1982

An Application of Information Theory in the Development of a Reliability Coefficient for Criterion-Referenced Mastery Tests

Richard E. Sherman
Loyola University Chicago

AN APPLICATION OF INFORMATION THEORY IN THE DEVELOPMENT OF A RELIABILITY COEFFICIENT FOR CRITERION-REFERENCED MASTERY TESTS

by

Richard E. Sherman

A Dissertation Submitted to the Faculty of the Graduate School
of Loyola University of Chicago in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

April
1982

ACKNOWLEDGMENTS

The author would like to acknowledge the invaluable assistance, support, and suggestions of his doctoral advisor Dr. Samuel T. Mayo, and the other members of his committee, Drs. Jack Kavanagh and Ronald Morgan. Gratitude is expressed to Science Research Associates, Inc., and specifically Ms. Rita Bode, for their cooperation in supplying the data used in this study. The author would further like to thank Mr. William Kadlec of Loyola University for his assistance in writing the computer programs required in the data analysis. Finally, the author would like to acknowledge the assistance of Mrs. Diane Lomedico and Ms. Beth McGarry in typing and revising the initial drafts and final copy.

VITA

The author, Richard Edward Sherman, is the son of Victor Edward Sherman and Ann (Wagner) Sherman. He was born December 10, 1946, in Chicago, Illinois.

His elementary education was obtained at St. Margaret Mary Catholic School in Chicago, Illinois, and secondary education at St. George High School, Evanston, Illinois, where he graduated in 1964.

In September, 1964, he entered Loyola University of Chicago, and in May, 1969, received the degree of Bachelor of Arts with a major in philosophy. In May, 1974, he was awarded the Master of Education in research methodology. In 1977, he was selected to be a member of Alpha Sigma Nu, the National Jesuit Honor Society. In June, 1979, he was retained as Coordinator of Research and Evaluation at Chicago's Alcoholic Treatment Center.

He has published two articles in collaboration with James H. Bryan, Ph.D., Northwestern University, which appeared in the 1980 volume of the _Learning Disabilities Quarterly_. He has had a third article accepted for future publication in the _Journal of Alcohol and Drug Education_.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ii

VITA iii

LIST OF TABLES viii

LIST OF FIGURES ix

Chapter

I. INTRODUCTION 1

Norm-Referenced (NR) Versus Criterion-Referenced (CR) Test Reliability 5

The Concept of Test Reliability . . 5

Some Theoretical Considerations of Classical Test Reliability . . . 7

A Theoretical Definition of Test Reliability 13

An Operational Definition of Classical Test Reliability . . . 15

Implications of Operational Definition of Test Reliability . 17

Reliability Estimates of Criterion-Referenced Mastery (CRM) Tests . 18

Implications of CRM Tests for Methods of Reliability Estimation 22

Summary 24

Information Theory as a Basis for Estimating Criterion-Referenced Mastery (CRM) Test Score Reliability 26

Method of Approach to the Solution of the Problem 26

Similarity between Concepts of Reliability and Information . . 28

Summary 32

II. REVIEW OF THE LITERATURE 34

Attempts to Estimate the Reliability of CRM Measures 35

Introduction 35

Suggested Alternative Reliability Estimates for CRM Measures . . . 38
Summary 53
Basic Aspects of Information Theory . . 56
Information Theory and Reliability 58
A Conceptual Definition of Information 60
Information and Uncertainty 62
Statistical Aspects of Information Theory 64
The Statistical Theory of Information 69
The Expected Amount of Information from a Set of Messages 76
Summary 79
Summary of Chapter 80

III. CONCEPTUAL BASIS OF METHODOLOGY 86

The Relationship between Hypothesis and Evidence 89
The Shared Information of Evidence and Hypothesis 93
The Degree of Hypothesis Confirmation by the Strength of Evidence 96
Reliability as Consistency of Mastery/ Nonmastery Classifications 101
Hypothesis Confirmation as "Degree of Covering" 102
Development of Problem Solution - Symbols and Definitions 107
Relationship of Evidence and Hypothesis to True Score and Total Score Variance 110
An Expression for "Degree of Covering" . 111
Development of Formula for Coefficient Iota (i) 113
Comment on the Mathematical Nature of the Formula for Iota (i) 118
Some Philosophical Implications of the Formula for Iota (i). 120
The Range of Possible Values of Coefficient Iota (i). 122
The Minimum Value of Iota (i) as a Measure of Evidential Strength . . . 123

The Maximum Value of Iota (i) as a Measure of Evidential Strength . . . 127

The Minimum Value of Iota (i) as a Mathematical Expression 132

The Maximum Value of Iota (i) as a Mathematical Expression 133

Summary 139

IV. ANALYSIS AND RESULTS 141

Data Base 141

Formulation of Parallel Test Forms . . . 141

Results of Analysis 149

Discussion of Results 151

Summary 166

V. SUMMARY AND CONCLUSIONS 172

Summary 172

Purpose of Study 172

Classical Theory of Test Reliability 173

An Operational Definition of Classical Test Reliability . . . 174

Issue of CRM Measures in Relation to Classical Test Theory 176

Estimation of the Reliability of CRM Classifications 178

Conceptual Similarity between Reliability and Information . . 179

Current Estimates of the Reliability of CRM Classifications 182

Basic Concepts of Information Theory 188

Statistical Definition of Information 189

Conceptual Basis of Methodology . . 191

Relationship between Evidence and Hypothesis 192

Reliability as Hypothesis Confirmation 194

The Concept of "Degree of Covering" 195

Development of Coefficient Iota (i) Ratio 196

The Range of Possible Values of Coefficient Iota (i) 200
Method of Analysis 203
Discussion of Results 204
Conclusions 208

BIBLIOGRAPHY 215

APPENDIX A 221

APPENDIX B 224

LIST OF TABLES

Table | | Page
4.1 Interpretation of Item Discrimination Values 143

4.2 Item Difficulties and Discriminations of 30 Item Pairs 145

4.3 Average Item Difficulties and Discriminations of Parallel Forms 147

4.4 Means, Standard Deviations, and KR-20 Reliabilities of Parallel Forms 148

4.5 Results of Analysis 150

4.6 Results of Subkoviak's Comparison of Four Suggested Reliability Coefficients 152

4.7 Comparison of Parameter Values of P_c and Iota 158

4.8 Amounts and Sources of Uncertainty and Information Involved in the Calculation of the Parameter Values of Iota 162

LIST OF FIGURES

Figure Page

2.1 An Example of the Game of "Twenty Questions" 67

3.1 An Illustration of the Relationship between Evidence and Hypothesis 95

3.2 An Illustration of the Domain and Range of Mastery/Nonmastery States and Classifications 99

CHAPTER I

INTRODUCTION

Many important decisions within the educational enterprise are based on information gained through the evaluation of test results. Such tests are designed with the intent of determining the degree to which a student's behavior has been affected, at least theoretically, by a particular type of learning experience within the school environment. Glaser and Nitko (1971) state that if such testing is to be justified, in terms of the time and expense required, test results must furnish relevant information on which to base decisions for "the development, operation, and evaluation of education".

However, the decisions made within education can be generally no more accurate than the information on which they are based. Hence, within any such endeavor, the matter of the accuracy of obtained measurement, or the degree of "experimental error" present in such measures, is of prime importance. The attempt to determine the degree of accuracy contained in a set of measurements is the concern of the topic of reliability. Although the methods of estimating reliability are varied and can be based on somewhat different conceptual definitions, the desired end product is always a numerical coefficient which is meant to serve as an indicator of the general degree of accuracy of a particular measurement instrument; be it test, inventory, or scale.

As was the case with the vast majority of other statistical techniques which emerged out of the early development of what might be termed traditional or classical test theory, procedures of reliability estimation were designed to be conceptually compatible with scores obtained from norm-referenced (NR) tests. The conceptual basis of NR testing is that individual performance or ability is evaluated on the basis of individual relative position within a range of test scores, produced by all similarly defined individuals who have taken the same test. It follows that there can only be variation in individual evaluations if there is variation in individual test performance. And, the greater the degree of individual test score variation, the more reliable the estimations of relative individual positions in the range of test scores. Thus, it is not surprising that traditional procedures of reliability estimation depend upon variation in test scores, and yield coefficients which increase in degree of estimated reliability as test score variation increases.

More recently however, a second type of evaluation procedure termed criterion-referenced (CR) or criterion-referenced mastery (CRM) testing has been developed. This latter approach to testing attempts to evaluate individual performance not on the basis of relative score location within a group of examinees, but rather in terms of individual performance in relation to a particular standard or criterion determined prior to testing. Therefore, individual performance is evaluated without reference to the performance of that individual's fellow examinees.

Both location within a distribution of scores and degree of score variation are thus unimportant in the case of individual evaluation on the basis of scores obtained from CR or CRM measures. As a result, those statistical procedures developed within the framework of norm-referenced (NR) testing are both conceptually and mathematically inappropriate for use with criterion-referenced (CR) and criterion-referenced mastery (CRM) test data. Overall, the purpose of the present research project is to develop a type of reliability estimate to be applied to scores obtained from CR or CRM tests.

This first chapter will be divided into two major sections. The first section will concentrate on the concept of test reliability as it has been traditionally applied in relation to norm-referenced (NR) testing, and, the conceptual and mathematical implications of this traditional approach in regard to criterion-referenced mastery (CRM) test data. The second section will present the basic concepts involved in the approach to be taken in the development of a reliability estimation procedure to be applied to CRM test data.

NORM-REFERENCED (NR) VERSUS CRITERION-REFERENCED (CR) TEST RELIABILITY

The Concept of Test Reliability

As expressed by Ebel (1968), according to traditional test theory the value of a reliability coefficient represents the proportion of the observed variance of scores yielded from a test, which is due to true score variance. That is, a test is the more reliable the less the error variance that is contained in the obtained outcomes of that test. This leads to an inverse relationship between the extent to which individuals' test results are the effect of the positions of those individuals on some hypothetical continuum, and the extent to which those test results are affected by extraneous, or error producing, variables.

For example, if a teacher attempts to evaluate the mathematical achievement of his or her students by means of a test, the hope is that the scores obtained on that test will be more the result of the true mathematical ability of those students, and less a result of various unrelated extraneous variables. The errors of measurement which result from these extraneous variables are assumed to be random, and can result from a number of unrelated factors. Kerlinger (1973) identifies some of the sources of errors of measurement as:

> the ordinary random or chance elements present in all measures due to unknown causes, temporary or momentary fatigue, fortuitous conditions at a particular time that temporarily affect the object measured or the measuring instrument, fluctuations of memory or mood, and other factors that are temporary and shifting. (p. 443)

Because the above sources of measurement error are random, and can be manifested in a particular score to varying degrees, any measure of the accuracy of a set of test scores will necessarily be an estimate. Hence, the numerical index previously mentioned, which is termed a reliability coefficient and is meant to serve as an indication of the degree of accuracy of a set of test scores, is an estimate.

It is true that various means have been developed with the purpose of yielding an estimate of the reliability of scores obtained from a particular test. Therefore, one might ask why another such approach need be developed. The task at hand is necessitated due to the fact that the "traditional" means of determining test reliability are inappropriate when applied in the case of criterion-referenced mastery tests. To see why this is the case, it is necessary to first examine the manner in which test reliability has been "traditionally" defined, both theoretically and operationally.

Some Theoretical Considerations of Classical Test Reliability

On the theoretical side of the issue, our discussion begins with consideration of the fact that any set of measures obtained from a particular instrument has a total obtained variance. It is this concept of obtained variance which is crucial to the problem at hand. Therefore, we will need to develop the concept of obtained variance to fully understand how it relates to a theoretical definition of reliability.

Now, theoretically, each individual score in a particular set of measures is assumed to consist of two components - a true component and error component. This relationship can be expressed by the following equation:

$$X_{O_i} = X_{T_i} + X_{E_i} \qquad (1.1)$$

where: X_{O_i} = an individual's obtained score

X_{T_i} = an individual's true score, which is a function of that individual's position in some hypothetical continuum.

X_{E_i} = that portion of an individual's obtained score which is due to random error - this effect can be either positive or negative.

The next step in calculating the obtained variance of a set of measures, would be to subtract the arithmetic average, or mean, of the set of measures, from each individual measure. In the case of our equation, in order to maintain algebraic equivalence, the mean would need to be subtracted from both sides of the equality, thus yielding:

$$X_{O_i} - \overline{X}_{O_i} = X_{T_i} + X_{E_i} - \overline{X}_{O_i} \qquad (1.2)$$

where: X_{O_i} = the mean of the set of obtained measures.

Since $\overline{X}_{O_i}$ represents the arithmetic average of a set of obtained scores, each of which is made up of a true score and an error score, this set of obtained scores could theoretically then be separated into a set of true scores and a set of error scores. It would then be a simple matter to compute the mean of the set of true scores and the mean of the set of error scores. Hence, we see that the mean of the set of obtained scores is itself made up of the combination of two means - the mean of the set of true scores and the mean of the set of error scores. Substituting this alternate expression for $\overline{X}_{O_i}$ into the right side of Equation 1.2, and arranging like terms, we have:

$$(X_{O_i} - \bar{X}_{O_i}) = (X_{T_i} - \bar{X}_{T_i}) + (X_{E_i} - \bar{X}_{E_i}) \tag{1.3}$$

where: $\bar{X}_{T_i}$ = the mean of the set of true scores.

$\bar{X}_{E_i}$ = the mean of the set of error scores.

The value $(X_{O_i} - \bar{X}_{O_i})$ is termed the deviation score of individual i. Such a score simply represents the distance in score units of an individual's obtained measure, from the mean of the entire set of obtained measures to which that particular individual's score belongs. The next step in computing the obtained variance would be to square each of these individual deviation scores. Thus, squaring both sides of Equation 1.3, we have:

$$(X_{O_i} - \bar{X}_{O_i})^2 = \left[(X_{T_i} - \bar{X}_{T_i}) + (X_{E_i} - \bar{X}_{E_i})\right]^2 \tag{1.4}$$

Multiplying out the right side of Equation 1.4, we have:

$$(X_{O_i} - \bar{X}_{O_i})^2 = (X_{T_i} - \bar{X}_{T_i})^2 \tag{1.5}$$

$$+ 2(X_{T_i} - \bar{X}_{T_i})\,(X_{E_i} - \bar{X}_{E_i}) + (X_{E_i} - \bar{X}_{E_i})^2$$

The calculation of the obtained variance is then actualized by summing these individual squared deviation scores across the n individuals in a particular group, and dividing these summed deviation scores by n. Performing these two operations on Equation 1.5, and separating the terms on the right side of the equation, the following equation results:

$$\frac{\sum (X_{O_i} - \bar{X}_{O_i})^2}{n} = \frac{\sum (X_{T_i} - \bar{X}_{T_i})^2}{n} \qquad (1.6)$$

$$+ \frac{2 \sum (X_{T_i} - \bar{X}_{T_i})\,(X_{E_i} - \bar{X}_{E_i})}{n}$$

$$+ \frac{\sum (X_{E_i} - \bar{X}_{E_i})}{n}$$

where: n = the number of individual scores in the particular group on which the obtained variance is calculated.

$\sum$ means "take the sum of".

With Equation 1.6, we then have the final formula for the calculation of the obtained variance of a set of scores or measures expressed in the left-hand side of the equation. Looking at the right-hand side of Equation 1.6, the first and third terms are also recognized as statistical expressions of variance. The first term represents the variance of the set of true scores for the particular group of individuals, while the third term represents the variance of the set of error scores for the same group. It is the middle term of the equation immediately above which at first presents some difficulty in interpretation. To explain the next step, it is necessary to here point out an assumption of the theory of reliability.

Test theorists assume that the correlation between true scores and error scores is zero. Stated conceptually, this assumption posits that there is no relationship between the true scores and error scores for either an individual or a group. Taken either individually or groupwise, even if the true score of an individual, or the set of true scores of a group, were known, this knowledge would be of no aid in predicting the error score or set of error scores that would be associated with the respective true score or set of true scores.

The statistical result of the above assumption is that if the correlation between two variables is zero, the sum of the cross products of individual scores from their respective group means will be zero, when those cross products are taken across the entire population. If this is the case, the numerator of the middle term on the right side of Equation 1.6, which contains such a sum of cross products, would be equal to zero, and hence, this middle term would drop out of the equation. It should be recalled that at this stage we are still speaking theoretically, and it is assumed that the middle term of Equation 1.6 will equal zero when measures are made either on the entire population of individuals in question, or, an infinite number of measures are obtained on a particular individual using the same instrument on each occasion.

It is not within the scope of this thesis to explore the validity of the assumption that the relationship between true scores and error scores is zero. However, the interested reader is directed to Magnusson (1967) for a more detailed discussion of this assumption and its additional implications.

Returning to the purpose at hand, with the cancellation of the middle term on the right side, Equation 1.6 becomes:

$$\frac{\sum (X_{O_i} - \bar{X}_{O_i})^2}{n} = \frac{\sum (X_{T_i} - \bar{X}_{T_i})^2}{n} + \frac{\sum (X_{E_i} - \bar{X}_{E_i})^2}{n} \qquad (1.7)$$

For purposes of brevity, Equation 1.7 can be stated thusly:

$$V_O = V_T + V_E \qquad (1.8)$$

where: V_O = variance of obtained scores.

V_T = variance of true scores.

V_E = variance of error scores.

From Equation 1.8, it can then be seen that the obtained variance of a set of measures for a particular group can itself be theoretically partitioned into two other variances - the variance of the true score for that same group, and the variance of the respective error scores.

A Theoretical Definition of Test Reliability

First, in theory, test reliability is defined to be the ratio of true score variance to observed score variance. This definition can alternately be interpreted as the proportion of observed score variance which is made up of true score variance, and can be expressed by the formula:

$$r_{tt} = \frac{V_T}{V_O} \qquad (1.9)$$

where: r_{tt} = the reliability coefficient.

Equation 1.8 can be algebraically manipulated to yield a second equivalent expression of reliability; that is:

$$r_{tt} = 1 - \frac{V_E}{V_O} \qquad (1.10)$$

Conceptually, Equation 1.10 defines reliability as unity minus the proportion of observed variance which is made up of error variance.

The range of numerical values of r_{tt} can be determined upon examination of Equations 1.9 and 1.10. Due to the nature of their respective statistical formulae, neither V_T nor V_E can be negative. Indeed, this is true of any measure of variance. Hence, from Equation 1.10, it is seen that the range of values for both V_T and V_E is from zero to the value of V_O.

Now, if all of the observed variance is made up of true score variance, V_T would equal V_O, and from Equation 1.9, we see that r_{tt} would equal unity. Therefore, unity represents the upper limit of the range of possible values of r_{tt}. Such a result agrees conceptually with our general understanding of the reliability of a set of measures. For, the reliability of an instrument which yields a set of measures, can be considered perfect, and possess an r_{tt} equal to one, only if that set of measures contains no errors and, therefore, V_E equals zero.

On the opposite extreme, if the set of measures obtained from a particular instrument are the result entirely of errors, V_E would equal V_O, and from Equation 1.10 it is seen that such an instrument would have an r_{tt}

equal to zero. Therefore, from a theoretical standpoint, r_{tt} can range from zero to unity.

A further point of interest, to which this discussion will return later, is that the theory of test reliability reveals an inverse relationship between the value of r_{tt}, and the value of V_E. That is, as the degree of error contained in the observations yielded by a particular instrument increases, and thus V_E increases, the reliability coefficient associated with that instrument, r_{tt}, decreases toward zero. Conversely, as the degree of error contained in such measurements decreases, r_{tt} increases toward unity. Such a case is, of course, compatible with the common sense notion by which the accuracy of an instrument should be judged.

An Operational Definition of Classical Test Reliability

The theoretical definition of reliability developed above could not be employed in practice since the value of an individual's true score as measured by a particular instrument is never known. It follows then, that an operational definition of test reliability is needed. Ebel (1972) defines test reliability operationally, as follows:

> The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores

> on an equivalent test obtained independently from the members of the same group. (p. 410)

Therefore, to actually estimate the degree of reliability of a particular instrument, two sets of scores obtained independently for the same group of individuals would first need to be procured. The reliability coefficient associated with the particular instrument would then simply be the correlation coefficient, which is the index of the degree of relationship between those two sets of scores. Two such sets of scores are obtained generally by three "traditional" means: 1) have the individuals retake the same instrument; 2) thru the administration of an "equivalent" form of the test; or, 3) subdivide the items on the particular test into two or more equivalent portions.

A sidenote of interest here is the slight difference, in semantics alone perhaps, between the theoretical and operational definitions of reliability. In theory, the degree of reliability possessed by an instrument depends upon the amount of error contained in the resulting measurements. When viewed operationally as the relationship between two sets of scores, reliability is best seen as synonymous with consistency. The assumption here then, is that the more consistent the repeated measures yielded by an instrument, the more accurate that instrument. The notion in this latter interpretation being, that if there is a relatively large amount of variability among repeated measurements of the same object by the same instrument, that instrument cannot be considered very dependable. The idea of reliability considered as consistency, will be of future importance.

Implications of Operational Definition of Test Reliability

This discussion now turns to consideration of the operational definition of reliability as the correlation coefficient between two independent and equivalent sets of scores of the same group. The fact that the coefficient of reliability has been traditionally considered as a correlation coefficient results in an effect which is the direct cause of the research problem at hand. In general, the relative size of any correlation coefficient is affected by the range of talent of the scores upon which that index is calculated. Range of talent is simply the distance, in score units, from the lowest score in a particular group to the highest score. Other things being equal, as the scores within each of two groups increase as to the degree to which they vary from one another, or as the variances of the sets of scores increase, the correlation coefficient calculated as the index of relationship between those two sets of scores will likewise increase.

As a practical example, the reliability coefficient estimated for a particular test from scores obtained from administration to a group of sixth grade students will, in general, be smaller than a reliability estimate for the same test, using the same method of test administration, but utilizing scores for fifth, sixth, and seventh grade pupils. What has happened, is that in the latter case, the range of talent has been increased from one grade level to three grade levels. The test remains the same, however the variance associated with the scores obtained from the test in the second case has increased. Hence, we see that the notion of test score variance is essential to both the theoretical and operational definitions of reliability, as those definitions have been understood traditionally.

Reliability Estimates of Criterion-Referenced Mastery (CRM) Tests

The problem under current investigation concerns the development of a reliability coefficient for CRM tests. One might well ask the question why such a pursuit is necessary if several means of estimating test reliability, herein referred to as "traditional" methods, are generally considered acceptable. An answer to this question can be achieved by consideration of the types of scores obtained from CRM tests.

While the various methods of statistical analysis in education and psychology were being developed, since approximately the turn of the present century, the major mode of testing within these two enterprises can be described as norm-referenced (NR). It should therefore be of little surprise that the more traditional means of statistical analysis should be most applicable to NR test data.

In general, NR tests are designed to yield scores which approximate the familiar bell-shaped, normal curve in their distribution. The range of such a group of scores would have a relatively small percentage of observations at the upper end of a score continuum, the majority of scores near the middle, and again, a small percentage at the lower end of the range. Items are chosen for such a test according to their ability to maximize variability between individual responses. Items which nearly all of the individuals taking such a test can be expected to answer either correctly or incorrectly, are considered to be of minor value. Hence, the emphasis of NR tests on maximizing score variability is seen to be compatible with the traditional operationally defined estimate of reliability. In fact, the vast majority of statistical techniques involve the analysis of the variance exhibited in a set of observations.

However, the emphasis in educational testing has currently shifted to what have been labeled as criterion-referenced (CR) tests. Glaser (1963) identified the emphasis of CR tests as the assessment of student behavior in terms of certain well-defined standards of performance. Scores on such tests should provide information pertinent to both the degree of proficiency a particular student has attained with respect to certain criteria of behavior, and the relative ordering of individuals taking the particular test.

Although various definitions of CR tests have followed, as Alkin (1974) states, they seem to share two general characteristics. First of all, test items are selected solely on their ability to elicit certain well-defined behaviors. The effects of the responses to a particular item on score variability within a group, of prime importance in the case of an NR test, is of little or no importance with a CR test.

Secondly, individual performance is assessed in light of a specified criterion. For example, it may be the case that an individual must be able to answer correctly

90 percent of the items on a particular test before a judgment can be made that that individual has successfully acquired the defined behaviors of interest. In contrast to an NR test, levels of performance may not be determined until after test scores have been collected. As a practical example, if a teacher has decided to "mark on the curve", the percentage of correct responses on a test which represents a performance level worthy of an "A", cannot be determined until after the test has been administered.

Therefore, as Millman and Popham (1974) assert, variability is an unnecessary characteristic of CR tests. The primary purpose of a CR measure is to assess the current status, either before or after some method of instruction, in regard to a particular domain of well-defined tasks. Such a set or class of specified tasks is considered a universe from which the items on a particular CR test represent a random or stratified random sample from that universe. An individual's score on such a test represents an estimate of the individual's true score on the entire universe of tasks. Hence, the familiar case results of attempting to estimate a parameter from a value obtained by random sampling. The degree to which these estimates vary from individual to individual is irrelevant.

Implications of CRM Tests for Methods of Reliability Estimation

As stated previously, the particular problem at hand concerns a reliability estimate made on a CRM test. Several definitions of mastery tests and testing have been offered (Bloom, 1968, 1973; Mayo, 1968; and Harris, 1974a). However, they all indicate that a CRM test is a CR test administered at the conclusion of a particular educational treatment, and is meant to determine the extent to which an individual has attained the tasks identified in the objectives of that particular treatment. A standard is set prior to testing, representing a cut-off point in respect to which decisions are made as to whether an individual has either mastered or not mastered the specified tasks. Upon being evaluated as having mastered the tasks specified by an educational treatment, an individual then would move on to the next higher level of tasks in such a program. If an individual fails to score at or above the cut-off point of the CRM test, he/she would then receive further instruction at that same level, and then be retested. This procedure can be repeated until an individual is adjudged to have mastered the tasks corresponding to a particular level of such an educational program. Hence, one can see the importance of being able to estimate the accuracy of the tests employed in the above situation; for the scores obtained through administration of a test-retest or equivalent forms format, provide perhaps the sole information upon which a mastery or nonmastery decision is made.

