DOCUMENT RESUME

ED 320 928 TM 015 143

AUTHOR	Huynh, Huynh; Saunders, Joseph C.
TITLE	Solutions for Some Technical Problems in Domain-Referenced Mastery Testing. Final Report.
INSTITUTION	South Carolina Univ., Columbia. Coll. of Education.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Aug 80
GRANT	NIE-G-78-0087
NOTE	387p.
PUB TYPE	Collected Works - General (020) -- Statistical Data (110) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC16 Plus Postage.
DESCRIPTORS	Ability Identification; Annotated Bibliographies; Bayesian Statistics; *Computer Assisted Testing; *Cutting Scores; *Decision Making; Elementary Secondary Education; Estimation (Mathematics); *Mastery Tests; Research Reports; Simulation; Statistical Data; Statistical Inference; Test Construction; *Testing Problems; Test Reliability; Test Use
IDENTIFIERS	Beta Binomial Test Model; Binomial Error Model; *Domain Referenced Testing; Nedelsky Method; South Carolina Statewide Testing Program; *Standard Setting

ABSTRACT

A basic technical framework is provided for the design and use of mastery tests. The Mastery Testing Project (MTP) prepared this framework using advanced mathematics supplemented with computer simulation based on real test data collected by the South Carolina Statewide Testing Program. The MTP focused on basic technical issues encountered in using test scores for making decisions regarding individual students. The initial overview of the project includes abstracts of the 17 papers included in this compilation. They are: (1) "A Nonrandomized Minimax Solution for Passing Scores in the Binomial Error Model" (H. Huynh); (2) "Bayesian and Empirical Bayes Approaches to Setting Passing Scores on Mastery Tests" (H. Huynh and J. C. Saunders); (3) "A Class of Passing Scores Based on the Bivariate Normal Model" (H. Huynh); (4) "An Empirical Bayes Approach to Decisions Based on Multivariate Test Data" (H. Huynh); (5) "A Comparison of Two Ways of Setting Passing Scores Based on the Nedelsky Procedure" (J. C. Saunders and others); (6) "Budgetary Consideration in Setting Passing Scores" (H. Huynh); (7) "Computation and Inference for Two Reliability Indices in Mastery Testing Based on the Beta-Binomial Model" (H. Huynh); (8) "Accuracy of Two Procedures for Estimating Reliability of Mastery Tests" (H. Huynh and J. C. Saunders); (9) "An Approximation to the True Ability Distribution in the Binomial Error Model and Applications" (H. Huynh and G. K. Mandeville); (10) "Adequacy of Asymptotic Normal Theory in Estimating Reliability for Mastery Tests Based on the Beta-Binomial Model" (H. Huynh); (11) "Considerations for Sample Size in Reliability Studies for Mastery Tests" (J. C. Saunders and H. Huynh); (12) "Statistical Inference for False Positive and False Negative Error Rates in Mastery Testing" (H. Huynh); (13) "Relationship between Decision Accuracy and Decision Consistency in Mastery Testing" (H. Huynh and J. C. Saunders); (14) "A Note on Decision-Theoretic Coefficients for Tests" (H. Huynh); (15) "Assessing Efficiency of Decisions in Mastery Testing" (H. Huynh); (16) "Assessing Test Sensitivity in Mastery Testing" (H. Huynh); and (17) "Selecting Items and Setting Passing Scores for Mastery Tests Based on the Two-Parameter Logistic Model" (H. Huynh). A discussion of the future of mastery testing is included, and nine appendices expand on the annotated papers. References follow the individual papers, many of which contain tables of study information. (SLD)

TM

SOLUTIONS FOR SOME TECHNICAL PROBLEMS IN DOMAIN-REFERENCED MASTERY TESTING

HUYNH HUYNH
JOSEPH C. SAUNDERS

FINAL REPORT
AUGUST, 1980

DEPARTMENT OF EDUCATIONAL
RESEARCH AND PSYCHOLOGY
COLLEGE OF EDUCATION
UNIVERSITY OF SOUTH CAROLINA
COLUMBIA, SOUTH CAROLINA 29208

This research was supported by the National Institute of Education, Department of Health, Education, and Welfare, Grant NIE-G-78-0087

SOLUTIONS FOR SOME TECHNICAL PROBLEMS IN DOMAIN-REFERENCED MASTERY TESTING

HUỲNH HUYNH

JOSEPH C. SAUNDERS

FINAL REPORT

AUGUST, 1980

DEPARTMENT OF EDUCATIONAL
RESEARCH AND PSYCHOLOGY
COLLEGE OF EDUCATION
UNIVERSITY OF SOUTH CAROLINA
COLUMBIA, SOUTH CAROLINA 29208

