speak of a sample containing N_i persons in each category.) Then if p_v is the probability of the sequence v occurring,

$$p_v = \Pi (p_i)^{N_i} \tag{20}$$

Obviously, p_v is our confidence, in the absence of received information, that a particular sequence will be v. That is, p_v is our a priori confidence in the hypothesis x = v. Then

$$\log p_v = \Sigma N_i \log p_i = N \Sigma w_i \log p_i \tag{21}$$

As N becomes very large, $w_i \to p_i$, and under the requirement of independence,

$$\log p_v \to N \Sigma p_i \log p_i \tag{22}$$

$$H_i = - \frac{1}{N} \log_2 p_v \tag{23}$$

As $N \to \infty$ the probability of all sequences becomes the same and p_v for any sequence approaches 2^{-NH}, where H is defined by (19). p_v is our a priori confidence that we can infer the sequence, and -NH is log p_v. Hence, from (12), if we can encode a whole sequence at once, NH tells us the number of standard items required to convey the sequence. 2^{-NH} expresses the a priori confidence. This interpretation is not especially useful for testing, however, because we cannot "encode" sequences of persons.

Finally, we may note that

$$p_v = \Pi p_i^{N_i} \tag{24}$$

and from (23)
$$H_i = - \log \left(\Pi p_i^{N_i} \right)^{1/N} \tag{25}$$

That is to say, - H is the log geometric mean of the p_i.

Cautions in employing H to measure uncertainty. The custom has arisen of referring to H as a measure of uncertainty. We can indeed interpret H or $\bar{t}$ as statements of the amount of information lacking. In these formulas, however, that information is expressed in terms of the number of standard items required to attain perfect confidence. Thus H and $\bar{t}$ are expressed on a scale of message space.

In evaluating a test, we might consider applying formulas based on H or $\bar{t}$ to assess uncertainty after testing. This measure tells us how much message space is required to eliminate residual uncertainty under certain conditions.

---

*The symbol i is used for the separate elementary inputs which constitute the complex input v.

The formula for $\bar{t}$ assumes that the test is being given sequentially, each person being given just the number of items needed to classify him at the desired confidence level. The formula for H assumes, in Shannon's words, that delay at the transmitter is possible; i.e., that many inputs (persons) can be sent through the channel at once. This has no meaning for the tester. While it is probably true that a test which reduces uncertainty as measured by $\bar{t}$ will also show similar effectiveness (relative to other tests) when evaluated by any other formula, we can give no exact interpretation to the formulas of this section except in terms of sequential testing.

Since the scale of message space is logarithmically related to the scale of confidence, we must ask which scale has "equal units." In terms of the gain from giving the test, the confidence scale is more directly interpretable than the log confidence (t) scale. Suppose we are hiring men who may later be judged successful (S) or unsuccessful (U) and for the sake of simplicity assume that every S man hired is worth $1,000 to the employer, and every U is worth $0. Assume further that among applicants $p_S = .125$. A priori confidence is .125 (H = 3) and in N decisions the company gains $125N. Now Test 1 selects applicants 25% of whom are S. Confidence at this stage is .25, H = 2; and the value of the men hired is $250N. Test 2, given in place of Test 1, identifies men 50% of whom will be S. Then confidence becomes .50, H = 1, and the value of decisions is $500N. The gain in "message space" is one bit from Test 1, and two bits from Test 2; but the dollar gain is $125 from Test 1, $375 from Test 2. It is clear that Test 2 is more than twice as useful to the tester as Test 1.

Because the H scale is not linearly related to practical gains, we shall turn in a later section to utility measures. In the remainder of the present section, we present developments we have arrived at through the Shannon model. The concepts involved are thought-provoking. The formulas themselves are not the most suitable basis for evaluating a test, and are ultimately to be replaced by utility measures representing the same concepts.

Information as the Reduction of Uncertainty

We conceive of a continuum from great uncertainty to complete certainty. At any time our degree of uncertainty can be located on this continuum in terms of H or some other uncertainty measure. H_x represents a priori uncertainty as we begin testing. After any stage of testing we have a residual uncertainty which is equal to or greater than zero.

Measure of Residual Uncertainty after Testing

We may assess this residual uncertainty by the same logic used to define H_x. After an item or series of items has been administered, each person is placed in one of many possible y categories, depending on his response (or configuration of responses).

The confidence with which we can classify a person whose response y is known has already been defined by equation (1) as $p_{x/y}$. Now what is the average number of standard items $H_{x/\alpha}$ required to place persons falling in the y category α into the proper x category with complete certainty? We can insert conditional probabilities in equation (19).

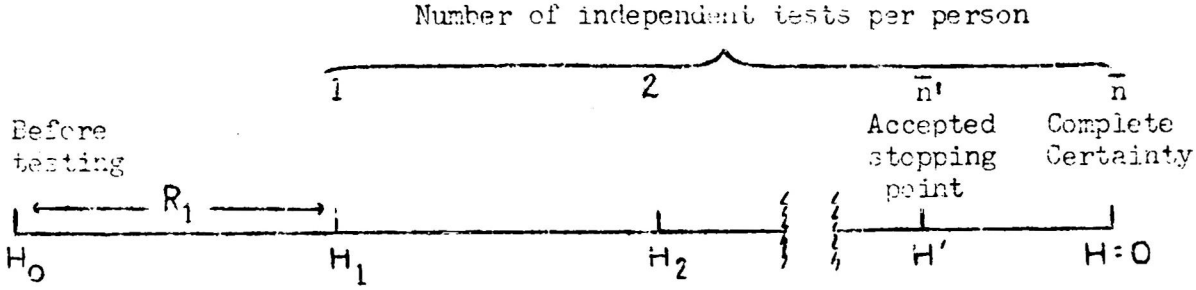


Figure 2. The confidence continuum viewed in terms of message space

$$H_{x/a} = - \sum_x p_{x/a} \log p_{x/a} \tag{26}$$

If $H_{x/y}$ represents the number of items required, averaged over all response categories,

$$H_{x/y} = - \sum_y \sum_x p_y \, p_{x/y} \log p_{x/y} \tag{27}$$

This is the residual uncertainty after administering whatever tests yielded responses y. Shannon calls this "equivocation," denoting it by $H_y(x)$. It is analogous to the standard error of estimate of conventional test theory.

We will find it useful hereafter, to refer to the uncertainty at various stages of testing as H_0, H_1,.... We shall ordinarily use H_0 for the uncertainty before testing, and H_1 for the uncertainty after administering a particular test. Subscripts can be defined in the context of any discussion.

The Confidence Continuum and the Measure R

The test moves us along the continuum of "information needed" from the point H_0 to the point H_1 (see Figure 2). Before testing, we required H_0 standard items per person to become certain; after testing, we need only H_1 such items. It is obvious that the test, then, gave us information equivalent in message space to $H_0 - H_1$ standard items. It is this difference which Shannon calls R, the "rate of transmission of information."

$$R = H_0 - H_1 = - \sum_x p_x \log p_x - (- \sum_y \sum_x p_y \, p_{x/y} \log p_{x/y}) \tag{28}$$

Shannon demonstrates these useful identities:

$$R = - \Sigma \, p_y \log p_y - (- \sum_x \sum_y p_x \, p_{y/x} \log p_{y/x}) \tag{29}$$

$$R = \Sigma\Sigma \, p_{xy} \log p_{xy} - \Sigma \, p_y \log p_y - \Sigma \, p_x \log p_x \tag{30}$$

A mnemonic device for these relations is Miller's logical diagram reproduced in Figure 3. In the notation of the diagram, equations (28) - (30) become

$$R = H_x - H_{x/y} = H_y - H_{y/x} = - H_{xy} + H_x + H_y \tag{31}$$

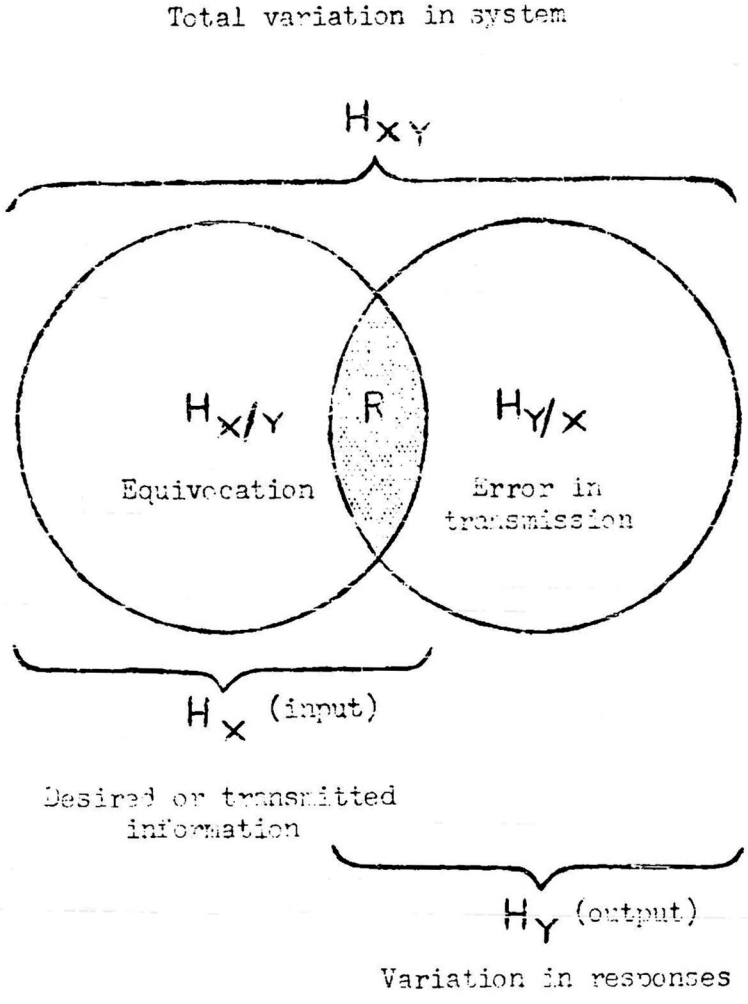


Figure 3. Relation between various sorts of variation or uncertainty (after Miller)

R represents the change in uncertainty as a result of testing, expressing the difference in terms of the number of standard items required after testing, compared to the number needed before testing. Thus, in a sense, R states how many standard items our test is equivalent to.

Exhaustiveness and Dependability

While R is a measure of the discriminating power of the item (or test) we may also be interested in the question, "How many independent items like this would be required to classify all persons with the desired degree of confidence?" If we have a series of items having the same transition matrix, but independent as specified by (5), then the average number of items of this type required for certainty is $\bar{n}$, where

$$\bar{n} = \frac{H_0}{H_0 - H_1} \tag{32}$$

This index is the "average sample number" expressed in terms of independent items all having the same R as the item under consideration. If we are willing to discontinue testing when certainty reaches C'

$$\bar{n}' = \frac{H_0 - H'}{H_0 - H_1} \tag{33}$$

where $H' = -\log C'$. This is a more general version of (32).

The Index of Exhaustiveness

The reciprocal of $\bar{n}'$ is a measure of exhaustiveness, which we shall designate $J_{x/y}$.

$$J_{x/y} = \frac{R}{H_0 - H'} = \frac{H_0 - H_1}{H_0 - H'} = \frac{\text{Information obtained}}{\text{Information desired}} \tag{34}$$

Here the "information desired" consists of the x_s classifications at some specified level of confidence. As before, "information" is measured in terms of message space. This index of exhaustiveness, ordinarily with H' as zero, has been used by some followers of Shannon as a measure of fidelity of transmission (e.g., 23).

The Index of Dependability

Whereas $J_{x/y}$ expresses the extent to which the information reported in y permits determination of x, it is possible to reverse the process. We might write a comparable ratio to specify the degree to which x determines y. We shall use the symbol K for this ratio, although it would be equally logical to use $J_{y/x}$.

The distribution of obtained data ("output") constitutes a series of reported individual differences. Some of these statements are error, introduced by "noise" in the channel. In a noise-free channel, all the individual differences in output would be determined by the transmitted signal, i.e., by the criterion information. A statement of the extent to which the output information is relevant to the criterion, then, is obtained if we measure the amount of information in the responses y, and then find out what fraction of that information is R.

The output uncertainty H_y is defined just as is H_x, in terms of the number of standard messages required to convey this information.

$$H_y = - \sum_y p_y \log p_y \qquad (35)$$

We may think of H_y as representing the amount of differentiation the test claims to make, and R as the amount that is criterion-relevant. Perhaps "relevance" (cf. Cureton, 12, p. 624) is a better term for the ratio of these than "dependability," but for the present we will employ the latter term as in our earlier report. The extent to which output variation is determined by the input is expressed in the dependability ratio K:

$$K = \frac{R}{H_y} = \frac{\text{Information obtained}}{\text{Information required to specify output}} \qquad (36)$$

As in (34), we could introduce H' into the denominator of (36) to allow for any desired degree of specification of the test score.

On the basis of our earlier report, Quastler developed the diagram reproduced here as Figure 4. In Quastler's notation, cr represents the criterion x, and te represents the test output y. I is a substitute for R, and the other changes in notation should be obvious. It is evident from this diagram that R tells how much the test and criterion have in common (as measured in terms of standard units). J and K, respectively, tell what relation R bears to the total desired information H_x and to the total reported information H_y.