At first glance, the notion of reliability as it concerns CRM testing does not seem different from the original theoretical definition provided earlier in this discussion. Indeed, Osburn (1968) has stated that reliability is the procedure for determining the accuracy of an estimate of a person's true score on a universe of items. And likewise, we have seen that the score on a CRM test can be considered to represent an estimate of an individual's true score on some universe of items, by means of a random sample of items from that universe. Of closer proximity to both the previously stated operational definition of reliability, and a mastery testing program, Millman (1974) defines reliability as the consistency of estimates regarding a tested individual's "level of functioning".

However, the problem created by applying the traditional means of estimating test reliability should already be apparent; that is, that test score variance is irrelevant to CR measures in general. And in the case of a CRM testing program, the number of items on a test can be relatively few. Hence, it would not be out of the ordinary that upon completion of an educational treatment in such a program, the majority of tested individuals might attain perfect or near-perfect scores. In this case, score variability would be quite low, or perhaps even nil. If traditional means of reliability, with their dependence on test score variance, were used in the above context of testing, a small range of talent would result in a calculated reliability coefficient of close to, or perhaps even equal to, zero. Therefore, a CRM test may be accomplishing its intended purpose of accurately and consistently estimating an individual's true score on a universe of items, yet yield a very low coefficient of reliability when the traditional means of calculation are used.

Summary

In summary then, it has been demonstrated that the traditional means of estimating test reliability are inappropriate when applied to CRM tests because of the likely lack of sufficient variability manifested by groups of scores obtained from such tests. On a more philosophical point, it may likewise be inappropriate to estimate the reliability of CR measures in general, by means dependent on score variance, when such variance has been shown to be irrelevant to the intended purpose of such tests.

The next section attempts to serve as an introduction to a suggested solution to the problem of estimating the reliability of CRM tests, which does not depend on test score variance.

INFORMATION THEORY AS A BASIS FOR ESTIMATING CRITERION-REFERENCED MASTERY (CRM) TEST SCORE RELIABILITY

Method of Approach to the Solution of the Problem

In a discussion of CR measures, Harris (1974) suggests two modes of problem-solving which are perhaps applicable to any area. First, one would attempt to identify and experimentally apply any already existing adaptable solutions. Economically speaking, in terms of both time and material resources, such a method should be that initially applied in any temporal sequence of problem-solving. Upon demonstrated failure of this first approach, the second mode of attack would be an attempt to create a new solution. One of the purposes of Chapter II will be to demonstrate that already existing solutions have been applied to the problem of estimating the reliability of CRM tests, and that for various reasons these attempts have proven inadequate. The purpose of this section is then, to introduce a new approach to the above problem, and demonstrate its conceptual appropriateness.

Robert L. Thorndike (1951) has stated that if one is interested in what purposes are to be served by measuring the reliability of a test, one must first analyze what is to be accomplished by such a test. This notion is referred to later by Stanley (1971) as the logical aspect of the study of reliability in educational measurement. A second aspect which Stanley mentions is a statistical one. From this latter perspective, methods of data collection and statistical analysis must be developed so that they are logically consistent with the inferences that are to be made with the calculated values. As seen in the previous section, score variance is irrelevant in the case of CRM tests in general. Therefore, a statistical analysis dependent on such variance, as is the traditional reliability coefficient, would seem to be logically inconsistent with the inferred purpose of CRM measures. The inference of interest concerns whether or not a particular individual has mastered the specified tasks related to a particular educational treatment. The extent to which that individual varies from his/her peers who have also taken the test, is of little or no concern.

As mentioned earlier in this discussion, test scores are the major source of information upon which educational decisions are based. Now, information in any situation is only as valuable as its accuracy and relevancy warrants. In a system of CRM testing the decision to be made is whether or not an individual student has mastered or attained the behaviors associated with a particular level of instruction. And, the degree of accuracy which accompanies such decisions is dependent for the most part upon the accuracy of the information on which they are based. It is a basic assumption of the approach taken within this paper, that if test scores can be considered as information, an index of the consistency of the information obtained from two independent and equivalent measures applied to the same group of individuals, is synonymous with the traditional notion of test score reliability.

Similarity between Concepts of Reliability and Information

It should be recalled, that the traditional operational definition of reliability is best interpreted in terms of consistency. Hence, an interpretation of test score accuracy in terms of consistency of the information provided by such scores seems to be clear of any conceptual difficulties regarding this point. If a decision is to be made as to the classification of a student as a master or nonmaster of a particular subject content, an estimate of the consistency of the information on which that decision is based, should bear directly upon the degree of accuracy of that decision. Accepting this line of reasoning, a statistical estimate of the information contained in a set of test scores, is of initial importance. Such a statistical expression of information is provided by the field of study termed information theory.

Every scientific process aims basically at the acquisition of information. Information theory assumes that it is valuable to be able to estimate the amount of information contained in a set of observations, termed messages, and provides a mathematical basis to do just that. Information is theoretically considered as something we have obtained from a source, which we did not know before. In an educational setting, the source is considered to be the individual student.

It should be mentioned that whether the information received in an act of communication is correct or incorrect, useful or useless, is irrelevant to a measure of the amount of information obtained. The relationship between the amount of information obtained in a message, and the correctness of that information, can be considered analogous to the relationship between reliability and validity in classical test theory. It is of course necessary in any situation to determine whether the information upon which decisions are to be based is correct, and in fact useful. However, just as a discussion of reliability can be conducted separate from consideration of test validity, an analysis of the consistency of information can proceed apart from attempts to determine the usefulness of that information. This is in no way meant to diminish the obvious importance of knowing whether or not the information obtained is useful. Instead, this researcher suggests that just as test reliability is considered to be a necessary but not sufficient condition for test validity (Gronlund, 1976, p. 106), information must be shown to be consistent before it can be examined for its usefulness.

Returning to the discussion of the nature of information, any act of communication provides information only to the extent that it reduces a condition of ignorance. In a CRM testing situation, the test administration is considered the act of communication and the scores obtained are assumed to provide the information which will remedy our ignorance as to whether or not a particular student has or has not mastered the behaviors relevant to a given level of instruction. The amount of information which can be obtained in a particular situation is determined exclusively by the amount of uncertainty, calculated a priori to the act of communication, concerning the state of affairs under consideration. As this quantity of uncertainty, which is a function of the number of alternatives present in a particular situation, is reduced, the information obtained is increased. The result then, is an inverse relationship between information and uncertainty. As the uncertainty is decreased by the types of responses observed in a particular situation, information increases. Uncertainty is in fact, potential information. The more the uncertainty associated with a situation, the greater the opportunity for information.

The above situation can be compared to the inverse relationship between error variance and the magnitude of the reliability coefficient in traditional test theory. The point of similarity here concerns the manner in which these quantities are viewed in light of educational decision-making.

Traditional test theory assumes that tests which yield generally more consistent results, are considered the more reliable in terms of judgments or decisions to be made, in part, on the basis of those results. In kind, the greater the extent to which the uncertainty contained in a testing situation is reduced, and hence, the more the amount of information which is gained - the greater should be the confidence placed in such test results when employed in a decision-making process.

Although the mathematics of these two approaches will be seen to differ, indeed they must if the obstacle of minimum score variance yielded by CRM measures is to be averted, the attempt of this chapter has been to demonstrate the conceptual similarities between these two approaches to the same problem.

Summary

At the beginning of this chapter it was stated that the purpose of this research was the development of a reliability coefficient to be applied to the decisions resulting from scores obtained from a criterion-referenced mastery (CRM) test. The first major section of this chapter outlined the traditional concepts and statistical definitions of test reliability. Included in this section was an argument as to why the traditional approach to test reliability can be considered inappropriate when applied to CRM measures. The second major section of this chapter has been intended to provide an introduction to the methodology which will be used to formulate a suggested solution to this problem. This new approach has been identified as one which will come from within the framework of information theory. In introducing the approach that this study will take in formulating a possible solution to the problem at hand, emphasis was placed on the attempt to demonstrate a similarity between the concepts of information and reliability. This similarity will be focused upon to a greater degree in the second section of Chapter II.

CHAPTER II

REVIEW OF THE LITERATURE

As was the case in Chapter I, this chapter will consist of two major sections. The first section deals with a review of previous attempts to develop a reliability coefficient, or its equivalent, for CR measures in general, and CRM measures in particular. In the second major section, a description of the statistical aspects and developments of information theory, as related to the present purpose, will be presented.

ATTEMPTS TO ESTIMATE THE RELIABILITY OF CRM MEASURES

Introduction

In a paper presented in 1970, Richard Cox argued that if the idea of CR measurement was to be accepted and be able to be applied to teacher-made tests, alternatives to the traditional statistical approaches to reliability, validity, and item analysis must be developed. Up to this point, statistical techniques were designed to be applied, in the main, to norm-referenced data, which analyzed a pupil's performance relative to the performance of his/her peers. Such statistical techniques seek to account for, or explain, the variance resulting from the responses of a number of individuals to a particular set of stimuli. Alternatives to these traditional means of statistical analysis are required for CR measures since, as seen in the previous chapter, individual performance is evaluated with respect to an a priori stated set of objectives. In the case of a criterion-referenced mastery (CRM) test, the variance yielded by a set of obtained scores may be relatively small or possibly even nil. The result of such a situation would be a low reliability coefficient, when calculated by traditional means, despite the fact that a test may be yielding accurate and consistent estimates of individuals' locations on a particular continuum.

Hambleton and Novick (1973) state that while NR measures aim at a "fixed quota" ranking of individuals, CR measures are in general "quota-free" in terms of selection. This can be seen to be simply another way of expressing the irrelevance of the relative performance of individuals when interpreting the results of a CR measure. In the case of a CRM measure, Hambleton and Novick go on to say that the primary problem is to determine whether a student's true mastery level is greater than the cut-off score specified for the test. The result would be a classification of individuals as either "masters" or "nonmasters" depending upon whether an individual's score was above or below the stated criterion level. Therefore, errors can be of two types; individuals can be incorrectly classified as "masters", or, incorrectly classified as "nonmasters". The need in such a situation is to minimize what Hambleton and Novick term as "threshold loss", or in other words, simply minimize the number of incorrect classifications.

Traditional correlational estimates of reliability will yield an estimate of the amount of error to be taken into consideration when interpreting scores obtained on a particular test. This error estimate is referred to as a standard error of measurement, and can be used to establish a confidence interval within which an individual's true score on the considered measure, can be said to fall with a particular probability.

Now, as stated here previously, Hambleton and Novick mention that whenever variance is restricted, as is the case with a CRM measure, correlational estimates of reliability will be necessarily low. However, the above authors find that a more serious objection to the use of correlational methods of reliability estimation with CRM measures stems from the standard error of measurement which results from this traditional technique.

If one accepts the premise that the reliability of a CRM measure depends upon the degree of "threshold loss" which results from decisions made on the basis of obtained test scores, the traditional correlational method of determining test reliability and an index of standard error is also inappropriate in the case of a CRM measure, because such an application represents an incorrect choice of loss function. For, a traditionally estimated index of standard error is in terms of squared-error loss in the score unit metric, and not in terms of the losses or incorrect decisions made when testees are classified on the basis of those test scores. Put simply, the units in which the standard error is expressed are score points, and do not serve to estimate the "threshold loss" that can be expected to occur through the formulation of incorrect decisions as to the "mastery" or "nonmastery" of individual testees.

This researcher agrees with the above authors that any proposed estimate of the error contained in a CRM measure must be in a dimension which reflects this "threshold loss". Such an approach would likewise appear to reflect Stanley's logical aspect of the topic of reliability referred to here earlier. Due to the type of inference to be made from a CRM measure, the reliability of scores obtained from such measures depends upon the consistency of individual decisions made on the basis of those scores, and not the consistency of the score values obtained.

Suggested Alternative Reliability Estimates for CRM Measures

With the above-noted restrictions in mind, attention is now directed toward suggested alternatives to the traditional approach to the reliability of CRM measures.

A method of estimating the reliability of CRM measures has been suggested by Carver (1970). This coefficient is based on the proportion of individual mastery decisions which remain consistent between parallel forms of a test. Calculation of the coefficient is quite simple, and can be readily obtained from a table of the following type:

		Form B	
		Master	Nonmaster
Form A	Nonmaster	b	a
	Master	c	d

$$r_{tt} = \frac{a + c}{N} \qquad (2.1)$$

where: $N = a + b + c + d$

Such an index possesses the difficulty in interpretation of any proportion or percentage - sample size. Indeed, Crehan (1974) in a discussion of various item-analysis techniques for CRM measures, refers to Carver's coefficient as "crude". This index would appear to best serve the purpose of a quick "thumb-nail" estimate of the consistency of decisions for teacher-made tests.

Livingston (1972) has proposed a reliability coefficient for CR measures which applies the principles of classical test theory. Livingston's index is based on the deviations of scores in a group from the chosen cut-off score, rather than the mean, which is, as seen in Chapter I, the case with a traditionally calculated reliability coefficient. The restriction on such a measure which perhaps comes most immediately to mind, is that the cut-off or criterion score, unlike the mean, is chosen. And the procedures by which this choice is made will almost certainly differ from one measure to the next.

In the case of a CRM measure, Livingston's index of reliability is subject to the problem of possible lack of score variance mentioned earlier. For, if all the examinees happened to score at the criterion level, the calculated r_{tt} would equal 0. Or, if all examinees obtained the same score, and that score was not equal to the criterion level, the resultant r_{tt} would equal 1.00. In either event, the estimated reliability coefficient of the measure would be of no aid in an analysis of the ability of such an instrument to estimate individual true scores, and yield accurate decisions as to mastery or nonmastery.

Swaminathan, Hambleton, and Algina (1974) note that it is to a certain extent conceptually appealing to think of the reliability of a CRM measure as the sum of the proportions of individuals assigned to the same category in a test-retest mastery/nonmastery decision framework. Such a measure would be expressed statistically as:

$$\sum_{i=1}^{k} p_{ii} \qquad (2.2)$$

where: k = the number of mastery states.

p_{ii} = the proportion of the total number of individuals who were assigned to category i on a first testing, and again to the same category on a second testing using a parallel form.

However, as the authors point out, such an estimate does not take into account the agreements which can be expected to occur by chance.

As an estimate of the reliability of CR measures, the above authors propose the use of a coefficient developed earlier by Cohen (1968, 1972). Cohen's K (kappa), as the coefficient is termed, is suggested as an index of the consistency of decisions formulated on the basis of results obtained from parallel forms of a CR test. The index is calculated by the formula:

$$K = (p_o - p_c) / (1 - p_c) \qquad (2.3)$$

where: p_o = the observed proportion of agreement.

p_c = the expected proportion of agreement.

The expected proportion of agreement, p_c, is calculated by:

$$p_c = \sum_{i = 1}^{k} p_{i.}\, p_{.i} \qquad (2.4)$$

where: k = the number of mastery states.

$p_{.i}$ = the proportion of examinees assigned to category i on the first testing.

$p_{i.}$ = the proportion of examinees assigned to category i on a second testing using a parallel form.

As is apparent from the formula, Cohen's K does include an estimate of the proportion of agreement which can be expected to occur by chance.

In addition, K has a range of +1 to -1, with +1 resulting only if there is exact agreement of the marginal proportions between the two testings. The coefficient approaches -1 as the differences between the marginal proportions become more and more extreme. It was demonstrated earlier, in Chapter I, that a traditional reliability coefficient cannot be negative. This presents no great difficulty in the interpretation of Cohen's kappa, since if K equals 0 or is negative, there would most certainly exist more disagreement in the decision process than would be tolerable.

Hence, although the index kappa, proposed by Swaminathan, Hambleton and Algina (1974), possesses the characteristic of being an estimate in the dimension of "threshold loss", it must be noted that the value of K is heavily influenced by certain factors within the decision process. These factors are for example, the manner in which the particular cut-off score was selected, test length, and the characteristics of the particular group in question. The authors quite readily recognize this, and offer that any decision-making reliability of this type is a measure of the consistency of the entire process. The test itself is but one form of input to the process. For that reason, if coefficient K were employed as an estimate of the accuracy of a mastery/nonmastery decision-making process, other information regarding the above factors would need to be reported as well to allow for a meaningful interpretation.

From the perspective of traditional test theory, it would be desirable to have an estimate of accuracy or consistency more specific to the effects of the test itself than to the influences of the particular situation as a whole. However, if one is to remain in the dimension of "threshold loss", the cut-off score and the manner in which it was determined are of prime importance.

As presented above, coefficient kappa (K) requires decision results from two test administrations. However, Huynh (1976) has provided steps by which kappa (K) can be estimated from a single test administration.

Huynh begins by making the familiar assumptions that the items on the test administered are homogenous in nature, that is, attempt to measure the same general type of behavior, have been selected from a larger domain or universe of similar items by a process of random selection, and there exists a cut-off score which provides the criterion on which mastery/nonmastery decisions are formulated. Recalling that coefficient kappa serves as an index of the consistency of mastery decisions, the obtained test mean and standard deviation are inserted into the Kuder-Richardson Formula 21. Now, as the reader is probably well aware, the KR_{21} formula yields a reliability coefficient of the traditional type on the basis of one test administration and the number of correct answers for each of the examinees. The problem of possible lack of variability again surfaces with use of the KR_{21}, and will be commented on shortly.

Upon obtaining a value from the KR_{21} formula, Huynh next proposes using this value to estimate the parameters γ and β of a beta-binomial or negative hypergeometric function. The beta-binomial is a univariate discrete density function (Mood, Graybill and Boes, 1974), where variable x can assume values (0, 1, ..., n). The beta form is used within Bayesian statistics to represent the distribution of prior information in a probability of success format which will in turn yield a posterior distribution with different indexing values (Cox and Hinkley, 1974). Hence, we have a mathematical model which employs a distribution based on test score data to develop a probability distribution of certain categories of success, in our case, mastery is assigned if an individual scores at or above the chosen criterion and denied if one's score is below the criterion.

This beta-binomial distribution is then used in both its univariate and joint density forms to yield estimates of the proportion of individuals classified as masters on both parallel forms and the proportion of individuals classified as masters on either form. The score values obtained on the single administration are combined with the designated cut-off score to yield these proportions, which are then substituted in the following equation to yield an estimate of kappa:

$$K = (p_{11} - p_1^2) / (p_1 - p_1^2) \qquad (2.5)$$

where: p_{11} = the proportion of individuals classified as masters on both parallel forms.

p_1 = the proportion of individuals classified as masters on either one or both forms.

As a practical limitation of this process Huynh notes, as anyone familiar with calculus is well aware, as the number of test items approaches ten or more, the calculations become increasingly tedious if done by hand. In such a situation it would be quite advisable to gain access to a computer.

Huynh goes on to discuss certain factors which influence the relative size of kappa. As expected, the designated value of the criterion score has its effect on the relative value of kappa. If the cut-off is either too small or too high, the proportion of consistent decisions will likely be close to 1. It is of course desirable to have as many consistent decisions as possible, however in either case that consistency is most probably due to the extreme value of the criterion than to the effects of the test itself. At any rate, within these two extremes, Huynh demonstrates that the relative values of kappa increase to a maximum and then decrease as they approach the opposite extreme.

In regard to test length, kappa is seen to increase as items of a homogenous nature are added. This is indeed what occurs in the case of a traditional reliability coefficient. However, as Huynh states, a simple formula does not yet exist which would estimate the increase in kappa as the items on the test were increased by a factor of n. This projection is provided for traditional reliability coefficients by the Spearman-Brown formula.

A final factor discussed, and one which was mentioned earlier, is the effect of test score variability on kappa. The sample data provided by Huynh illustrate a positive relationship between score variability and the relative size of kappa. Therefore, as score variability decreases the relative size of kappa will likewise tend to become smaller. Huynh states that kappa is essentially correlational in nature, and as seen in Chapter I, with a measure of this sort restricted score variability will generally serve to minimize the values of indices of this type.

Therefore, as with the use of coefficient kappa by Swaminathan, Hambleton and Algina (1974) as an index of reliability obtained from the administration of parallel forms, Huynh's kappa as calculated from a single administration is in the dimension of "threshold loss" as suggested by Hambleton and Novick (1973). Hence, Huynh's estimated kappa is situation specific in the same sense as is that calculated from the administration of parallel forms. As a result, for a particular set of data there is no unique value for coefficient kappa, since the value of kappa will change as the mastery criterion level changes. And, as was the case with kappa as proposed by Swaminathan et al., if the value of kappa is to be meaningfully interpreted, situational factors such as test length, score variability, the value of the criterion score, as well as the methods by which it was determined, and the characteristics of the examinees, must also be reported.

An approach to estimating the consistency of mastery/nonmastery decisions from a single administration of a CR measure, which is quite similar to Huynh's suggestion, has been forwarded by Subkoviak (1976). Subkoviak begins by defining "the coefficient of agreement for an individual i as the probability that i is assigned to the same mastery state on parallel tests X and X'." This coefficient of agreement, symbolized as P_c, is the sum of the probabilities of consistent mastery/nonmastery decisions over the two test administrations for individual i, when the criterion score is equal to c. The "coefficient of agreement P_c for a group of N persons" is operationally defined as the mean of these individual coefficients; that is,

$$P_c = \frac{\sum_{i=1}^{N} P_c^i}{N} \qquad (2.6)$$

where: P_c = the coefficient of agreement on the parallel forms for the group of N persons.

P_c^i = the coefficient of agreement for individual i.

c = the value of the criterion score.

N = the number of individuals.

The calculation of P_c^i depends upon the estimation of the probability that individual i's score on test X is greater than or equal to the criterion value. This latter probability, is expressed as $P(X_i \geq c)$, and defined as:

$$P(X_i \geq c) = \sum_{X_i = c}^{n} \binom{n}{X_i} p_i^{X_i} (1 - p_i)^{n - X_i} \qquad (2.7)$$

where: p_i = the probability of a correct item response for person i.

X_i = the score of individual i on test X.

n = the number of items on test X.

c = the value of the criterion score.

Subkoviak employs the proportion of test items answered correctly on test X by individual i, as an estimate of p_i. In this approach, Subkoviak makes the assumptions that the scores for each individual i on tests X and X' are independently distributed and identically binomial in form. For these scores to be binomially distributed, the items must be scored either right or wrong, it must be reasonable to assume that the items themselves are independent of one another in terms of responses, and the probability of a correct response remains constant across all items within each individual i.

As may have been noted already by the reader, there are general similarities between the approach of Subkoviak and that of Huynh. Indeed, while Huynh assumes that the distribution of scores on parallel tests is beta-binomial in form, Subkoviak posits that this distribution is a simple binomial. Since these two distributions are of the same family, it is no surprise that, as Subkoviak states, P_c is a function of coefficient kappa.

There is one difference however between the two estimates which is of interpretive interest. Subkoviak's sample data indicates that as the value of C is changed from a relatively low value to one which is relatively high, P_c ranges from close to 1.00 at the low end, decreases to a minimum somewhere between the two extremes, and then increases back to near 1.00 as C approaches its high extreme. It will be recalled that Huynh's coefficient kappa behaves in an exactly opposite fashion.

This comparative difference should really come as no surprise, since P_c is an index of the proportion of mastery/nonmastery decisions which are consistent between parallel forms. And, as Huynh mentions, when the criterion score is either very low or very high, one can expect consistent mastery/nonmastery decisions. However, as also previously mentioned, either case is of dubious practical worth. Nevertheless, one must keep this difference in mind when comparing estimates on the basis of these two methods.

Since Subkoviak's coefficient of agreement is, as coefficient kappa, in the dimension of "threshold loss", it is to a high degree situation specific. Therefore the factors which were suggested as needing to be reported along with the value of kappa, would likewise need to be reported with the value of P_c. In light of the comparison immediately above, the value of the criterion score and the number of items would be of especial interest.

A comparison of the Swaminathan et al. (1974), Huynh (1976), and Subkoviak (1976) methods for estimating the reliability of CRM tests has been carried out by Subkoviak (1978). This investigation compared the various estimates of P_c yielded by these three techniques. It should be recalled that although coefficient kappa received the major emphasis in the Swaminathan et al., and Huynh approaches, the proportion P_c is estimated in both cases.

Subkoviak estimated P_c from the three above procedures on tests of 10, 30, and 50 items in length; and, with criterion levels of 50%, 60%, 70%, and 80%. Each index produced estimates which were reasonably close to the parameter value of P_c over the various conditions. The Swaminathan et al. procedure yielded estimates possessing a relatively higher standard error. In terms of a recommendation, Subkoviak mentions that the Huynh procedure requires only one testing, has a mathematically sound base, "and produces reasonably accurate estimates, which appear to be slightly conservative for short tests".

As Subkoviak states, the data used in the study referred to immediately above is not of the mastery test type. Scholastic Aptitude Test item responses from 1586 students served as the data base, with items being deleted "on the basis of content, difficulty, and discrimination" to create forms with the varying numbers of items. Such items are clearly more heterogeneous in nature than the items generally found on CRM measures. It should be recalled that these various procedures are based on mathematical distributions which assume homogeneity of item content. The more heterogeneous the items on a test, the greater the likelihood for an increase in score variance.

It remains to be seen what effects restricted score variance will have on these various estimates.

Summary

Two general approaches to the problem of estimating the reliability of CRM measures have been discussed. The approach taken from the perspective of classical test theory encounters the operational difficulty of the possibility of limited test score variance. However, even if this obstacle were to be overcome, there are numerous conceptual problems. The error term associated with such classical or traditional estimates, is in the dimension of squared-error loss. Such an error term does not fulfill Stanley's logical aspect of reliability in that it is inconsistent with the type of inference which is to be made from the information contained in the results of CRM measures. The decision to be made from such information is whether or not an individual has mastered a particular content area. An estimate based on the variance of scores among individuals is irrelevant in a case where the decision to be made is whether or not a particular individual's obtained test score has correctly placed him or her, above or below a specified criterion level.

An estimation of the accuracy with which individuals have been classified as masters or nonmasters must be concerned with the number of false "positives" and false "negatives" in relation to a chosen criterion level or cut-off score. Hambleton and Novick (1973) have referred to this dimension as "threshold loss". In terms of Stanley's logical aspect of reliability, the notion of "threshold loss" seems conceptually consistent with the types of inferences which are made from CRM test data.

Three estimates within the dimension of "threshold loss" were reviewed and, were seen to yield relatively accurate estimates of the proportion of consistent decisions between two parallel test forms. While the Swaminathan et al. procedure required the results from two testings, the Huynh and Subkoviak approaches were able to estimate the proportion of consistent decisions on the basis of the data obtained from a single test administration.