This research was supported by the National Institute of Education, Department of Health, Education, and Welfare, Grant NIE-G-78-0087

CONTENTS

APPENDICES

ACKNOWLEDGEMENTS

The research reported herein was supported by the National Institute of Education. During two funding years, extending from September, 1978, through August, 1980, Carlyle Maw and Lawrence Rudner, in their capacity as Project Officers, were very helpful in seeing that the research project was carried out smoothly. Their assistance is gratefully acknowledged.

Both Anthony Nitko and Elizabeth Haran acted as consultants; in the process they provided valuable comments and suggestions on several research papers. Joseph Ryan provided the necessary environment in which the Nedelsky study was conducted and Garrett Mandeville labored through the computer program which was instrumental in the study on ability estimation. The Office of Research of the South Carolina Department of Education, Paul Sandifer, Director, supplied data from the Statewide Testing Program which were used to validate several theoretical procedures. Roan Garcia-Quintana assisted in the selection and compilation of the data. To all, our special note of thanks.

A researcher's life would be easy if solutions were always found at a moment's thought and if theoretical procedures always worked well in real situations. Unfortunately, this was not the case for us. The many long hours at the office during weekdays and frequent absences during weekends were endured graciously by our spouses, Sarah Seaman-Huynh and Nancy Saunders. Without their patience and care, and sharing of the upsets and frustrations, the project could not have reached a satisfactory conclusion.

The final report was typed by Michele Davis Bergen, Dianne Suber, and Farzana Karim. Their super typing deserves more than a mere expression of our appreciation. Finally, a note of acknowledgement is due to the Computer Service Division of the University of South Carolina, which made available hardware and software to facilitate the completion of the project.

August 31, 1980

Huynh Huynh
Joseph C. Saunders

ABSTRACT

In recent years, there has been considerable interest in the precise assessment of instructional outcomes. The inadequacy of norm-referenced devices has been recognized. In addition, there has been a movement toward gearing educational tests to the specific educational outcomes that instructional programs are intended to reflect. These tests are often referred to as criterion-referenced, domain-referenced, or mastery tests.

A mastery test is typically designed to reflect specific educational objectives and is normally used to make decisions regarding student achievement. Such tests also form an integral part of any program evaluation, where the focus is on the number of students judged as competent in a given domain of performance. Other situations in which institutional decisions about individuals are required include: testing for certification in a profession; testing for minimum competency, such as for high school graduation; and the assessment of basic skills.

This study provides a basic technical framework for the design and use of mastery tests. The topics discussed are (a) appropriate ways to select test items, (b) practical methods for extracting the best information from test data, (c) efficient procedures for using data to make decisions, and (d) means for relating test scores to the instructional outcomes being evaluated. Statistical procedures and computer programs have been developed to help testing practitioners deal with these issues in a simple and convenient way.

The solutions reported in this study are directed toward the improvement of educational testing in the context of instruction.

An Overview of the Mastery Testing Project

AN OVERVIEW OF THE MASTERY TESTING PROJECT

Huynh Huynh
Joseph C. Saunders

I. BACKGROUND

Recent developments and interest in adaptive instruction and mastery learning call for new testing procedures focusing on the evaluation of individual performance in terms of some competency criterion. Given that a domain of behaviors is uniquely defined by the mastery of some unit of instruction, a test is deliberately constructed to produce scores that reflect the degree of competency in those behaviors. At the end of the period of instruction, the test is administered to the individual student, and on the basis of the observed test score he or she is classified in one of several achievement categories. In typical instructional situations there are two such categories, usually labeled mastery and nonmastery.

Using test scores to make decisions about individual students is a daily activity in any effort to evaluate instructional programs. When the objectives are clearly specified, an obvious concern of the evaluator is the number of students or trainees who have mastered any or all the objectives as a result of participating in the program. The classification of students actually serves a dual purpose: first, it pinpoints the objectives that a disproportionate number of students have failed to master, thus encouraging a closer

---

The Mastery Testing Project was supported by Grant NIE-G-78-0087 with the National Institute of Education, Department of Education, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE position or policy and no official endorsement should be inferred. Requests for reprints of the papers described in the Publication Series in Mastery Testing *should be addressed to Huynh Huynh, College of Education, University of South Carolina, Columbia, South Carolina, 29208.*

look at the instructional strategies for those objectives; second, it identifies individual students who have not mastered some of the objectives and for whom special provisions need to be made to facilitate their attainment of these objectives.