Significance of the Formulas

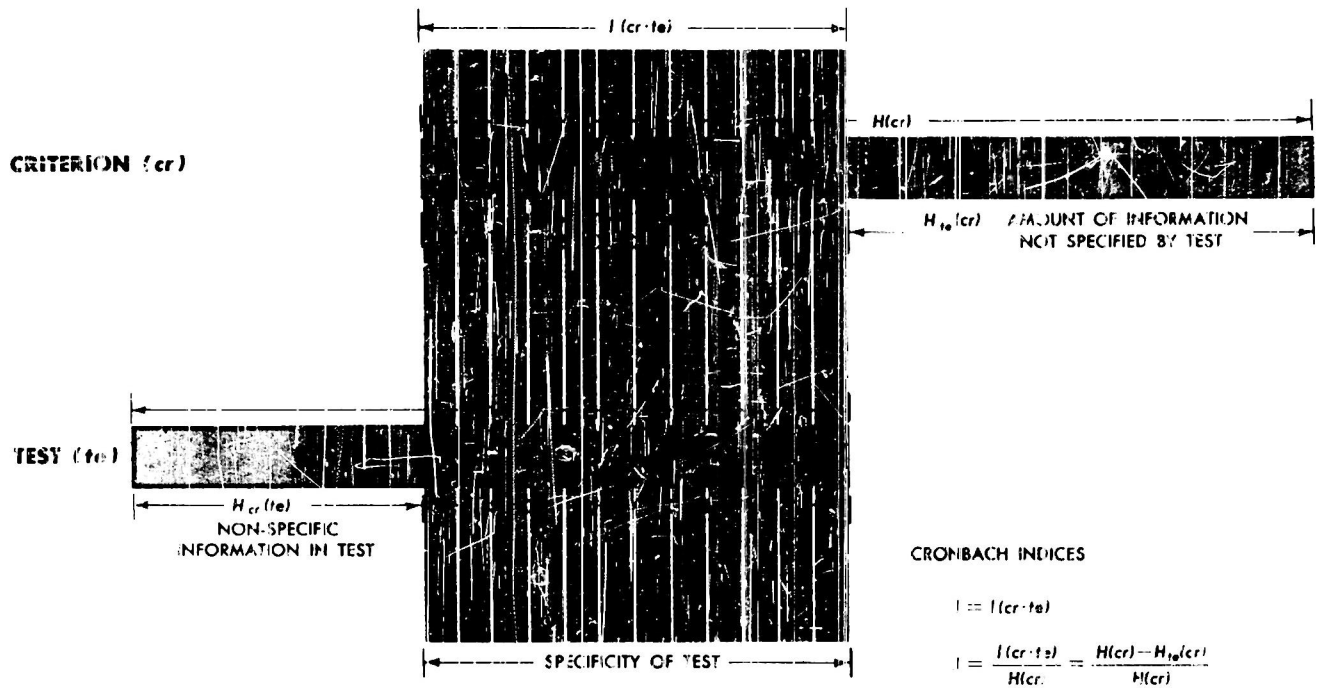
Identifying three aspects of the information-carrying capacity of a test in itself implies that it is insufficient to evaluate the test by a single index such as R. It may at times be far preferable to have a test with high K -- for example, a pathognomonic sign which makes highly trustworthy discriminations -- even though the test does not answer many questions that interest us and so leaves us with considerable residual uncertainty. At other times, we might be content to have a very low degree of dependability, provided that by sifting a large amount of this undependable information we could predict the criterion more exhaustively. An example is observation of a pupil's oral report to a class; this provides cues regarding his tensions, postural habits, interests, articulation habits, knowledge, grammatical ability, and so on. The cues are unreliable, but nonetheless useful as a general survey of the pupil.

Figure 5 shows the various patterns which can occur. When the test and criterion are defined so that the number of categories in each is about the same, the diagrams in the top row are likely to apply. Whether $H_x = H_y$, of course, depends on the frequencies in the categories. The top diagram shows conditions found in reliability studies, or in comparisons of predicted classification to actual classification. The second row represents the case where the criterion divides people finely into many categories. This occurs,

---

Figure 4 (next page). Exhaustiveness and dependability. Reproduced by permission from (24).

SCALE (Bits)

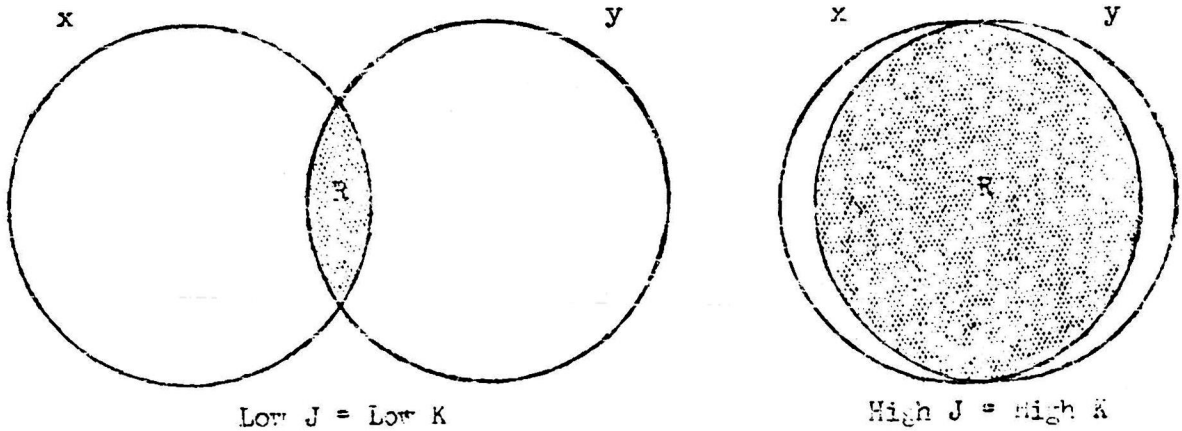


CRONBACH INDICES

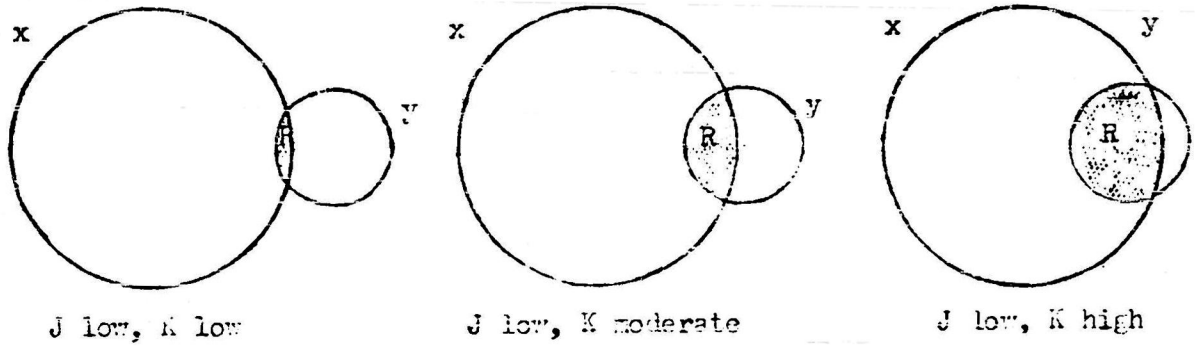
$I = I(cr \cdot te)$

$I = \frac{I(cr \cdot te)}{H(cr)} = \frac{H(cr) - H_{te}(cr)}{H(cr)}$

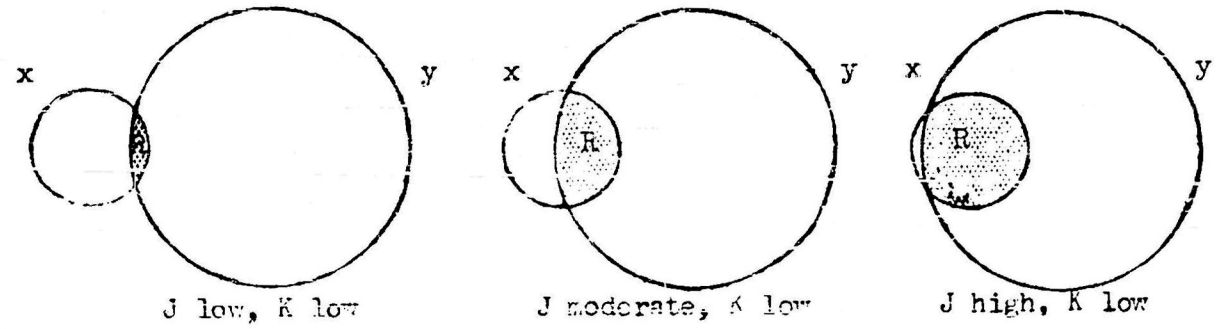
$K = \frac{H(te) - H_{cr}(te)}{H(te)} = \frac{I(cr \cdot te)}{H(te)}$



$H_x = H_y$; Test and criterion differentiate equally finely



$H_x > H_y$; Criterion finely differentiated, test coarsely differentiated



$H_x < H_y$; Criterion coarsely differentiated, test finely differentiated

Figure 5. Possible relations among H_x, H_y, R, J and K

for instance, where we desire precise measurements. It also occurs if the criterion to be described is a complex configuration, many such configurations being possible. The third row represents the relatively simple criterion combined with complex (finely divided or multidimensional) test information.

Under some circumstances, having a given H_x, it is wise to seek a test which will have a similar H_y; but for other testing problems we might be wiser to make H_y larger or smaller than H_x. In general, dependability is more to be valued in making final, irreversible judgments. If judgments are tentative and it is practical to reverse them on the basis of later evidence, dependability can be sacrificed for exhaustiveness.

Responses to a typical test will have a very high uncertainty H_y. If there are n items, each with c alternatives, then there are c^n possible configurations of responses. The uncertainty is reduced if we score all items as right or wrong, for there are then only 2^n configurations. Adding scores into a total reduces the number of response categories still further because variations in pattern are ignored. Each such successive reduction reduces H_y. When H_y is reduced, either R or $H_{y/x}$ is reduced, or both. That is, the variation now being ignored may be criterion relevant or may be irrelevant. Such scoring procedures as those mentioned above ordinarily increase dependability, since it is unlikely that R is reduced more rapidly than $H_{y/x}$. But so long as <u>any</u> of the variation eliminated is criterion relevant, both R and J are reduced, perhaps to an important extent.

An increasing interest is being shown in methods of drawing inferences from test data which use more information per item. According to the foregoing argument, any such more complex analysis should permit an increase in validity. Perhaps the gain in validity may be too small to have practical value, but some gain may be expected. Among the devices which promise to use more of the information in the test responses are

(a) Considering <u>which</u> wrong alternative a person selects, when he makes an error on an ability test

(b) Considering configurations of responses to various items

(c) Asking the person to mark more than one alternative per item, as in the Troyer-Angell self-scorer or Coombs' recent proposal of directions to "mark all wrong answers" (1, 9).

<u>Non-symmetric validity relations.</u> The introduction of J and K suggests the importance of regarding validity relationships as possibly non-symmetrical. A test which permits accurate prediction in one direction (y to x) may be quite inaccurate in predicting in the reverse direction (x to y). Suppose we have two types of classification, x and y, neither of which is necessarily regarded as the criterion. For example, we might wish to know whether a person's placement in certain interest categories predicts his personality structure; or whether the information about his personality predicts his interest classification. Either direction of interest is legitimate. The joint distribution might be of the following form, where the cell entry represents joint probability p_{xy}.

		Personality structure (x)				
		A	B	C	D	E
Interest Category (y)	a	O	O	x	x	x
	b	O	x	O	O	O
	c	x	O	O	O	O

In this relationship, we can predict interest classification with certainty from the personality data, but the reverse inference cannot be made as confidently. Interest information is determined by the personality data ($J_{y/x}$ is 1.00) but the interests do not specify the personality ($J_{x/y} < 1.00$). Measures such as r and χ^2 treat both directions of inference symmetrically. In contrast, non-symmetric measures like J and K -- or the curvilinear correlation eta for ordered categories -- treat the inferences separately.

The following discussion, repeated from our preliminary report, seems an important consequence of the above argument. Perhaps most attempts at qualitative assessment of performance or personality run afoul of non-symmetric relations. Suppose a given test (or set of tests) offers c possible configurations and there are C possible criterion configurations. When C > c, prediction from test to criterion will be more hazardous than inferring the test pattern from the criterion (i.e., postdiction). Now it might be contended that a complex test affords a practically infinite number of possible response patterns and that the number of criterion structures could not be greater. But each personality structure might interact with any of an infinite variety of situations to produce the criterion behavior, unless we can specify the criterion situation in advance. Inferring personality structure from behavior will probably always be more secure ("exhaustive") than predicting behavior from structure.

An important proposal for test development follows from this. When a test is intended for a specific selection task where there are only a few clearly defined criterion categories (i.e., success or failure to attain a proficiency standard in learning to type), accuracy of prediction from test to criterion should be the first concern of the investigator. Since he is likely to use multiple tests, H_y will exceed H_x and it will probably be easier to infer from test to criterion than the reverse. This is obvious with relation to the school-grade criterion, where it is easier to predict that a student will fail, knowing test scores and the curriculum to be studied, than to tell what requisite he was weak in, knowing only that he failed.

On the contrary, when the criterion is highly complex, as in the case where one can achieve "success" in a variety of ways, or where the situation in which success is demonstrated varies, it will generally be easier to "postdict" from performance to the test. Until preliminary research has developed a sound basis for making postdictions, it is hopeless to try to use the test predictively. If the complex criterion behaviors are pooled (perhaps by some rater) into a single index of success, unless we know the basis on which the complex behavior is implicitly evaluated and weighted, it is hopeless to try to predict the simplified criterion.

To determine which tests are promising for further study and development, postdiction research should often be the first attempt at validation. This applies especially to personality tests. To implement this suggestion requires an adequately complex criterion which provides information about the situational pressures and demands under which the person works. Then, unless enough cases are available for likelihood-ratio analysis or the like, the assessor would make "postdictions" using his theory regarding the tested behavior. He might report "In view of the criterion behavior of this person I expect him to show good performance on test A but poor performance on test B." Such inferences can be validated against the test record. Interest of course centers on $J_{test/criterion}$. This type of inference will be difficult, but far less difficult than the task typically posed to assessors in studies like that of Kelly and Fiske (20). Their validity is judged by $J_{criterion/test}$, even though the criterion is derived in unknown ways from a very complex situation such as the performance of a clinician in internship.

Naturally, when the end-goal of test development is prediction, one does not cease research when a test has shown postdictive validity. Our proposal is only that research should employ hypotheses wherever possible that have a good chance to be verified, and that in many fields where prediction is difficult because $H_x > H_y$, formal postdiction studies would be the appropriate first line of attack.

Relations Involving a Fallible Criterion

We may consider a more complex case which shows how error of measurement may be taken into account in assessing validity. Quite often, the available transition matrix of $p_{y/x}$ is not based upon the desired or criterion information. Instead, a fallible criterion x is used, and we desire to predict some true criterion χ. x is viewed as generated from χ. It may be possible to estimate (or hypothesize) some rate of error in generating x from χ. Then it should be possible to investigate how much information y gives about χ, and what fraction of the desired information is obtained.

The problems of inference involving three variables would require an extensive digression. The interested reader should consult McGill (21). We shall sketch very superficially the concepts for this analysis. The correction for criterion fallibility is a fairly simple case of multivariate information analysis.

If we adapt the Miller diagram to three variates we can sketch Figure 6. A large number of identities can be constructed by the reader. The diagram is interpreted like Figure 3. $R_{\chi xy}$ is the common information in all the variables. $R_{xy} - R_{\chi xy}$ is the information common to x and y but irrelevant to χ. And so on. The reader should not assume that the circles have any specified relative size, or that the areas shown are proportional to the various H and R.

In the case we are presently interested in, we may assume that the error in the fallible criterion ($H_{x/\chi}$) is independent of the error in predicting the test response from the true criterion ($H_{y/\chi}$). Let us assume further that y and x do not overlap in any way independent of χ (i.e., that they contain no common factor irrelevant to χ). This condition specifies that $R_{xy} = R_{\chi xy}$.