However, the three above techniques were seen to possess the shared disadvantage of being situation specific. The reliability estimate calculated by each of these approaches on a particular set of data would not be unique, but would change as the criterion level or cut-off score was altered. Therefore, if such a reliability estimate is to be interpreted meaningfully, the calculated value should be reported along with the cut-off score and how it was determined, characteristics of the examinees, and test length.

A further disadvantage of techniques of reliability estimation within the dimension of "threshold loss", and one so far not discussed is that they treat all errors equally. That is, if an individual is incorrectly classified as a master, it would not matter whether the person's true score were one point below the criterion level or several points below. The severity of the error would be treated equally in both cases. That is to say that errors are in terms of misclassifications; distance does not enter into the problem.

BASIC ASPECTS OF INFORMATION THEORY

The purpose of this section is to serve as a description of the basic conceptual and statistical aspects of information theory. The literature in this area is both vast and diverse, and the presentation here is designed to provide only those preliminary aspects on which the methodology of Chapter III is based.

The field of statistics is concerned with the measurement and analysis of a number of concepts, for example; variance, deviation, average, relationship, and error, which likewise possess a conceptual meaning in our common everyday experience. What the study of statistics does of course, as is the case with any scientific enterprise, is to impose an exact and rigorous definition on those concepts. That is, science in general looks at the factors which appear to regulate and determine the nature of our common sense world, and attempts to rigorously define and measure those factors so as to arrive at an objective analysis, estimate or prediction of their nature or effects. The field of information theory reflects this scientific study of an influential aspect of our everyday experience.

Any inquiry is marked by the desire to gain information of some type. Whether that inquiry is in the form of research of the printed word, the experimentation and study of animal and human behavior, or the simple questioning of those believed to have desired answers, the goal is to become more informed than we were previously. All such forms of inquiry are in fact modes of communication. In particular, that communication can be between the psychologist and man's mental faculties, the physician and the body, or the educator and the learner; in general, it is between man and the world. Since both layman and scientist alike seek information daily, it would therefore seem desirable to possess the means of determining how much information had been gained in a particular communicative act. This quantification of transmitted information is the basic goal of information theory.

Information theory was formulated to solve the basic problems of communication engineering; that is, "How does one measure the amount of information in a message to be transmitted?"; and, "How much information was actually communicated?" By the nature of these questions, it should be of no surprise that the initial work in this field was performed by electrical engineers.

What is being attempted herein then, is to take a procedure developed basically in electronics and apply it to the explanation of educational and psychological phenomena.

There is nothing new of course, in the application of a framework in one field of study as a model for the description of concepts in another field. However, to do so properly, the aspects of the borrowed framework must exhibit a degree of similarity with the phenomena which its application seeks to explain. As an example, the mathematical properties of the normal distribution have up to now been seen to be similar to certain hypothesized aspects of various human characteristics as possessed within a population. Hence, if information theory is to be seen to offer a suitable alternative to traditional reliability estimation, certain conceptual similarities must be demonstrated to exist between the two areas.

Information Theory and Reliability

The primary concern of information theory is to quantify the amount of information transmitted from sender to receiver. Whether that information is true or false, as well as matters of human value, are not considered. Information so measured is thus seen to possess a certain similarity with reliability in terms of the latter's relationship to validity.

As stated earlier, the degree of reliability attributed to the measurements obtained from an instrument depends, in theory, upon the amount of test score variance which is due to error. Operationally, the issue of reliability is handled in classical test theory by the analysis of the consistency of obtained measurements from one application of the instrument to a second independent and equivalent application on the same group. As such, reliability's concern is with the accuracy or consistency of measurements and not with <u>what</u> is being measured. This latter task is the topic of validity.

Reliability is best viewed as a necessary but not sufficient condition for validity (Gronlund, 1976). That is, accurate measurements of something can be obtained, without that something measured being relevant to the purposes to which those measurements are intended. On the other hand, before it can be asked whether or not a set of obtained measurements is relevant to a particular purpose, the question of the accuracy of those instruments must be satisfied. In short, reliability concerns the measurements themselves, validity applies to the uses to which those measurements are to be put.

The concept of information shares the concept of reliability's concern with the measurements themselves. Test scores can be viewed as messages from testees to the examiner, regarding level of achievement in a particular subject area. Just as reliability considers the accuracy of those scores apart from the question of whether indeed the items on which those scores are based, do in fact measure aspects of the subject area intended, information theory is concerned solely with the amount of information transmitted by those messages.

A Conceptual Definition of Information

An introduction to the conceptual definition of information can be perhaps best begun by examining its relationship to the term entropy in physics. All physical systems are to varying degrees incompletely defined, to the extent that, certain variables of a macroscopic nature can be measured, while particular aspects of a more microscopic nature within the system remain unknown. For example, physicists agree that the hydrogen isotope, tritium, has a nucleus composed of two neutrons and a single proton. However, the complete number and types of subatomic particles which make-up a neutron or proton are not known. Within such systems, a good deal of information regarding detailed structure is missing. The amount of uncertainty which remains within a system, is labeled entropy.

The application to education appears clear. Educators and psychologists seek to define and measure certain human characteristics, for example, intelligence, creativity, aptitude, achievement, anxiety, and, make decisions based in part or completely, upon the information provided by those measurements. Nevertheless, a great deal remains uncertain regarding what underlying factors are connected to those "macroscopic" variables in terms of cause and effect relationships. For example, a group of characteristics collectively defined as intelligence are measured and, as a result, children are labeled mentally retarded, learning disabled, average or genius, to a great extent on the basis of those measurements. However, it remains a matter of debate not only what caused individuals to possess varying degrees of such characteristics, but in part also, what are the effects of being more or less intelligent.

Entropy then, measures the lack of information in a system. A reduction in entropy is sought through communication with the world, basically, through experimentation if one is pursuing the problem from a scientific perspective. Since, as more information is obtained through such communications the amount of entropy is reduced, there exists an inverse relationship between the two concepts.

To sum up this discussion of the conceptual nature of information thus far then, information is obtained from some source and insofar as that this information was not previously possessed, the uncertainty regarding a situation is to some extent reduced. Information so obtained is considered apart from its being true or false, useful or useless. And, the amount of information provided by an act of communication depends upon the extent to which uncertainty is reduced regarding a particular state of nature.

Information and Uncertainty

Next, it should be noted that the uncertainty contained in a particular action is a function of the number of possible outcomes. For example, if we desired to predict the result of first, the roll of a fair six-sided die, and secondly, the toss of a fair coin, there would be a greater amount of uncertainty regarding the outcome of the first action relative to that of the second. This is the case simply because there are more possible alternatives available in the former case.

Indeed, it would be impossible to gain information from a message if some uncertainty as to the nature of the response did not exist beforehand. And, due to the inverse relationship between information and uncertainty, the greater the amount of uncertainty contained in the possible outcomes of a communicative action, the greater the potential information. Therefore, if the mathematical means to quantify information are to be developed, such mathematical statements must be a function of the number of possible outcomes.

Hence, it can be seen that the conceptual notion of information within information theory is not far different from its everyday usage. Further information is not obtained by asking a question or performing an experiment of which the outcome is known a priori, and indeed, the expected outcome occurs. Information is possible only in a questioning format in which the result is uncertain. In fact, the more improbable the result, the more the information that is gained. In a sense, the more surprising the nature of an outcome, the more informed the receiver has become.

This last statement offers a hint as to the approach that will need to be taken in the mathematical quantification of information. Information will not only be a function of the number of possible situation outcomes, but most of all, of the probability of occurrence of those various outcomes.

Statistical Aspects of Information Theory

A review of the development of the statistical basis of information theory is begun with mention of two early papers. In 1924, Nyquist, an engineer at Bell Laboratories, published a paper concerning the factors affecting telegraph speed, in which he proposed that the efficiency with which messages are transmitted, is a function of the logarithm of the number of possible levels of current. Later, Hartley (1928), also working at Bell, concurred that a measure of information needed to be both a function of the number of alternative outcome sequences, and, logarithmic in form.

However, a detailed statistical model by which information could be measured was not formulated until Claude E. Shannon and Warren Weaver (1949) published a work entitled, "The Mathematical Theory of Communication". The work of Shannon and Weaver suggested applications outside the field of engineering, and resulted in a number of attempts to employ information theory in the solution of various psychological problems.

As often seems to be the case when unbounded enthusiasm accompanies the wide-spread acceptance of a new solution to old problems, some of these early applications were, as Attneave (1959) offers, "successful and illuminating, some were pointless, and some were downright bizarre". With the hope that the present application will not be placed in one of the last two of Attneave's categories, this discussion now looks at why a statistical statement of information has been held to be logarithmic in nature.

An example often used to illustrate the basic statistical nature of information is the old game "Twenty Questions". Here, there are a number of categories, one of which contains the item or answer sought. By means of a series of questions, capable of being answered either "yes" or "no", the categories are eliminated until the correct one is discovered. As an illustration, an example employed by Attneave (1959) will be used.

Suppose that the questioner is thinking of a particular square on a chessboard and it is the task of the inquirer to simply find out which it is. Even though there are 64 possible squares, one could readily determine the correct location by asking six questions of the form:

1.) Is it one of the 32 on the left half of the board? (Yes)

2.) Is it one of the 16 in the upper half of the 32 remaining? (No)

3.) Is it one of the 8 in the left half of the 16 remaining? (No)

4.) Is it one of the 4 in the upper half of the 8 remaining? (No)

5.) Is it one of the 2 in the left half of the 4 remaining? (Yes)

6.) Is it the upper one of the 2 remaining? (Yes)

Figure 2.1 depicts how the area of uncertainty was systematically reduced until the correct square was identified. Of course, the questions could have been differently constructed and would have been equally efficient, as long as the remaining area of uncertainty was reduced by one-half. If not, however, more than six questions will often be needed to determine the correct square.

The next step is to numerically express, and quantify the information contained in the above example. The six questions will result in a different series of "yes" and "no" responses as the square which we seek varies about the board. Now suppose 1 is allowed to signify "yes", and 0, "no". In such a system, based on the same six questions, each square's identity will be represented by a unique six digit number. Each of these digits is binary in nature in that only one of two values can be assumed. With such a system, a number one digit in length would be required to eliminate the uncertainty contained in 2 alternatives, two

Figure 2.1

An Example of the Game of "Twenty Questions"*

2	2	2	2	1	1	1	1
2	2	2	2	1	1	1	1
2	2	2	2	1	1	1	1
2	2	2	2	1	1	1	1
3	3	4	4	1	1	1	1
3	3	4	4	1	1	1	1
3	3	Ⓧ	5	1	1	1	1
3	3	6	5	1	1	1	1

* Each square contains the number of the question which eliminated it as the square which the questioner was thinking of.

digits would be needed for 4 or 2^2 alternatives, and as seen in the above example, six digits are needed for 64 or 2^6 alternatives.

Within information theory, the binary digit has been contracted to "bit", and is used as the unit measure of information and uncertainty. Therefore, asking someone to locate a particular square on a chessboard of 64 equally likely locations, represents a question having 6 bits of uncertainty, and in turn, contains 6 bits of information in its solution.

Thus, the amount of uncertainty contained in a number, n, of such alternatives, or the amount of information required to remove that uncertainty, can be expressed by the equation:

$$n = 2^{U(x)} \tag{2.8}$$

where: n = the number of alternatives.

x = some random variable (in this case, a square on a chessboard).

$U(x)$ = amount of uncertainty in x.

An equivalent expression, solving for U(x), would be:

$$U(x) = \log_2 n^* \qquad (2.9)$$

where: log is taken to the base 2.

In the above example, it was implied that each of the 64 alternatives had an equal probability, that is, 1/64, of being the one for which we were searching. One can reasonably ask how the situation changes when the existing alternatives have unequal probabilities associated with their respective chances of occurrence. Indeed, as the statistical theory of information is developed further, it will be seen to be unnecessary to assume that the possible outcomes of a message have equal probabilities of being sent.

The Statistical Theory of Information

Before proceeding, however, comment should be made concerning the notation used in this discussion. It appears to be the case with statistics texts in general, that notation differs, in varying degrees from source to source. The situation is similar within the field of information theory. Therefore, it may prove helpful to the reader to mention that the notation used herein is adopted from Garner (1962).

[*] Unless otherwise stated, when the terms "log" or "logarithm" are used within the present paper, the base 2 is implied.

To begin, assume that a message has been sent from some source, and that the message had been selected from a set of n possible alternatives, which is represented thusly:

$$X = \left[x_1, x_2, \ldots, x_i, \ldots, x_n \right] \qquad (2.10)$$

where: X = the set of n possible alternatives.

x_i = the ith alternative in set X.

Next, assume that each message in the set of alternatives X, has a particular probability of being sent. This set of probabilities can then be written as:

$$P(X) = \left[p(x_1, p(x_2), \ldots, p(x_i), \ldots, p(x_n) \right] \qquad (2.11)$$

where: $P(X)$ = the set of probabilities of occurrence associated with the alternatives in set X.

$p(x_i)$ = the probability of occurrence of the ith alternative in set X.

and, $$\sum_{i=1}^{n} p(x_i) = 1$$

The greater the probability of a message being sent, the less the information that it conveys. Such a relationship agrees with the common usage of the term "information".

If we ask someone a question, and receive the answer we expected, not much, if any, information is gained. However, if an answer is received that is to a certain extent quite surprising, we would most likely feel that a good deal of information had been gained. This relationship can be expressed as:

$$I(x_i) < I(x_k) \text{ iff } p(x_i) > p(x_k) \tag{2.12}$$

where: iff is read "if and only if".

$I(x_i)$ = the information associated with alternative x_i.

$p(x_i)$ = the probability of the occurrence of x_i.

and, similarly for $I(x_k)$ and $p(x_k)$.

It is then the probability of a message's occurrence which determines its information value. And, the information associated with the same message may vary from situation to situation or from source to source, simply because the associated probability may differ. In one situation, a particular response may be highly probable, while in a different situation, that same response would be highly improbable.

Hence, information is a function not of what is said, but rather of what could have been said, and wasn't.

As Shannon and Weaver (1949) state:

> The concept of information applies not to the individual messages (as the concept of meaning would), but rather to the situation as a whole, the unit information indicating that in this situation one has an amount of freedom of choice, in selecting a message, which it is convenient to regard as a standard or unit amount. (p. 9)

Next, consider the case of a source which sends two messages; the first from a set X, and the second from set Y, where:

$$X = \left[x_i, x_2, \ldots, x_i, \ldots, x_n \right] \tag{2.13}$$

and,

$$Y = \left[y_i, y_2, \ldots, y_j, \ldots, y_n \right] \tag{2.14}$$

The messages in both X and Y, have probabilities associated with the likelihood of their transmission in the same manner as the example immediately above. It was stated in that example that the amount of information contained in a particular message was in some way a function of the probability of its being sent. This relationship can be represented thusly:

$$I(x_i) = f\left[p(x_i) \right] \tag{2.15}$$

where: $\mathcal{f}\left[p(x_i)\right]$ = a yet to be defined function of the probability of message x_i.

With the case of two messages, x_i and y_j, being sent, the assumption is made that the amount of information conveyed by both is equal to the amount conveyed by x_i, plus the amount conveyed by y_j, given that x_i has been selected. An expression of the amount of information conveyed by the two messages, as a function of the probabilities of x_i and y_j given x_i, would be:

$$I(x_i \text{ and } y_j) = \mathcal{f}\left[p(x_i)\right] + \mathcal{f}\left[p(y_j/x_i)\right] \qquad (2.16)$$

where: $\mathcal{f}\left[p(y_j/x_i)\right]$ = some function of the probability of y_j being sent given that x_i had previously been sent.

The selection of messages x_i and y_j can also be viewed as the selection of a single ordered pair from the Cartesian product space of sets X and Y. The information conveyed by messages x_i and y_j would, in this case, be a function of the probability that the ordered pair (x_i, y_j) would be selected. Such a relationship can then be expressed as:

$$I(x_i \text{ and } y_j) = \mathcal{f}\left[p(x_i \text{ and } y_j)\right] \qquad (2.17)$$

where: $f\left[p(x_i \text{ and } y_j)\right]$ = some function of the probability of the pair (x_i, y_j) being selected.

Therefore, it follows from equations 2.16 and 2.17, upon restating the right side of 2.16 in terms of function f, that:

$$f\left[p(x_i \text{ and } y_j)\right] = f\left[p(x_i) + p(y_j/x_i)\right] \qquad (2.18)$$

Now, it is an assumption of probability theory, that:

$$p(x_i \text{ and } y_j) = p(x_i)\ p(y_j/x_i) \qquad (2.19)$$

So, if the right side of equation 2.19 is substituted for its equivalent in the left side of equation 2.18, the result is:

$$f\left[p(x_i)\ \ p(y_j/x_i)\right] = f\left[p(x_i) + p(y_j/x_i)\right] \qquad (2.20)$$

Finally, if "a" is allowed to represent the quantity, $p(x_i)$, and, "b" to represent the quantity, $p(y_j/x_i)$, an equation of the following form results:

$$f(ab) = f(a) + f(b) \qquad (2.21)$$

where: $a = p(x_i)$

$$b = p(y_j/x_i)$$

A common function which satisfies this condition of equality is the logarithm. Hence, the appropriateness of the choice of the log function as a measure of information is further evidenced.

Up to now, this discussion has dealt with $I(x_i)$ as a function of the probability of the occurrence of x_i, but has not as yet defined that function further than determining that it should be logarithmic in nature. That is, a measure of the information conveyed by message x_i would be:

$$I(x_i) = \int \left[p(x_i) \right] = k \log p(x_i) \qquad (2.22)$$

where: k = some constant.

It should be recalled from our earlier discussion that as the probability of a message's occurrence increases, the information that message conveys will decrease. Therefore, the function "$\int$" must be of such a nature that $I(x_i)$ will become increasingly positive as $p(x_i)$ becomes increasingly negative, and vice versa. That is, the function should reflect the fact that the amount of information, $I(x_i)$, obtained from a message, x_i, will increase as the probability, $p(x_i)$, of that message being received decreases. This function will express this relationship if "k" is allowed to be a negative number, so as to reflect such an inverse relationship, and for simplicity's sake the value "-1" is chosen.

The result then, is an equation by which a measure is obtained which seeks to quantify the information received thru the occurrence of a message, x_i. This equation is:

$$I(x_i) = - \log p(x_i) \qquad (2.23)$$

The Expected Amount of Information from a Set of Messages

Before concluding this chapter, there is one final concept which must be introduced, as it will prove to be of major importance in Chapter III. This concept is the expected amount of information that is conveyed by an entire set of messages.

The application of this notion to an educational situation seems relatively straightforward. In the vast majority of cases, a test is made up of a number of items. It can be seen by the above discussion that there may indeed be a way in which the amount of information conveyed by the response to a particular item might be measured. However, if the amount of information conveyed by the responses on the test as a whole are desired, a method by which those individual item measures can be combined, is required.

When statisticians speak in terms of an expected value, some type of long term average is being considered. The situation here is no different. Since the probability associated with a particular message was initially required to determine the information conveyed by that message, it would be a simple matter to "weight" the quantified information, expressed in bits, by multiplying that quantity by its probability of occurrence. On the basis of probability theory, this weighted value can then be summed with similarly weighted values for the other messages in a given set, to obtain the expected information contained in a set of messages. Within information theory, this expected information of a set of messages is defined as uncertainty. And, the uncertainty contained in a particular message set can be estimated by the equation:

$$U(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \qquad (2.24)$$

where: $U(X)$ = the uncertainty contained in message set X.

$p(x_i)$ = the weight given to message x_i.

As was the case with the representation of information, the mathematical expression for the uncertainty contained in a set of messages is compatible with the everyday use of the term.

Suppose, for example, that one is confronted with a situation which has four possible outcomes, and one, and only one, of these outcomes will occur. If asked to predict which alternative will result, the maximum amount of uncertainty would be contained in the choice if each of the alternatives were equally likely to occur. On the other hand, if one or more alternatives were more likely to occur than the others, the maximum uncertainty would be reduced. Indeed, if one of the alternative probabilities is allowed to approach one, the uncertainty will in turn approach zero. Obviously, if it is certain what will occur in a particular situation, there is no uncertainty involved in predicting the outcome.

This common sense notion is reflected in the statistic representing uncertainty. The uncertainty or expected information contained in a set of alternative messages will be at a maximum when each alternative has the same probability of being sent. At the other extreme, the uncertainty or expected information will decrease toward zero, as the probability associated with a particular alternative approaches one.

Summary

This second section of Chapter II has been designed to sketch the origins and basic conceptions and statistical definitions of information theory. It was noted at the beginning of this discussion that since every scientific endeavor seeks information of some type, it would seem desirable to have some way in which the amount of information transmitted in a particular situation could be estimated. Information theory was seen to provide such a technique, which preserved the common sense notions of information and the reduction of uncertainty.

Finally, it should be stressed that information theory is far more complex and statistically diverse than the preceding discussion may lead one to suspect. Although only univariate and bivariate applications were touched upon, multivariate procedures have also been developed. However, the basic concepts and definitions presented above will prove of sufficient aid in the development of the methodology of Chapter III.

SUMMARY OF CHAPTER

This chapter consisted of two major sections. The first section was designed to serve as a review of previously suggested reliability estimation procedures to be applied to the results of CRM measures. The section began with the assertion by Cox (1970) that if CR measurement was to be accepted and applied to teacher-made tests, alternatives to the various traditional statistical concepts such as reliability must be developed. To substantiate this position, reference was made to Hambleton and Novick (1973) who demonstrate that the loss function inherent in NR reliability estimates depends upon score variability, and is inappropriate for use with CRM measures. These authors suggest that the concept of "threshold loss", based on the number of incorrect mastery/nonmastery classifications, is a more appropriate perspective from which to consider the errors associated with CRM measures. Thus, the need for the development of statistical techniques specific to the nature of CRM data was seen to have been previously recognized, and a number of reliability estimation procedures have been suggested to meet this need.

Five of these previously suggested estimation procedures were reviewed, all of which have been considered to provide loss functions within the dimension of "threshold loss". The first, formulated by Carver (1970), has been described as providing only a "thumb-nail" estimate of the consistency of mastery/nonmastery classifications, while the second (Livingston, 1972) was seen to be grounded in classical test theory and was therefore dependent upon a certain degree of score variability. The third coefficient reviewed is based upon Cohen's coefficient kappa (1968, 1972), and has been suggested by Swaminathan et al. (1974). This third coefficient is designed to provide an estimate of the consistency of mastery/nonmastery classifications obtained from parallel forms of a CRM measure.

In discussion of the Swaminathan et al. procedure, it was noted that the calculated values which result are to a certain extent "situation specific". That is, that the nature of the mastery/nonmastery classifications obtained will be to some extent dependent not only on the location of the mastery criterion score, but also upon the manner in which the criterion was chosen. In short, since the calculated coefficient value is dependent upon the location of the mastery criterion, the same CRM measure can have a number of associated degrees of reliability depending upon the location of the mastery criterion. This notion of a CRM measure being "situation specific" was seen to run counter to the more traditional position that a particular test should have a single degree of reliability associated with the scores which it yields. However, since the nature of the obtained mastery/nonmastery classifications will necessarily change if the mastery criterion changes, the fact that an index of reliability is sensitive to such "situational" changes should be viewed as a desirable property of such a coefficient.

The two further coefficients reviewed were also seen to be "situation specific" as well as yielding estimates within the dimension of "threshold loss". Huynh (1976) has developed a procedure of estimating coefficient kappa from a single test administration, involving the use of the Kuder-Richardson formula 21 as a means of estimating parameters which are then inserted into a beta-binomial distribution to provide the estimate of the kappa value. One practical disadvantage of the Huynh procedure is that the involvement of calculus makes accessibility to a computer almost mandatory, and therefore makes the possibility of its use by classroom teachers quite unlikely. Huynh's coefficient was also seen to be dependent upon the existence of score variability, and to approach a maximum value of 1.00 as the involved test became either too easy or too difficult for the classified examinees.

The final coefficient reviewed was that suggested by Subkoviak (1976) which also requires only a single test administration, and is designed to provide an estimate of the proportion P_c of consistent mastery/nonmastery classifications over two test administrations. The procedure developed by Subkoviak was seen to be quite similar to that suggested by Huynh, as evidenced by the fact that P_c is a function of coefficient kappa.

The first major section of this chapter concluded with a reference to Subkoviak's 1978 study which provides a comparative analysis of the final three procedures discussed above. All of the indices were found to yield accurate estimates of the parameter P_c, with the Swaminathan et al. procedure having a relatively higher standard error. Subkoviak concluded his study by recommending the Huynh procedure in that it was seen to be "mathematically sound" and required only a single testing. However, Subkoviak notes that the test items which yielded the scores used in the analysis were fairly heterogeneous in nature, and it was therefore somewhat doubtful if these procedures would behave similarly with an increased homogeneity of item content.

The second major section of this chapter was designed to provide an introduction into the basic concepts and statistical definitions of information theory. The purpose of this introduction was to provide a familiarization with the methodology to be used in Chapter III.

It was noted that since any scientific investigation is intended to in some way communicate information of a particular sort, it would seem advantageous to be able to measure the amount of information transmitted in the data obtained from such an activity. Information theory is concerned with answering the questions associated with this type of communication. The activity to be considered here is of course the administration of a CRM measure, with some degree of information being communicated from the mastery/nonmastery classifications which result. In an attempt to justify the application of information theory in the solution of the problem at hand, it was noted that emphasis will be placed on the similarities between the concepts of information and reliability.

The information obtained through a particular communication was seen to some extent reduce the uncertainty which existed prior to the communication. It was next noted that the amount of pre-existing uncertainty was a function both of the number of possible outcomes which could be communicated, and, the relative probabilities of occurrence associated with those outcomes. The more improbable a possible outcome the more uncertainty that is associated with it, and likewise, the greater the amount of information that is communicated if that outcome actually occurs. The game of "twenty questions" was then used to illustrate the development of a statistical definition of uncertainty. A statistical definition of information was also introduced, and was found to be based on logarithms to the base 2.

The final statistical concept introduced was that of the expected amount of information contained in an entire set of messages. A statistical definition of this concept of expected information was presented, and was defined to be the amount of uncertainty contained in a given set of messages. It was mentioned that this statistical definition of uncertainty would be of particular importance in the methodology of the next chapter.