Thus, using test scores to make decisions is an integral part of the educational enterprise. In various stages of educational testing development, this effort has been known as criterion-referenced, domain-referenced, or mastery testing. Though these terms have different interpretations, it seems important to note that they often refer to different aspects of the same process. Consider, for example, the case in which test items are deliberately constructed (or selected from an item bank) to reflect specific educational objectives; the resulting test scores are referenced to these objectives for interpretation and are then used to assess the competency or mastery of the individual student with respect to each of the objectives.

Criterion-Referenced and Domain-Referenced Testing

Though the term criterion-referenced is used by most testing practitioners (e.g., those working at school districts), the term domain-referenced has been used in the report to make it clear that test items are referenced directly to specific educational objectives. The term mastery, on the other hand, is used to draw attention to the fact that test scores are used to make certain decisions regarding the individual student. It may also be noted that it would be difficult to make meaningful decisions on the basis of test scores unless the test items can be directly referenced to a well-defined domain of performance. (This domain may be defined by a single objective or by several objectives; in these cases the test is typically labeled objective-referenced.) When a student is judged to be a master on the basis of a high test score, what in fact has been mastered? In order to answer this question, the objectives or domain of performances on which the student is to be judged must be specified in advance. If this line of reasoning is correct, then the process of mastery testing embodies the concept of domain-referenced testing.

AN OVERVIEW

Minimum Competency Testing and Basic Skills Assessment

The procedures associated with mastery testing resemble those used in minimum competency testing or in basic skills assessment. In attempting to reverse the decline in the level of student achievement over the last decade, several states have implemented statewide programs testing for minimum competency in the basic skills. Many of these programs aim to insure that high school graduates possess a minimum level of academic achievement or have acquired the skills required to function effectively as adults in American society. Minimum competency testing, in this sense, acts as a high school exit examination or what has been called a certification examination. When used in this manner, minimum competency examinations do not have the positive connotation of some other basic skills assessment programs. The latter programs are specifically designed for a continuous monitoring of the acquisition of basic skills (namely, reading, writing, and mathematics) across succeeding grade levels. The results of these continuous monitoring programs are used to diagnose a student's deficiencies in the basic skills and to provide for instructional remediation.

Although sometimes differing in their ultimate purposes, mastery testing, minimum competency testing, and the monitoring of basic skills are similar in many aspects of test development and other technical problems. The selection or construction of test items relies heavily on a thoughtful specification of the educational objectives or domain of skills to which scores are to be referenced via performance on the test items. The specifications for the items themselves must, in most instances, be worked out in considerable detail so that there will be a high degree of congruence between the test items and the corresponding educational objectives. Technical aspects held in common include issues such as setting passing scores (or performance standards), assessing decision reliability, assessing errors of classification, determining test length, selecting items to maximize the accuracy of classifications, referencing test items to segments of the curriculum or currently adopted textbooks, constructing alternate forms, and studying bias in decisions based on test scores.

II. TECHNICAL PROBLEMS IN MASTERY TESTING

For a period of two years (September 1, 1978, through August 31, 1980), the National Institute of Education provided financial support for the work of the principal investigator concerning some of the above-mentioned technical issues in mastery testing. This research has dealt with the following questions.

(1) What are some of the optimum ways to approach the issue of setting test passing scores in both large testing programs and in a typical classroom situation? How should passing score judgments based on the content of the test items be processed?

(2) In which ways should the concept of reliability in mastery testing be formulated? How can reliability indices be approximated when repeated testing of the same examinees is not feasible? Which inferential procedures are appropriate for studies regarding estimates of reliability?

(3) How should the rate of misclassification be assessed for domain-referenced tests? What are the sampling characteristics of the estimates?

(4) What approaches should be used to study the consequences of making passing decisions on the basis of test scores? Which models would be useful in forecasting the budgetary consequences associated with the selection of a particular passing score?

(5) How should decisions based on test data be evaluated in terms of efficiency or cost-effectiveness?

(6) What are appropriate ways to assess the sensitivity of a test within the context of instruction?

(7) What are some of the scoring rules based on decision theory which may be useful in the context of mastery testing?

(8) What are the appropriate procedures by which items can be selected from an item bank to form a test which must meet specific requirements regarding reliability or decision accuracy?