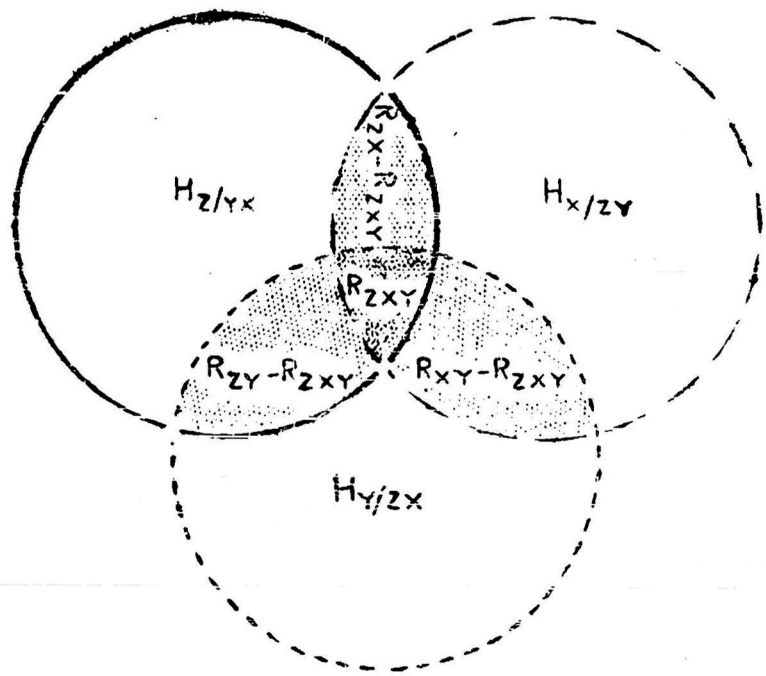


Figure 6. Information elements in the three variate case.

We also require that $R_{\chi y} = R_{\chi xy}$; this states that error $H_{\chi\tilde{x}}$ is random. Then Figure 6 takes the form shown in Figure 7, the circles of Figure 6 being altered in shape to conform to the new conditions. We can read off various identities (which would otherwise be derived from definitions involving $P_{x/\chi}$, $P_{y/x}$, etc., and our assumptions).

The desired information is H_χ. But the desired information in x is only $R_{x\chi}$. If x is fallible, our exhaustiveness is not represented by $R_{xy} / H_x - H'$. Instead we may be interested in either of two questions:

(1) What fraction of the desired information (measured in message space) about the true criterion did we obtain? This is answered by

$$J_{\chi/y} = \frac{R_{\chi xy}}{H_\chi - H'} \tag{37}$$

If, contrary to assumption, y and χ should contain some common factor not in x, $R_{\chi y} > R_{\chi xy}$; but if data exist to measure $R_{\chi y}$ directly, multivariate treatment is not called for.

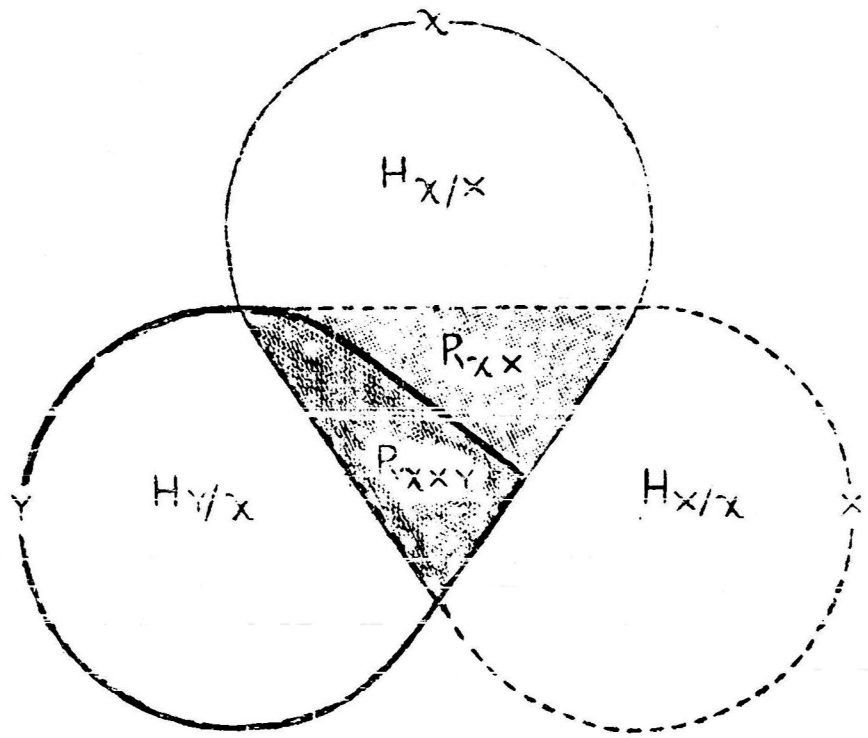


Figure 7. Relation of true criterion χ, fallible criterion x, and test y.

(2) What amount of relevant information did we obtain relative to the amount of desired information in the fallible criterion? Let this exhaustiveness index be $J_{x_\infty/y}$.

$$J_{x_\infty/y} = \frac{R_{xy}}{H_x - H_{x/\chi} - H'} = \frac{R_{xy}}{R_{\chi x} - H'} \qquad (38)$$

Formula (38) is closely comparable to the concept of "efficiency" employed by Fisher to evaluate the amount of information elicited by a statistical procedure. Fisher regards the procedure as a device ("communication channel," we might say) for gaining information about parameters of a population distribution, employing sample data. He defines intrinsic accuracy as equivalent to "the amount of information in a single observation belonging to such a distribution" (16, p. 709). This clearly resembles the "average information per item," R_{xy}. "The efficiency of a statistic is the ratio of the intrinsic accuracy of its random sampling distribution to the amount of information in the data from which it has been derived" (16, p. 714). This

latter phrase describes a concept very close to R_{zx}, the amount of relevant information in a fallible criterion where all error is random, and the efficiency ratio is comparable to $J_{x_\infty/y}$.

Formula (38) is a correction for criterion attenuation. It is thoroughly comparable to the formulas customarily used in correlational work. Similar formulas can be worked out to correct for test unreliability. It has been noted by Garner and Hake (18) that K has some of the properties of a contingency coefficient. Our statements in a preceding paragraph show that we may regard J and K as comparable to eta, in that (unlike the contingency coefficient) eta is directional. If we regard $J_{x/y}$ and $K_{x/y}$ ($J_{y/x}$) as suitable measures of relationship, then we have in (38) a way of correcting them by comparing them to their maximum values.

Since we question the use in test analysis of measures based on H, these comments lead primarily to the suggestion that other measures of relationship for categorical data can be corrected for attenuation in a similar manner. Correction for attenuation is used to judge whether a given relationship is high relative to what a fallible criterion permits. It is used to judge whether an unreliable test is sufficiently saturated with valid variance to justify lengthening the test to obtain greater reliability. While the assumptions of randomness of error are often untenable, and although the sampling error of corrected coefficients is often very large, nearly all interpretation of correlations involves at least a crude application of attenuation formulas, and such formulas for categorical data would indeed be useful.

Application of the Formulas to Ordered Scales

While we have so far discussed the measures of uncertainty and information in terms of a distribution of persons into a set of unordered categories, Shannon demonstrates that they can be written in different form when categories are ordered and there is a known frequency distribution on some underlying measurement scale. In particular, the rectangular and normal distributions are of interest.

The Rectangular Distribution

Suppose that x is a continuous variate having a rectangular distribution in the population under study. Then in any score interval Δx, p_x is a constant equal to Δx divided by the range. We shall denote the range by 2m; $p_x = \frac{\Delta x}{2m}$.

$$H_x = - \sum_x p_x \log p_x = -\sum_x p_x \log \Delta x + \sum_x p_x \log 2m \qquad (39)$$

$$H_x = - \log \Delta x + \log 2m \qquad (40)$$

By dividing the scale into smaller and smaller units, we increase the differentiation in the criterion scale, i.e., the information desired. As Δx is allowed to become infinitely small, H_x becomes infinitely large.

Instead of inquiring how many standard items are required to provide infinitely fine differentiation, we may set some particular degree of resolution as desirable. That is, we may say that we wish to differentiate persons with certainty into intervals of size u, and that differentiation beyond that fineness does not interest us. This is equivalent to saying that we will accept a residual uncertainty H" where, since u is the a posteriori range,

$$H'' = - \log \Delta x + \log u \tag{41}$$

H" is conceptually like H', but is specified in a different way. Therefore, the information desired is

$$H_x - H'' = \log 2m - \log u \tag{42}$$

It will be noted that if the range 2m is expressed as a multiple of u -- i.e., if u is taken as the unit of measurement -- the desired information becomes log 2m. In general, whether categories are ordered or not, if they are equally probable the uncertainty is the logarithm of the number of categories.

The Normal Distribution

If the distribution of x is normal,

$$f(x) = \frac{1}{2\pi\sigma} \; e^{-(x^2/2\sigma^2)} \tag{43}$$

In any differential element whose midpoint is x_i and whose width is Δx, the proportion p_i is $f(x_i)\Delta x$.

$$-\log p_{x_i} = \log 2\pi\sigma + \frac{x_i^2 \log_2 e}{2\sigma^2} - \log \Delta x \tag{44}$$

$$H_x = \sum_{-\infty}^{\infty} - p_{x_i} \log p_{x_i} = \Sigma p_{x_i} \log 2\pi\sigma + \Sigma p_{x_i} \frac{x_i^2 \log_2 e}{2\sigma^2} - \Sigma p_{x_i} \log \Delta x \tag{45}$$

$$H_x = \log 2\pi\sigma + \frac{\log_2 e}{2\sigma^2} \Sigma p_{x_i} x_i^2 - \log \Delta x \tag{46}$$

But $\Sigma p_x x^2$ is the average x^2, and as Δx becomes very small this approaches σ_x^2.

$$\lim_{\Delta x \to 0} H_x = \log 2\pi\sigma + \frac{1}{2} \log_2 e - \log \Delta x \tag{47}$$

$$\text{Lim} \; H_x = \log \sqrt{2\pi e}\sigma - \log \Delta x \tag{48}$$

This equation holds if the differential elements are sufficiently small that f(x) as given by the normal curve is very close to f(x) given by the discontinuous curve based on differential elements.

If we require differentiation between elements of width u, but do not require differentiation into infinitesimal elements, this can be taken into account. Where u is sufficiently small that we may regard the distribution as rectangular within the interval, the desired information is

$$H_x - H'' = \log \sqrt{2\pi e}\,\sigma - \log u \qquad (49)$$

and if σ is expressed in the units u,

$$H_x = \log \sqrt{2\pi e}\,\sigma = \log \sqrt{2\pi e} + \log \sigma \qquad (50)$$

It is of interest to inquire how sensitive (50) is to distribution shape. In a rectangular distribution of range 2mu, $\sigma = mu/\sqrt{3}$. Then from (42)

$$H = \log 2m$$

But (50) would estimate

$$H = \log \sqrt{\frac{2\pi e}{3}}\, m = \log 2m + \log \sqrt{\frac{\pi e}{6}}$$

$$= \log 2m + \log 1.19$$

$$= \log 2m + .25$$

Hence, as distributions become more platykurtic than the normal, formula (50) overestimates H by an amount not exceeding .25. This error is independent of m. (50) may be applied to non-normal distributions with greatest confidence when σ is large relative to u.

R Under Assumptions of Normality

If we have the ordered quasi-normal variates x and y, we may set up the usual bi-variate distribution or scatter diagram (Fig. 8). The error (within rows) has a variance $\sigma^2_{y.x}$ which may be different for different values of x. Likewise, the equivocation (within columns) has a variance $\sigma^2_{x.y}$ which need not be uniform over columns. The sigmas are the usual standard errors of estimate of test from criterion, and criterion from test, respectively.

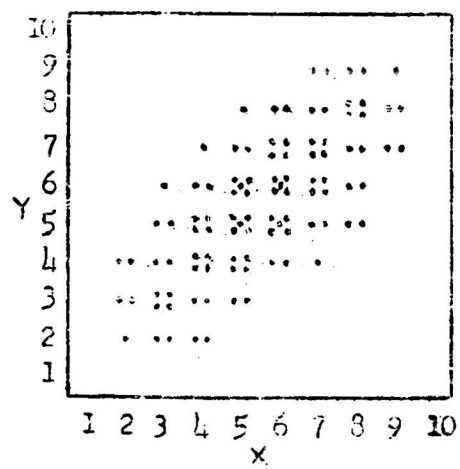


Figure 8. Bivariate correlation surface.

Assuming normal distribution within arrays,

$$H_{x/y} = \log \sqrt{2\pi e} + \int p_y \log \sigma_{x.y}\, dy \tag{51}$$

$$H_{y/x} = \log \sqrt{2\pi e} + \int p_x \log \sigma_{y.x}\, dx \tag{52}$$

From (28),

$$R = - p_x \log p_x\, dx - \log \sqrt{2\pi e} - p_y \log \sigma_{x.y}\, dy \tag{53}$$

If the distribution of x and y are normal, and the equivocation uniform over arrays,

$$R = \log \sigma_x - \log \sigma_{x.y} = \log \frac{\sigma_x}{\sigma_{x.y}} \tag{54}$$

or

$$R = \log \frac{\sigma_y}{\sigma_{y.x}} \tag{55}$$

If we employ variance instead, using V_E for error in transmitting y, and V_I for error in inference or estimation of x from y,

$$R = \frac{1}{2} \log \frac{V_x}{V_I} = \frac{1}{2} \log \frac{V_y}{V_E} \tag{56}$$

By the usual definition of correlation (r),

$$r_{xy}^2 = 1 - \frac{V_E}{V_y} \text{ or } 1 - \frac{V_I}{V_x} \tag{57}$$

$$R = - \frac{1}{2} \log (1 - r^2) = - \log \sqrt{1 - r^2} \tag{58}$$

The radical in (58) is the familiar coefficient of alienation. Thus we find that information measure is closely related to conventional test theory, provided we have a normal bi-variate distribution of test against criterion.

The rate of information which one test yields regarding an equivalent test, and regarding the underlying true score, may be considered by referring to Figure 9. We now let χ represent the true score, and x and y be fallible measures of it. x and y are independent in the sense that errors $H_{x/\chi}$ and $H_{y/\chi}$ are independent. Figure 9 describes the reliability relations. To make x and y equivalent, assume $H_x = H_y$ and $R_{\chi x} = R_{\chi y}$. Then it can be shown that

$$R_{xy} = R_{\chi xy} = - \log \sqrt{1 - r_{xy}^2} \tag{59}$$

$$R_{x\chi} = R_{\chi y} = - \log \sqrt{1 - r_{xy}} \tag{60}$$

The radical in (60) is of course the standard error of measurement.