CHAPTER III

CONCEPTUAL BASIS OF METHODOLOGY

The development of information theory is seen to have progressed with different emphases, when the work done in the United States is compared to that accomplished in Europe (Weltner, 1973; Hintikka and Suppes, 1970). It has been noted that information theory had its statistical beginnings in America, basically through the work of C. E. Shannon. And it is an emphasis on the statistical aspects of the theory which characterizes the major portion of the work done in the United States.

However, this emphasis on the purely statistical has presented problems in the application of information theory to the data of the social sciences. This was a difficulty briefly hinted at by Attneave (1959). It was noted that a mood bordering on unbounded enthusiasm welcomed the arrival of information theory. And why not? There may be nothing more basically appealing to a scientist or philosopher than to be able to measure the information contained in a set of data or a logical proposition. Nevertheless, as Attneave states, this enthusiasm soon became somewhat subdued when many of the early applications of information theory proved to be either worthless or "downright bizarre".

This problem was due basically to the fact that, as with statistical theory in general, information theory is based on formulas which are made up of symbols which are nonlinguistic in nature. And, information theory lacked the work necessary in the area of semantics so that these nonlinguistic symbols could be interpreted and meaningfully applied to the perspective of the language of the social sciences. Earlier, it was mentioned here that adapting a framework from one field, to be used as an explanatory model in another, is a frequently used problem-solving technique of the sciences. However, a set of transformation rules must be developed and utilized if the application of a framework as a model is to fulfill the intended purpose of explaining and predicting phenomena in the more unfamiliar field. Work in semantics and inductive logic was required to develop such rules for the meaningful application of information theory to the social sciences.

It is the development of information theory in the areas of semantics, inductive logic, and epistemology, which characterizes the direction of research of this field in Europe. Considering the relative traditions of Europe and America in regards to linguistics, theory of knowledge, and formal logic, it does not seem particularly surprising that European authors would be doing the majority of work on the semantic perspective of information theory.

The purpose of this chapter will be to present and describe one suggested bridge between these two necessary aspects of information theory, as an application to be used in the estimation of the reliability of CRM instrument classifications. To do so meaningfully in light of the previous discussion, it will first be necessary to make mention of the logical and semantic basis of the herein suggested solution. This will be attempted by first examining the conceptual informational relationship between obtained evidence and a tested hypothesis; which will be followed by a suggested manner in which the informational strength of obtained evidence can estimate the degree to which a tested hypothesis has been confirmed.

Examination of the relationship between a tested hypothesis and the evidence which results will begin with consideration from a conceptual perspective. An attempt will be made through this discussion to illustrate that the extent to which a particular hypothesis is confirmed, to be termed "degree of covering", depends upon the strength of evidence obtained from a testing of that hypothesis. It will then be argued that the concept of the reliability of mastery/nonmastery classifications resulting from the scores obtained from a CRM measure can be expressed in terms of this general model. In particular, the relationship between evidence and hypothesis will be presented as an analogy of the relationship between true score variance and total score variance as expressed within traditional test theory.

A statistical definition of "degree of covering" will next be presented, with this definition serving as a basis for the CRM reliability coefficient then developed. A discussion of both the mathematical nature and the philosophical implications of the developed coefficient will follow. Finally, the range of possible values of the coefficient will be examined.

The Relationship between Hypothesis and Evidence

Scientific investigations involve, in a majority of situations, the testing of null hypotheses. Evidence is obtained, hopefully relevant to the specific null hypothesis being tested, which provides information, in the form of observations of some type, which form the basis for a decision concerning the rejection or non-rejection of that null hypothesis.

In the situation at hand, a conceptual hypothesis is generated concerning the ability of two tests, constructed with the intent to be parallel in form, to yield consistent decisions as to the classification of the testees as masters or nonmasters, in regard to the achievement of a specific set of objectives. The evidence required to make a judgment regarding such a hypothesis would be an observation of the consistent mastery and nonmastery decisions yielded by the two tests. Upon determining the extent to which the evidence implies the hypothesis, one could make a decision regarding the acceptance or rejection of the tested hypothesis. Generally, this is of course, the role of inductive inference upon which the current notion of the scientific method depends. As specifically stated, such an analysis of the evidence at hand would provide an estimate of the degree to which the results of the two tests under consideration yielded consistent mastery/ nonmastery decisions.

In terms of the concepts under present study, a hypothesis, as stated scientifically, is a conjecture which contains within its expression, some amount of uncertainty. We ask questions, and forward suggested solutions, precisely because we are uncertain about the correct or best answer. And herein lies perhaps the best opportunity to visualize the similarity between the concept of uncertainty and the statistical definition of variance.

Traditional statistical techniques, whether it be for example, regression analysis, analysis of variance, discriminant analysis, or cluster analysis, all attempt to explain or account for observed variance of some type. It has been observed that groups, individuals, or objects vary in the extent of their estimated possession of some characteristic or attribute of interest, and it is uncertain as to what degree and in what direction. Therefore, an attempt is made to gather relevant evidence, seek to identify the sources and extent of the observed variance, and on the basis of the information obtained, make a decision about the hypothesis.

Much of what has been said immediately above is probably not new to the reader. However, the current research seeks to suggest an alternative solution to a traditional statistical technique. And, the author believes that such a situation necessitates the attempt to demonstrate the conceptual compatibility of this newly suggested solution, with that accomplished by traditional techniques in similar circumstances.

It has been the desired purpose of this section to illustrate the similarity between the concepts of uncertainty and variance. As noted earlier in Chapter I, theoretically, the reliability of measures obtained from an instrument is considered to be the proportion or degree of observed score variance which is due to true score variance. In effect, to what extent is the variance of obtained scores on some instrument due to the true position of the individuals taking the test, on a continuum of degree of possession of the attribute or characteristic which the test seeks to measure.

As also noted earlier, an operational estimate of reliability is the correlation of the set of observations obtained from the administration of two independent and equivalent measures to a particular group. That is, the reliability coefficient, as traditionally defined, is an estimate of the degree of total score variance which is shared by the two instruments. In this degree of shared variance lies the information for a decision regarding the reliability of the measurements obtained from the instruments.

The author will now attempt to demonstrate that the ratio of true score variance to total score variance has an analogy in the concepts of information and uncertainty in information theory.

The Shared Information of Evidence and Hypothesis

The discussion of Chapter III has so far centered on the relationship between evidence and hypothesis, as specifically applied to an estimation of test score reliability, within the framework of inductive inference. And, if the notions of evidence and hypothesis are to be expressed in the language of information theory, because of the nature of the statistical statements involved, a logical relation between evidence and hypothesis, based on a probability measure, must be determined. In 1970, Risto Hilpinen published a study, "On the Information Provided by Observations", which offers the basis for just such a measure.

As Hilpinen states, from the viewpoint of inductive logic, "probability is a logical relation between two sentences". In application to the problem at hand, assume that the hypothesis under study is represented by sentence "H", and the evidence on which the credibility of that hypothesis is to be decided is termed sentence "E". On this basis, a probability statement designed to express the degree of credibility of H on the basis of E would be "P(H/E) = R". In this relationship, R, as any probability estimate, is a real number within the closed interval (0, 1), and represents a "justified degree of belief" in H, on the basis of E. (See Figure 3.1)

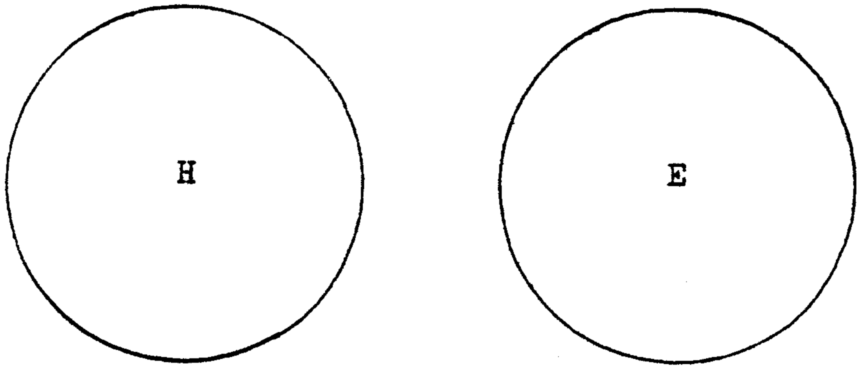
Now, as noted, a reliability coefficient is an estimate of the ratio of true score variance to total score variance. Such a coefficient represents the degree or proportion of the total score variance which is "shared" by true score variance. And, that coefficient provides us with a "justified degree of belief" on which a decision concerning the accuracy of the obtained measures can be based.

Within the social sciences, evidence consists of observations of some type. In the case at hand, mastery/ nonmastery decisions are made on the basis of scores obtained from a first test administration, and these same types of decisions are made for the same group of students on the basis of the independent administration of a second test designed to be equivalent to the first. Such evidence could then be used to evaluate the hypothesis that the two tests yield consistent mastery/nonmastery decisions.

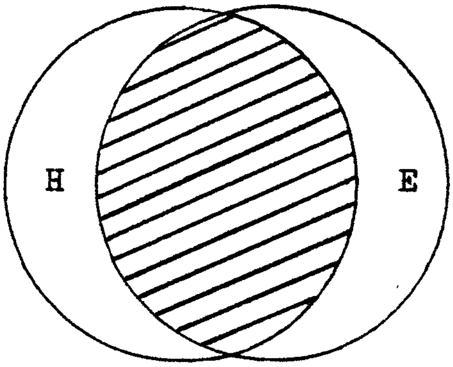
In other words, what would be of aid in such a situation, would be an index of the degree to which the extent of observed consistent mastery/nonmastery decisions,

Figure 3.1

An Illustration of the Relationship between Evidence and Hypothesis



Situation in which the information provided by the evidence is completely independent from the uncertainty contained in the tested hypothesis.



Situation in which the information provided by the evidence "covers" the uncertainty contained in the tested hypothesis, to the degree represented by the shaded area.

"confirms" the hypothesis that two instruments do in fact yield such decisions.

Evidence gathered with the purpose of obtaining information, if relevant to the hypothesis under study, will to some extent relieve the uncertainty which is contained in that hypothetical statement. Hence, basis is provided for combining the concepts of evidence and hypothesis in inductive inference with the concepts of information and uncertainty in information theory. What is further needed to apply this theoretical relationship to an observable and practical situation, is a semantic interpretation of information theory, to be added to the statistical definitions.

The next section presents such a semantic interpretation.

The Degree of Hypothesis Confirmation by the Strength of Evidence

The logical relation between two sentences, formulated on the basis of the probability of those sentences, as outlined in the cited work of Risto Hilpinen, had an earlier application in two papers authored by Hakan Tornebohm of the University of Gothenburg, Sweden.

In 1966, Tornebohm published a paper titled "Two

Measures of Evidential Strength", which developed and described two techniques designed to estimate the degree by which a hypothesis was confirmed on the basis of obtained evidence. Tornebohm followed this paper with a 1968 article, "On the Confirmation of Hypotheses About Regions of Existence", which presented suggestions for the application of the earlier described measures. The current author posits that Tornebohm's work provides the semantic aspects which makes possible the application of information theory to the task at hand. Therefore, an outline of his notion of "degree of covering" as presented in the two abovementioned papers must be considered preliminary to the consistency coefficient that will be developed.

Tornebohm begins by assuming that we have a state space of objects, R. In seeking to find the position of an object in R, a measurement instrument, Z, is employed. The result of a particular measurement represents a vector. The state space of all such vectors is designated M. This state space M thus contains a finite number of cells, each cell corresponding to the vector which results from the measurement of an object in state space R, by an instrument in Z. If it can be assumed that the vectors obtained from Z are independent of one another, Z is seen to produce a functional relationship between R and M, in which every element in R can have one and only one image in M. In such a relationship, R is considered the domain, and M the range. Such a relationship is illustrated by Figure 3.2. Within this framework, a hypothesis H could be formulated concerning the conjecture that the images of the cells in M, are indeed the region of existence of the objects in R.

This model can easily be applied to the situation in which testees are classified either masters or nonmasters on the basis of obtained test scores.

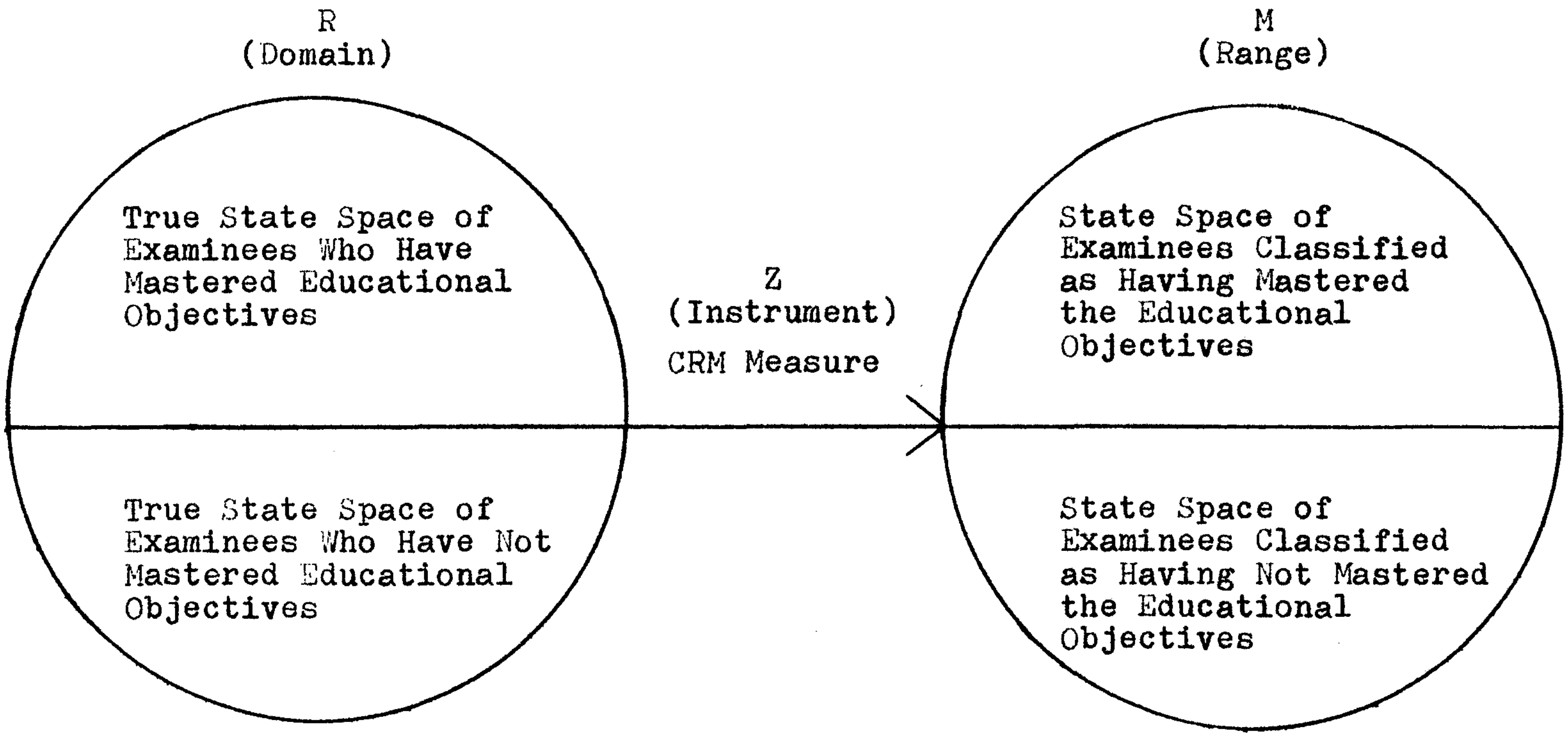
In such a specific application, a set of individuals who have received a particular treatment or mode of instruction, represent the objects in state space R. Upon completion of the treatment or instructional program, a CRM measure is administered to the individuals in this group, and, mastery/nonmastery decisions are made on the basis of scores obtained on that measurement instrument. This, of course, corresponds to the formulation of an image in M composed of the vectors resulting from the measurement of the objects in R by instrument Z.

It should be recalled that the items on a CRM are meant to represent a random sample from a larger domain of items. And, the percentage of items an individual

Figure 3.2

An Illustration of the Domain and Range of

Mastery/Nonmastery States and Classifications


R
(Domain)
M
(Range)
True State Space of Examinees Who Have Mastered Educational Objectives
Z
(Instrument)
CRM Measure
State Space of Examinees Classified as Having Mastered the Educational Objectives
True State Space of Examinees Who Have Not Mastered Educational Objectives
State Space of Examinees Classified as Having Not Mastered the Educational Objectives

correctly answers on that measure represents an estimate of the percentage of items in the entire domain which the individual can correctly answer. Thus, the familiar situation is noted of locating a person's true position on a continuum, by a random sample of behaviors that continuum is assumed to reflect.

Hence, it follows that R represents the true state space of such examinees, as they exist on some continuum of achievement. The scores on a CRM measure, corresponding to the vectors produced by instrument Z, then result in the assignment of these examinees to either a mastery or nonmastery region of existence on that continuum. These assigned regions of existence correspond to the state space M.

The hypothesis of interest then becomes one concerning the extent to which the assigned regions of existence in M are images of the true regions of existence of those individuals in R. And, a statement of reliability regarding the accuracy of mastery/nonmastery classifications is arrived at. As is the case with traditional reliability, this problem will be approached operationally from the perspective of consistency of classifications.

Reliability as Consistency of Mastery/Nonmastery Classifications

An outline of Tornebohm's model continues with a symbolic statement of the hypothesis under study. In this specific application, individual c_i, as existing in state space R, is either in the region of mastery or that of nonmastery on the ability continuum of interest. The division between these two regions is the selected cut-off score or percentage.

Now, let U represent the region of masters in R, and let U' represent the region of nonmasters in that same state space. The hypothesis of interest can then be written as:

$$H = \underset{U}{\wedge} c \ \& \ \underset{U'}{\wedge} \bar{c} \qquad (3.1)$$

where: c = a designated master.

$\bar{c}$ = a designated nonmaster.

"$\underset{U}{\wedge} c$" = "c is a cell in U"

and,

"$\underset{U'}{\wedge} \bar{c}$" = "$\bar{c}$ is a cell in U'".

Equation 3.1 appears here exactly as stated by Tornebohm (1968).

The type of evidence by which the hypothesis H is tested are of the kind that measurements on the group of interest, made by means of some instrument, point to values in the M state space. These values in M-space in turn produce images in R space. And, upon the nature of these reverse images, decisions can be made as to an individual object's region of existence.

The application to a mastery/nonmastery testing situation easily follows. Individuals in a group are tested by a CRM measure, a set of scores corresponding to the values in M-space result, and on the basis of a chosen cut-off score, these values are applied back to R-space and locate each individual in either a mastery or nonmastery region of existence. Keeping in mind Ebel's operational definition of reliability, the hypothesis of interest in this case will concern the extent to which images produced in the mastery and nonmastery regions of existence remain consistent from CRM measure to a second independent and equivalent such measure.

Hypothesis Confirmation as "Degree of Covering"

Returning to the development of an index of hypothesis confirmation, Tornebohm lists three necessary definitions. They are as follows:

Def. 1: $I(H) = -\log p(H)$

Def. 2: $I(H/E) = I(HE) - I(E)$

Def. 3: $I(H) > 0 \rightarrow Dc(H/E) = \left(\frac{I(H) - I(H/E)}{I(H)}\right)$

The first definition is familiar from the discussion of the basic concepts of information theory in Chapter II. This is simply a measure of the amount of information in a hypothesis H. However, it should also be recalled from Chapter II, that when information is expressed in terms of an expected value, the measure becomes one of uncertainty, since by definition, expected information and uncertainty are synonymous. It will prove valuable later in this discussion to speak of the information contained in a hypothesis H as the amount of potential information, or uncertainty, which can in turn be shared by the evidence collected.

The second definition can be considered to be a measure of the amount of information that hypothesis H adds to evidence E. Or alternately, I(H/E) represents the amount of information contained in H that remains after the information common to E, I(E), is subtracted from the amount of information in both H and E, I(HE). Again, it is conceptually helpful to think in terms of the expected information in H as uncertainty. In this event, I(H/E) corresponds to the amount of uncertainty which remains in hypothesis H after the shared information communicated by evidence E is subtracted out.

The third definition provides an index for the degree of confirmation of a hypothesis by evidence, Dc(H/E). Tornebohm refers to this index as an estimate of "the degree of covering". The definition begins by assuming that the uncertainty contained in H is greater than zero. Some uncertainty must exist in a situation before any information can be obtained concerning it.

Now, if Dc(H/E) is to be used as an index of evidential strength, there exist certain conditions which it should satisfy. To more easily facilitate the determination of whether Dc(H/E) fulfills these conditions, the expression for I(H/E) in Definition 2 will be substituted into Definition 3 to yield the following equation:

$$Dc(H/E) = \left(\frac{I(H) - I(HE) + I(E)}{I(H)}\right) \qquad (3.2)$$

A first condition which a degree of evidential strength should fulfill, is that if the evidence logically implies the hypothesis the "degree of covering" should be at a maximum. Evidence would logically imply a hypothesis only if there was a perfect overlap in these two measures of information; that is, if all of the expected information in the hypothesis was communicated by the evidence. In this case, I(H) would equal I(E), and -I(HE) would equal -(E). Upon substituting this value for I(HE) into 3.2, it can be seen that the ratio, Dc(H/E), would equal 1. And, it should be noted that this is the maximum value a degree of evidential strength should be able to assume, since no more information can be transmitted by evidence than there exists uncertainty in the hypothesis.

Secondly, an index of the degree to which evidence confirms a hypothesis should be at a minimum when the information contained in the evidence is completely independent of the expected information contained in the hypothesis. In this respect, it was noted in Chapter II that if two sources of information were independent of one another, their combined information was equal to the sum of their individual measures of information. This is a familiar notion from probability theory, and in this case denotes that there is no overlap in information between the two sources.

Now, if the above were the case, -I(HE) would equal $-\Big(I(H) + I(E)\Big)$. And upon substitution into equation 3.2, the numerator can be seen to cancel to zero, which would of course cause the "degree of confirmation" to likewise equal zero.

Therefore, Tornebohm's index of degree of evidential strength does indeed assume a minimum and maximum under the appropriate conditions.

As it will prove to be of importance, it should be noted that if the evidence E is fully implied by the hypothesis H, then I(E/H) would equal zero. That would of course be the desired case, since the obtained evidence cannot add any information to the expected value of information already contained in the hypothesis. With that in mind, equation 3.2 can be simplified thusly:

$$Dc(H/E) = \left(\frac{I(H) - I(HE) + I(E)}{I(H)}\right)$$

$$I(H)Dc(H/E) = I(E) - \left(I(HE) - I(H)\right)$$

$$I(H)Dc(H/E) = I(E) - I(E/H)$$

$$I(H)Dc(H/E) = I(E)$$

and,

$$Dc(H/E) = \frac{I(E)}{I(H)} \qquad (3.3)$$

Thus, Dc(H/E) as the ratio of the information in evidence E to the information in hypothesis H, is the degree to which the information contained in H is conveyed by the evidence E. It may again be conceptually easier to think of this relationship as the degree to which the expected information or uncertainty in the stated hypothesis H, is "covered" by the information transmitted from the evidence E.

The next major section will apply Tornebohm's "degree of covering" ratio as a model in the development of a suggested reliability coefficient for CRM measures.

Development of Problem Solution - Symbols and Definitions

The development of the operational form of a suggested reliability coefficient for CRM tests begins by assuming that the individuals in a group of interest have been evaluated as being either masters or nonmasters regarding achievement of some subject area content, on the basis of scores obtained from the administration of a Test A. After some passage of time, this same group is again individually adjudged to be masters or nonmasters of the same subject area content, on the basis of scores obtained from a Test B. It is also the case that Tests A and B are designed with the intent of being equivalent measures of the same set of stated objectives. And finally, the administration of the two tests are considered to be independent of one another.

Such a situation of course, corresponds to that required by Ebel's operational definition of test score reliability, with the difference in this case being that instead of dealing with score values, the results under study are classification decisions concerning regions of existence. This would need to be the case if a coefficient is to result, which will be within Novick's dimension of "threshold loss".

If the above described test-retest design is executed, the following sets of observations will result:

N = number of students taking both tests.

$N(U_A)$ = number of classified masters on Test A.

$N(U_B)$ = number of classified masters on Test B.

$N(U'_A)$ = number of classified nonmasters on Test A.

$N(U'_B)$ = number of classified nonmasters on Test B.

$N(U_o)$ = number of consistently classified masters on the two testings.

$N(U'_o)$ = number of consistently classified nonmasters on the two testings.

On the basis of these classifications, the following proportions can be generated:

let; $x_o = \frac{N(U_o)}{N}$ and, $x'_o = \frac{N(U_o')}{N}$

and; $x_A = \frac{N(U_A)}{N}$ and, $x_B = \frac{N(U_B)}{N}$

and, $x'_A = 1 - x_A$

$x'_B = 1 - x_B$

Thus:

x_o = the proportion of students in the group who are consistently designated masters on the two testings.

x_o' = the proportion of students in the group who are consistently designated nonmasters on the two testings.

x_A = the proportion of students who are designated masters on Test A.

x'_A = the proportion of students who are designated nonmasters on Test A.

x_B = the proportion of students who are designated masters on Test B.

x'_B = the proportion of students who are designated nonmasters on Test B.

It will also be necessary to the following discussion to let:

c_i = an individual classified consistently as a master on Tests A and B.

and,

$\bar{c}_j$ = an individual classified consistently as a nonmaster on Tests A and B.

In regard to the notion of "threshold loss", the extent to which the two independent and equivalent CRM instruments yield consistent mastery/nonmastery decisions is of prime interest. What is desired in such a situation is the degree to which the evidence obtained lends support to the hypothesis that the two testings yield consistent classification decisions. Thus, the notion of CRM instrument reliability appears analogous to the relationship between evidence and hypothesis expressed by Tornebohm's "degree of covering".

Relationship of Evidence and Hypothesis to True Score and Total Score Variance

If the hypothesis concerns the degree to which independently administered equivalent CRM instruments yield consistent mastery/nonmastery decisions, the degree of confirmation will reflect the extent to which the information contained in the evidence removes the uncertainty contained in the hypothesis. What the author believes to be of especial importance, in determining the appropriateness of the herein suggested solution, is the similarity between the above relationship of evidence to hypothesis, and that between true score variance and total score variance in traditional test theory.