(9) What procedures are appropriate in formulating decisions based on multivariate test data?

III. PUBLICATION SERIES IN MASTERY TESTING

As the Mastery Testing Project concludes, seventeen papers have been written. All have been distributed nationally through the Publication Series in Mastery Testing and are abstracted as follows.

Research Memorandum 78-1

Computation and Inference for Two Reliability Indices in Mastery Testing Based on the Beta-Binomial Model

Huynh Huynh

Presented at the 17th Annual Southeastern Invitational Conference on Measurement in Education, University of North Carolina at Greensboro, December 8, 1978. Journal of Educational Statistics, Fall, 1979.

Abstract: In mastery testing the raw agreement index and the kappa index may be secured via one test administration when the test scores follow beta-binomial distributions. This paper reports tables and a computer program which facilitate the computation of those indices and of their standard errors of estimate. Illustrations are provided in the form of confidence intervals, hypothesis testing, and minimum sample sizes in reliability studies for mastery tests.

Research Memorandum 78-2

A Nonrandomized Minimax Solution for Passing Scores in the Binomial Error Model

Huynh Huynh

Psychometrika, June 1980.

Abstract: A nonrandomized minimax solution is presented for mastery scores in the binomial error model. The computation does not require prior knowledge regarding an individual examinee or group test data for a population of examinees. The optimum mastery score minimizes the maximum risk which would be incurred by misclassification. A closed-form solution is provided for the case of constant losses, and tables are presented for a variety of situations including linear and quadratic losses. A scheme which allows for correction for guessing is also described.

Research Memorandum 79-1

Accuracy of Two Procedures for
Estimating Reliability of Mastery Tests

Huynh Huynh
Joseph C. Saunders

Presented at the annual conference of the Eastern Educational Research Association, Kiawah Island, South Carolina, February 22-24, 1979. A short version of this paper will appear in Journal of Educational Measurement (in press).

Abstract: The beta-binomial estimates for the raw agreement index p and the kappa index in mastery testing are compared with those based on repeated testings in terms of bias and sampling stability. Across a variety of test score distributions, test lengths, and mastery scores, the beta-binomial estimates tend to underestimate the corresponding population values. The percent of bias, however, is negligible (about 2.5%) for p and moderate (about 10%) for kappa. Both beta-binomial estimates are almost twice as stable as those based on repeated testings. Though the beta-binomial estimates presume equality of item difficulty, the data presented indicate that even gross departures from equality do not affect the performance of the estimates.

Research Memorandum 79-2

Bayesian and Empirical Bayes Approaches
to Setting Test Passing Scores

Huynh Huynh
Joseph C. Saunders

Presented at the symposium "Psychometric approaches to domain-referenced testing" sponsored jointly by the American Educational Research Association and the National Council on Measurement in Education at their annual meetings in San Francisco, April 8-12, 1979.

Abstract: The Bayesian mastery scores as proposed by Swaminathan et al. and the empirical Bayes mastery scores derived from Huynh's decision-theoretic framework are compared on the basis of approximate beta-binomial and real CTBS test data. It is found that the two sets of mastery scores are identical or almost identical as long as the test score distribution is reasonably symmetric or when the true criterion level is high. Large discrepancies tend to occur when this level is low, especially when the test scores concentrate at some extreme scores or are fairly bumpy. However, in terms of mastery/nonmastery decision, the Huynh procedure provides the same classifications as the Bayesian method in practically all situations. Moreover, the former may be used for tests of arbitrary length and has been generalized to more complex testing situations.

Research Memorandum 79-3

Budgetary Consideration in Setting Mastery Scores

Huynh Huynh

Presented as part of the symposium "Setting standards: Theory and practice" sponsored jointly by the American Educational Research Association and the National Council on Measurement in Education at their annual meetings in San Francisco, April 8-12, 1979.

Abstract: A general model along with four illustrations is presented for the consideration of budgetary constraints in the setting of cutoff scores in instructional programs involving remedial actions regarding poor test performers. Budgetary constraints normally put an upper limit on any choice of cutoff score. Given relevant information, this limit may be determined. Alternately, ways to assess the budgetary consequences associated with a given cutoff score are provided. Such information would be useful in any final decision regarding the cutoff score.

Research Memorandum 79-4

A Class of Mastery Scores Based on the Bivariate Normal Model

Huynh Huynh

Proceedings of the 1979 meeting of the American Statistical Association (Social Statistics Section).