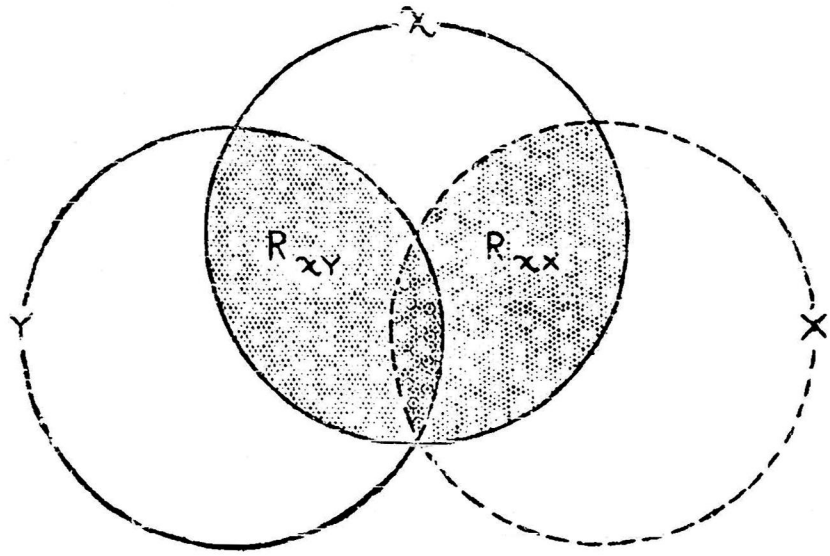


Figure 9. Relation of two fallible measures and the underlying true score.

In deriving R we have made the assumption that the distribution within arrays is continuous and normal. This assumption is in conflict with the assumption made in deriving (49), that the distribution within differential elements is rectangular. The latter assumption is acceptable for the distribution of x or y, but gives us a contradiction as $\sigma_{y.x}$ or $\sigma_{x.y}$ becomes smaller and smaller. Formulas (54) to (60) apply exactly only if we regard x and y as truly continuous, capable of being infinitely finely divided. But this means that H_x and H_y are infinite (as u is infinitesimal). Therefore J and K cannot be defined under a strict assumption of normality. J and K are meaningful for the ordered case provided $H_{x/y}$ and $H_{y/x}$ are determined from the actual discontinuous distribution of persons into cells u units wide. Formula (49) and its derivatives may not be used for this.

The <u>necessary</u> conditions for (54) and subsequent formulas are

(1) that the shape of the distribution within arrays has the same shape as the marginal distribution, and

(2) that the dispersion within arrays is uniform. Normality is not required. These conditions do imply that u becomes indefinitely small; i.e., continuity is required.

It will be observed that as r increases toward 1, R becomes indefinitely large. This is possible, of course, under the assumption that the information desired as infinite. The statement of R as $\log \sigma_x / \sigma_{x.y}$ clarifies the meaning of R. Before testing, we locate S within a band of uncertainty described by σ_x. After testing we have the interval of uncertainty described by $\sigma_{x.y}$. This describes uncertainty in just the way we do when we append a standard error (e.g., $2.7 \pm .35$) to a score. The ratio shows

how we have divided our original uncertainty. Cutting it in half gives one bit of information; cutting in eighths is equivalent to three bits. Three "one-bit" items independent in the sense of condition (5) are able to reduce the error of estimate to one-eighth σ_x.

Summary of Major Implications

In our examination of the uncertainty measure H, and the measure of rate of change, R, we found that these measures do not have the characteristics desirable for the evaluation of tests. The formulas apply precisely only when patterns of inputs can be encoded simultaneously, a condition not fulfilled in our testing problem. As approximations, the formulas express the number of standard independent items required per person to classify him, before and after testing. But this has meaning only if tests are used sequentially, with different persons given different numbers of items. Hence, there will be few occasions to treat data by the formulas of this section.

The sequential conception of testing has great potentiality. Since we can make decisions about some persons after fewer items than are required for others, we may be able to attain greater efficiency by sequential testing. This is practicable only where the cost of administering a sequential plan is little greater than the cost of administering uniform-length tests to all persons. Evans has shown why sequential testing can be recommended for performance testing.

The sequential concept can be applied also to diagnosis or evaluation of a single person. Suppose a decision as to college admission requires information that the subject has attained a certain level of proficiency in each of eight areas (English, mathematical comprehension, etc.). The first day could be used for a test of items having properly chosen difficulty in every area. The tests would be scored promptly. Some students would earn such high scores that a decision to treat them as passing in all areas could be made with great confidence. Perhaps some could be rejected with confidence on the basis of the short tests. It is more likely that for most individuals we can find some areas where he is definitely passable, and others where it is uncertain how he should be classified. In those areas only, further testing on the second day is wise. A more adequate decision can be reached by thorough testing in these dubious areas than if the time over both days were divided equally over all areas. In general, educational and psychiatric diagnosis employ such sequential testing, and a statistical model based on sequential analysis should be used to evaluate the effectiveness of the procedures.

Attention was drawn in this section to the possibility of evaluating separately the rate of information, exhaustiveness, and dependability. These concepts have not come to attention in analyses of test validity where both test and criterion were taken as continuous. The concepts are particularly helpful in thinking of categorical, multidimensional data. Exhaustiveness and dependability are not necessarily equal. It is to be expected that personality test responses can be "postdicted" from practical criteria better than we can predict from test to criteria. Such studies are recommended in the early stages of test validation. In some practical situations, tests

with high exhaustiveness are preferable to tests of high dependability; in other situations, dependability is the desideratum.

Equations for correcting information measures for unreliability have been developed. Similar correction formulas can be developed for other measures of association between categorical variates.

The conceptions summarized above are not bound to the Shannon formulas, and we can expect them to be useful in any system of test analysis.

III. MEASURES OF UNCERTAINTY AND INFORMATION IN TERMS OF CORRECT DECISIONS

A Measure of Average Uncertainty

In this section, we shall examine a set of information formulas based on a somewhat different attack than Shannon's. These formulas are more suitable for the tester than Shannon's. They are expressed in terms of average confidence rather than average message space, and they assume that decisions are made about one person at a time. The formulas are, however, much less general than utility formulas.

Since one of the formulas to be developed in this section has been proposed by other writers as a way of judging tests, we are able to extend the meaning of their work. By placing their formula in the same perspective as Shannon's and later relating it to utility theory, we clarify its meaning and its limitations.

We may begin as in Section II, defining confidence.

$$\text{Conf}\ (a, y_s) = \Pr\{x_s = a \mid y = y_s\} = p_{a/y_s} \tag{1}$$

When we make many decisions, we have some degree of confidence in each, i.e., some probability of being correct. What is the average goodness of our decisions, or our average confidence? Each time we classify a person correctly we gain something, regardless of errors in classifying others. If we add the confidence values for successive decisions, we obtain the expected number of correct decisions. The mean confidence (the mean probability of a correct decision) then is a measure of the goodness of decisions.

Call C_x the proportion of correct decisions regarding x to be expected. To assess a priori confidence, assume that N persons are assigned to criterion classifications by chance, Np_x persons being assigned to category x. p_x is the expected frequency of category x. The probability of a "hit" when a person is assigned to category x is p_x, and

$$\lim_{N \to \infty} C_x = \frac{1}{N} \sum_x p_x (Np_x) = \sum p_x^2 \tag{2}$$

Let U be the proportion of misclassifications in a very large sample of chance assignments.

$$U_x = 1 - \sum p_x^2 = \sum_x p_x q_x \tag{3}$$

If compound categories, such as red-even, are built up, we can express the confidence of correct assignment simply, provided the category sets are independent. If the compound category xX includes any person who is both x and X, the two basic category systems are independent when $p_{xX} = p_x p_X$. Then

$$C_{xX} = C_x\ C_X \tag{4}$$

That is to say, the probability of hits for assignments to the complex categories is the product of the average probabilities for the separate category systems.

Previously Published Formulas for Measuring Discriminating Power

Other writers have regarded a test as a tool for discriminating between persons (15; 29; 3, p. 1241). A discrimination is a statement such as "A is unlike B." If there are N persons, we have $\frac{N(N-1)}{2}$ possible discriminations to be made. But if there are k possible scores or categories, and k < N, the number of discriminations cannot be greater than $\frac{N}{2}(N - \frac{N}{k}) = \frac{N^2}{2}(1 - \frac{1}{k})$. The actual number of discriminations made by the test is $\frac{1}{2}\sum_{y=a}^{k} N_y (N - N_y)$ or $\frac{N^2}{2}(1 - \Sigma p_y^2)$.

Ferguson suggested evaluating the test by its discriminating power relative to the maximum for the given k, and offers the index

$$\delta = \frac{k}{k-1}(1 - \Sigma p_y^2) \tag{5}$$

Thurlow takes the ratio relative to the limit set by N, and suggests the index

$$\delta' = \frac{N}{N-1}(1 - \Sigma p_y^2) \tag{6}$$

If we let k or N increase indefinitely, either δ or δ' reduces to U_y, as derived by substituting the p_y in (3).

Thus the discrimination indices are essentially measures of U_y, the uncertainty or variation in responses, just as U_x as expressed in (3) is a measure of the variation in the criterion.

Geometric Interpretation as a Dispersion Measure

Formula (3) has an interesting and possibly useful geometric significance. We may regard the x categories as defining a set of orthogonal vectors of unit length. Category a may be represented by the vector 1, 0, 0, 0, ...; b is represented by the vector 0, 1, 0, 0, ...; etc. These vectors define a coordinate system, and each person has a true location in this system. A person belonging in category a is correctly located when he is assigned at point (1, 0, 0, 0, ...). Figure 10 shows three coordinates, and also the point (p_a, p_b, p_c). The point located by (p_a, p_b, p_c) is the centroid of all points, when persons are properly located. Every possible distribution of a sample of persons is defined by some point in the plane x + y + z = 1.00. In Figure 11, therefore, we view this plane directly.

When a priori, we know p_a, p_b, p_c, but know nothing about the location of individuals, we might consider all persons as located at the point M (p_a, p_b, p_c). Then a person whose true category is a is misplaced from his proper position by the distance MA.

$$\overline{MA}^2 = (1 - p_a)^2 + (p_b)^2 + (p_c)^2 \tag{7}$$

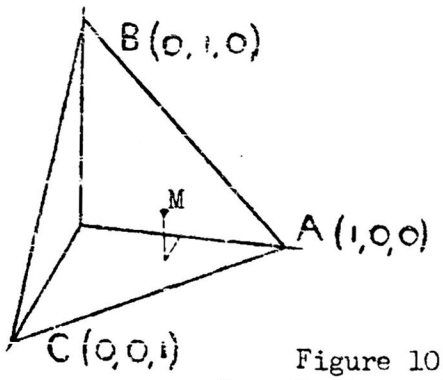


Figure 10.
Coordinate system

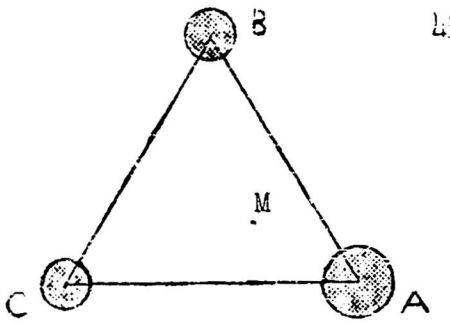


Figure 11.
Plane viewed directly

The heavy circles in Figure 11 show how persons would be distributed if each was assigned to his true category. There are Np_a persons displaced by the distance MA. Averaging the squared displacements over persons in all categories, we find

$$p_a \overline{MA}^2 + p_b \overline{MB}^2 + p_c \overline{MC}^2 = 1 - \sum_a^c p_x^2 \qquad (8)$$

That is to say, if we regard all persons as located at the mean of the distribution to which they are known to belong then the average uncertainty is the mean squared error. This is a measure of dispersion, the multivariate analog to variance. Ferguson also has noted this analogy to variance. Essentially, in setting up this geometric model, we have assumed that any misclassification is as serious as any other; i. e., that assigning an a to category b is as serious as assigning a b to category a or c., etc. This amounts to assuming a particular evaluation matrix (see Section IV).

Residual Uncertainty after Testing

Previous authors are perhaps unwise in referring to U_y as a measure of "the discriminating power" of a test, just as it is unwise to refer to H_y as "the information yielded." U_y and H_y are comparable. H_y has been interpreted as a statement of the variation in the output messages. U_y indicates the amount of reported discrimination, but these discriminating statements will not all be valid, as Thurlow has pointed out.

A test would ordinarily permit us to classify people into x categories with greater confidence than we had a priori. Each person in a given y category could be assigned to an x category, basing inferences on the known $p_{x/y}$. Assuming a chance assignment as before, if N_y people give response y, $p_{x/y}$ of them would be assigned to category x. For any y, say α, the average confidence is denoted by $C_{x/\alpha}$. From (2),

$$C_{x/\alpha} = \sum_x p_{x/\alpha}^2 \qquad (9)$$

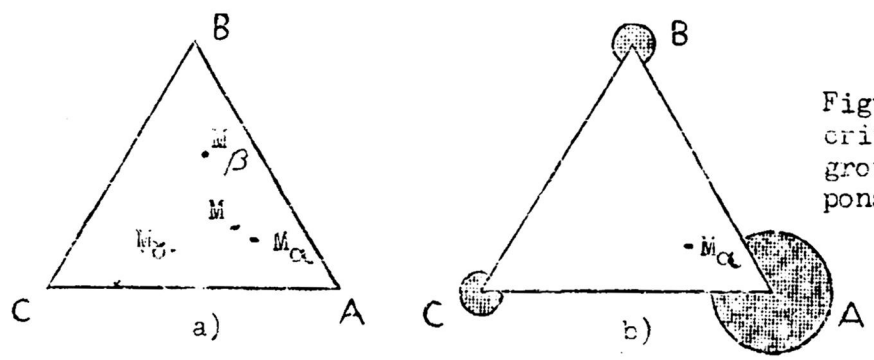


Figure 12. Centroids of criterion distributions of groups giving various responses, and distribution for Group α

And over all y's, the average a posteriori confidence is

$$C_{x/y} = \sum_y p_y \sum_x p^2_{x/y} = \sum\sum_{yx} p_y\, p^2_{x/y} \tag{10}$$

Here we shall use C_1 to indicate average confidence in classifying on the basis of item 1. We can rewrite (10) thus:

$$C_1 = \sum\sum_{yx} p_{xy}\, p_{x/y} = \sum\sum_{yx} \frac{p^2_{xy}}{p_y} = \sum\sum_{xy} p_y\, p^2_{x/y} \tag{11}$$

We obtain residual uncertainty U_1 by subtracting C_1 from unity.

We may discuss this in terms of the geometric model. For each y category, we have a set of $p_{x/y}$ which define the centroid of the criterion distribution for persons giving response y. Figure 12(a) shows the original M of Figure 11, together with the centroids M_α, M_β, M_γ. Figure 12(b) shows the distribution of persons giving response α. The dispersion within any group is smaller than the a priori dispersion. U_1 is the weighted average of these within-group (residual) dispersions.

Gain In Average Confidence (Information)

We employ ΔC to represent the gain in average confidence (i.e., in expectation of correct decisions) as a result of testing.

$$\Delta C = C_1 - C_0 = U_0 - U_1 \tag{12}$$

$$\Delta C = \sum_y \sum_x p_y\, p^2_{x/y} - \sum_x p^2_x = \sum_x \sum_y p_y\, (p^2_{x/y} - p^2_x) \tag{13}$$

Hence ΔC is the increase in confidence for any category, when y is known, weighted and summed over all categories and all values of y. We can rewrite ΔC in a form allowing us to state ΔC_{12}, the distance between any two states.

$$\Delta C = \sum_x \left(\sum_y \frac{p^2_{xy}}{p_y} - p^2_x \right) = \sum_x \sum_y \left(\frac{p^2_{xy} - p^2_y\, p^2_x}{p_y} \right) \tag{14}$$

$$\Delta C_{12} = \sum_x \left(\sum_{y_2} \frac{p^2_{xy_2}}{p_{y_2}} - \sum_{y_1} \frac{p^2_{xy_1}}{p_{y_1}} \right) \tag{15}$$

In the model studied in this section, ΔC represents the ability of the test to reduce uncertainty or to convey information. It is analogous to R of Section II.