It was noted in Chapter I that Glass and Stanley (1970) compare score variance to the notion of uncertainty - a comparison which now appears clearly appropriate. One cannot allocate, partition, or account for more variance than already exists in a set of scores. And within traditional hypothesis testing, decisions are made as to whether to confirm or reject a hypothesis on the basis of the results of such allocation, partitioning, or accounting for of total score variance. In the same vein, it is not possible for evidence to convey more information than there exists uncertainty in the hypothesis.

In the case of the situation under current study, the uncertainty existent in a hypothesis concerning consistent classifications by independent and equivalent CRM instruments, is to some degree "covered" by the information conveyed by the extent of such consistent classifications in the evidence gathered. When Tornebohm's "degree of covering" is applied as a model, an index of the extent of overlap between the information in the evidence and the uncertainty in the hypothesis is obtained. This ratio would seem to be analogous to the theoretical definition of reliability as a ratio expressing the degree of overlap between true score variance and total score variance.

An Expression for "Degree of Covering"

Returning to the development of a suggested reliability coefficient it is a basic definition of information theory that the amount of information provided by a single consistent mastery/nonmastery classification, would be:

$-\log p(c_i)$ or, $-\log p(\bar{c}_j)$

Or alternately:

$-\log \frac{N(U_o)}{N}$ or $-\log \frac{N(U'_o)}{N}$

This measure of information would imply that the amount of information conveyed would be the same for each consistent master and the same for each consistent nonmaster.

Now, if it can be assumed that the individual masters and nonmasters are so designated independently of one another - and this should certainly be the case in a CRM decision framework - then the information conveyed by the evidence obtained from one testing would be the sum of the information conveyed by the individually classified masters and nonmasters. And, the total amount of information provided by the evidence relevant to such a framework would be that conveyed by the evidence from the combination of the two testings. However, since these two testings are to be considered independent of one another, the information provided by the total evidence would equal the sum of the information conveyed by each of the individual tests. Therefore, the equation for Dc(H/E) would in this case become:

$$Dc(H/E) = \frac{I(E_A) + I(E_B)}{I(H)} \tag{3.4}$$

where, $I(E_A)$ = the information contained in the evidence from Test A.

$I(E_B)$ = the information contained in the evidence from Test B.

Development of Formula for Coefficient Iota (i)

Since the individual consistently classified masters and nonmasters can be assumed to be independent of one another, the information contained in either Test A or Test B will be the sum of the information transmitted by each of these individual classifications. This being the case, the numerator of 3.4 becomes:

$$I(E_A) + I(E_B) = \left[\sum_{i=1}^{\overline{N}(U_o)} I(c_i) + \sum_{j=1}^{N(U'_o)} I(\overline{c}_j)\right]$$

$$+ \left[\sum_{i=1}^{\overline{N}(U_o)} I(c_i) + \sum_{j=1}^{N(U'_o)} I(\overline{c}_j)\right] \tag{3.5}$$

Recalling that each consistent master yields the same amount of information, and that it is the same case for each consistent nonmaster, 3.5 reduces to:

$$I(E_A) + I(E_B)$$

$$= \left[N(U_o) \quad I(c_i) \; + \; N(U'_o) \quad I(c_j) \right]$$

$$+ \left[N(U_o) \quad I(c_i) \; + \; N(U'_o) \quad I(c_j) \right] .$$

$$= \left[N(U_o) \left(-\log \frac{N(U_o)}{N} \right) + N(U'_o) \left(-\log \frac{N(U'_o)}{N} \right) \right]$$

$$+ \left[N(U_o) \left(-\log \frac{N(U_o)}{N} \right) + N(U'_o) \left(-\log \frac{N(U'_o)}{N} \right) \right]$$

$$= 2 \left[N(U_o) \left(-\log \frac{N(U_o)}{N} \right) + N(U'_o) \left(-\log \frac{N(U'_o)}{N} \right) \right]$$

$$= -2N \left[\frac{N(U_o)}{N} \left(\log \frac{N(U_o)}{N} \right) + \frac{N(U'_o)}{N} \left(\log \frac{N(U'_o)}{N} \right) \right]$$

$$= -2N \, (x_o \log x_o + x'_o \log x'_o) \qquad (3.6)$$

Hence, equation 3.6 reflects the equivalency of the two measures, in that each conveys the same amount of information, as well as their independence, in that the amount of information conveyed by the total evidence is equal to the sum of the information sources.

Now, let us examine the expected information or uncertainty contained in the hypothesis as related to the observed situation. This measure will, first of all, need to take into account the potential information contained in both testing situations. Secondly, in keeping with the traditional notion of hypothesis testing, I(H) should also be a function of sample estimates of the population proportions of consistent masters and consistent nonmasters. With these restrictions in mind, I(H) can take the following form:

$$I(H) = -\log p(H)$$

$$= \left[N(U_A)\ I(c_i) + N(U'_A)\ I(c_j) \right]$$

$$+ \left[N(U_B)\ I(c_i) + N(U'_B)\ I(\bar{c}_j) \right]$$

$$= \left[N(U_A) \left(-\log \frac{N(U_o)}{N} \right) + N(U'_A) \left(-\log \frac{N(U'_o)}{N} \right) \right]$$

$$+ \left[N(U_B) \left(-\log \frac{N(U_o)}{N} \right) + N(U'_B) \left(-\log \frac{N(U'_o)}{N} \right) \right]$$

$$= -N \left\{ \left[\frac{N(U_A)}{N} \left(-\log \frac{N(U_o)}{N} \right) + \frac{N(U'_A)}{N} \left(-\log \frac{N(U'_o)}{N} \right) \right] \right.$$

$$\left. + \left[\frac{N(U_B)}{N} \left(-\log \frac{N(U_o)}{N} \right) + \frac{N(U'_B)}{N} \left(-\log \frac{N(U'_o)}{N} \right) \right] \right\}$$

$$= -N \left[(x_A \log x_o + x'_A \log x'_o) + (x_B \log x_o + x'_B \log x'_o) \right] \qquad (3.7)$$

The results of equations 3.6 and 3.7 can now be substituted into equation 3.4 to obtain the following:

$$\frac{I(E_A) + I(E_B)}{I(H)}$$

$$= \frac{-2N(x_o \log x_o + x'_o \log x'_o)}{-N\left[(x_A \log x_o + x'_A \log x'_o) + (x_B \log x_o + x'_B \log x'_o)\right]}$$

$$= \frac{2(x_o \log x_o + x'_o \log x'_o)}{(x_A \log x_o + x_B \log x_o + x'_A \log x'_o + x'_B \log x'_o)}$$

$$= \frac{2(x_o \log x_o + x'_o \log x'_o)}{(x_A + x_B)(\log x_o) + (x'_A + x'_B)(\log x'_o)} \quad . \qquad (3.8)$$

The ratio as expressed in Equation 3.8, as a particular application of Tornebohm's concept of an index of "degree of covering" will be designated as coefficient iota (i).

The form of the denominator of coefficient iota deserves some comment. One might reasonably ask why, in determining a measure of expected information, the sample proportions of consistent masters and consistent nonmasters were employed, instead of the proportions of masters and nonmasters on the two tests. This form would after all, yield an index which would appear to be more consistent with the notion of uncertainty as statistically defined in Chapter II.

There are two reasons why the existing form of the denominator of coefficient iota was chosen - the first is mathematical, while the second is of a philosophical nature.

Comment on the Mathematical Nature of the Formula for Iota (i)

When considering the degree to which obtained evidence confirms a hypothesis of region of existence, it is in fact being assumed that a certain proportion p, of the N subjects in the population of interest, possess a property C. Faced with the inability to obtain information from each of the individuals in the population, a random sample is drawn from that population, and it is found that a certain proportion, s, of the individuals or objects in this sample, are observed to possess the property C. The proportion s is then used to obtain a measure of the information contained in the evidence. However, if this measured information is to be related to the expected information conveyed by the hypothesis to yield an index of the "degree of covering", the two information measures must have some basis of commonness. In short, there must be some way of knowing if the obtained information is relevant to the hypothesis being tested.

Tornebohm (1966) has demonstrated through the application of probability calculus that a ratio such as coefficient iota, which serves as a measure of evidential strength, will yield a measure of the commonness of the sample structure to the population structure. This measure of commonness is then shown to reach a maximum when the sample structure is used to estimate the population structure. That is, on the basis of this measure of commonness it can be asserted that upon obtaining "a random sample of size n from a population of size N ... it is most likely that the sample comes from a population such that those subsets which are like the sample are the most common kind of subsets".

This is of course the desired characteristic of any sampling procedure. But what is of importance to the purpose at hand, is that if a measure of the degree of evidential strength is to have this property, an estimate of the uncertainty or expected information contained in the hypothesis must include an estimate of the degree to which the characteristic of interest is manifested in the population. If the characteristic of interest is the consistency of region classification, the informational structure of the hypothesis must be formulated on the basis of an estimate of the frequency of that characteristic in the population. If not, evidence of consistency, as obtained from the sample, will lack a maximum degree of commonness when related to such a hypothesis. Simply stated, such evidence, when used to test a hypothesis which does not reflect an estimate of the property of interest, will lack a certain degree of relevance when compared to the situation in which the tested hypothesis includes an estimate of the studied characteristic.

As applied to the particular situation at hand, if the expected information contained in the hypothesis was formulated on the basis of the proportions of masters and nonmasters on each of the two testings, any evidence concerning the consistency of such classifications from one testing to the next, will lack a certain degree of relevance. Tornebohm's argument demonstrates that this degree of relevance, as reflected by a measure of commonness, is maximized when the hypothesis is stated in terms of the characteristic of interest.

Some Philosophical Implications of the Formula for Iota (i)

The second argument is rooted in the current generally accepted approach to hypothesis testing. Within the social sciences, hypotheses can never be proven either true or false. Hypotheses are either substantiated or rejected by obtained evidence within some chosen level of probability. Such an approach to research assumes that knowledge is advanced by means of a succession of formulated and tested hypotheses. Since none of these hypotheses can be held either totally true or false, each, at best, can be considered to be partially true. On this basis, one hypothesis succeeds another because sample evidence indicates that it possesses a higher degree of partial truth than its predecessor. Now, according to Stanley (1971), the logical perspective to the problem of reliability dictates that the method of data collection and statistical analysis must be logically consistent with the inference to be made. The hypothesis to be considered in the problem under current discussion is of course, the extent to which a particular CRM instrument yields consistent mastery/nonmastery decisions. As related to the notion of partial truth, evidence additional to that already obtained may lead us to change our position of belief as to the degree of consistent mastery/nonmastery decisions, and as to whether that estimated degree of consistency is acceptable to the purposes to which the test results are to be put.

If the chosen mode of statistical analysis is to be logically consistent with the inferences by which such a series of hypotheses advance, that analysis should result in a quantification of the uncertainty contained in a particularly stated hypothesis. Such a quantification must be a function of all the variables upon which these inferences are to be based if all the information available is to be taken into account. In the case of the type of reliability here being examined, this notion of logical consistency demands that the uncertainty contained in a particular hypothesis be a function of both the proportions of masters and nonmasters resulting from the two testings, and the proportions of consistent mastery/nonmastery decisions between the two tests. This requirement is satisfied by the form of coefficient iota as stated in Equation 3.8.

Before proceeding to Chapter IV, which will concern an application of coefficient iota on sample data, one further topic needs to be discussed. That topic concerns the range of possible values which can be assumed by coefficient iota.

The Range of Possible Values of Coefficient Iota (i)

As the reader is well aware, the range of possible values of a traditional reliability coefficient is from 0 to 1. And since coefficient iota is also a type of ratio, it would seem desirable to demonstrate that iota likewise assumes such a range of values, and in addition, that it assumes the extremes of that range under conditions which are conceptually compatible with the notions of consistency and reliability.

An analysis of the range of coefficient iota will be approached from two perspectives: the first from a consideration of iota as a measure of evidential strength developed within the framework of information theory; and, secondly from the aspect of iota as a mathematical expression.

The Minimum Value of Iota as a Measure of Evidential Strength

The very worst case from a consistency of classification point of view, would be if there were no consistent masters and no consistent nonmasters among the individuals classified by the results of two testings designed to be equivalent. This would necessarily result from the case where all individuals who were classified as masters on Test A were classified as nonmasters on Test B, and all those classified as nonmasters on Test A were classified as masters on Test B. Obviously, this is the most extreme example of inconsistency, and it would be expected that an index of consistency would be equal to zero under such circumstances.

If such a situation were to occur in reality there would of course be no need to calculate a coefficient of consistency since the evidence obtained in the form of test scores would indicate that whatever the two tests measure are independent of one another. Additionally, the talents and abilities sampled by the two tests are probably to some extent independent of the information and instruction conveyed in the learning component in question. In this case, no information was transmitted by the evidence in regard to the hypothesis being tested, and any measure of the degree to which such evidence confirms a hypothesis should be expected to be at its absolute minimum.

Tornebohm (1966, 1968) and Hilpinen (1970) provide examples of the manner in which the minimum value of evidential strength, such as coefficient iota, can be determined. Recall that the matter of present interest is the degree to which a particular hypothesis is substantiated by the evidence obtained, or, in another sense, the extent to which a particular hypothesis explains such obtained evidence. From the perspective of information theory, it would first be of interest to determine the amount of information which the hypothesis adds to the information provided by the evidence. This measure of "relative" information can be expressed as:

$$I(H/E) = I(HE) - I(E) \qquad (3.9)$$

where: $I(HE)$ = the information contained in both H (hypothesis) and E (evidence);

$I(E)$ = the information conveyed by E;

and, $I(H/E)$ = the amount of information H adds to the information conveyed by E.

A measure of evidential strength, or an index of the degree to which a particular hypothesis H is confirmed by the evidence E can then be expressed thusly:

$$D_c(H/E) = \left(\frac{I(H) - I(H/E)}{I(H)}\right) \tag{3.10}$$

where: $D_c(H/E)$ = the degree to which the information transmitted by H is shared by the information conveyed by E.

The value of I(H/E) in 3.9 can next be substituted into 3.10 to obtain:

$$D_c(H/E) = \left(\frac{I(H) - I(HE) - I(E)}{I(H)}\right) \tag{3.11}$$

Now, by definition, if the information carried by hypothesis H is totally independent of the information conveyed by evidence E, then:

$$I(HE) = I(H) + I(E) \tag{3.12}$$

The expression 3.12 indicates that if H and E carry relative amounts of information of a type which are independent of one another, then the information conveyed by a combination of H and E is simply equal to the sum of the information transmitted by each of the separate messages. This can be the case if and only if H and E are independent of one another in information carried and there is no overlap in the type of information transmitted. A much similar concept is a basic definition of probability theory.

If such independence between H and E was indeed the case, we would expect a measure of evidential strength to be at its minimum. That is, the information conveyed by evidence would be required to substantiate hypothesis H. Indeed, if the value of I(HE) in 3.12, given that H is independent of E on the basis of the information conveyed, is substituted into 3.11, the value of $D_c(H/E)$ becomes:

$$\begin{aligned} D_c(H/E) &= \left(\frac{I(H) - \left(I(H) - I(E)\right) - I(E)}{I(H)}\right). \qquad (3.13) \\ &= \left(\frac{I(H) - I(H) + I(E) - I(E)}{I(H)}\right) \\ &= \frac{0}{I(H)} \\ D_c(H/E) &= 0 \end{aligned}$$

Therefore, the minimum of a measure of evidential strength (coefficient iota) when considered from the perspective of information theory, is 0.

This minimum value of coefficient iota would be assumed when evidence E, in the form of obtained test scores, fail to convey any information concerning the substantiation of the hypothesis H that the two tests, which have yielded those scores, are equivalent in terms of the mastery/nonmastery classifications that result. Again, such would be the case if absolutely no examinees were consistently classified as either masters or nonmasters. Evidence of this type would be completely independent or unrelated to the tested hypothesis, with a result being that coefficient iota would assume a value consistent to that expected of a traditional reliability coefficient under the same conditions.

The Maximum Value of Iota as a Measure of Evidential Strength

In regard to the maximum value which coefficient iota can assume, it can readily be seen by inspection of 3.8 that iota can never be greater than 1. Such is the case since there can never be more consistent masters or nonmasters than there are masters and nonmasters on either of the individual testings considered individually. In addition, from the standpoint of information theory, it would be logically impossible for a hypothesis H to account for more information than that which is carried by the obtained evidence E, when considered on the basis of that information alone.

Conceptually, a measure of evidential strength would assume its maximum value when a given hypothesis H accounts for the total amount of information conveyed by evidence E. It seems reasonable to expect that a ratio of this type would assume a value of 1 at its maximum. Coefficient iota does indeed do so under two somewhat different sets of circumstances which will be considered separately.

The first case is the simplest and can be confirmed by mere inspection. Assume that all examinees classified as masters and nonmasters by Test A were to an individual similarly classified as such by the results of Test B. Likewise, assume that neither proportion of masters or nonmasters was equal to 0 or 1. In this case, since all mastery/nonmastery decisions are consistent from Test A to Test B, we would expect the degree of consistency to be perfect, and the index of the degree of consistency to assume a value of 1. In other words, all the information contained in the evidence E would be accounted for by the hypothesis H. Such a situation would result in the following equalities:

$x_o = x_A = x_B$, where $x_o \neq 0$ or 1;

and,

$x'_o = x'_A = x'_B$, where $x'_o \neq 0$ or 1.

And if these resultant equalities are substituted into Equation 3.8, it can readily be seen that coefficient iota would reduce to 1.

The second set of circumstances for which one would expect coefficient iota to be at its maximum is if all examinees are classified as masters on the basis of the results of both Test A and Test B, or all examinees are consistently classified as nonmasters by the two testings.

Returning to Equation 3.10, the value I(H/E) assumes under the conditions described immediately above is again of interest. First of all, recall that I(H/E) is defined as the amount of information that hypothesis H adds to the evidence E. In the case of either total examinee mastery of both Test A and Test B or total examinee nonmastery, the evidence E logically confirms the hypothesis H that the two tests yield consistent mastery/nonmastery decisions. From another point of view, since there were no inconsistent mastery/nonmastery decisions, or, no variance, there was no uncertainty contained in the evidence. Therefore, the hypothesis H could not add any information to the evidence E, and I(H/E) would equal 0. Inserting this value for I(H/E) into Equation 3.10, it can be seen that D_c(H/E) becomes:

$$D_c = \left(\frac{I(H) - 0}{I(H)}\right) \qquad (3.14)$$

$$D_c = \frac{I(H)}{I(H)}$$

$$= 1$$

It has been determined then, that coefficient iota, as a measure of evidential strength, has a maximum of 1 in the case where the evidence E logically confirms the hypothesis H, and, a minimum of 0 when the evidence E is logically independent of the hypothesis H. This range of values has been identified on the basis of that which would be expected of a measure of evidential strength when considered from the perspective of information theory. Such a result would be consistent with the range of values assumed by a traditional reliability coefficient, however the task remains to determine mathematically whether this is actually the case.

Before moving on to a mathematical consideration of the extremes of the range of values of coefficient iota, however, a point should be mentioned that is somewhat obvious. The sets of circumstances which are seen to result in coefficient iota being equal to 0 or 1 have practical implications which would render the calculation of any consistency coefficient unnecessary. In the case of the total lack of even a single consistent mastery/nonmastery decision from one testing to another, it would be self-evident to the examiner that either the two tests lacked even the slightest degree of equivalence, or else something had gone terribly wrong within the teaching/learning component itself. On the other hand, for either the case of total consistent mastery or total consistent nonmastery, it would readily be revealed to the examiner that in the former case the tests were too easy, or in the latter case that the two tests were too difficult. As always, the practical aspects of the individual situation must be considered. It may be possible that the examiner may be content with total consistent mastery if he/she is convinced that the two tests do a valid job of measuring the material covered in the specific teaching/learning component. However, even in the case of total consistent nonmastery, the practical aspects of the individual situation would have to be considered before making the decision that the tests were too difficult for those examinees who will be taking the tests. Nevertheless, this information would be directly revealed by the test classifications themselves and the calculation of an index of consistency or reliability would provide no further information. In short, the situations considered above are those situations in which the value of such an index would not need to be calculated.

The Minimum Value of Iota as a Mathematical Expression

Consideration of the range of values of coefficient iota from a mathematical perspective will also begin with examination of the set of circumstances corresponding to a complete lack of consistency in the mastery/nonmastery classifications yielded by the results of two testings. As the reader will recall, in such a case there are neither any consistent mastery decisions nor any consistent nonmastery decisions. This situation would result in both x_o and x'_o as they are found in Equation 3.8, being equal to 0.

It will be necessary to further discussion to note that the logarithm to the base 2 of 0 is $-\infty$. This value does not present any immediate difficulties however, since it can be seen by inspection of Equation 3.8, that upon substitution of 0 for the values of both x_o and x'_o, the numerator of coefficient iota becomes 0 while the denominator tends to $-\infty$. This would of course then result in the value of coefficient iota being equal to 0 under such circumstances. Therefore, it is seen from a second perspective that coefficient iota becomes 0 when calculated on the basis of mastery/nonmastery decisions which are completely inconsistent from one testing to another.

The Maximum Value of Iota as a Mathematical Expression

However, the value which coefficient iota assumes when there results either total consistent mastery or total consistent nonmastery is not as readily apparent. When total consistency is the result, either x_o or x'_o will equal 1, and the remaining value will necessarily equal 0. Having previously noted that the logarithm to the base 2 of 0 is $-\infty$, it remains necessary to note that the logarithm to the base 2 of 1 is 0.

As a means of demonstrating the behavior of coefficient iota under the conditions of total consistency, it will be arbitrarily chosen that x_o will equal 1 which necessarily determines that x'_o must equal 0. The results of the following proof would be the same if x'_o had been chosen to equal 1. Given that x_o equals 1, the following equalities would necessarily result from the definition of coefficient iota:

$$x_o = 1;$$

$$x_A = 1;$$

$$x_B = 1;$$

$$x'_o = 0;$$

$$x'_A = 0;$$

$$x'_B = 0;$$

$\log x_o = 0;$

and,

$\log x'_o = -\infty;$

In effect then, when either x_o or x'_o is equal to 1, the values of the remaining proportions in the equation for coefficient iota are necessarily determined and are no longer free to vary. When the above values are correspondingly substituted into Equation 3.8, the following results:

$$\text{Iota (i)} = \frac{2(x_o \log x_o + x'_o \log x'_o)}{(x_A + x_B)(\log x_o) + (x'_A + x'_B)(\log x'_o)}$$

$$= \frac{(2)(1(0) + 0(-\infty))}{(1+1)\ (0) + (0+0)\ (-\infty)}$$

$$= \frac{0}{0}$$

And, such a ratio is considered to be indeterminate in form.

It is somewhat of a misnomer however, that ratios of this type are labeled indeterminate when the variables of the function involved yield such a value. For this does not mean that such a function has a value when, in the case of coefficient iota either x_o or x'_o equal 1, but it can not be determined what that value is. In fact, if such functions have a limit, that is approach a particular value as, in this case, the value of the function tends to 0/0, it can indeed be determined what that limit is if it exists.

A method often employed in such situations is relatively simple and is referred to as L'Hospital's rule. This technique is frequently applicable to situations involving a ratio of two functions, say f and g of some variable x, wherein that ratio becomes indeterminate in form as x approaches a value c. Examples of indeterminate forms would be 0/0, ∞/∞, or $-\infty/-\infty$. In order to apply L'Hospital's rule to determine if such a ratio approaches a real value as x approaches c, the following five assumptions must be fulfilled (Fobes and Smyth 1963):

(1) Both f and g are continuous in the neighborhood of c.

(2) The derivatives of f and g, designated f' and g', exist in that neighborhood.

(3) The limit of f(x) as x approaches c is equal to the limit of g(x) as x approaches c, which is equal to 0 (zero), or

$$\lim_{x\to c} f(x) = \lim_{x\to c} g(x) = 0.$$

(4) The derivative of g(x), or g'(x), does not equal 0 (zero in the neighborhood of c).

(5) And, if the above assumptions hold, and

$\lim_{x\to c} \frac{f'(x)}{g'(x)}$ exists and equals a number r, then

$$\lim_{x\to c} \frac{f(x)}{g(x)} = \lim_{x\to c} \frac{f'(x)}{g'(x)} = r.$$

To begin application of L'Hospital's rule, the coefficient as it appears in Equation 3.8 is expressed as:

$$\frac{f(x_o)}{g(x_o)} = \frac{2(x_o \log x_o + x'_o \log x'_o)}{(x_A + x_B)(\log x_o) + (x'_A + x'_B)(\log x'_o)} \qquad (3.15)$$

This discussion will consider the above equation as a ratio of two functions of x_o. It is possible to consider only x_o since, as previously mentioned, as $x_o \longrightarrow 1$, the remaining proportions in the equation must by definition approach certain values, until when x_o does equal 1, the remaining values in the equation are necessarily determined. Therefore under these circumstances the value of coefficient iota is determined by the value of only one variable, x_o.

It can be seen from equation 3.15 that both $f(x_o)$ and $g(x_o)$ are continuous in the neighborhood of $x_o = 1$, since the logarithm to the base 2 of any value greater than 0 and less than 1, will result in a real number. Therefore the first assumption necessary for the application of L' Hospital's rule is fulfilled.

Prior to taking the derivatives of $f(x_o)$ and $g(x_o)$ as stated in the second assumption, it will be of aid to note that the derivative of $\log x_o$ when taken with respect to x_o, where the logarithm is to the base 2, is:

$$D_{x_o}(\log x_o) = \frac{\log_{10}e}{\log_{10}2}(1/x_o)\, D_{x_o}(x_o) \qquad (3.16)$$

where: $D_{x_o}(\log x_o)$ = the derivative of $\log x_o$ with respect to x_o;

and,

$$\frac{\log_{10}e}{\log_{10}2} = \frac{.43429}{.30100}$$

In the following discussion, the constant .43429/.30100 will be expressed as b.