Abstract: This study touches some aspects of the determination of mastery scores on the basis of the bivariate normal test model. The loss ratio associated with classification errors is assumed to be constant, and the referral success function ranges in the normal ogive family. Alternately, the model also provides a fairly simple way to assess the loss consequences associated with each mastery score. Such information is deemed useful to the test user who may wish to examine these consequences before making a final choice of cutoff score. It is also noted that the model provides a latent trait analysis for testing/measurement situations involving instructed and noninstructed groups, or pretest and posttest data.

Research Memorandum 79-5

An Approximation to the True Ability Distribution in the Binomial Error Model and Applications

Huynh Huynh
Garrett K. Mandeville

Abstract: Assuming that the density p of the true ability θ in the binomial test score model is continuous in the closed interval [0,1], a Bernstein polynomial can be used to uniformly approximate p. Then via quadratic programming techniques, least-square estimates may be obtained for the coefficients defining the polynomial. The approximation, in turn, will yield estimates for any indices based on the univariate and/or bivariate density funccion associated with the binomial test score model. Numerical illustrations are provided for the projection of decision reliability and proportion of success in mastery testing.

Research Memorandum 79-6

Statistical Inference for False Positive and False Negative Error Rates in Mastery Testing

Huynh Huynh

Psychometrika, March 1980.

Abstract: This paper describes an asymptotic inferential procedure for the estimates of the false positive and false negative error rates. Formulae and tables are described for the computation of the standard errors. A simulation study indicates that the asymptotic standard errors may be used even with samples of 25 cases as long as the Kuder-Richardson Formula 21 reliability is reasonably large. Otherwise, a large sample would be required.

Research Memorandum 79-7

An Empirical Bayes Approach to Decisions Based on Multivariate Test Data

Huynh Huynh

Presented at the annual meeting of the Psychometric Society, Iowa City, Iowa, May 28-30, 1980.

Abstract: A general framework for making mastery/nonmastery decisions based on multivariate test data is described in this study. Over all, mastery is granted (or denied) if the posterior expected loss associated with such action is smaller than the one incurred by the denial (or grant) of mastery. An explicit form for the cutting contour which separates mastery and nonmastery states in the test score space is given for multivariate test scores which follow a normal distribution with a constant loss ratio. For the case involving multiple cutting scores in the true ability space, the test score cutting contour will resemble the boundary defined by multiple test cutting scores when the test reliabilities are reasonably close to unity. For tests with low reliabilities, decisions may very well be based simply on a suitably chosen composite score.

Research Memorandum 80-1

A Comparison of Two Approaches to Setting Passing Scores Based on the Nedelsky Procedure

Joseph C. Saunders
Joseph P. Ryan
Huynh Huynh

Presented at the annual conference of the Eastern Educational Research Association, Norfolk, Virginia, March 5-8, 1980. Applied Psychological Measurement (in press).

Abstract: The Nedelsky procedure has been proposed as a method for setting minimum passing scores for multiple-choice tests, based on an analysis of item content. Two versions of the procedure are compared. Two groups of judges, one using each version, set passing scores for a classroom test. Comparisons are based on (1) the distributions of passing scores, (2) the consistency of pass-fail decisions between the two versions, and (3) the consistency of pass-fail decisions between each version and the passing score established by the test designer. In addition, the relationship between the passing score set by a judge and that judge's level of achievement in the content area is investigated.

Research Memorandum 80-2

Adequacy of Asymptotic Normal Theory in Estimating Reliability for Mastery Tests Based on the Beta-Binomial Model

Huynh Huynh

Abstract: Simulated data based on five test score distributions indicate that a slight modification of the asymptotic normal theory for the estimation of the p and kappa indices in mastery testing will provide results which are in close agreement with those based on small samples. The modification is achieved through the multiplication of the asymptotic standard errors of estimate by the constant $1+m^{3/4}$ where m is the sample size.

Research Memorandum 80-3

Considerations for Sample Size in Reliability Studies for Mastery Tests

Joseph C. Saunders
Huynh Huynh

Presented at the annual conference of the Eastern Educational Research Association, Norfolk, Virginia, March 5-8, 1980.