Uncertainty may be decomposed as in analysis of variance to show various sources of discrimination. For N large, the criterion permits $\frac{N^2}{2}(U_x)$ discriminations. When a test is given to persons in category a, error introduces differences in response. There are $\frac{N_a^2}{2}(U_{y/a})$ added discriminations, where

$$U_{y/\alpha} = 1 - \sum_{y} p^2_{y/\alpha} \tag{16}$$

Summing over all x categories, the total error discriminations are

$$\frac{N^2}{2} U_{y/x} = \sum_{x} \frac{N_x^2}{2} \left(1 - \sum_{y} p^2_{y/x} \right) \tag{17}$$

$$\sum_{x} \frac{N^2}{2} U_{y/x} = \frac{N^2}{2} \sum_{x} p_x^2 \left(1 - \sum_{y} p^2_{y/x} \right) \tag{18}$$

Then the total number of discriminations is

$$\frac{N^2}{2} U_x + \Sigma \frac{N_x^2}{2} U_{y/x} = \frac{N^2}{2} \left(1 - \sum_{x} \sum_{y} p^2_{xy} \right) \tag{19}$$

The right hand member of (19) is exactly the number of discriminations (which we might designate $\frac{N^2}{2} U_{xy}$) that we would have if we could record the complete matrix of true responses and errors, thus dividing people among yx categories. Dividing through by $\frac{N^2}{2}$, we obtain

$$U_{xy} = 1 - \Sigma \Sigma p^2_{xy} = U_x + U_{y/x} \tag{20}$$

Figure 3 of Section II is a diagram of precisely this type of additive relation. Furthermore, we could write

$$U_{xy} = U_y + U_{x/y} \tag{21}$$

The total discriminations in the system may be divided into true and error discriminations, or into obtained discriminations and equivocation (i.e., desired discriminations not obtained). The relations implied by Figures 4, 5, and 6, apply to U as well as to H.

The measure of information gained, given in (13), is close to Thurlow's proposal for judging the number (or proportion) of valid discriminations made by a test when categories are unordered (29, p. 304-5). ΔC is easier to compute than Thurlow's index, however, and ΔC has several properties not found in Thurlow's measure. Thurlow's index for unordered categories is applied only to the case where there is an obvious one-to-one correspondence between the x and y categories. Moreover, some questions may be raised regarding Thurlow's method of counting correct discriminations. Ferguson and Bechtoldt make no suggestion for considering the validity of discriminations.

Gain in certainty can be translated into the terms of analysis of variance. The mean square dispersion between groups is $p_\alpha (MM^2_\alpha) + p_\beta (MM^2_\beta) + p_\gamma (MM^2_\gamma)$. This is equal to $C_1 - C_0$ or $U_0 - U_1$. Thus U_0, the a priori uncertainty, equals the dispersion between response groups plus the total dispersion within response groups (residual uncertainty).

Information Yielded by a Series of Independent Items

When two independent items are combined, what is their total information yield? As in Section II, we define independence by the conditions:

$$\left.\begin{aligned} p_{y_1 y_2} &= p_{y_1}\, p_{y_2} \\ p_{y_1 y_2/x} &= p_{y_1/x}\, p_{y_2/x} \end{aligned}\right\} \tag{22}$$

After one item, a certain number of persons are classified in any category x. We may denote by $C_{a(1)}$ the confidence in classifying a person as a, on the basis of item 1. We shall let C_{1a} be the proportion of all persons who are correctly classified in category a. Then $C_{1a} = p_a C_{a(1)}$. Similarly for all x categories, so that

$$C_1 = \Sigma\, C_{1x} \tag{23}$$

Equation (11) demonstrated that C_1, over all x's, is $\Sigma\Sigma\, p_{xy}^2/p_y$. For category a,

$$C_{1a} = \sum_{y_1} p_{ay_1}^2/p_{y_1} = \sum_{y_1} \frac{p_{y/a}^2\, p_a^2}{p_{y_1}} \tag{24}$$

If C_{12a} is the similar proportion based on independent items 1 and 2 together,

$$C_{12a} = \sum_{y_1}\sum_{y_2} \frac{p_{ay_1y_2}^2}{p_{y_1y_2}} = \Sigma\Sigma \frac{p_{y_1/a}^2\, p_{y_2/a}^2\, p_a^2}{p_{y_1} p_{y_2}} \tag{25}$$

Using (24),

$$C_{12a} = \frac{C_{1a}\, C_{2a}}{p_a^2} \tag{26}$$

We may generalize (25) to t items thus:

$$C_{(12\ldots t)a} = \Sigma\Sigma \ldots \Sigma \left(\frac{p_{y_1/a}^2}{p_{y_1}}\right)\left(\frac{p_{y_2/a}^2}{p_{y_2}}\right) \cdots \left(\frac{p_{y_t/a}^2}{p_{y_t}}\right) p_a^2 \tag{27}$$

$$C_{(12\ldots t)a} = \prod_i \sum_{y_i} \left(\frac{p_{y_i/a}^2}{p_{y_i}}\right) p_a^2 \tag{28}$$

If each item has the same ability to reduce uncertainty,

$$C_{(12\ldots t)a} = \left(\frac{C_{1a}}{p_a^2}\right)^t p_a^2 = p_a^2 \left(\frac{C_{1a}}{C_{oa}}\right)^t \tag{29}$$

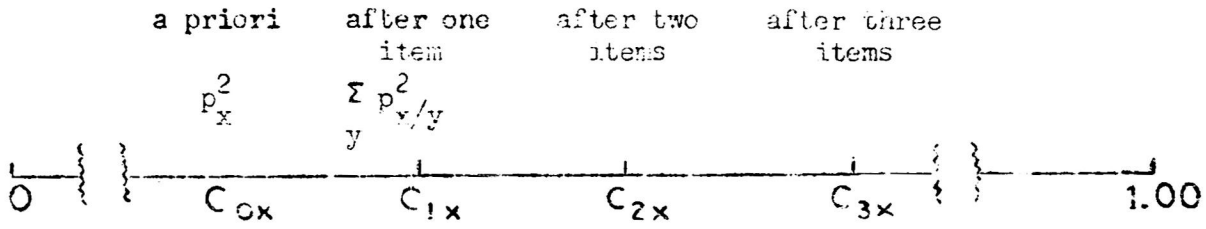


Figure 13. Expectation of correct classification within any category x.

As in Section II, we may inquire how many uniform independent items are required to raise to 1.00 the average *a posteriori* confidence in classifying a person as a. Following (1), we employ a conditional notation to write

$$\text{Conf } (a,y_s) = p_{a/y_s} = C_{(12\ldots t)a/p_a} \tag{30}$$

Dividing through (29) by p_a, and setting confidence equal to 1, we have

$$1 = \left(\frac{C_{ia}}{p_a^2}\right)^{n_a} p_a \tag{31}$$

$$n_a = \frac{-\log p_a}{\log \frac{C_{ai}}{p_a} - \log p_a} = \frac{-\log p_a}{\log \text{Conf } (a,y_s) - \log p_a} \tag{32}$$

This formula is similar to (32) of Section II, and demonstrates that the model of the present section yields the same measure of number of independent items required in sequential testing as does Shannon's, so long as we consider persons in any one criterion category. The weighted average of n_a from (32) here is not the same as $\bar{n}$ from (32, Sec. II).

While (29) can be summed over x categories to give the total *a posteriori* confidence after t items, this relation is not simple in form unless $C_{(12\ldots t)}$ is the same for all categories. Figure 13 shows the way in which confidence increases with added items. This may be compared with Figure 2. Note that here, instead of each item adding the same or less than preceding items, we find that each item adds more to average confidence (C) than did its predecessor, until certainty is reached. No practical implications should be drawn from this rather startling result until formulas are developed to take into account correlations among successive items. Our independence conditions require that items have a negative correlation, with x held constant, and this is not usual in practice.

Exhaustiveness and Dependability

In the system here discussed, we can write measures of exhaustiveness and dependability.

$$E\ (\text{Exhaustiveness}) = \frac{\text{Information obtained}}{\text{Information desired}} = \frac{C_1 - C_0}{1 - C_0} \qquad (33)$$

If we wish to know how completely the test responses are specified by the criterion, we have the comparable formula,

$$D\ (\text{Dependability}) = \frac{\Sigma p^2_{y/x}\, p_x - \Sigma\, p^2_y}{1 - \Sigma p^2_y} \qquad (34)$$

All the argument of pages 31 - 33, regarding the significance of J and K for validity studies may be carried forward to E and D.

Significance of Confidence Formulas

In summarizing this section, we should note that all the implications of Sections I and II are consistent with the equations developed here. Thus nothing of the practical meaning of the Shannon formulation--insofar as it applies to testing--is lost by adopting a measure of average confidence.

The actual formulas presented here may be of some direct use, even though Section IV will emphasize the reasons for seeking a treatment of test data which evaluates confidence on a basis other than chance assignment of men to categories. Such a method, while logically superior to the one developed in Section III, will often be difficult to treat mathematically. In working out functional relationships to determine how much confidence is increased by some change in testing procedure, it may be desirable to employ the formulas of this section as first approximations to results from utility analysis.

Our results carry implications for the "discrimination" formulas others have offered. The formulas of Ferguson, for instance, should be regarded as a measure only of claimed or attempted discriminations, not as an index of the useful information in the test. In any event, these formulas are appropriate only if all misclassifications are equally serious--as is probably not true for ordered categories or scores. While our formulas may provide the best way to take validity into account in discrimination formulas, our doubts regarding the average certainty formulas suggest that the discrimination formulas are also somewhat unsuitable for test analysis.

IV. AN INTRODUCTORY STATEMENT OF UTILITY THEORY, GIVING ITS IMPLICATIONS REGARDING THE INFORMATION FORMULAS

It is our purpose here to indicate the nature of utility theory. We are not yet prepared, however, to give a general statement of the utility theory with a well-tested notation, systematic formulas, and the like. Section V applies utility theory to a specific limited problem, and this section will prepare the reader for that example.

The concepts introduced in utility theory also permit us to judge the value of the formulas based on H and C, presented in Sections II and III.

Utility Theory

A utility theory is an attempt to state the benefit derived from an operation or procedure, by comparing the goodness or utility of decisions based on the procedure with the utility of decisions made without the procedure. Utilities are a key concept in economic theory, and in theory of games and strategy (30, 31).

Basic Data Required

To judge the utility of a test (or of a procedure for making decisions from the test), we require three matrices or tables of data.

Transition matrix. The first matrix is a transition matrix relating the criterion scores x to the obtained scores y. The entries in this matrix may be written in the form of either joint or conditional probabilities. We shall here employ conditional probabilities; Table 1 (p. 16) showed such a matrix. Data for a transition matrix come from an empirical trial where test and criterion data are obtained for each person. The margin of the transition matrix gives the distribution of x. One problem is that this matrix is based on sample data. Population data are assumed in the way we have computed utilities below.

Table 3. Specimen Evaluation Matrix

		Assigned Classification (x')				
		a'	b'	c'	. . . . .	k'
Criterion Classification (x)	a	$e_{aa'}$	$e_{ab'}$	$e_{ac'}$		
	b	$e_{ba'}$	$e_{bb'}$	$e_{bc'}$		
	c	$e_{ca'}$	$e_{cb'}$	$e_{cc'}$		
	.					
	.					
	.					
	k					

Evaluation matrix. The second matrix is an evaluation matrix, telling what value we assign to each successful or unsuccessful classification. The test information (y_s) is used to assign person S a classification x'. The evaluation matrix consists of values $e_{xx'}$. For example, $e_{ab'}$ states the value (gain or loss) when a person whose criterion classification is a is assigned to category b. A specimen evaluation matrix is given in Table 3.

Evaluations must be determined by an accounting process of some sort. Brogden (5) and Brogden and Taylor (6) have discussed the desirability of introducing such utilities into test analysis. When it is impractical to fix evaluations by accounting studies, they must be specified by the judgment of some person concerned in the decision. Some setting of utilities or risks is required whenever decisions are made. Often the value judgments are made implicitly. Each conventional method of assessing tests embodies some particular weighting of risks or errors; the user of the formula may not realize what is thus assumed. Utility theory only makes these value judgments explicit so that they can be reviewed openly and altered when necessary.

Interpretation matrix. The third matrix is the interpretation matrix. The interpretation matrix specifies the "strategy" or decision function to be used in assigning persons to the x' categories. The rule might be that all persons giving response α will be placed in category a; or that some percentage of them selected at random will be called a's, and the rest b's; etc. Each entry (see Table 4) takes the form of a probability $w_{x'/y}$. Different interpretation rules give different benefits. For any given transition matrix and evaluation matrix, there is a best strategy. The determination of cutting scores in Section V illustrates how we choose a best interpretation matrix or decision procedure for a given body of data.

Calculation of Utility

If the interpretation matrix shows how to convert y to x', and the transition matrix shows how x depends on y, then we may construct a new matrix in which each element is $p_{xx'}$, the joint probability of x and x'.

$$p_{xx'} = \sum_y p_x \, p_{y/x} \, w_{x'/y} = \sum_y p_{xy} \, w_{x'/y} \qquad (1)$$

Table 4. Specimen Interpretation Matrix

Response Category (y)	Assigned Classification (x') a'	b'	c' . . . . . k'
α	$w_{a'/\alpha}$	$w_{b'/\alpha}$	$w_{c'/\alpha}$
β	$w_{a'/\beta}$	$w_{b'/\beta}$	$w_{c'/\beta}$
γ	$w_{a'/\gamma}$	$w_{b'/\gamma}$	$w_{c'/\gamma}$

The evaluation matrix assigns a value to each xx', and the value of all classifications in the xx' cell is

$$V_{xx'} = e_{xx'} (Np_{xx'}) = N \sum_{y} e_{xx'} p_{xy} w_{x'/y} \qquad (2)$$

The sum of all $V_{xx'}$ over all x and x' gives the total utility of decisions. This sum may be divided by N to obtain an average utility per person, V. We are using V rather than U as our symbol here because U was employed previously in this report for Uncertainty.

$$V = \Sigma\Sigma\Sigma\, e_{xx'} p_{xy} w_{x'/y} \qquad (3)$$

The utilities may be interpreted directly. They may also be compared to the gain over some other strategy, or over a chance decision. This gain may be divided by the maximum possible gain to obtain a relative utility, RV.

$$RV = \frac{V_1 - V_0}{V_{max} - V_0} \qquad (4)$$

Here, V_{max} is the value obtained when every person in a particular category x is assigned to that x' for which $e_{xx'}$ is greatest, i.e., when we make the most profitable assignment possible.