With this added notation, the ratio of the derivatives of $f(x_o)$ and $g(x_o)$ when taken with respect to x_o result in:

$$\frac{f'(x_o)}{g'(x_o)} = \frac{2(x_o\ (1/x_o)\ (b) + \log x_o)}{x_A\ (1/x_o)\ (b) + x_B\ (1/x_o)\ (b)}$$

$$= \frac{2(b + \log x_o)}{b\ (1/x_o)\ (x_A + x_B)} \qquad (3.17)$$

It can thus be noted that the derivatives of $f(x_o)$ and $g(x_o)$ exist in the neighborhood of $x_o = 1$ since all the values involved yield real numbers in both of the functions. Likewise, it is noted that the derivative of $g(x_o)$ does not equal 0 in the neighborhood of $x_o = 1$. And, it has been previously noted above that the limits of both $f(x_o)$ and $g(x_o)$ approach 0 as x_o approaches 1. Thus it has been shown that the first four assumptions necessary for the application of L'Hospital's rule are fulfilled in the case of coefficient iota. It remains then, to examine the ratio of the limits of these two functions as x_o approaches 1 to determine if the fifth and final assumption is satisfied. On doing this, the following is obtained:

$$\frac{\lim\limits_{x_o \to 1} f'(x_o)}{\lim\limits_{x_o \to 1} g'(x_o)} = \frac{2(b + \log 1)}{b(1/1)\ (x_A + x_B)} \tag{3.18}$$

$$= \frac{2b}{b(x_A + x_B)}$$

And, since all individuals were classified as masters on both Test A and Test B, it has been noted earlier that x_A and x_B would both be equal to 1, Equation 3.18 thus becomes:

$$\lim_{x_o \to 1} \quad \frac{f'(x_o)}{g'(x_o)} = \frac{2b}{b(1+1)} \tag{3.19}$$

$$= \frac{2b}{2b}$$

$$= 1$$

Hence, the ratio of the limits of the derivatives of $f(x_o)$ and $g(x_o)$, as x_o approaches 1, is seen to exist and is equal to 1. It thereby follows from the fifth assumption of L'Hospital's rule, that the limit of coefficient iota, as either x_0 or x'_0 approach 1, and the remaining proportion necessarily approaches 0, exists and is equal to 1. This is of course the result that would be desired of an index of consistency when applied to a situation in which mastery/nonmastery classifications are totally consistent from one test administration to the next, and the two tests are designed to be equivalent in regard to the talent and ability sampled.

It has been demonstrated then, that coefficient iota does assume a range consistent with that which would be expected of an index of evidential strength developed within the conceptual framework of information theory. This demonstrated range of values is also consistent with the traditional concept of reliability and consistency. In addition, the extreme values of this range are assumed under those conditions which are also compatible with the traditional notion of reliability and consistency.

Summary

An attempt was made in this chapter to demonstrate both the conceptual and statistical similarities between a ratio expressing the "degree of covering" of the uncertainty contained in a hypothesis H by the amount of information conveyed in obtained evidence E, and, the ratio of true score variance to total score variance. This latter ratio being the theoretical definition of a traditional reliability coefficient. The notion of "degree of covering", as expressed in two articles by Hakan Tornebohm (1966, 1968), was then applied as a model in the development of a coefficient designed to serve as an index of the degree to which two tests, designed to be equivalent, yield consistent mastery/nonmastery decisions. This suggested index was designated coefficient iota (i). Finally, it was determined from both conceptual and mathematical perspectives that coefficient iota assumes a range of values from 0 to 1, and assumes the extremes of this range under conditions compatible with the traditional concepts of reliability and consistency.

It will be the purpose of Chapter IV to apply coefficient iota to sample data, and analyze its behavior in a manner similar to the comparative study undertaken by Subkoviak (1978).

CHAPTER IV

ANALYSIS AND RESULTS

Data Base

The data base of the analysis consisted of the responses of 2182 eighth and ninth grade students to the items on a CRM mathematics instrument, published by Science Research Associates (SRA), Inc. This instrument consists of 120 items, evaluating the mastery of 40 objectives, with each objective being represented by three items. The objectives range in difficulty from the addition of three positive integers, to determining the volume of three-dimensional solids.

Formulation of Parallel Test Forms

Item difficulties and item discriminations were obtained for each of the 120 items on the test instrument. Out of these 120 items, parallel forms were created at each of 30, 20, and 10 item-length levels. This was accomplished by selectively deleting items from the total of 120 on the basis of content, difficulty, and discriminating power.

In terms of content, it was assured that each objective having items in the reduced total from which the parallel forms were created, was represented by one, and only one, item on each of the forms. Thus the parallel forms at each item-length level would evaluate the same objectives. In addition, the pairs of items per objective were selected on the basis of similarity in difficulty and discriminating power.

The index of item difficulty is simply the percentage or proportion of examinees who answered the items correctly. Such an index, therefore, gives a ready indication of how easy or difficult the item was for the entire group. In creating the parallel forms, it was considered necessary that each item of the pair chosen to represent a particular objective, have similar item difficulty values.

The third criterion used in creating the parallel forms was that of the discriminating power of an item, measured on the basis of the item's index of discrimination. The formula used is that developed by Johnson (1951), and is as follows:

$$D_j = \frac{R_U - R_L}{1/2\ T} \qquad (4.1)$$

where:

D_j = the index of discrimination of item j.

R_U = the number of examinees having total test scores in the upper half of the group, and answered the item correctly.

R_L = the number of examinees having total test scores in the lower half of the group, and answered the item correctly.

T = the total number of examinees in the group.

As noted by inspection of the above formula, items with negative discrimination values would most certainly be poor ones. Such is the case since this would reveal that more examinees in the lower half of the group answered the item correctly, than examinees in the upper half. In terms of how high an item discriminator should be, Ebel (1972, p. 399) offers the below evaluation criteria.

Table 4.1

Interpretation of Item Discrimination Values

Index of Discrimination	Item Evaluation
0.40 and up	Very good items
0.30 to 0.39	Reasonably good, but possibly subject to improvement
0.20 to 0.29	Marginal items, usually needing and subject to, improvement
Below 0.19	Poor items, to be rejected, or improved by revision

In selecting items for inclusion in the construction of the parallel forms, Ebel's criteria were used as a guideline, as well as, the similarity between the values of the indices of discrimination for the items in each pair.

The three criteria of content, difficulty, and discriminating power, were thus used to delete items from the total of 120 to create parallel forms of 30 items each. The same process was then used to create the two smaller parallel forms of 20 and 10 items each. The construction of these parallel forms was accomplished by the same procedure as that used by Subkoviak (1978).

Table 4.2 lists the indices of difficulty and discrimination for each of the items in the 30 item pairs from which the parallel forms were created. The pairs which were used in the creation of the 20 and 10 item parallel forms are indicated. The average item difficulties and discriminations for the various parallel forms are reported in Table 4.3.

Further descriptive information regarding the parallel forms is also provided by the respective means, standard deviations, and KR-20 reliabilities reported in Table 4.4. The value of these statistics are based on the entire

Table 4.2

Item Difficulties and

Discriminations of 30 Item Pairs

Item Pair #	Form	Item Difficulty	Item Discrimination
1**	A	.621	.555
1**	B	.630	.562
2**	A	.412	.566
2**	B	.460	.594
3**	A	.334	.421
3**	B	.443	.451
4*	A	.582	.298
4*	B	.424	.420
5**	A	.430	.506
5**	B	.434	.464
6*	A	.758	.389
6*	B	.766	.378
7**	A	.638	.454
7**	B	.661	.449
8*	A	.692	.419
8*	B	.671	.429
9*	A	.582	.324
9*	B	.438	.392
10*	A	.426	.510
10*	B	.384	.503
11	A	.442	.291
11	B	.441	.281
12**	A	.356	.477
12**	B	.336	.445
13	A	.314	.246
13	B	.178	.226
14*	A	.449	.442
14*	B	.670	.388
15**	A	.468	.411
15**	B	.637	.395

* Used in creation of 20 item parallel forms

** Used in creation of both 20 and 10 item parallel forms

Table 4.2 (Continued)

Item Difficulties and

Discriminations of 30 Item Pairs

Item Pair #	Form	Item Difficulty	Item Discrimination
16	A	.675	.492
16	B	.315	.292
17*	A	.758	.373
17*	B	.630	.478
18*	A	.512	.440
18*	B	.486	.369
19	A	.741	.415
19	B	.476	.347
20**	A	.564	.415
20**	B	.513	.425
21	A	.406	.250
21	B	.814	.322
22*	A	.238	.258
22*	B	.343	.321
23	A	.326	.252
23	B	.317	.226
24	A	.509	.263
24	B	.385	.249
25	A	.659	.282
25	B	.455	.301
26**	A	.418	.378
26**	B	.415	.396
27*	A	.494	.367
27*	B	.378	.319
28	A	.311	.282
28	B	.386	.409
29	A	.239	.248
29	B	.215	.234
30*	A	.631	.408
30*	B	.349	.342

* Used in creation of 20 item parallel forms

** Used in creation of both 20 and 10 item parallel forms

Table 4.3

Average Item Difficulties and

Discriminations of Parallel Forms

Form	Item Length	Average Item Difficulty	Average Item Discrimination
A	30	.497	.382
B	30	.468	.380
A	20	.518	.420
B	20	.503	.426
A	10	.467	.469
B	10	.491	.468

Table 4.4

Means, Standard Deviations, and
KR-20 Reliabilities of Parallel Forms

Statistics	Form	Test Length		
		10	20	30
Mean	A	4.60	10.21	14.69
	B	4.85	10.38	14.51
Standard Deviation	A	2.89	5.03	6.78
	B	2.88	5.05	6.78
KR-20 Reliability	A	.702	.802	.837
	B	.698	.805	.833

2182 students in the population.

Results of Analysis

To summarize thus far then, the creation of the parallel forms made available a distribution of scores for the responses of the 2182 students on each form, at each level of 10, 20, and 30 items. As also was the procedure of Subkoviak (1978), mastery criterion of 50%, 60%, 70%, and 80% correct were considered for each of the pairs of parallel forms, at each item-length level. Twelve values of coefficient iota were then obtained through calculations over the entire population, at each item-length by mastery criterion level. These parameter values are recorded in the third column of Table 4.5. The remainder of Table 4.5 reports the results of the final step in the analysis. At each item-length by mastery criterion level, 50 random samples of 30 students each were selected from the population of 2182 student test scores. This sampling consisted of mastery/nonmastery decisions based upon the respective criterion level. Coefficient iota was calculated for each sample drawn at each item-length by mastery criterion level. The fourth column of Table 4.5 reports the means of coefficient iota for the 50 random samples at each level, as well as the standard deviation of the sampling distribution for each of the item-length by mastery criterion

Table 4.5

Results of Analysis

Mastery Criterion	Item Length	Population Parameter	Sample Mean	Standard Error
50%	10	.88	.88	.06
	20	.87	.86	.07
	30	.89	.87	.05
60%	10	.87	.85	.06
	20	.88	.87	.07
	30	.88	.88	.08
70%	10	.88	.88	.08
	20	.88	.88	.08
	30	.86	.86	.08
80%	10	.87	.86	.10
	20	.84	.84	.09
	30	.82	.84	.13

levels.

Discussion of Results

On inspection of Table 4.5, comparison of the individual sample means with their respective parameter values, would indicate that coefficient iota estimates are unbiased. The sample values of the standard deviation, or estimates of the standard error of coefficient iota, are provided basically for discussion purposes. These values must of course be considered relative to sample size. The values of the estimate of the standard error obtained from each sampling distribution could be reduced simply by in-. creasing sample size. However, consideration of the behavior of the values of the estimates will enter into later discussion.

Insight into the nature of the estimates obtained from the various sampling distributions of coefficient iota may be best served by consideration of the results reported by Subkoviak (1978) in his comparison of four types of suggested reliability coefficients for criterion-referenced mastery tests. The results of this are reproduced here as Table 4.6.

It is necessary to further discussion to recall that the procedures considered by Subkoviak all concern the

Table 4.6

Results of Subkoviak's Comparison of Four Suggested Reliability Coefficients

Mastery Criterion	Test Length	Parameter	Swaminathan		Marshall		Subkoviak		Huynh	
			Mean	St. Error	Mean	St. Error	Mean	St. Error	Mean	St. Error
50%	10	.67	.68	.08	.74	.08	.66	.06	.66	.06
	30	.79	.79	.07	.82	.04	.81	.04	.80	.03
	50	.83	.84	.06	.84	.03	.84	.03	.83	.02
60%	10	.72	.72	.07	.75	.05	.69	.06	.67	.06
	30	.84	.83	.06	.84	.03	.84	.04	.82	.03
	50	.87	.87	.06	.87	.03	.88	.03	.86	.02
70%	10	.80	.79	.08	.79	.03	.79	.05	.76	.06
	30	.88	.88	.06	.88	.03	.89	.04	.88	.03
	50	.91	.91	.05	.91	.03	.93	.03	.91	.02
80%	10	.88	.87	.06	.85	.04	.90	.05	.86	.05
	30	.94	.93	.05	.93	.03	.95	.03	.94	.02
	50	.96	.96	.08	.96	.02	.97	.02	.96	.02

proportion (P_c) of students in a population who are classified as either consistent masters or consistent nonmasters, on the basis of scores obtained from a test-retest situation. The parameter values reported in Table 4.6 then, are the population values of P_c at each item-length by mastery criterion level. The size of the population in the Subkoviak study was, as mentioned previously, 1586 students. The Swaminathan procedure is the actual value of P_c obtained from a sample, while the remaining three procedures are different types of estimates of P_c obtained from a single testing. The sampling procedure of drawing 50 random samples of 30 students each, at each level, was the same as that of the present study.

The parameter values in Table 4.6 are seen to increase markedly as either the mastery criterion or the item-length levels are increased. This is of course to be expected with such a proportion. As the mastery criterion becomes more extreme in either direction, classification will become more consistent. The mean score value for the parallel forms in the Subkoviak study were approximately 50% of the total, as they were in the present study. Therefore, as the mastery criterion increases, the proportion of consistent nonmasters increases, which has the overall effect of increasing P_c. The test essentially becomes too difficult for the students. An increase in item length will result in an increase in P_c, simply because a more representative sampling of the students' level of ability is being obtained.

In comparison, the parameter values of coefficient iota in Table 4.5 are relatively stable over changes in either mastery criterion level or item-length. However, it must be recalled that iota is neither the value of P_c obtained from a sample or population of student mastery/nonmastery classifications, nor an estimate of P_c. It is true that the formula for coefficient iota involves the proportion of consistent masters and nonmasters in a sample or population, however the formula takes more than the value of P_c into consideration.

Iota is an estimate of the extent to which a certain amount of obtained information relieves or "covers" a certain amount of given uncertainty. The uncertainty created in this instance evolves from the hypothesis that two parallel test forms, of the same item-length, will yield consistent mastery/mastery and consistent nonmastery/nonmastery decisions, on the basis of a chosen mastery criterion level. In estimating the extent to which the obtained information covers the uncertainty created by this particular hypothesis, it is seen here as necessary to consider not just the information provided by the value of P_c, but also the probability of the individual decisions which determined the value of P_c. Consideration of this second factor of probability, while being basic to information theory, also points out a similarity between coefficient iota and traditional reliability coefficients.

As noted in Chapter I, the traditional theory of reliability depends to a great extent on the variability of test scores. And, it was seen in Chapter I also, that the expected lack of variability in test scores resulting from criterion-referenced tests often made the use of traditional reliability coefficients impossible in such cases. Additionally, the concept of the extent to which individual test scores vary from one another was posited to be logically inconsistent, when used in estimations of the accuracy of mastery/nonmastery criterion-referenced decisions. Nevertheless, basic to the conceptual nature of coefficient iota as a proposed estimate of reliability to be used with mastery/nonmastery decisions on the basis of results from criterion-referenced mastery tests, is the theoretical similarities between variance and uncertainty.

One of the simplest traditional measures of score variance is the standard deviation. It can be easily seen from the formula for the standard deviation, that not all scores will contribute the same amount of information, if you will, to the value of the statistic. Those scores which fall on either extreme of the score distribution, will of course have more extreme deviations from the mean of the distribution, and will contribute more to the sum of squares which will yield the value of the standard deviation. And, if your interest is in the extent to which individual scores vary from one another, this is exactly the way things should be. It would not make sense to give an extremely deviant score the same weight in the determination of score variance, as a score which occurs near the mean of the distribution.

A traditional reliability coefficient attempts to estimate the extent to which total observed test score variance can be explained, or accounted for, by the variance of true test scores. In the same manner, information theory is concerned with the extent to which the uncertainty existent in a particular situation can be relieved by obtained information. However, just as individual test scores do not contribute equivalently to a measure of variance, individual events do not contribute equivalently to a measure of uncertainty.

One of the basic theoretical concepts of information theory is that the more improbable an event, the greater the information that is conveyed by that event's occurrence. This is analogous to the notion that the more extreme a test score within its distribution, the greater its contribution to a measure of test score variance. The operational definition that the amount of information conveyed by an event with a particular probability of occurrence, is recalled to be the logarithm to the base 2 of the event's probability. The values in the Table of Appendix B can be seen to clearly reflect this theoretical concept. A lower probability of occurrence results in a greater amount of measured information.

An example of how this relationship affects coefficient iota estimates, and one of the differences between these estimates and estimates of P_c, can be illustrated by discussion of the values in Table 4.7.

It can be seen from comparison of the values of P_c and iota, across item-length and mastery criterion levels, that while P_c increases as the mastery criterion level increases, iota tends to decrease. And, as would be expected, it can also be seen from comparison of the values of x_o and x'_o, that as the mastery criterion increases the majority of consistent classifications are nonmastery/nonmastery. The question may be raised as to why iota does not likewise increase.

Table 4.7

Comparison of Parameter Values of P_c and Iota*

Mastery Criterion	Item Length	Parameter Values of Indices							
		P_c	Iota	x_o	x'_o	x_A	x'_A	x_B	x'_B
50%	10	.88	.88	.44	.45	.48	.52	.51	.49
	20	.87	.87	.46	.41	.51	.49	.54	.46
	30	.89	.89	.42	.47	.47	.53	.47	.53
60%	10	.88	.87	.33	.55	.39	.61	.41	.59
	20	.89	.88	.35	.54	.39	.61	.41	.59
	30	.90	.88	.29	.61	.34	.66	.33	.67
70%	10	.90	.88	.25	.66	.28	.72	.31	.69
	20	.90	.88	.24	.66	.28	.72	.30	.70
	30	.91	.86	.18	.73	.23	.77	.22	.78
80%	10	.92	.87	.18	.75	.20	.80	.23	.77
	20	.91	.84	.14	.77	.17	.83	.20	.80
	30	.94	.82	.08	.86	.11	.89	.11	.89

* Values of iota calculated from the proportions in this Table may differ from those reported in Table 4.5 as a result of rounding errors.

It is of value here to recall from Chapter II one of the criticisms which has been raised regarding many suggested reliability coefficients for CRM tests. The criticism in question is that all inconsistent mastery/nonmastery classifications are treated equally in the estimation of test reliability. In considering what is implied here, recall the example of various deviation distances and their contribution to the value of a standard deviation. In this case, extreme deviations can occur in two directions, above the mean and below it. The value of the standard deviation is said to be sensitive to such extreme scores. Such extreme scores contribute more information to the calculation of the standard deviation than do scores relatively closer to the mean of the distribution. This relationship between extreme scores and the calculated value of the standard deviation is analogous to the relationship between the probability of consistent mastery/mastery and nonmastery/nonmastery decisions, and the calculated value of iota.

The latter relationship referred to immediately above, can be revealed through examining the parameter values of the proportions in Table 4.7. It can be noted that at the 50% mastery criterion level the proportions of consistent masters (x_o) and the proportions of consistent nonmasters (x'_o), differ from the respective proportions of masters (x_A and x_B) and nonmasters (x'_A and x'_B) on the two test forms on an average of .058. It should also be noted that at the 50% criterion level the reported proportions are near the midpoint of the range of possible probabilities.

On proceeding to the 60% and 70% mastery criterion levels it is seen both, that the proportions of consistent nonmasters are increasing while the proportions of consistent masters is decreasing, and that the absolute differences in the proportions are decreasing. At the 60% criterion level the average absolute difference between the corresponding proportions is .055, while at the 70% level the average absolute difference is .047. While the respective absolute differences between the involved proportions have been decreasing, the reported value of iota has also been decreasing, although very slightly over the lower three criterion levels.

However, this trend becomes more pronounced at the 80% criterion level. Here the average absolute difference between the respective proportions decreases to .037, while iota is seen to decrease at a greater degree than at the three previous levels. Such a trend would indicate that the mastery/nonmastery classifications made at the 80%

criterion level are not as reliable as the classifications made at the 50%, 60%, and 70% levels. This is opposite to the conclusion that would be arrived at if P_c, or any one of its estimates were used as the coefficient of reliability. This would indicate a quite serious limitation of these measures as estimates of the reliability of CRM tests.

In discussing this limitation further, the question must be addressed as to why iota values decrease as the proportions involved in the ratio become closer in value. An initial conclusion might very well be to suspect that the value of iota would increase toward its maximum value of 1.00, as the absolute values of the proportions tended to become more similar. It might seem that the coefficient iota ratio in such a case would approach unity. However, iota is a ratio of obtained information to the uncertainty present, and the influence of probability on these two quantities must be considered. Again, it is best to refer to a table in explaining why iota behaves as it does.

The entries in Table 4.8 are the amounts of information and uncertainty, measured in bits and identified by source, which are used in the calculation of iota at the various item-length by mastery criterion levels. Column 3 reports the total amount of uncertainty present at a

Table 4.8

Amounts and Sources of Uncertainty and Information

Involved in the Calculation of the Parameter Values of Iota*

Mastery Criterion	Item Length	Uncertainty Created by Masters on Forms A & B	Uncertainty Created by Nonmasters on Forms A & B	Total Uncertainty	Information Obtained from Consistent Masters	Information Obtained from Consistent Nonmasters	Total Information
50%	10	1.17	1.16	2.33	1.04	1.04	2.08
	20	1.18	1.22	2.40	1.03	1.05	2.09
	30	1.18	1.15	2.33	1.05	1.02	2.07
60%	10	1.28	1.04	2.32	1.06	0.95	2.01
	20	1.21	1.07	2.28	1.06	0.96	2.01
	30	1.20	0.94	2.14	1.04	0.87	1.91
70%	10	1.18	0.85	2.03	1.00	0.79	1.79
	20	1.19	0.85	2.04	0.99	0.79	1.78
	30	1.11	0.71	1.82	0.89	0.67	1.56
80%	10	1.14	0.65	1.79	0.89	0.62	1.51
	20	1.05	0.61	1.66	0.79	0.58	1.37
	30	0.80	0.39	1.19	0.58	0.37	0.95

* Values of iota calculated from this Table may vary from those reported in Table 4.5 as a result of rounding errors.

particular level, resulting from the proportions of masters and nonmasters on the parallel test forms. The specific amounts of uncertainty resulting from these two sources are reported in Columns 1 and 2. At a particular level therefore, the figure in Column 3 would correspond to the denominator of the coefficient iota ratio.

Column 6 reports the total amount of information that is obtained from the consistent masters and the consistent nonmasters on the parallel forms. Columns 4 and 5 separate this total information into the two sources. Thus, Column 6 reports the value of the numerator of the coefficient iota ratio at a particular level.

On inspection of Table 4.8, it can be seen that at all three item-length levels of the 50% mastery criterion, the amounts of information obtained respectively from the consistent masters and the consistent nonmasters "cover" to approximately the same extent the uncertainty present from the corresponding sources. This stands to reason, since it was seen from Table 4.7 that the proportions of consistent masters, consistent nonmasters, and masters and nonmasters on the individual parallel forms, were quite similar to one another at this criterion level. Therefore, there is little difference in the extent to which the two sources of information "cover" the respective uncertainty associated with each.

As the mastery criterion level increases however, there begins to be a discrepancy in the extent to which the two sources of information "cover" the corresponding uncertainty. In particular, the bits of information obtained from the proportions of consistent nonmasters to a better extent cover the bits of uncertainty present, on the basis of the proportions of nonmasters on the two test forms. On the other hand, the bits of information obtained from the consistent masters start to do a poorer job of "covering" the uncertainty associated with this source. The combined result is that at the 30 item-length level of the 80% mastery criterion, despite the fact that the measures of information and uncertainty associated with the nonmasters are practically equal, the measures of information and uncertainty associated with the masters differ to such a degree that the resulting value of the coefficient iota ratio is at its least in regard to the levels measured. As mentioned previously, this would lead to the conclusion that the mastery/nonmastery classifications are the least reliable at this level.

The question which initiated this discussion was, as the mastery criterion level increases, why do iota estimates decrease while P_c and its estimates increase? As has been seen, this would lead to quite opposite decisions regarding the reliability of the mastery/nonmastery classifications which result from the scores on the parallel test forms. It is now clear why this is the case.

The value of P_c and its estimates increase as the mastery criterion level increases simply because all misclassifications are weighted equally, and as the mastery criterion approaches 100% there are generally an increasing proportion of consistent nonmasters, and fewer misclassifications. This would be equivalent to the calculation of the standard deviation with all scores being weighted equally in terms of their deviation distances from the mean of the distribution. This is of course not the case. The standard deviation, as a measure of variance, does weight the scores in the distribution differently in regard to their relative distance from the mean, and is most sensitive to scores at the extreme ends of the distribution.

It is clear on the basis of the above analysis, that while coefficient iota weights misclassifications differently, P_c and its estimates do not. And as also seen, this difference in approach can lead to quite varying conclusions. In that the approach taken by coefficient iota is analogous to that used in traditional test theory, it can be concluded that iota estimates adequately fulfill the need for a reliability coefficient for CRM tests. Indeed, it has been demonstrated that P_c and its estimates may lead to inaccurate decisions at particular mastery criterion levels.

Summary

The data base of the analysis consisted of the responses of 2182 students on a mathematics mastery evaluation instrument. From these 120 items, two parallel forms were created at each of 30, 20, and 10 item-length levels. The items making up the parallel forms were paired-off on the basis of an item analysis which focused on similarity of content, item difficulty, and item discriminating power. Descriptive information obtained for each pair of parallel forms demonstrated that they were quite similar in terms of mean, standard deviation, and KR-20 reliability.

The analysis began with a calculation of the parameter values of iota at each of the three item-length levels for each of four mastery criterion levels - 50%, 60%, 70%, and 80%. These values were reported in Table 4.5, and it was noted that the parameter values of iota varied only to a slight degree over the first three mastery criterion levels. At the 80% criterion level however, it was observed that the values began to decrease.