Abstract: In most reliability studies, the precision of a reliability estimate varies inversely with the number of examinees (sample size). Thus, to achieve a given level of accuracy, some minimum sample size is required. An approximation for this minimum size may be made if some reasonable assumptions regarding the mean and standard deviation of the test score distribution can be made. To facilitate the computations, tables are developed based on the Comprehensive Tests of Basic Skills. The tables may be used for tests ranging in length from five to thirty items, with percent cutoff scores of 60%, 70%, or 80%, and with examinee populations for which the test difficulty can be described as low, moderate, or high, and the test variability as low or moderate. The tables also reveal that for a given degree of accuracy, an estimate of kappa would require a considerably greater number of examinees than would an estimate of the raw agreement index.

Research Memorandum 80-4

A Note on Decision-Theoretic Coefficients for Tests

Huynh Huynh

Abstract: A modification is suggested for the decision-theoretic coefficient δ proposed by van der Linden and Mellenbergh. Under reasonable assumptions, the modified index varies from 0 to 1 inclusive. It is argued that in many practical applications of mastery testing, coefficients such as δ are not readily available, and consistency of decisions may serve as evidence of the quality of the decision-making process.

Research Memorandum 80-5

Assessing Efficiency of Decisions in Mastery Testing

Huynh Huynh

Abstract: Two indices are proposed for assessing the efficiency of decisions in mastery testing. The indices are generalizations of the raw agreement index and the kappa index. Both express the reduction in the proportion of average loss (or the gain in utility) resulting from the use of test scores to make decisions. Empirical data are presented which show little discrepancy between estimates based on the beta-binomial and compound binomial models for one index.

Research Memorandum 80-6

Selecting Items and Setting Passing Scores for Mastery Tests Based on the Two-Parameter Logistic Model

Huynh Huynh

Presented at the Informal Meeting on Model-Based Psychological Measurement sponsored by the Office of Naval Research, Iowa City, Iowa, August 17-22, 1980.

Abstract: Three issues in mastery testing are considered, using a minimax decision framework, based on the two-parameter logistic model. The issues are: (1) setting passing scores, (2) assessing decision efficiency, and (3) selecting items to maximize decision efficiency. The losses or disutilities under consideration have a constant or normal ogive form. It is found that, in the context of minimax decisions, the item selection procedure based on maximum information may not provide the best decision efficiency.

Research Memorandum 80-7

Assessing Test Sensitivity in Mastery Testing

Huynh Huynh

A preliminary version of this paper was presented as part of the symposium "Approaches to test design for the assessment of the effectiveness of educational programs" sponsored by the American Educational Research Association at its annual meeting in Boston, April 7-11, 1980.

Abstract: This paper addresses the concept of test sensitivity within the context of mastery testing. It is argued that correlation-based indices may not be appropriate for the assessment of test sensitivity. Global assessment of test sensitivity may be carried out via indices such as p-max or δ-max. Local measures of sensitivity may be described via a two-parameter logistic model. Procedures are described to check the tenability of test sensitivity on the basis of observed test data.

Research Memorandum 80-8

Relationship between Decision Accuracy and Decision Consistency in Mastery Testing

Huynh Huynh
Joseph C. Saunders

Abstract: In mastery testing, decision accuracy refers to the proportion of examinees who are classified correctly, in one of several achievement categories, by test data. Decision consistency expresses the extent to which decisions agree across two test administrations. Based on twelve cases involving a wide range of α_{21} reliabilities, it was found that decision accuracy and decision consistency were almost perfectly related.

IV. CONCLUDING REMARKS

As the readers of this summary may note, the work of the Mastery Testing Project has focused on the very basic technical issues encountered in using test scores for making decisions regarding individual students. The work blended mathematical rigor with the ambiguity typically encountered in the reality of testing. Oftentimes, advanced mathematics was used, supplemented with computer simulation based on real test data collected from the South Carolina Statewide Testing Program. It is hoped that the many results reported herein will contribute to the best use of testing in the educational enterprise.

PART ONE

SETTING PASSING SCORES

A NONRANDOMIZED MINIMAX SOLUTION FOR PASSING SCORES IN THE BINOMIAL ERROR MODEL

Huynh Huynh

University of South Carolina

Psychometrika, June 1980.

ABSTRACT

A nonrandomized minimax solution is presented for passing scores in the binomial error model. The computation does not require prior knowledge regarding an individual examinee or group test data for a population of examinees. The optimum passing score minimizes the maximum risk which would be incurred by misclassifications. A closed-form solution is provided for the case of constant losses, and tables are presented for a variety of situations including linear and quadratic losses. A scheme which allows for correction for guessing is also described.