Table 5 draws attention to basic similarities between these types of measures and measures of preceding sections.

Table 5. Comparable Concepts in Three Systems

	Measure of Section III	Measure of Section I	Measure of Section II
Quality of decisions at end of testing	V_1	H_1	U_1
Gain over a priori situation	$\Delta V = V_1 - V_0$	$R = H_0 - H_1$	$\Delta C = U_0 - U_1$
Gain relative to possible gain	$RV = \frac{V_1 - V_0}{V_{max} - V_0}$	$J = \frac{H_0 - H_1}{H_C}$	$E = \frac{U_0 - U_1}{U_C}$

It will be noted that we are able to take into account, in utility measure, the implications of preceding sections. For instance, the suggestion that the test be evaluated in terms of gains over the best a priori decision (p. 13) is embodied in the formula ΔV. The fact that exhaustiveness is not synonymous with rate of information is represented in the distinction between ΔV and RV. Dependability will require more indirect treatment in utility theory than in the information theories.

Review of the Average Confidence Formulas

Every method of evaluating a test in some way specifies an interpretation plan. For instance, correlational analysis evaluates the goodness of decisions when persons are assigned on the basis of a regression line. Also, every method embodies some set of evaluations. (In correlation, for example, the mean square error is determined; thus, an error of two points is counted as four times an error of one point.) Other formulas may be viewed as special cases within utility theory.

The formulas of Section III, based on average certainty, count all hits as equal in value, and all errors as equal in value. Thus the evaluation matrix is

		x'		
		a'	b'	c' ...
	a	1	0	0
x	b	0	1	0
	c	0	0	1
	⋮			

This is clearly not always the most suitable evaluation matrix for a given problem.

The interpretation matrix assumed in Section III for determining the expected number of hits sets each $w_{x'/y}$ equal to the corresponding $p_{x/y}$. This distributes persons having a given response according to the known expectancy of each x category in the group. This is not ordinarily the best strategy available. Our number of hits is always greatest if we assign all persons giving response α to that x' for which the corresponding $p_{x/\alpha}$ is greatest. Some other strategy may be superior to this when evaluations are introduced. We cannot expect, in general, that the chance assignment by $w_{x'/y} = p_{x/y}$ will give the greatest utility the test can provide.

We conclude that C, ΔC, and related formulas are not satisfactory measures of the goodness of a test. This criticism also applies to the "discriminating power" formulas reviewed in Section III.

Review of the Average Log Confidence Formulas

We can examine the formulas based on H from the same point of view. — NH is the logarithm, it will be recalled, of the probability of correctly classifying an indefinitely large configuration of persons, drawn at random from the population specified by the p_x. Thus 2^{-NH} may be regarded as a utility measure making the same assumptions as are involved in C, except that the evaluation and interpretation matrices now are to apply to the whole very long sequence. The evaluation matrix looks like the one shown above, but now is based on a sequence v instead of a single person. Here, then, a stands for a particular sequence. The evaluation matrix is:

		v' a'	b'	c'	...
	a	1	0	0	
v	b	0	1	0	
	c	0	0	1	
	⋮				

The interpretation matrix implied is not open to the criticism advanced in connection with average confidence formulas. In interpreting a single response or finite set of responses there usually will be a set of $w_{x'/y}$ differing from the corresponding $p_{x/y}$ which gives greater average confidence than the $w_{x'/y} = p_{x/y}$. This is, however, not true for the infinite sequences of independent messages because all p_v are equal, and every $p_{v/y}$ is 1 or 0, or equal to p_v. When evaluations are as shown above, interpreting an indefinitely long sequence of responses y by assigning persons to the configurations v' so that $w_{v'/y}$ equals the corresponding $p_{v/y}$ is as good as any other strategy.

It is highly unlikely that the evaluation matrix shown will ever fit the testing problem. Questions were raised above as to the use of equal values for all hits and equal values for all errors. Apart from this, the matrix shown above counts an assignment as of zero value unless every element in the sequence is correct. That is, in the testing problem, 2^{-NH} measures utility only if we regard it as of zero value to classify N-1 persons properly so long as the Nth is wrongly classified.

Since we find that the Shannon H and R measures may be used only approximately in evaluating sequential testing, may not be interpreted in terms of coding in our situation, and may not be translated reasonably into utilities, we are abandoning attempts to derive useful formulas for test analysis from the Shannon treatment. We do not regard it as sound to follow Hick, for example, in his proposal (19, p. 162) to treat information measures formally to determine the best design for a test.

We have, of course, profited from the conceptual leads suggested by the information analogy, but believe these can best be formulated in utility measure.

V. ANALYSIS OF A PSYCHIATRIC SCREENING TEST BY UTILITY THEORY

This study of the value of a psychiatric screening test is a demonstration of some consequences of an approach through utility theory. Utility theory, as outlined in Section IV, studies the value of a test by taking into account the probabilities of making correct inferences from the test to the criterion and the value (benefit or cost) of each correct and incorrect inference. If we apply this approach to typical test data, we identify several points not apparent when a single validity coefficient is reported for a test.

A test yields a score or pattern of responses. This "output" is translated by an interpretation formula of some sort into a proposed decision. In order to judge the worth of the test, we must have an indication of the correct decision for each person in the validation sample. For each possible pairing of actual decision with correct decision, we also take into account an evaluation which expresses the seriousness of any error or the gain from any correct decision. The net value of the test is expressed in terms of the value of the decisions based on the test, as compared with the decisions that might have been made without the test.

It was recognized during World War II that the correlation coefficient is not well adapted to reporting the worth of a psychiatric screening test(28). The practice was then introduced of reporting the effectiveness of a screening test in a more complicated form. For a test used to screen recruits, the typical report of that period states for each cutting score the resulting number of hits (correct decisions), misses (deviates reported as normals) and false positives (normals reported as deviates). This rather complex report permits any user to evaluate the test for his purposes.

To illustrate a simple utility analysis, we shall examine here the NDRC screening test, developed to detect probable psychiatric casualties among Naval recruits. Such a test was needed because of the importance of identifying men who would probably become psychiatric casualties, and because the large intake of men overburdened psychiatrists who might otherwise have made this judgment in an individual interview. The screening test was introduced as an adjunct to the psychiatric interview in order to reduce the number to be interviewed carefully. Validity data for the test, comparing the test score with a psychiatrist's judgment on each man, are reported in the Summary Technical Report of NDRC (27). (The table of data we employ comes from a secondary source, and probably contains some very small inaccuracies.) Table 6 shows the distribution of test scores for the two criterion groups.

One emphatic caution needs to be expressed regarding the present analysis. We have data only for a dichotomous criterion, and our method of analysis assumes a dichotomy. But not all those judged "normals" by the psychiatrist are equally free from troublesome symptoms, and not all "deviates" are equally undesirable to the service. Our analysis necessarily fails to give the test credit for any power it has to make practically significant discriminations within criterion groups. Our present demonstration serves only to show the general method. It presents the conclusions of an analysis with a discrete two-point criterion but this analysis is undoubtedly oversimple for ultimate evaluation of the screening test.

Table 6. Raw Data Indicating the Validity of the NDRC Screening Test

Score	Frequency of score among men judged by psychiatrists to be: Deviates	Normals	Total
19		1	1
18	1		1
17		1	1
16		1	1
15		1	1
14	2	1	3
13	1	1	2
12		1	1
11	5	1	6
10	3	5	8
9	3	4	7
8	1	8	9
7	2	18	20
6		24	24
5	3	41	44
4	1	56	57
3	4	80	84
2	1	96	97
1	2	97	99
0	1	21	22
Total	30	458	488

Correlational Validity

The conventional method of summarizing validity in a single index would be to compute point-biserial r. For these data, $r_{pt\text{-}bis} = .354$. Such an index is not very satisfactory, however, because the result would be altered if the scale of measurement of the test were transformed, for instance, into normalized scores (cf. Brogden, 4). Such a transformation would not alter the effectiveness of the test as a screening device, and so the change in coefficient would be misleading. In contrast to this coefficient, the value derived by utility analysis does not change when the measuring scale for the test is transformed.

A second correlational index of some merit would be the phi coefficient, obtained after dichotomizing the test at some cutting point. The value of phi varies as the cutting point changes. The only unique value of phi is the maximum, obtained by cutting so as to divide the test in categories having the same proportions as the criterion categories. According to Table 6, this would call deviates all persons with a score of 10 or over, and 5/7 of those at 9. For this particular cutting score, the fourfold contingency table is as shown below and phi is .437.

		Assigned Category (x')		
		Deviate	Normal	
True Category (x)	Deviate	14 1/7	15 6/7	30
	Normal	15 6/7	442 1/7	458
		30	458	488

We may point out two features of correlational validity.

1. It is expressed on a scale ranging from -1 to +1, on which zero represents the correlation that would be expected if persons were divided into categories by chance.

2. No explicit assumption is made regarding the comparative seriousness of misses and false positives. The extent to which phi is reduced by an additional error of either type depends on values in the fourfold table. For most tables, phi essentially counts each type of error as equally serious.

Utility Analysis

Introduction of Evaluations

In practice, a cutting score is fixed by considering the relative seriousness of misses as opposed to false positives (or, in selection generally, of incorrect "accept" decisions, as opposed to incorrect rejections). If false positives are very damaging, we set a high cutting score; if we would rather risk false positives than misses, we set a low cutting score. The optimal cutting score depends upon the values placed on these errors.

We prepare an evaluation matrix of this form:

		Assigned Category (x')	
		Deviate	Normal
True Category (x)	Deviate	e_{dd}	e_{dn}
	Normal	e_{nd}	e_{nn}

Evaluations may be positive or negative. Here, e_{dd} is the value associated with a correct decision deviate-called-deviate. e_{dn} is the value associated with a "miss" (this value is negative, because this decision is costly), and e_{nd} is the value of a "false positive."

Since we will evaluate the test by the gain in goodness of decisions, we may add or subtract any constant to both entries in a row with no change in our end results, so long as the criterion proportions remain fixed. In the first row, we find it helpful to subtract e_{dn} from both entries; in the second, we subtract e_{nd}. This yields the equally useful matrix:

		Assigned Category (x')	
		Deviate	Normal
True Category (x)	Deviate	$e_{dd} - e_{dn}$	0
	Normal	0	$e_{nn} - e_{nd}$

A useful summary of such an evaluation matrix for a dichotomy is the evaluation ratio.

$$E. R. = (e_{dd} - e_{dn})/(e_{nn} - e_{nd})$$

If we can count e_{dd} and e_{nn} as zero, E. R. is the ratio of the cost of misses to that of false positives.

Determining entries for the evaluation matrix. Values such as e_{dd}, e_{nd}, etc., are always taken into account in setting cutting scores even when the test user does not recognize it, for he must decide on the acceptable ratio of misses to false positives, and this implies a valuation. The evaluation ratio is very similar to the risks specified in making decisions on the basis of an experiment. There one is asked to state what risk he is willing to run of accepting a false hypothesis (Type I error) and what risk of rejecting a true hypothesis (Type II error). These risks are derived by the user on the basis of some notion of the practical cost of the two sorts of error. The weighting of the risks needs to be brought out explicitly, for otherwise it cannot be examined and criticized.

Without better data than we now have, values in the evaluation matrix will always involve some arbitrary judgments. To arrive at $e_{dd} - e_{dn}$, we estimate the gains and losses (a) if a deviate is called deviate, and (b) if he is called normal on the basis of the test. Suppose recruits called deviates are dropped from service; suppose normals are sent into basic training. The deviate-called-normal brings certain benefits and costs. We expect him to perform duties as a sailor which have value. He will continue until he gets into disciplinary trouble, is picked up on sick call, or breaks down in combat. Now we must assess:

Gain from work capably performed until discharge

Cost of training

Costs arising from symptoms of his disorder (time required to investigate disciplinary problems, time of medical staff, cost of his errors, loss in efficiency of his unit due to his failure to perform normally, etc.)

Cost of treatment for which the service takes responsibility (pension for service-incurred disability, etc.)

The fact that these (and other) costs are likely to outweigh the benefits from his service is the reason for desiring a psychiatric screening device. In the absence of accounting figures, these values are estimated, just as they are implicitly estimated in setting cutting scores at present. The sum of gains-minus-costs for this deviate-called-normal gives us e_{dn}.

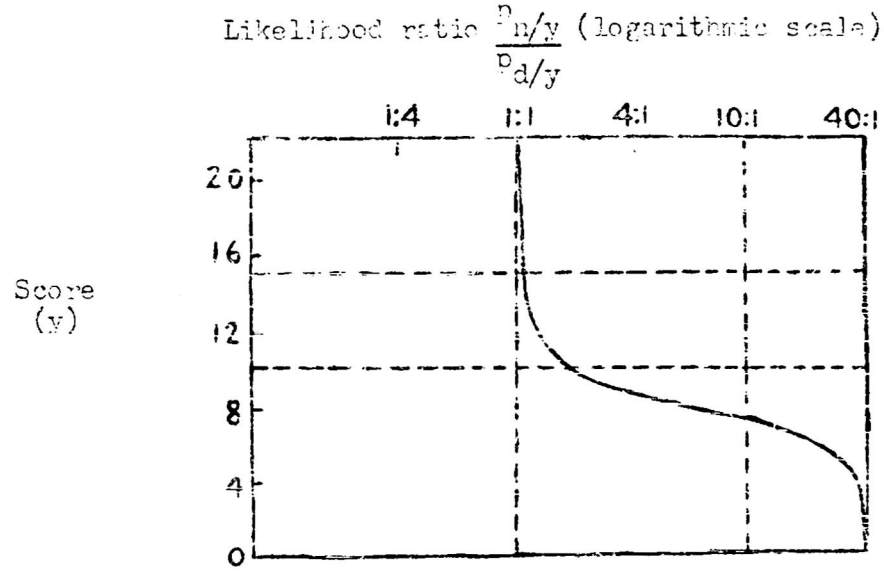


Figure 14. Likelihood ratios at various cutting scores (smoothed)

For the deviate-called-deviate, we may take e_{dd} as zero. He is rejected, and there is neither gain nor cost from him. A more precise analysis, as Brogden shows (5), should take the cost of testing into account in each cell, but this factor would be a distraction in this introductory report. When e_{dn} and e_{dd} are fixed, $e_{dd} - e_{dn}$ is determined. As we have set e_{dd} equal to zero, $e_{dd} - e_{dn} = - e_{dn}$.

A similar analysis would be made for normals to get $e_{nn} - e_{nd}$.