The next step in the analysis was to draw 50 random samples of 30 student test scores at each of the 12 item-length by mastery criterion levels. Coefficient iota was computed for each of the selected samples, with the mean and standard deviation of each of the sets of random samples also being reported in Table 4.5. It was observed that the 12 sample means appeared to be unbiased estimates of the respective parameter values, with the largest absolute difference being .02.

To aid in the analysis of results, the findings of the study by Subkoviak (1978) were cited in Table 4.6. This study, as recalled, involved the comparison of the index P_c and three estimates of P_c, as coefficients of the reliability of the mastery classifications obtained from CRM tests. P_c was noted to be the proportion of students in a group who were consistently classified as either masters or nonmasters in a test-retest situation. It was observed from the values in Table 4.6, that the values of P_c and its estimates increase markedly as either the mastery criterion or item-length level increase. This was seen to be expected since the proportion of consistent nonmasters will increase as the test becomes increasingly more difficult to master.

Prior to a comparison of P_c and its estimates and iota, the conceptual differences between iota and P_c and estimates were analyzed. One basis of difference was seen to arise from the relation of each to the concept of variance within traditional test theory. It was recalled to be a basic tenet of information theory that the more improbable an event, or the more deviant it is from the norm, the greater the uncertainty associated with it. Similarly, in the case of the standard deviation as a measure of variance, the more extreme the score, or the more deviant it is from the norm, the greater that score's contribution to the final value of the index. P_c and its estimates, on the other hand, give equal weight to each mastery or nonmastery classification. On the basis of this difference, it would seem that the theory upon which coefficient iota is based is more consistent with traditional test theory.

Table 4.7 reported the parameter values of P_c for the present data at each of the mastery criterion by item-length levels. Comparison of these values revealed that the values of P_c increased as the mastery criterion increased, as was the case in Subkoviak's study, while the values of iota began to decrease slightly. This was the case despite the fact that the respective proportions involved in the numerator and denominator of the coefficient iota ratio became increasingly similar in value. It was further noted that separate analyses of these two trends would result in quite different conclusions concerning the reliability of the same mastery/nonmastery classifications.

Further analysis of this difference in conclusions was conducted on the basis of the individual proportions after being converted into amounts of uncertainty and information. These amounts, measured in bits, in addition to their sources were reported in Table 4.8. This Table displayed the two amounts of information which combined to equal the value of the numerator of coefficient iota at each level, and the two amounts of uncertainty the sum of which yielded the denominator of the ratio. On the basis of this data, the influence of involving the additional factor of the logarithm to the base 2 into the calculation became evident.

In the case of the nonmasters, the trend of increasing similarity between the values of x'_O and x'_A and x'_B, as reported in Table 4.7, was repeated in terms of the resulting amounts of uncertainty and information in Table 4.8. As the mastery criterion increases from 50% to 80%, the amounts of information resulting from the consistent nonmasters do an increasingly better job of "covering" the uncertainty present from the proportions of nonmasters on each of the respective parallel forms. If the value of coefficient iota was based solely on the ratio of these two values, iota would indeed approach unity as the mastery criterion increased. However, the uncertainty and information resulting from the respective proportions of masters on the two parallel forms, and the proportions of consistent masters must also be taken into account.

It was seen in Table 4.8 that unlike the information resulting from the consistent nonmasters, the information obtained from the consistent masters began to do an increasingly poorer job of "covering" the uncertainty present as the mastery criterion increased. The reason that this is the case is basic to information theory, and is analogous to what is involved in the calculation of the standard deviation. As the mastery criterion increases, the test becomes increasingly more difficult and the occurrence of a consistent master becomes more improbable, and its occurrence results in an increasing amount of information. In comparison, a consistent nonmaster is much more likely and results in relatively little information. This relationship was seen as the same as that involved in the calculation of the standard deviation, in that extreme scores are more improbable than those relatively closer to the mean, and contribute a greater share of information to the calculated value.

Such is not the case with P_c and its estimates however, which give all misclassifications equal weight. These measures do not take into account all the available information, and in addition, were seen to lead to a decision regarding the reliability of mastery/nonmastery classifications quite contrary to that reached on the basis of the coefficient iota estimates. It was concluded therefore that, in that coefficient iota estimates are calculated in a manner consistent with the approach taken to the estimation of reliability within traditional test theory, such estimates meet the need for an index of the reliability of the mastery/nonmastery classifications resulting from scores obtained from CRM tests.

CHAPTER V

SUMMARY AND CONCLUSIONS

This final chapter is divided into two major sections. The first provides a brief summary of the first four chapters. The second section is devoted to the final conclusions of this study.

SUMMARY

Purpose of Study

At the beginning of Chapter I, it was stated that most decisions within the field of Education are based on the evaluation of the results of tests, which are administered with the purpose of determining whether some learning experience of interest has had any effect on a particular ability level of the students involved. Quite obviously, a great deal of time, money, and effort goes into this testing process. As a result, in order to justify this process there must be some evidence that the information obtained from such testing is accurate.

The issue of estimating the accuracy of test results within traditional test theory is the concern of the topic of reliability. It was additionally noted that although a number of methods of estimating reliability existed, each attempted to estimate reliability of test results on the basis of the value of a numerical coefficient. The purpose of the present study was then identified to be an attempt to develop a coefficient of reliability for criterion-referenced mastery (CRM) tests.

Before being able to explore the need for the development of such a reliability coefficient, it was seen as necessary to examine further the concept of test reliability, and also the notion of a CRM test. It was decided upon to deal first with the issue of reliability.

Classical Theory of Test Reliability

Test reliability was conceptually defined, as stated by Ebel (1968), as the proportion of observed score variance which can be accounted for by true score variance. That is, student scores on a particular test can vary from one another for a variety of reasons, only one of which is the actual ability levels of the individual students on the theoretical construct being measured. Unfortunately, a number of extraneous variables also exert influence over the obtained values of test scores.

Observed score variance is therefore seen to be made up of two main parts, true score variance and error variance. And the less error present in the obtained scores, the more accurate the scores and the closer the resulting proportion of true score variance to observed score variance would be to unity. However, due to the fact that some error always exists in measurements of this type, and because this error is due to chance elements, in reality all measures of reliability must be considered as estimates.

In order to adequately address why a need existed for the development of a reliability coefficient to be used specifically with CRM tests, it was necessary to summarize the manner in which an operational definition of test reliability had been developed within traditional test theory.

An Operational Definition of Classical Test Reliability

After a detailed analysis of the nature of the theoretical definition of test reliability as the ratio of true score variance to observed score variance, the need for an operational definition became evident. Since it will never be the case that true score values will be known, the variance of these scores can not be calculated. Therefore, an expression was required that could be used operationally applied to distributions of test scores.

Ebel (1972) was cited as providing a definition of the traditional operational definition of test reliability as the correlation coefficient derived from two sets of scores, obtained independently, on equivalent test forms given on each occasion to the same group of examinees. It was noted that this can be accomplished in any one of three different ways: 1.) having the examinees retake the same instrument; 2.) administer equivalent or parallel forms of the same test; or, 3.) sub-divide the items of a particular test into two or more equivalent forms. It was further noted that the operational definition of reliability approached the issue conceptually from the standpoint of consistency, the notion being in this case that the less the variability from one equivalent measure to the next, the more reliable or accurate is the measuring instrument.

An implication of this operational definition of test reliability that was of immediate interest was next discussed. The issue related to the reliability coefficient as an index of correlation. One of the factors which influences the relative size or magnitude of a correlation coefficient is the range of talent, or put simply, the score distance from the lowest score in the distribution to the highest score. All other factors being equal, the larger the range of talent, the larger the value of the correlation coefficient. The result of this relationship is that the relative size of any correlation coefficient is dependent upon score variance. Although another reason would be cited later, this result provided the operational reason as to why traditional reliability coefficients should not be used in the case of CRM tests.

Issue of CRM Tests in Relation to Classical Test Theory

In order to fully understand the above difficulty, the basic differences between CRM and NR tests were examined. It was stated that during the period of time when traditional methods of test analysis were being developed, the major mode of testing was NR. As a result, the methods of analysis which were developed focused upon the characteristics and objectives of the NR approach to testing.

NR tests have the basic objective of yielding a distribution of scores which would approach the familiar bell-shaped normal curve. Relatively small percentages of scores would be located at the extreme ends of the curve, while the bulk of the scores would be located near the center of the distribution. Those items which are of optimum use to this type of test have the characteristic of maximizing the variability between individual responses. Items which are either too easy or too difficult for the tested group are undesirable. It was readily seen therefore, that the concept of a reliability coefficient as an index of correlation is quite compatible for use with NR tests, which have the basic objective of maximizing score variability.

Possibly as an offshoot of the general move toward accountability, it was stated that the emphasis in educational testing has recently shifted from NR tests toward what have been termed CRM tests. The basic conceptual difference between these two types of measures is that, while NR tests judge individual performance in relation to the performance of the group as a whole, CRM tests judge individual performance in relation to a specified set of standards or objectives. In evaluating the individual results of a CRM measure, it is of no interest how others in the group did. All that is of issue is the degree to which the individual answered correctly the items on the CRM test which were designed to measure performance in relation to a specified set of objectives. A judgment as to the degree to which the individual had "mastered" the measured objectives could then be made on the basis of the number of items answered correctly.

Hence, as Millman and Popham (1974) have stated, score variability is an unnecessary characteristic of CRM tests. In fact, if a particular group of students have done an exceptional job of mastering a specified criterion, there may be little, if any, score variance. And, if there is little score variability, a measure of test reliability derived from a correlation coefficient would be quite small, perhaps even zero, despite the fact that the CRM test might be doing an extremely accurate job of measuring the objectives of interest. Thus, it can be seen that not only is it possible for a traditional reliability coefficient to yield inaccurate information concerning the consistency of the results of CRM tests, the approach itself is conceptually inconsistent with the purpose of a CRM measure.

Estimation of the Reliability of CRM Classifications

The next step in this discussion was to develop an approach to reliability which would be both operationally and conceptually consistent with the purpose of CRM tests. A number of sources were cited which identified the purpose of CRM measures as the determination of the degree to which the examinees had mastered the objectives of interest. This determination is made on the basis of whether the individual examinees scored above or below a pre-chosen cut-off or criterion score. Those examinees scoring at or above the criterion level are labeled "masters" while those scoring below the cut-off point are classified as "nonmasters". The values of the raw scores of the individuals are of interest only in leading to mastery/nonmastery classifications. And, an individual mastery/nonmastery classification is in no way affected by the number of other examinees who were classified as either masters or nonmasters.

Of course it would be desirable to know the extent of the accuracy of such mastery/nonmastery classifications, but to approach this issue from the point of view of the variability of raw scores avoids addressing the nature of CRM test results. As Thorndike (1951) has stated, methods of reliability must first address what is to be accomplished by the type of measure of interest. Traditional methods of estimating test reliability do not satisfy this condition in the case of CRM measures. Consistency is still a viable concern of what is to be accomplished by such measures, but this issue must be approached within the framework of the mastery/nonmastery classifications and not from the standpoint of the variability of raw scores.

Conceptual Similarity between Reliability and Information

Basic to the approach taken within this paper is the notion that mastery/nonmastery classifications, and the raw scores from which they are derived are a form of information. Such information, for example, can lead to a decision as to whether a particular examinee has mastered a specific level of material and should then proceed to the next higher level. It was thus assumed that an index of the consistency of information obtained from two independent CRM measures, based upon the mastery/nonmastery classifications, would be synonymous with the traditional notion of test reliability. It was then noted that statistical expressions of information have been developed within the field of information theory, and that the approach to the development of a reliability coefficient for CRM measures would be taken from this perspective.

The common sense notion of information is that this is something obtained from a message source which relieves to some extent the uncertainty that was previously associated with some matter of interest. The theoretical basis of information theory was seen to be much the same. The difference is that information theory provides mathematical expressions for information and uncertainty which allow for the quantification of the extent to which information relieves the uncertainty present in a particular situation.

From this standpoint, two similarities between the concepts of information and reliability were identified.

First of all, information can be either useful or useless to one's particular needs or intent. In the same way, test results can have a high degree of reliability while having a relatively low degree of validity. Therefore, just as the issue of test reliability can be considered apart from the issue of test validity, the extent to which the information obtained relieves the uncertainty present, can be considered apart from the usefulness of that information.

The second similarity, and the most important one, is the relationship between information and uncertainty as compared to that between true score variance and observed score variance. The amount of true score variance contained in a set of test scores cannot exceed the amount of obtained score variance, and as had been previously seen the conceptual definition of reliability was the ratio of true score variance to observed score variance. In the same context, it would not be possible to obtain more information from a message or set of messages, than the amount of uncertainty present. And, just as the reliability of a set of test scores is traditionally considered to be the extent to which true score variance "covers" the amount of obtained score variance, test reliability could also be considered as the extent to which the information obtained "covers" the uncertainty present. The concept of information is seen to be analogous to that of true score variance, just as uncertainty is analogous to the concept of obtained score variance. The task that remained was the development of a reliability coefficient which estimates the degree of consistency of obtained information, and which is also logically consistent with the decision-making process involved in CRM testing.

Current Estimates of the Reliability of CRM Classifications

In order to put the present study within a frame of reference of what has already been suggested in terms of reliability coefficients for CRM tests, the first section of Chapter II was devoted to a review of the major indices which have appeared. It had been earlier stated that Stanley (1971) has mentioned that there are two aspects to the issue of reliability, one is logical and the other is statistical. It had already been noted that traditional reliability coefficients are logically inconsistent with the purposes of CRM measures. And before reviewing the above-mentioned suggested coefficients, attention was devoted to a statistical inconsistency of the traditional approaches to reliability which has been identified by Hambleton and Novick (1973).

These authors state that one of the basic differences between NR and CRM measures is that the former rank individuals according to a "fixed quota", while the latter are "quota free" in terms of selection and classification. For example, if a normal distribution of scores from a NR measure is assumed, and one standard deviation above the mean is decided upon as the cut-off point at which those students scoring above will receive a grade of "A", there will never be more than 15.87% of the students receiving such a grade. On the other hand, there is no such restriction on the percentage of students in a tested group who can be classified as "masters" on the basis of a CRM measure. Now, it was also mentioned previously, that all values obtained from reliability coefficients are estimates. However, such values when correctly reported also include some type of estimate of error contained in the estimates. For traditional reliability coefficients, this error estimate is the standard error of measurement, which can be used to construct a confidence interval around an individual's obtained score. It is the application of the standard error of estimate to CRM measures which the authors believe is the most serious objection to the application of traditional estimates of reliability on the results of such measures.

This objection was seen to stem from the fact that the use of the standard error of estimate with the results of CRM measures would result in an incorrect choice of "loss function". In short, error estimates obtained from the standard error of measurement are in the metric of score units. However, although raw scores serve to yield mastery/nonmastery classifications, an error estimate in terms of an interval of raw score units would not readily yield the information as to whether a mastery/nonmastery misclassification had resulted. Hambleton and Novick term such misclassifications as "threshold loss", and state that any reliability coefficient to be used with CRM measures must yield an error estimate which reflects this loss in information due to misclassification. Keeping this and the previous objections regarding traditional reliability estimates in mind, a review was provided of the major estimates of CRM reliability which have been suggested.

The first coefficient presented was an index suggested by Carver (1970) which was simply the proportion of consistently classified masters and nonmasters obtained for the same group of examinees, on the basis of scores obtained from equivalent forms of a CRM measure. It was noted that this index has been described in the literature as being "crude" (Crehan, 1974), and should be used only for quick "thumb-nail" estimates of consistency. An estimate based on the procedures of classical test theory which has been proposed by Livingston (1972), was next reviewed. Livingston's index is based on score deviations not from the mean of the distribution, but rather from the value of the cut-off score. It was noted, that like traditional estimates of reliability, such a measure would also be rendered useless in the case of restricted score variability.

An estimate of reliability that has been given considerably more attention in the literature has been the index kappa (K) developed by Cohen (1968, 1972). Cohen's K has the advantage over Carver's proportion of consistent mastery/nonmastery classifications of incorporating into the analysis the proportion of consistent classifications which can be expected to occur by chance. Swaminathan, Hambleton, and Algina (1974) who have suggested kappa's use in estimating the reliability of CRM mastery/nonmastery classifications, were seen to note that reported values of kappa can be substantially influenced by test-length and the particular value of the cut-off score chosen. The influence of these factors were noted through the findings both of the present report and Subkoviak's 1978 study.

It was also noted at this point that although kappa estimates mastery/nonmastery classification consistency within the dimension of "threshold loss", it would be desirable to have such an estimate that was relatively insensitive to changes in test-length and criterion score location.

The next index reviewed is an estimate of kappa derived from a single CRM testing which has been developed by Huynh (1976). Huynh's index begins with a KR_{21} value obtained from the test results, and uses this value to estimate the parameters of a beta-binomial distribution. This distribution provides the mathematical model from which the estimate of kappa is derived. It was noted that due to the nature of the calculations, whenever test-length approached 10 or more items, a computer would almost certainly be required for convenience.

The final index reviewed is that submitted by Subkoviak (1976). This author's "coefficient of agreement" was seen to be based on a sum over the population of examinees, of the individual probabilities that each individual i had scored at or above some chosen cut-off score. In similarity to Huynh's estimate, Subkoviak's coefficient of agreement was likewise "situation specific" in that reported values would depend upon the factors of test-length and cut-off score.

A second study by Subkoviak (1978) was then cited which compared the Swaminathan et al., Huynh, and Subkoviak procedures for estimating CRM test reliability. As reported, this study compared these three procedures on the basis of their estimation of P_c at each of three different item-length levels, for each of four mastery criteria. It was found that each procedure produced estimates of P_c which were reasonably and consistently close to the population parameter. As a recommendation, Subkoviak noted that the Huynh procedure required only one testing, was mathematically sound, and produces "reasonably accurate estimates".

In a summary of this section, it was stated that while the indices compared by Subkoviak meet the criteria of being within the appropriate dimension of "threshold loss", they each have two major disadvantages. First of all, the techniques are highly "situation specific". This is illustrated in the Subkoviak (1978) study which reveals a quite marked change in the values of P_c and its estimates as either the test length or particularly the criterion level changes. Secondly, for each technique errors in classification are treated equally. This would become a major point in the discussion of the results of the analysis involving coefficient iota.

Basic Concepts of Information Theory

Following the above review, the discussion shifted to an introduction of the basic concepts of information theory, upon which the present methodology was based. Information theory was seen to be a statistical approach to the quantification of the amount of information obtained from some form of communicative act. The primary concern of information theory is to quantify the amount of information transmitted from a sender to a receiver. And, just as validity is an issue of itself apart from reliability, the usefulness of the information obtained is also an issue apart from the quantification of the information. In the present context, test scores would be viewed as messages from testees to an examiner regarding the level of achievement of a particular subject matter.

Conceptually stated, when information is received from a particular source, the uncertainty contained in the situation of concern is to some degree relieved. Indeed, information is not possible if some degree of uncertainty does not exist a priori in regard to the outcome of the sent message. The amount of uncertainty present in a particular context was then seen to depend in part on the number of outcomes that were possible. However, as was illustrated in the presentation of the development of an operational definition of information, what is perhaps most important is the probability of occurrence associated with the individual possible outcomes.

Statistical Definition of Information

In that information is seen as something derived from a message transmitted from a sender to a receiver, it was seen as not surprising that early work in the quantification of information was conducted within the field of electrical engineering. Much of this early work was conducted in the 1920's. However, a detailed statistical model was not formulated until Shannon's and Weaver's 1948 publication.

It had been previously hypothesized that any measure of information or uncertainty must be logarithmic in nature. A practical illustration of the basis for this assertion was provided through the game of "Twenty Questions". In this instance all alternatives are considered to possess equal probability of occurrence, with the questioner selectively reducing the number of alternatives through a series of inquiries which can be answered either "yes" or "no", until the correct choice remains. In order to correctly proceed, each asked question must reduce the remaining alternatives by half, until only two remain. The number of questions required to complete this process corresponded to the amount of uncertainty contained in the original question, as measured in units termed "bits". It was then demonstrated that a general measure of uncertainty could be expressed as the logarithm to the base 2 of the number of possible alternatives. The value thus obtained would be the amount of uncertainty measured in bits, and the solution to the particular question would contain exactly that number of bits of information.

Consideration was next directed to the situation wherein the possible alternatives do not have equal probabilities of occurrence. The basic concept here, was that the least likely that a particular alternative was to occur, the greater the amount of information that would be conveyed if it did occur. It was seen here, that the concept of information does not apply to the individual messages themselves, but rather to the situation as a whole. Further evidence for the logarithmic nature of a function of information was provided, and this combined with the notion of probability of occurrence of the alternatives and the inverse relationship between probability of occurrence and obtained information to lead to an operational definition of information.

Before concluding Chapter II, the concept of uncertainty as expected information was introduced. This expected value was expressed as a sum of the information that would be provided by each of the possible alternatives, with each alternative being weighted by being multiplied by its respective probability of occurrence. This expected information can also be expressed as the amount of uncertainty contained in a particular message set.

Conceptual Basis of Methodology

It was noted at the beginning of Chapter III, that the statistical framework of information theory will be used in a conceptual approach, involving the estimation of the degree to which a tested hypothesis has been confirmed, on the basis of the informational strength of the obtained evidence. The hypothesis in the situation at hand, is of course, that two tests constructed to be equivalent in form, will yield consistent decisions regarding the classification of examinees as either masters or nonmasters. The evidence that would be used to estimate the degree to which this hypothesis has been confirmed would be a sample observation of the extent of the consistent mastery/nonmastery decisions yielded by the two instruments. The conceptual basis of the suggested methodology then, was that the concept of information and uncertainty as expressed in the relationship between evidence and hypothesis, form an analogy to the ratio of true score variance to observed score variance.

Relationship between Evidence and Hypothesis

The next step in the methodology was to express the conceptual relationship between the notions of evidence and hypothesis, in the form of a statistical expression. And, in order for this statistical expression to be compatible with the concepts of information theory, and also with the traditional concepts of hypothesis testing, it was seen as necessary that this statistical expression be in the form of a probability measure. A study by Hilpinen (1970) was cited as the model for this statistical expression.

Based on the definition that, "probability is a logical relation between two sentences", Hilpinen first posits that the hypothesis under study can be expressed as sentence "H", and the evidence upon which the credibility of "H" is decided is defined as sentence "E". Using these definitions, Hilpinen predicates a probability statement designed to express the degree of credibility of "H" on the basis of "E" as, "P(H/E) = R". In this relationship, "R" is an estimate of probability, and represents the "justified degree of belief" in "H", on the basis of "E".

This relationship between hypothesis and evidence was seen to be analogous to the expression of a reliability coefficient as an estimate of the ratio of true score variance to observed score variance. Such an estimate can also be interpreted as "justified degree of belief".

In the instance under present consideration, the evidence consists of consistent mastery/nonmastery decisions made on the basis of test scores, as interpreted in light of some cut-off score criterion, and the same types of decisions on the same group of examinees derived from scores obtained from a second administration of the same test or a test designed to be equivalent to the first. Such evidence is then used to test the credibility of the hypothesis that the two tests yield consistent mastery/nonmastery decisions. And, just as a reliability coefficient estimates the ratio between true score variance and observed score variance, it would be advantageous to have an index which reflects the degree to which the above type of evidence "justifies" or "confirms" the hypothesis that the two instruments yield consistent mastery/nonmastery decisions.

There is uncertainty involved in the statement of any hypothesis, and the evidence gathered to test a hypothesis contains some amount of information concerning the reliability of that hypothesis. It was next seen as necessary to develop a statistical formula which would express the relationship between hypothesis and evidence, in terms of uncertainty and information.

Reliability as Hypothesis Confirmation

In beginning the development of this statistical formula, reference was made to two articles by Tornebohm (1966, 1968). Tornebohm's technique of estimating the degree to which a hypothesis is confirmed on the basis of obtained evidence, expressed in the above two articles as "degree of covering", was identified as the basis upon which a reliability coefficient for CRM tests would be developed.

Tornebohm's model was seen as beginning with the assumption that there exists a state space of objects, termed R, and that there is a desire to find the location of these various objects as they occur in R. As applied to the current study, this is the true state space of a group of individuals who have been exposed to some educational activity, and either have or have not, on the basis of a pre-chosen criteria, mastered the content of that activity. In order to estimate the true location of these individuals, that is, as being either in the state space of masters or the state space of nonmasters, a measurement instrument Z is used. The administration of Z to each individual in R thus results in a vector representing that particular measurement. The state space of all such vectors formed by the administration of Z to the individuals in R creates a second state space designated as M. The instrument Z thus produces a functional relationship between R and M, which creates M as an image of the state space R. However, the degree to which M will be an accurate image of R will of course to a great extent depend upon the accuracy of the instrument Z. A hypothesis regarding the reliability of mastery/nonmastery classifications obtained from a measurement instrument, could therefore be expressed as the extent to which the assigned mastery/nonmastery regions of the examinees as determined by their test scores, reflects their true mastery/nonmastery states.

The Concept of "Degree of Covering"

In developing a statistical index of the degree to which a hypothesis is confirmed by obtained evidence, reference was again made to the work of Tornebohm. The index developed by this author incorporates the concepts of information theory, and yields a value referred to as an estimate of "degree of covering". The index is quite similar, as first stated, to an expression of conditional probability. This index was later simplified to a ratio of the information received in evidence E, to the expected information, or uncertainty, contained in hypothesis H. If the information provided by the obtained evidence exactly covers the uncertainty or expected information existent in the hypothesis, the value of the ratio will equal a maximum of 1. On the other hand, if the information received from the obtained evidence to no extent covers the uncertainty contained in the hypothesis, then it is seen that the numerator of the ratio cancels to 0, resulting in the minimum of the range of values of the index. Therefore, the index has the closed interval of 0 to 1 as a range of possible values.

The next step was to use Tornebohm's index of hypothesis confirmation as a model in the development of a reliability coefficient for CRM measures.

Development of Coefficient Iota (i) Ratio

At this point in Chapter III the frame of reference involved in CRM testing was recalled. It was also noted that in regard to Ebel's operational definition of reliability, the only difference between CRM testing and NR testing is that in the latter case results are in the form of score values while in the former, the results of concern are classification decisions. Reliability coefficients applied to either case would need to take into account these characteristics if such estimates were to be consistent with Novick's concept of "threshold loss".