1. INTRODUCTION

Much interest has been generated in recent years on the setting of passing (mastery or cutoff) scores. Situations in which passing scores are needed include (a) entrance requirements for an instructional program, (b) advancement of students from one instructional unit to the next, presumably more complex unit, (c) certification

This paper has been distributed separately as RM 78-2, December, 1978.

for occupations and the professions, and (d) minimum competency testing legislated in several states. Most procedures for setting passing scores fall into three broad categories: comparisons with the performance of other individuals (e.g., using norm-referenced data), an examination of item content (e.g., such procedures as the Nedelsky scheme), and a consideration of the consequences incurred by misclassifications. A fairly comprehensive review of some of these procedures may be found in Meskauskas (1976) and in Hambleton, Swaminathan, Algina, and Coulson (1978).

Misclassifications may be characterized by their probabilities of occurrence and losses. The papers by Fhanér (1974) and by Wilcox (1976) consider the selection of passing scores and of test length which would set maximum tolerable limits for the percents of false positive and false negative errors in decision. Both papers rely on the concept of indifference zones centered around the minimum true ability for mastery, and the procedures so presented may be generalized to include the case of arbitrary but constant losses. As subsequently described, the Fhanér-Wilcox presentation may be framed within the minimax context in statistical decision theory.

A simultaneous consideration of false positive errors, false negative errors, and losses--often referred to as the decision-theoretic approach to setting passing scores--is presented in a number of sources including Swaminathan, Hambleton, and Algina (1975); Huynh (1976, 1977); and van der Linden and Mellenbergh (1977). These papers take into account knowledge concerning the true ability of the examinees, and therefore may be applicable when passing scores are to be set for a group of examinees. The procedure advanced by Swaminathan et al. (1975) is based on the assumption of exchangeability of prior information as described in Lindley and Smith (1972) and implemented in Novick, Lewis, and Jackson (1973). It requires specification of how much prior information is exchangeable. On the other hand, solutions proposed by Huynh (1976, 1977) may be classified as Bayes or empirical Bayes. The first qualifier applies to the case of the individual examinee, when the prior distribution regarding his ability must be available. This distribution may be assessed via procedures described in Novick and Jackson (1974) and implemented via the CADA system (Novick, Isaacs, and DeKeyrel, 1977). The second category, empirical Bayes, may be used when test data are available for a group of examinees.

The empirical Bayes approach seems appropriate where past data or data collected in field testing are used for setting passing scores for future examinees who will take the same test or alternate forms of the same test. There are, however, situations in which such group data or prior information about the individual examinee may not be appropriate. This is the case of individualized instructional programs. Here decisions regarding mastery or nonmastery for an individual examinee ought to be based solely on the subject's test score, not on the performance of other examinees who happen to be in the same situation.

The present paper focuses on a minimax approach to setting passing scores. This procedure does not require specification of prior information regarding the ability of an individual examinee or group of examinees. Using this procedure, a passing score may be established prior to any administration of the test. Section 2 of this paper presents the overall minimax framework for binary classifications. In subsequent sections, various illustrations are provided, based on the binomial error model.

2. BASIC ELEMENTS OF THE MINIMAX PROCEDURE

The true ability of a given examinee is defined as θ with range Ω. For the binomial error model (Lord & Novick, 1968, chap. 23), θ is the proportion of items in a large item pool that the examinee is expected to answer correctly, and Ω is the interval $[0,1]$. If a test is administered to the examinee, it is assumed that his observed test score x is distributed according to a conditional density $f(x|\theta)$. In subsequent discussions, the notation $P(A|\theta)$ denotes the conditional probability that x is in A given that the true ability is θ.

A referral task (Huynh, 1976) shall be assumed to exist. The task is operationally defined via a nondecreasing function $s(\theta)$ which specifies the probability that an examinee with true ability θ will succeed in performing the task. The referral task may be real or hypothetical. For example, if the test scores reflect achievement in the current instructional unit, then the next, presumably more advanced, unit may serve as the referral task. This may be the case, for example, if instructional units are hierarchically sequenced according to the level of complexity (Huynh and Perney, 1979). In other situations, such as minimum competency testing, a consensus on what constitutes an acceptable level of performance may be conceptualized as a referral task. To be specific, let it be agreed that in order to qualify as a true master, an examinee must have a true ability of at least θ_o. Then the referral success function may be taken as $s(\theta) = 0$ for $\theta < \theta_o$ and $s(\theta) = 1$ for $\theta \geq \theta_o$. The constant θ_o is referred to as a criterion level by Hambleton and Novick (1973) and a true mastery score by Huynh (1976).