Likelihood Ratios as a Basis for Cutting Scores

We compute a likelihood ratio (L.R.) for each test score as a convenient way to examine Table 6. First we divide the frequency of normals at any score by the frequency of deviates. If y is a test score,

$$L.R. = p_{ny}/p_{dy} = p_{n/y}/p_{d/y}$$

At score 10, the likelihood ratio is 1.67. These likelihood ratios are unreliable, especially where the frequency of deviates is low. We smooth the curve on the assumption that this will yield more accurate estimates of population values than the points based on the sample. Points for the smoothed curve are obtained by pooling adjacent scores, when the frequency per score category is small. Thus, the ratio at score 10 can be estimated by pooling data for 9, 10, and 11 given in Table 6; the ratio is $\frac{4 + 5 + 1}{3 + 3 + 5}$ or .91. The logs of the ratios (or ratios plotted on semilogarithmic paper) are then plotted against score. The logarithms make for better smoothing. The best fitted smooth curve gives the final estimate of likelihood ratios. Theoretically, the best curve may have one or more dependable maxima or minima; if so, we may require cutting scores on both sides of the maximum (minimum). In this test, where no such maximum exists, a single optimum cutting score can be determined for each E.R.

Choice of Cutting Score

Figure 14 shows that, in a further sample, at every score level we may expect normals to exceed deviates. In Table I, deviates exceed normals at some scores (e.g., 11); this is almost certainly due to a sampling fluctuation. Figure 14 gives a likelihood ratio at each score; thus, according to the chart, we may expect L.R. to be 1.6 for persons earning score 10 in the population. If we classify persons at score 10 as normals, we will be right 1.6 times for every error. If we call all 10's deviates, we make 1.6 errors (false positives) for every hit.

The best cutting score we can select is that at which L.R. equals E.R.* At any score y such that $L.R._y > E.R.$, people should be called normals; and where $L.R._y < E.R.$, they should be classified as deviates.

We can show this by considering three evaluation ratios: 1:1, 2:1, 1.6:1. If L.R. is 1.6 at 10, our probability of misclassification** and our utility per decision regarding persons who receive score 10 are as shown in Table 7.

Table 7. Expected Probability of Errors and Utility per Decision for Persons Whose Score Is 10

	$p_{xx'}$		Utility ($\Sigma\Sigma p_{xx}e_{xx'}$) when E.R. is		
	Deviates called normals	Normal called deviates	1:1	1.6:1	2:1
10's are called normal		.615	.615c	.615c	.615c
10's are called deviates	.385		.385c	.615c	.770c

Here c is a positively valued constant, equal to $e_{nn} - e_{dn}$. It is apparent that if E.R. > 1.6 (i.e., when identifying deviates is relatively important), we gain most by calling 10's deviates. When identifying deviates is relatively unimportant compared to avoiding false positives, we gain most by calling all 10's normal. If E.R. = 1.6, persons above 10 should be called deviates and persons below 10 called normals; persons at 10 may be assigned either way.

According to the figure, the optimal cutting score for these data shifts very markedly as E.R. shifts from 1:1 to 1.5:1. Over a range of E.R. from 1.5:1 to 30:1, the optimal cutting point shifts slowly. The curve changes slope again, and there is no score for which L.R. > 39:1.

The selection of a cutting score may be very sensitive to the E.R. We shall show later that one pays a substantial price for using a non-optimal cutting score. Hence, the price of misses and false positives must be judged carefully.

*A proof for this relation has been developed by Goldine Gleser.

**In the population represented by the smoothed curve.

Optimal strategy at high and low E.R. Figure 14 showed that there is no optimal cutting score for evaluation ratios 1:1 or lower, nor for 40:1 and higher. This means that if $e_{nn} - e_{nd} \geq e_{dd} - e_{dn}$, the greatest possible net gain (unless a better test is used) will result when we call all men normals. If $40(e_{nn} - e_{nd}) \leq e_{dd} - e_{dn}$, the wisest decision is to call all men deviates. (This is probably a decision we cannot allow, for we must accept some men into the service even if it is "a losing proposition.") Our analysis to this point has considered all strategies allowable. We shall later have to introduce limitations on the strategies, but for the present we shall continue with no limits on strategy.

Comparison with Conclusions of Correlational Analysis

It seems paradoxical to argue that it may be wise to call all men normal, when deviates are known to be present and our test has some validity. Let us examine this. If every man in this sample is called normal, we have 30 misses. Suppose instead we call the top 30 men deviates, as in our earlier table for finding phi. Then we have 15 6/7 misses and 15 6/7 false positives; the change in number of errors is small, but we do have more errors than before. A cut at this point would be wiser than passing every man only if false positives are cheaper, i.e., more desirable than misses. A "valid" test may have no utility, at least for a discrete criterion. It is therefore apparent that the conclusion as to whether a test should be used cannot be based solely on a correlational validity coefficient.

Judgment by utility analysis adds to that offered by correlational analysis because the latter compares decisions based on the test only with decisions based on chance. If 30 persons were called deviates by chance, we would have 28 false positives and 28 misses, and the test did indeed improve on this. But a test is used because we want to improve on the best decisions we could make without the test. The best a priori decision (when no data on individuals are at hand) is whatever decision yields maximum utility with the expected probabilities in the criterion categories. If we call everyone normal, our errors cost us less than any chance assignment, until E.R. rises to N_n/N_d (458/30 = 15.3:1). At E.R. higher than this, our optimum a priori solution is to call everyone a deviate.

For tests having a finite range, there will almost always be a maximum and minimum E.R. beyond which we should not use the test in selection. This principle has not been pointed out in earlier studies because models for studying test validity have generally assumed normal distributions which have unlimited range. Further study is required to determine what may occur when the criterion range expressed in utility units is essentially unlimited but the test distribution has finite limits. It appears that when both test and criterion are unlimited in range, a test having validity above chance will always have some utility (though perhaps not greater than the cost of testing).

Utility Curves for Various Strategies

Decisions with the Test Compared to a priori Decisions

In order to demonstrate the relation of strategy to utility, we shall study two strategies for interpreting the test results, along with the three conceivable a priori strategies.

Strategies to be compared. The a priori strategies are as follows:

a. A priori chance. Thirty persons at random called deviates.

b. All persons called deviates. This is an optimum a priori strategy for all high E.R.

c. All persons called normals, the optimum a priori strategy for all low E.R.

The a posteriori strategies are:

d. Thirty highest scoring persons called deviates (best cutting score for E.R. 1.6:1).

e. Persons above 4 called deviates (best for E.R. 30:1).

In obtaining utility curves, we must plot evaluations rather than evaluation ratios. Because many sets of e_{dd}, etc., yield the same E.R., we need to select certain values. In the absence of actual utility data, it is most convenient to employ a scale with the following definition:

e_{dn} and e_{nd} are set equal to zero. $e_{dd} = \text{E.R.} \cdot e_{nn}$

The utility for any strategy is then the weighted number of hits it allows.

$$V = N_{dd}\ \text{E.R.}\ e_{nn} + N_{nn}\ e_{nn}$$

Here N_{dd} is the number of deviates called deviates, etc. We shall plot V/e_{nn}, recognizing that we have no fixed unit of measurement. The units on the scale V/e_{nn} are not to be regarded as equal from one E.R. to another. V/e_{nn} allows us to compare strategies only within any E.R.

Results. Figure 15 presents utility functions for the strategies, V/e_{nn} being plotted against E.R. The line labelled Max shows the value V would take if the test were able to make perfect classifications. In Figure 16, we plot the same data, giving a different set of cross-sections of the three dimensional surface relating V, E.R., and strategy. Here E.R. is fixed, to show V as a function of strategy. The same data are shown three dimensionally in Figure 17. The interpretations to be derived from Figures 15 - 17 are as follows:

a. Among the a priori strategies either b or c has greater utility than a, for every E.R. except 15.3:1. At this point, all classification schemes not employing test data are equally good.

b. As E.R. becomes very small, b crosses above every other strategy; as E.R. becomes very large, c crosses above every other strategy. Thus at extreme E.R. the test gives us no gain in utility.

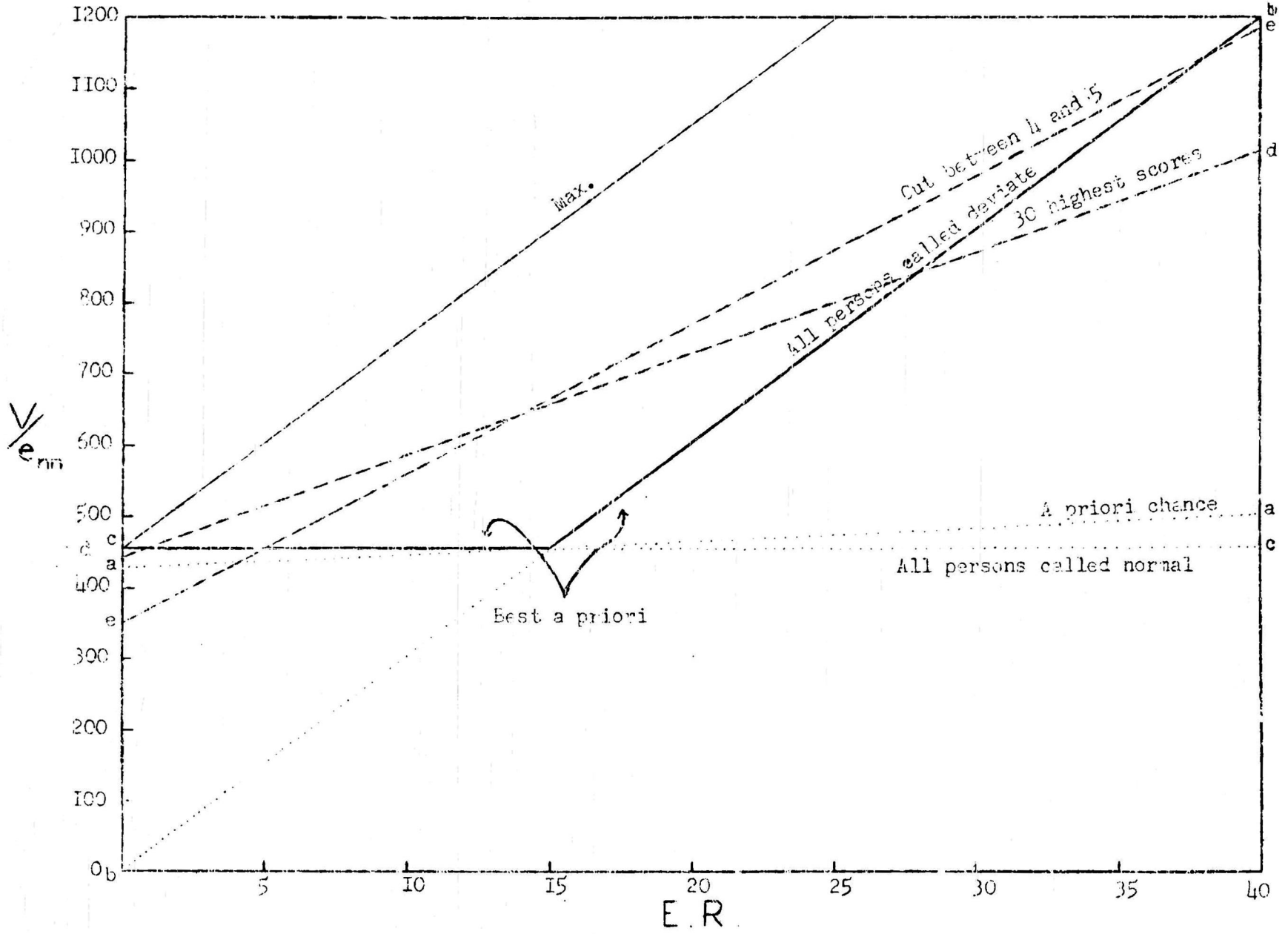


Figure 15. Utilities of various strategies as a function of evaluation ratio.

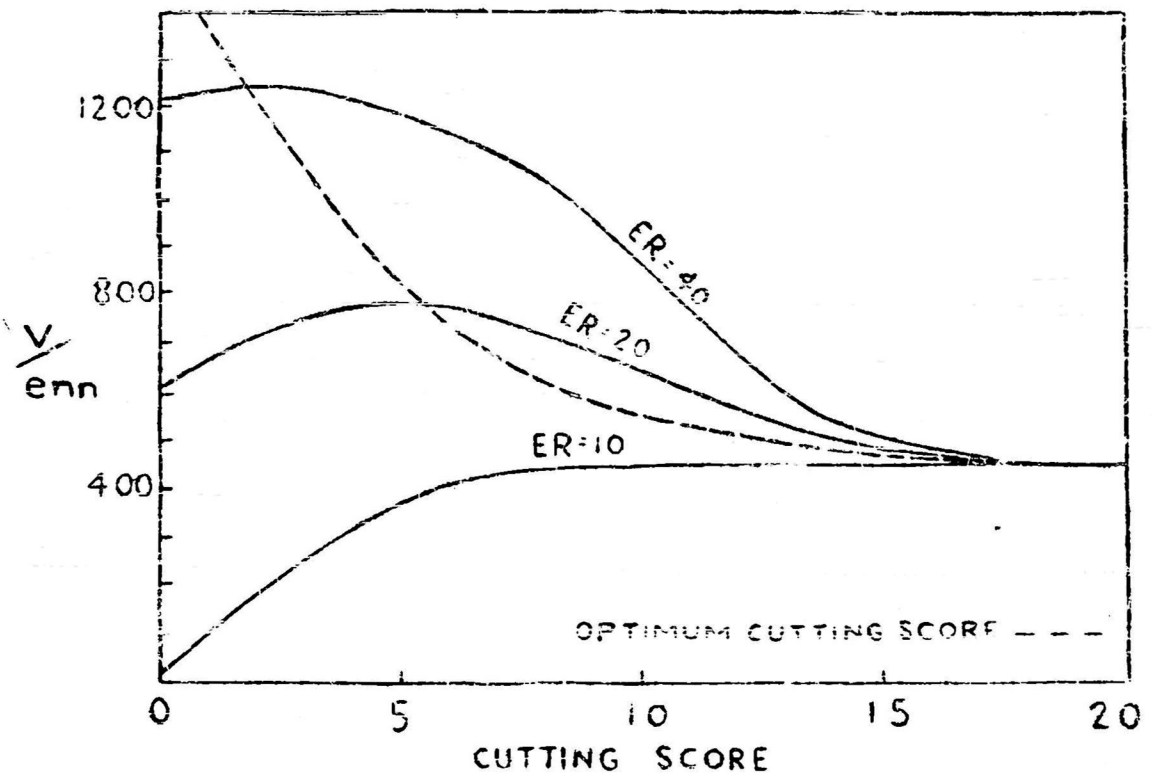


Figure 16. Utilities at various cutting scores for each E.R.

c. As E.R. becomes smaller and smaller, a value of E.R. is reached where decisions based on a particular cutting score give less utility than the a priori chance classification a. For some strategies (including d) this occurs only for negative E.R.

d. Each cutting score gives greater utility than alternative strategies in that section of the chart where E.R. is close to the L.R. corresponding to that score.

All these confirm our previous discussion.

Figure 18 reorganizes some of the same information. Consider how much gain over the best a priori strategy the test provides, and divide by the possible gain. This gives a relative utility (RV) for each strategy.

$$RV = \frac{V - V_o}{V_{max} - V_o}$$

V_o is the maximum utility obtainable before testing (strategy b or c); V_{max} is N_d E.R. + N_n e_{nn}. RV is comparable to the concept of exhaustiveness in Sections II and III.