The first step in the development of the desired coefficient was to define a number of necessary terms. Assuming a group of students had been exposed to some educational experience, and then tested and retested with CRM instruments A and B, which are designed to be parallel, the following six proportions would be needed: 1.) the proportion of students classified as masters on Form A (x_A); 2.) the proportion of students classified as nonmasters on Form A (x'_A); 3.) the proportion of students classified as masters on Form B (x_B); 4.) the proportion of students classified as nonmasters on Form B (x'_B); 5.) the proportion of the entire group of students who are consistently classified as masters on both Forms A and B (x_o); and, 6.) the proportion of the entire group of students who are consistently classified as nonmasters on both Forms A and B (x'_o). Symbols were also defined for an individual examinee classified consistently as a master on both Forms A and B (c_i), and for an individual classified consistently as a nonmaster on both Forms A and B (c_j).

The goal of this section then, was the development of an index which would estimate the degree to which the evidence obtained from the mastery/nonmastery classifications made on the basis of the score results of Forms A and B, support the hypothesis that the two CRM measures yield consistent mastery/nonmastery decisions. Tornebohm's index of "degree of covering" thus was seen to be an appropriate model to apply to this situation. It was also seen as important to demonstrate the conceptual compatibility between the relationship of evidence to hypothesis, and, that of true score variance to observed score variance as reflected in traditional test theory.

Beginning with the application of the statistical definition of information to the expressions for an individual consistent master and an individual consistent nonmaster, amounts of information obtained independently from Forms A and B were defined as the summation across all such consistent classifications for each type of classification. The numerator of Tornebohm's "degree of covering" model thus, was in this case simply the addition of these two independent amounts of information. This then was a statistical expression of the amount of information received from the evidence provided by the CRM test classifications on the two parallel forms, and would therefore represent the numerator of the desired reliability coefficient.

The next step was to develop a statistical definition for the expected information or uncertainty contained in the hypothesis. It was necessary that this definition express the total amount of expected or potential information contained in the test-retest situation. Remaining consistent with the concepts of information theory, this was done by expressing, for each of the parallel forms, the contained expected information as a sum of the uncertainty resulting from the proportions of masters and nonmasters. As stated, this amount of expected information or uncertainty was expressed for each of the parallel forms. And again, because the two sources of expected information are assumed to be independent, these two quantities can be added to obtain an expression for total amount of expected information in the test-retest situation.

Finally, after the cancellation of a like term and further simplification, the development of the desired coefficient was completed upon the designation of the above referred to statistical expression as the denominator of the index. This index, as an estimate of the degree to which the evidence obtained from the administration of CRM measures which are designed to be parallel, relieves the uncertainty created by statement of a hypothesis that these parallel forms yield consistent mastery/nonmastery classifications, was designated as coefficient iota (i). Attempts were then made from both mathematical and philosophical perspectives to justify the form of coefficient iota.

The Range of Possible Values of Coefficient Iota (i)

It was noted at the beginning of this section that it would be conceptually advantageous if coefficient iota would be found to have a range of possible values consistent with that of traditional reliability coefficients. Additionally, these minimum and maximum values should be assumed under conditions similar to those which yield minimum and maximum values for such traditional coefficients. Analysis of these minimum and maximum values for iota was approached separately from two different perspectives: first, within the framework of information theory; and secondly, on the basis of iota as a mathematical expression.

From the perspective of information theory, the range of possible values of iota was considered on the basis of the coefficient being a measure of evidential strength. In this respect, the values which are entered into the coefficient iota ratio were considered solely on the basis of their being amounts of information and uncertainty, derived from particular sources. Based on the concept of evidential strength, it was reasoned that the range of possible values for iota should be at a minimum when, in a particular test-retest situation, there are neither any consistent masters nor any consistent nonmasters. In such a case, no information was transmitted by the evidence in regard to the hypothesis being tested. On the other hand, the range of possible iota values should be at a maximum when all the examinees are classified consistently as either masters or nonmasters. In this second case, the information provided by the evidence would totally cover the uncertainty contained in the hypothesis.

In analyzing iota's range of possible values from this perspective, the work of Tornebohm and the work of Hilpinen were again cited. On the basis of an examination of the quantities of information which are represented in an index of degree of covering, it was determined that iota did indeed assume a value of 0 at its minimum, and a value of 1 at its maximum. It was therefore concluded that coefficient iota, as a measure of evidential strength, has a maximum of 1 in the case in which the evidence E logically confirms the hypothesis H, and a minimum of 0 in the case in which the evidence E is logically independent with the expected information contained in hypothesis H.

The analysis of the ratio of possible values from a mathematical perspective, considered the actual manner in which the various quantities involved in the coefficient iota ratio are calculated. It was possible to determine by inspection that under the conditions which would be the case when iota assumes the minimum in its range of values, that the ratio would reduce to $0/-\infty$, which would of course further reduce to 0. Therefore, it was relatively easy to determine from this second perspective that the minimum of iota's range of values was 0, as desired. The examination of the maximum value of iota as a mathematical expression, was not as straightforward.

Thus, it was noted that there are generally two conditions under which iota may assume a maximum. In the first case it is necessary that on the basis of Form A, there are some examinees classified as masters and some as nonmasters, and that all of these examinees are classified in the same relative manner on the basis of scores obtained from Form B. In such a case, there is perfect consistency in classification on the basis of the two test forms. Again on the basis of inspection, it was relatively simple to determine that under such circumstances the iota ratio would reduce to 1/1. Thus, under such circumstances, the maximum value of iota would indeed be 1. However, there exists a second set of conditions under which iota assumes a maximum, which does not readily submit to conclusion from inspection.

This last condition arises when the examinees are consistently classified as either all masters or all non-masters. The factor which makes interpretation of the coefficient's maximum value under these conditions difficult, is that the ratio reduces to 0/0, a form which is considered to be indeterminate. However, upon application of the methods of calculus, it was found that the coefficient iota ratio approaches a limit of 1 under these conditions.

Therefore, it was demonstrated that coefficient iota assumes a range of values that is consistent with both an index of evidential strength, and a traditional reliability coefficient.

Method of Analysis

The data base upon which the sample analysis using coefficient iota was conducted, consisted of the responses of 2182 eighth and ninth grade students on a mathematics mastery instrument. Out of the instrument, which consists of 120 items evaluating 40 objectives, parallel test forms were created at each of 30, 20, and 10 item-length levels.

The items which make up these parallel forms were paired off on the basis of content, difficulty, and discriminating power.

At each of these item-length levels then, mastery criterion levels of 50%, 60%, 70%, and 80% were considered. Thus, 12 item-length by mastery criterion levels were created. The first step in the actual analysis consisted of computing the population values of iota at these various 12 levels. The next step in the analysis was then to select from the population 50 random samples of 30 students each, at each of the 12 item-length by mastery criterion levels, and to compute iota for each of the drawn random samples. The means and standard deviations of the 50 iota values computed at each of the 12 levels, as well as the parameter values of iota for each of the levels were then reported.

Discussion of Results

Results of the present analysis were compared to those reported by Subkoviak (1978) in his study of four types of suggested reliability coefficients for CRM measures. One of these coefficients, that developed by Swaminathan et al., (1975), was the value of the proportion of examinees who were consistently classified as either masters or nonmasters (P_c) as calculated from a sample. This value is then considered as an estimate of the population value of P_c. The remaining three coefficients, those developed by Subkoviak (1976), Marshall and Haertel (1976), and Huynh (1976), are all estimates of P_c based on a single testing.

The major difference of note at this point between coefficient iota and the measures reported on in the Subkoviak study is, that while iota estimates remained relatively stable across changes in both item-length and mastery criterion, the latter measures varied quite markedly. The measures in the Subkoviak study were explained to vary in the manner in which they do, precisely because they are estimates of P_c. Therefore, it is quite logical to assume, that as the criterion level changes to either extreme, P_c will necessarily begin to approach unity. For the tests are either becoming too difficult or too easy for the examinees, and most will be either consistently classified as nonmasters or consistently classified as masters. Iota however, although it involves proportions of consistent masters and nonmasters, takes more into consideration than P_c.

In expressing this difference, the analogy between variance and uncertainty was again focused upon. The standard deviation, as a measure of variance, does not take all scores in a distribution into equal account in formulating an estimate of variance. The relative amount of information contributed by an individual score depends upon its relative distance from the mean of the distribution. Compatible with this approach is the fact that within information theory, the relative amount of information provided by an event depends upon its relative probability of occurrence. The further toward the extremes of a distribution, the greater the amount of variance an individual score contributes to the total variance of the distribution. Similarly, the more improbable the likelihood of an event's occurrence, the greater its contribution to the total uncertainty contained in the situation as a whole. The models are analogs of one another.

This aspect of the nature of iota was illustrated by analysis of the individual probabilities which are involved in the coefficient's formula. Initially, it was noted that the respective proportions of masters and nonmasters on the two parallel forms were approaching the proportions of consistent masters and nonmasters, as the criterion level increased. Thus, on this basis, it might be assumed that iota, like P_c and its estimates, should also approach unity. However, when these proportions were expressed in terms of bits of uncertainty and information in Table 4.8, the reason for the difference in the trend of iota values became evident.

As the proportions of masters and nonmasters on the two Forms approach certainty, as they do when the mastery criterion level increases, there is less and less uncertainty involved in these classifications. And, this situation is reflected in the above-mentioned Table. Nevertheless, there is seen to be a decline in the degree to which the information provided by the consistent mastery/nonmastery decisions covers the existent uncertainty. This was seen to be a result of the fact that, although the information received from the consistent nonmasters does an increasingly better job of covering the relative uncertainty associated with those types of classifications, the information received from the consistent masters is seen to do an increasingly poorer job of covering the amounts of uncertainty associated with this latter source. The overall result is that at the 80% criterion level, the value of coefficient iota indicates that the parallel test Forms yield less reliable mastery/nonmastery classifications than at the three lower criterion levels. This is a conclusion exactly opposite to what would have been concluded on the basis of P_c.

CONCLUSIONS

The purpose of this dissertation has been the development of an index of the degree of reliability of the mastery/nonmastery classifications yielded from the scores obtained from CRM measures. In that the equivalence of parallel forms is extremely important in a mastery instructional context, the analogy between reliability and consistency found in traditional test theory, was seen to be especially applicable to the problem at hand. However, as expressed by Stanley (1971), both logical and statistical aspects should be considered in evaluating issues of reliability. And, on the basis of these issues, it was demonstrated that traditional reliability coefficients run into difficulties in regard to both of the above when applied to CRM measures.

While a number of authors have recognized these difficulties, and various estimates of reliability have been developed for CRM measure classifications, there remains considerable discussion as to their relative merits. This researcher is of the position that the technique developed herein, and labeled coefficient iota, satisfactorily addresses the above issues, and therefore merits consideration and further investigation as a possible approach to be adopted in the estimation of the reliability of CRM classifications. Indeed, coefficient iota has been seen to avoid certain disadvantages, and perhaps even errors in interpretation, which are encountered when using indices of CRM reliability which are based on the proportion P_c. Discussion of these disadvantages will focus on three specific points.

First of all, in discussing the disadvantages of previously suggested CRM measure reliability coefficients, it was noted that the values obtained from these coefficients tend to fluctuate, sometimes markedly, as the mastery criterion level changes. In this way, such measures are considered to be "situation specific". That is, a CRM measure would not have a single reported degree of reliability associated with its results, as is the case with NR test scores. Rather, it is necessary to report a number of coefficient values, one for each criterion and item-length level. As was noted in Chapter IV, this was seen to be at least in part due to the fact that these CRM reliability estimates are based on the proportion of consistent mastery/nonmastery classifications in the sample (P_c).

It is not being suggested here that only a single value of a reliability coefficient should be reported for a CRM measure. This researcher agrees that a CRM measure cannot have a "single" degree of reliability, since this property of the measure will likely vary as the criterion level is changed. However, on the basis of the sample results of this study, it appears that coefficient iota values may vary to a less degree across changes in criterion and item-length levels than do the indices reported upon by Subkoviak (1978). As a direction for possible further investigation, coefficient iota should be applied to samples of mastery/nonmastery classifications which are based on scores which exhibit a more rapid fluctuation of P_c across changes in these levels. Indeed, it may prove of interest to also apply the coefficient iota technique to NR test scores to observe the manner in which these values compare to those yielded by classical reliability measures.

The second major disadvantage of the CRM reliability coefficients which were reviewed in Chapter II is that the mathematics involved would render them virtually unusable by most classroom teachers. In fact, even if one were familiar with the calculus involved, once tests consist of about 10 items or more in length, access to a computer is almost necessary. In comparison, about all that is required to make use of coefficient iota is the ability to calculate a proportion, and access to a table of log values as is reproduced here in Appendix B.

The third point to be discussed here is not simply a disadvantage of the coefficients reviewed in Chapter II, but rather, the seeming likelihood that the values which they yield can lead to errors in the interpretation of the reliability of the examined CRM measure classifications. It was noted in Chapter IV that because these coefficients are based on the proportion P_c, they will necessarily appear to become more reliable as the measures become either too difficult or too easy for the group being tested. All one would apparently need to do to obtain more reliable mastery/nonmastery classifications is to either increase or decrease the criterion cut-off score. This is not the case with coefficient iota.

Interpretation of the results in Table 4.7 would seem to indicate that the involved parallel forms yield quite reliable mastery/nonmastery classifications at each of the 50%, 60%, and 70% criterion levels. And, the degree of reliability is approximately the same at each of these levels. If one were attempting to decide which criterion level to use, the choice could be made solely on the basis of how difficult a measure was desired. Any of the three cut-off points could be chosen based on evidence that highly reliable classifications are likely for each.

However, as the criterion level rises above 70%, it is indicated that the classifications obtained become less reliable. As mentioned previously, this is the opposite conclusion that would be reached from the estimates obtained from coefficients based on the proportion P_c. And, the position is taken here that this property of the coefficients based on P_c runs contradictory to the classical concept of reliability.

The present researcher has attempted to stress, it is hoped not overly so, the analogy between the concepts of uncertainty and variance. It was noted in Chapter I that in the case of NR measures, reliability is dependent upon variability. To be specific, as variability increases, and other things remain the same, reliability will likewise tend to increase. From the standpoint of uncertainty as an analog of variance, this relationship is not maintained in the case of CRM reliability estimates based on P_c. As uncertainty decreases as the measures become either too easy or too difficult for the population being tested, the values obtained from these coefficients would lead to the conclusion that the measures become more reliable. But do they really?

One might reasonably counter this criticism by arguing that reliability is defined within classical test theory as the degree of consistency of a set of measures. And certainly, CRM measures that are either relatively too difficult or too easy will tend to yield consistent mastery/nonmastery classifications. Therefore, such classifications should be considered to have a relatively higher degree of reliability as compared to situations in which P_c is less.

This argument however, overlooks one of the basic aspects of the concept of reliability as expressed by Stanley (1971). He states that in considering reliability, "one must first determine what is to be accomplished and what purposes are to be served by a measure of reliability" (p. 359). The purpose of a CRM measure is to provide evidence, or information, in regard to the mastery of a particular set of instructional objectives. Establishing either a relatively high cut-off criterion, resulting in a situation in which most of the examinees are classified as nonmasters, or a relatively low cut-off criterion, resulting in a situation in which most of the examinees are classified as masters, would not seem to provide a substantial amount of information for the purpose at hand. And, this is the conclusion that would be arrived at on the basis of the trend in coefficient iota values across criterion levels.

It is concluded therefore, that the findings of this study indicate that coefficient iota not only avoids some of the disadvantages of CRM reliability estimates thus far suggested, but also to a greater extent addresses the empirical utility of the consistency of mastery/non-mastery classifications. It is clear from the present literature that considerable debate remains regarding both the appropriateness and utility of the types of CRM reliability estimates that have up to now appeared. The present author believes, that although further investigation is required, coefficient iota deserves consideration as a means of estimating the reliability of CRM measure classifications.

BIBLIOGRAPHY

Algina, J., & Noe, M.J. An investigation of Subkoviak's single-administration reliability estimate for criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Alkin, M. "Criterion-referenced measurement" and other such terms. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles: Center for the Study of Evaluation, U.C.L.A., 1974.

Attneave, F. Application of information theory to psychology: A summary of basic concepts, methods and results. New York: Holt, 1959.

Bloom, B.S. Taxonomy of educational objectives, handbook I; Cognitive domain. New York: David McKay, 1956.

Bloom, B.S. Learning for mastery. Evaluation Comment, 1968, 1, 26-32.

Bloom, B.S. Recent developments in mastery learning. Educational Psychologist, 1973, 10, 53-57.

Bormuth, J.R. On the theory of achievement test items. Chicago: The University of Chicago Press, 1970.

Brennan, R.L., & Kane, M.T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.

Carver, R.P. Special problems in measuring changes with psychometric devices. In Evaluation research: Strategies and methods. Pittsburgh: American Institutes for Research, 1970.

Cohen, J.A. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cohen, J.A. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, *70*, 213-220.

Cohen, J.A. Weighted Chi-Square: An extension of the kappa method. *Educational and Psychological Measurement*, 1972, *32*, 61-74.

Cox, D.R., & Hinkley, D.V. *Theoretical statistics*. London: Chapman and Hall, 1974.

Cox, R.C. *Evaluative aspects of criterion-referenced measures*. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

Crehan, K.D. Item analysis for teacher-made mastery tests. *Journal of Educational Measurement*, 1974, *11*, 255-262.

Ebel, R.L. Content standard test scores. *Educational and Psychological Measurement*, 1962, *22*, 15-25.

Ebel, R.L. The value of internal consistency in classroom Examinations. *Journal of Educational Measurement*, 1968, *5*, 71-73.

Ebel, R.L. *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, N.J.: Prentice Hall, Inc., 1972.

Ebel, R.L. Evaluation and educational objectives. *Journal of Educational Measurement*, 1973, *10*, 273-279.

Fobes, M.P., & Smyth, R.B. *Calculus and analytic geometry*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1963.

Garner, W.R. *Uncertainty and structure as psychological concepts*. New York: Wiley, 1962.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist*, 1963, *18*, 519-521.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Glass, G.V., & Stanley, J.C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1970.

Gronlund, N.E. Measurement and evaluation in teaching (3rd ed.). New York: Macmillan Publishing Co., Inc., 1976.

Hambleton, R.K., Hutton, L., & Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. Journal of Experimental Education, 1976, 45, 57-64.

Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C.W. Problems of objectives-based measurement. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles: Center for the Study of Evaluation, U.C.L.A., 1974 (a).

Harris, C.W. Some technical characteristics of mastery tests. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles: Center for the Study of Evaluation, U.C.L.A., 1974 (b).

Hartley, R.V.L. Transmission of information. Bell Systems Technical Journal, 1928, 7, 535-563.

Hays, W.L. Statistics for the social sciences. New York: Holt, Rinehart & Winston, Inc., 1973.

Hilpinen, R. On the information provided by observations. In J. Hintikka, & P. Suppes (Eds.), Information and inference. Dordrecht, Holland: D. Reidel Publishing Co., 1970.

Hintikka, J., & Suppes, P. (Eds.). Information and inference. Dordrecht, Holland: D. Reidel Publishing Co., 1970.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Kerlinger, F.N. Foundations of behavioral research. New York: Holt, Rinehart & Winston, Inc., 1973.

Kuder, G.F., & Richardson, M.W. The theory of the estimation of test reliability. *Psychometrika*, 1937, *2*, 151-160.

Livingston, S.A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13-26.

Lovett, H.T. Criterion-referenced reliability estimated by analysis of variance. *Educational and Psychological Measurement*, 1977, *37*, 21-29.

Magnusson, D. *Test theory*. Reading, Massachusetts: Addison-Wesley, 1967.

Marshall, J.L., & Haertel, E.H. *The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests*. Unpublished manuscript, University of Wisconsin, 1976.

Mayo, S.T. The methodology and technology of educational and psychological testing. *Review of Educational Research*, 1968, *38*, 92-101. .

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205-216.

Millman, J. Criterion-referenced measurement. In W.J. Popham (Ed.), *Evaluation in education*. Berkeley, California: McCutchan Publishing Co., 1974.

Millman, J., & Popham, W.J. The issue of item and test variance for criterion-referenced tests: A clarification. *Journal of Educational Measurement*, 1974, *11*, 137-138.

Mood, A.M., Graybill, F.A., & Boes, D.C. *Introduction to the theory of statistics*. New York: McGraw-Hill, Inc., 1974.

Nitko, A.J. Problems in the development of criterion-referenced tests: The I.P.I. Pittsburgh experience. In C.W. Harris, M.C. Alkin, & W.C. Popham (Eds.), *Problems in criterion-referenced measurement*. Los Angeles: Center for the Study of Evaluation, U.C.L.A., 1974.

Nyquist, H. Certain factors affecting telegraph speed. *Bell Systems Technical Journal*, 1924, *3*, 324-346.

Osburn, H.G. Sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Popham, W.J. Indices of adequacy for criterion-referenced test items. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

Popham, W.J., & Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Shannon, C.E., & Weaver, W. (Eds.). The mathematical theory of communication. Urbana, Illinois: University of Illinois Press, 1949.

Skager, R.W. Generating criterion-referenced tests from objectives-based assessment systems: Unsolved problems in test development, assembly, and interpretation. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles: Center for the Study of Evaluation, U.C.L.A., 1974.

Stanley, J.C. Reliability. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Subkoviak, M.J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Subkoviak, M.J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, 111-116.

Swaminathan, H., Hambleton, R.K., & Algina, J.A. Reliability of criterion-referenced tests: A decision-theoretic approach. Journal of Educational Measurement, 1974, 11, 263-267.

Swaminathan, H., Hambleton, R.K., & Algina, J.A. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-98.

Thorndike, R.L. Reliability. In E.F. Lindquist (Ed.), Educational measurement (1st ed.). Washington, D.C.: American Council on Education, 1951.

Tornebohm, H. Two measures of evidential strength. In J. Hintikka, & P. Suppes (Eds.), Aspects of inductive logic. Amsterdam: North-Holland Publishing Co., 1966.

Tornebohm, H. On the confirmation of hypotheses about regions of existence. Synthese, 1968, 18, 28-45.

Weltner, K. The measurement of verbal information in psychology and education. New York: Springer-Verlag, Inc., 1973.

Woodson, M. I., & Charles, E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64.

APPENDIX A

Richard E. Sherman
2930 N. Commonwealth Ave.
Apt. 509
Chicago, Illinois 60657
25 September, 1980

Science Research Associates, Inc.
155 North Wacker Drive
Chicago, Illinois 60606

To Whom It May Concern:

I am writing to request permission for the use of test score data gathered by your corporation.

The data requested needs to be of a criterion-referenced nature, and would be desirably have been obtained from either a test of arithmetic or reading skills.

It would also be necessary to have the data collected over a rather large sample of students having taken the same items.

If this data is made available, I intend to use these scores in the analysis section of the doctoral dissertation which I am currently writing.

The topic of my dissertation is the development of reliability coefficient for criterion-referenced mastery tests.

In the analysis section of the dissertation I intend to use the requested sample of data as a bank from which to draw random samples to estimate the standard error of the statistic.

In addition, upon its completion, I would forward a copy of my dissertation to your corporation.

Your consideration of my request is greatly appreciated.

Sincerely,

Richard E. Sherman

Richard E. Sherman

Richard E. Sherman
2930 N. Commonwealth Avenue
Apt. 509
Chicago, Illinois 60657
20 October, 1980

Ms. Rita Bode
Science Research Associates, Inc.
155 North Wacker Drive
Chicago, Illinois 60606

Dear Ms. Bode:

I am writing this letter in regard to the issue of confidentiality of the source of the scores contained on the computer tape which I requested in my letter to you dated 25 September, 1980.

You have my full assurance that both the individuals and school districts from which these scores were obtained will remain anonymous.

Indeed, I am aware of the fact that such identifying information will be removed from the tape which I would receive.

Additionally, since my dissertation is of a statistically theoretical nature, there would be no need to report such information.

Your continued consideration of my request is greatly appreciated.

Sincerely,

Richard E. Sherman

Richard E. Sherman

APPENDIX B

Values of $-\log_2 p$ for Selected p(Probability) Levels

p Level	-log p	p Level	-log p	p Level	-log p
.01	6.640	.35	1.515	.68	.556
.02	5.645	.36	1.474	.69	.535
.03	5.060	.37	1.434	.70	.515
.04	4.645	.38	1.396	.71	.494
.05	4.322	.39	1.358	.72	.474
.06	4.058	.40	1.322	.73	.456
.07	3.837	.41	1.286	.74	.434
.08	3.644	.42	1.251	.75	.415
.09	3.474	.43	1.218	.76	.396
.10	3.322	.44	1.184	.77	.377
.11	3.184	.45	1.152	.78	.358
.12	3.059	.46	1.120	.79	.340
.13	2.943	.47	1.089	.80	.322
.14	2.836	.48	1.059	.81	.304
.15	2.737	.49	1.029	.82	.286
.16	2.643	.50	1.000	.83	.269
.17	2.556	.51	.971	.84	.252
.18	2.474	.52	.943	.85	.234
.19	2.396	.53	.916	.86	.218
.20	2.322	.54	.888	.87	.201
.21	2.251	.55	.863	.88	.184
.22	2.184	.56	.836	.89	.168
.23	2.120	.57	.811	.90	.152
.24	2.059	.58	.786	.91	.136
.25	2.000	.59	.761	.92	.120
.26	1.943	.60	.737	.93	.105
.27	1.888	.61	.713	.94	.089
.28	1.836	.62	.690	.95	.074
.29	1.786	.63	.666	.96	.059
.30	1.737	.64	.644	.97	.044
.31	1.690	.65	.622	.98	.029
.32	1.644	.66	.600	.99	.014
.33	1.600	.67	.578	1.00	0.000
.34	1.556				

APPROVAL SHEET

The dissertation submitted by Richard E. Sherman has been read and approved by the following committee:

Dr. Samuel T. Mayo, Director
Professor, Foundations of Education, Loyola

Dr. Jack A. Kavanagh
Associate Professor, Chairperson,
Foundations of Education, Loyola

Dr. Ronald R. Morgan
Associate Professor, Foundations of Education,
Loyola

The final copies have been examined by the director of the dissertation and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the dissertation is now given final approval by the Committee with reference to content and form.

The dissertation is therefore accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

4/20/82
Date

Samuel T. Mayo
Director's Signature