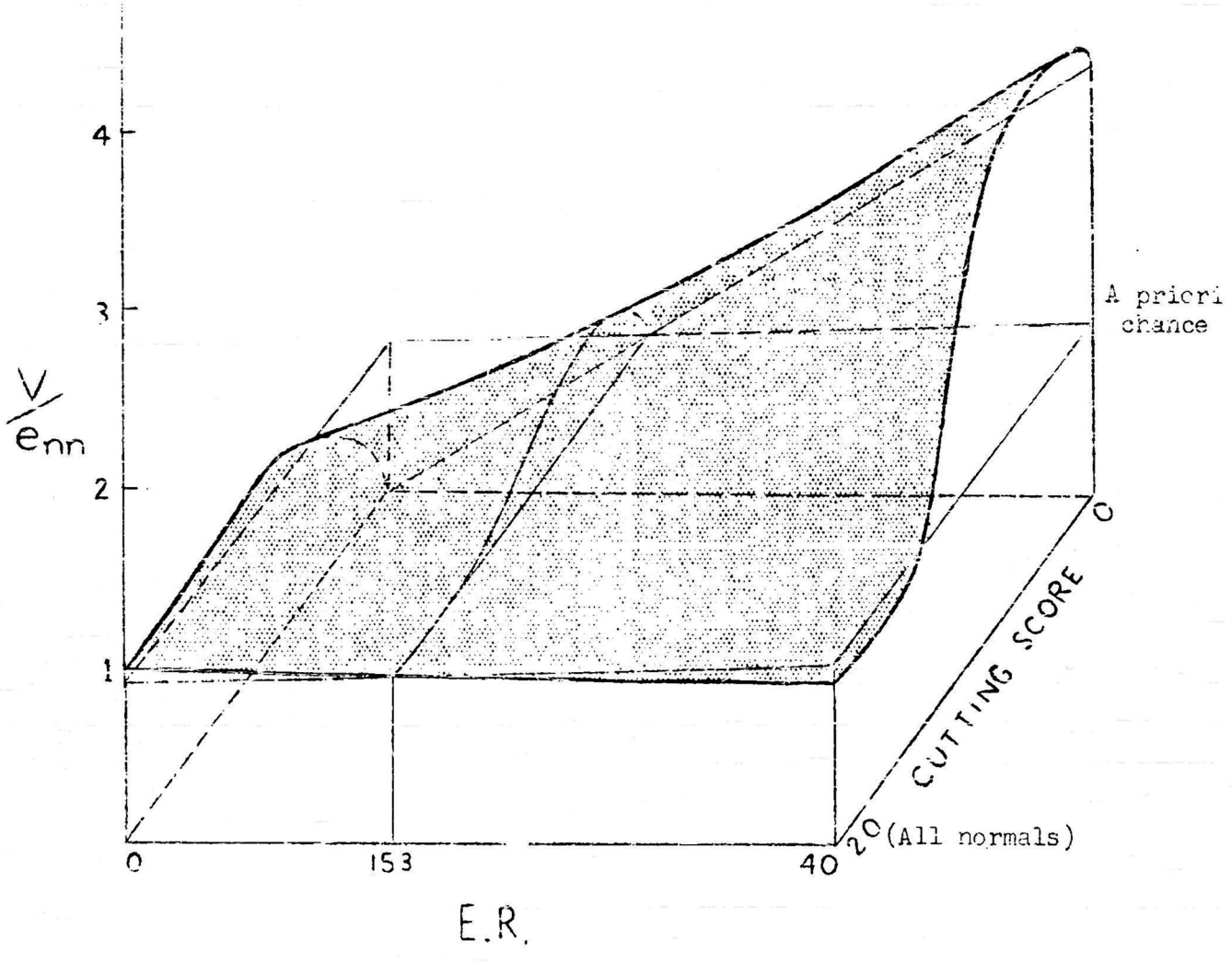


Figure 17. Utility as a function of E.R. and cutting score.

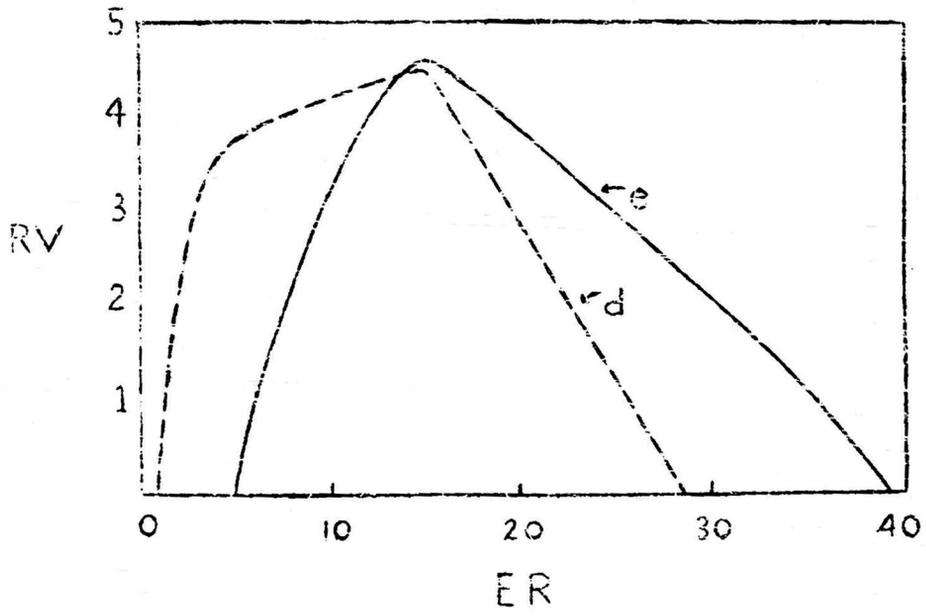


Figure 18. Improvement relative to possible improvement for three strategies

Figure 18 plots relative utilities for strategies d and e. It is apparent that at extreme E.R. the screening test does not improve the net worth of decisions, and that a given cutting score will have quite different exhaustiveness depending on the E.R. Each strategy has its greatest relative utility when E.R. corresponds to the proportion of normals to deviates in the population.

Misses are usually far more expensive in psychiatric screening than false positives, at least so long as we are dealing with unselected recruits rather than specialists. In view of the trouble a misfit or breakdown can cause, E.R. may well be 20:1 or 40:1. Our judgment of the value of the test will be much less favorable if E.R. is very high than if E.R. is moderate. Decisions about E.R. are of major importance in test evaluation. It appears that investigators are unwise to dismiss this problem casually by assuming misses equal in seriousness to false positives (cf. 2).

Limited Strategies

Instead of having the unlimited strategies so far considered, a tester may be restricted by practical circumstances. He would often be required to accept some minimum number of men, and then strategy b, and possibly also c would be disallowed. Our formulation can handle this problem, and the results are informative.

Suppose that at least 400 men out of 488 must be passed. Then c is the best a priori strategy for E.R. < 15.3. The best a priori strategy if E.R. > 15.3 is to reject just as many men as we are allowed to reject. The a priori utility so attained will be less than when the tester is allowed to

reject everyone; as a consequence the test used with any particular cutting score can contribute more at any high E.R. than it did when strategies were unlimited. On the other hand, cutting scores below 6, which are the most promising for high E.R., are not now allowed.

When we are required to reject 6.35% (30/488) of the recruits, neither more nor less, we may use only strategy a or d. Then and only then does this test permit improvement on the allowable a priori decisions, for all E.R. greater than 1.

Reduction of Costs

An approach is available which makes tests useful even where they might otherwise not contribute to utility. Suppose, using the test for screening, we estimate E.R. to be 40:1. Then the test is not more advantageous than rejecting all men, supposing that this is allowed. If we can reduce E.R., however, the test may nonetheless be useful.

This ratio, $(e_{dd} - e_{dn}):(e_{nn} - e_{nd})$, can be reduced in these ways:

by raising e_{nn}

by raising e_{dn} (decreasing the cost of misses)

by lowering e_{dd}

by lowering e_{nd} (increasing the cost of false positives)

If we put the men called normal by the test through a further screening procedure having some validity, we divide the apparent normals into accepts-still-called-accept, and accepts-now-called-reject. This detects some deviates who would originally have been missed. Thus we raise e_{dn}, the average value of the group of deviates called normal by the original test. The second screen lowers e_{nn} and e_{dn} by the cost of the screening procedure, which must therefore not be too great. No strategic change to raise e_{nn}, to lower e_{dd}, or to lower e_{nd} suggests itself.

The relative utility curves (Figure 18) show that there exists some E.R. for which RV is maximal. Theoretically, at least, for any testing situation it is possible to adjust E.R. to the value where the test is maximally effective, as judged by RV. It is ordinarily more reasonable, however, to devise a test to fit the situation.

The foregoing paragraph suggests that when the cost of a psychiatric interview for everyone is prohibitive, and when E.R. is high, it is advisable to use the psychiatrists' limited time to rescreen the men accepted, not the men judged to be deviates by the test. A test might be of no utility when employed to make a final accept - reject decision, if the cost of misses is very high. But by employing a second screen to reduce E.R. by cutting the cost of misses, we move into a region of Fig. 15 where the test can make a positive contribution.

This is consistent with the general principle that a fallible test may be profitably used in a tentative decision which can be reversed on the basis of further evidence, even though the test is too undependable to use in final decisions.

Conclusions

This paper has demonstrated how utility analysis would investigate whether the NDRC screening test contributes to the goodness of decisions in psychiatric screening of recruits. We show that

a. The optimal cutting score can be rigorously determined, if data on costs of various decisions are obtained.

b. A test which allows better-than-chance decisions may not permit improvement over the best a priori decisions, unless allowable strategies are markedly restricted.

c. For any particular cutting score, there are some evaluation ratios at which the test makes no contribution.

d. A test may have value for preliminary screening even though it does not have utility if employed as a final screen.

These conclusions are based on a treatment of the criterion as a dichotomy.

Our purpose has been to illustrate how utility theory makes explicit many assumptions and relations not readily recognized heretofore. We have seen that taking evaluations into account permits sounder analysis than correlational treatment, or an assumption that false positives and misses have equal value. Furthermore, in judging the usefulness of a test, one must consider how it is to be employed.

The emphasis on a "dollar criterion" makes clear that studies of test validity must in the future give considerable attention to cost data, in order to determine evaluations as exactly as possible. Seat-of-the-pants estimates of evaluation ratios -- or even worse, blind assumptions that all errors have equal cost -- can lead to quite costly errors of judgment in deciding whether to use a test, whether to use it for preliminary or final screening, and what cutting score to employ.

References

1. Angell, G. W. The effect of immediate knowledge of quiz results on final examination scores in freshman chemistry. J. educ. Res., 1949, 42, 391-394.

2. Barry, J. R., and Raynor, G. H. Psychiatric screening of flying personnel: research on the Cornell Index. USAF School of Aviation Medicine, Project No. 21-0202-0007, Report No. 2, Randolph Field, Texas, 1953.

3. Bechtoldt, H. P. Selection. In S. S. Stevens (ed.). Handbook of experimental psychology. New York: Wiley, 1951. Pp. 1237-1266.

4. Brogden, H. E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. J. Educ. Psychol., 1946, 37, 65-76.

5. Brogden, H. E. When testing pays off. Personnel Psychol., 1949, 2, 171-183.

6. Brogden, H. E., and Taylor, E. K. The dollar criterion--applying the cost accounting concept to criterion construction. Personnel Psychol., 1950, 3, 133-154.

7. Coombs, C. H. Mathematical models in psychological scaling. J. Amer. Statist. Assn., 1951, 46, 480-489.

8. Coombs, C. H. A theory of psychological scaling. Ann Arbor, Mich: Univer. of Mich. Press, 1952. Engineering Research Bulletin No. 34.

9. Coombs, C. H. On the use of objective examinations. Educ. Psychol. Measmt., 1953, 13 (2), 308-310.

10. Conrad, H. Information which should be provided by test publishers and testing agencies on the validity and use of their tests. Proceedings, 1949 invitational conference on testing problems. Princeton: Educational Testing Service, 1950. Pp. 63-68.

11. Cronbach, L. J. A generalized psychometric theory based on information measure. Urbana, Ill.: Bureau of Research and Service, College of Education, University of Illinois, 1952. (Mimeographed)

12. Cureton, E. E. Validity. In E. F. Lindquist (ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951. Pp. 621-694.

13. Dailey, C. A. The practical utility of the clinical report. J. consult. Psychol., 1953, 17, 297-302.

14. Evans, R. N. A suggested use of sequential analysis in performance acceptance testing. Urbana, Ill.: Univer. of Ill., 1953. Technical report under ONR contract N6ori-07142. (Mimeographed)

15. Ferguson, G. A. On the theory of test discrimination. Psychometrika, 1949, 14, 61-68.

16. Fisher, R. A. Theory of statistical estimation. Proc. Cambridge Philos. Soc., 1925, 22, 700-725.

17. Gage, N. L. Judging interests from expressive behavior. Psychol. Monogr., 1952, 66, No. 18 (Whole no. 350).

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

19. Hick, W. E. Information theory and intelligence tests. Brit. J. Psychol. (Statist. Section), 1951, 157-164.

20. Kelly, E. L., and Fiske, D. W. The prediction of performance in clinical psychology. Ann Arbor, Mich.: Univer. of Mich. Press, 1951.

21. McGill, W. J. Multivariate transmission of information and its relation to analysis of variance. Cambridge: Mass. Inst. of Tech., 1953. Human Factors Operations Research Laboratories Report No. 32, Research Laboratory of Electronics.

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

23. Pollack, I. The assimilation of sequentially-encoded information. I: Methodology and an illustrative experiment. Washington: Bolling Air Force Base, 1952. Human Resources Research Laboratories Memo Report No. 25.

24. Quastler, H. (ed.) Information theory in biology. Urbana, Ill.: Univer. of Ill. Press, 1953.

25. Shannon, C. E. Communication in the presence of noise. Proc. Inst. Radio Eng., 1949, 37, 10-21.

26. Shannon, C. E., and Weaver, W. The mathematical theory of communication. Urbana, Ill.: Univer. of Ill. Press, 1949.

27. Shipley, W. C., and Graham, C. H. Final report in summary of research on the personal inventory and other tests. Applied Psychology Panel, Project N-113, Report No. 10. OSRD Rep. No. 3963; Publ. Bd. No. 12060. Washington: U. S. Dept. Commerce, 1946.

28. Stalnaker, J. M. Personnel placement in the armed forces. J. Appl. Psychol., 1945, 29, 338-345.

29. Thurlow, W. R. Direct measures of discriminations among individuals performed by psychological tests. J. Psychol., 1950, 29, 281-314.

30. Von Neumann, J., and Morgenstern, O. Theory of games and economic behavior. Princeton: Princeton Univer. Press, 1947.

31. Wald, A. Statistical decision functions. New York: Wiley, 1950.