The examinee will be classified in either the mastery status (action a_1) or the nonmastery status (action a_2) on the basis of the test score x and by relying on some decision rule c. Given a specific true ability score θ, test scores may take a variety of values in a certain range. Hence, for each examinee, actions a_1 and a_2 may both have positive probabilities of being chosen. These probabilities sum to one since either a_1 or a_2 must be taken. The performance of the examinee on the referral task may be deemed success (true state b_1) or failure (true state b_2). If the true state is b_1, then action a_1 should be taken. For b_2, a_2 should be selected. For these two cases, each course of action taken is the best, hence no (opportunity) losses are involved. On the other hand, the combination (a_1,b_2) constitutes a false positive decision, and (a_2,b_1) a false negative classification. Let the loss associated with (a_1,b_2) be $C_f(\theta)$ and that incurred by (a_2,b_1) be $C_s(\theta)$. These losses are functions of a particular true ability θ. At this true ability, b_1 occurs with probability $s(\theta)$ and b_2 with probability $1 - s(\theta)$. Hence, the loss is expected to be $C_f(\theta)\cdot[1-s(\theta)]$ for taking action a_1, and $C_s(\theta)\cdot s(\theta)$ for taking action a_2.

Consider the decision rule denoted by c. This rule partitions the range of the test scores into two disjoint subsets: A_1 (for action a_1), and A_2 (for action a_2), each with a conditional probability of $P(A_1|\theta)$ and $P(A_2|\theta)$, respectively. For an examinee with true ability θ, the expected loss associated with c is

$$L(c,\theta) = C_f(\theta)\cdot[1-s(\theta)]\cdot P(A_1|\theta) + C_s(\theta)\cdot s(\theta)\cdot P(A_2|\theta). \tag{1}$$

Let

$$M(c) = \sup_{\theta\in\Omega} L(c,\theta). \tag{2}$$

Then the minimax decision rule c_o is the one which corresponds to the minimum (if it exists) of M(c) when c ranges in the space consisting of all possible decision rules. This paper, however, will restrict itself to the case of nonrandomized decision rules.

More details regarding the minimax principle and its relationship with Bayesian decision procedures (as implemented in Huynh [1976], for example) may be found in Ferguson (1967). The reader may note that, in a number of situations, there exists a (least favorable) prior distribution on the true ability such that the corresponding Bayes solution is exactly the same as the minimax decision rule.

The remaining portion of this paper will deal only with the binomial error model when it is used with a 0-1 form for the referral success function. The binomial error model appears to be applicable when the test given to each examinee can be thought of as a random sample of items drawn from a large item pool. On the other hand, the 0-1 form for $s(\theta)$ implies a consensus on a minimum level of mastery on the true ability continuum.

3. THE BINOMIAL ERROR MODEL WITH 0-1 REFERRAL SUCCESS

Consider the case where $s(\theta) = 0$ for $\theta < \theta_o$ and $s(\theta) = 1$ for $\theta \geq \theta_o$. In the simple context of mastery testing, the inequality

"$\theta < \theta_o$" describes a true nonmastery state whereas the inequality "$\theta \geq \theta_o$" indicates a true mastery state. In other words, θ_o is the minimum true ability that an examinee must have in order to qualify for true mastery in the domain of content under consideration. It follows that the expected loss associated with the decision rule c as specified in (1) becomes

$$L(c,\theta) = \begin{cases} C_f(\theta)P(A_1|\theta) & \text{if} \quad \theta < \theta_o \\ C_s(\theta)P(A_2|\theta) & \text{if} \quad \theta \geq \theta_o . \end{cases} \tag{3}$$

Now let

$$L_1(c) = \sup_{\theta<\theta_o} C_f(\theta)P(A_1|\theta)$$

and

$$L_2(c) = \sup_{\theta\geq\theta_o} C_s(\theta)P(A_2|\theta) ;$$

then

$$M(c) = \max \{L_1(c), L_2(c)\} .$$

Suppose that for a fixed θ, the distribution of x follows the binomial density function $f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. This is called the binomial error model (Lord & Novick, 1968). Such a distribution belongs to the monotone likelihood ratio family (Ferguson, 1967, chap. 5). Under fairly general conditions regarding $C_f(\theta)$ and $C_s(\theta)$, the search for a nonrandomized minimax rule c_o may be confined to the class of partitions of the test score range $A_1 = \{x; x \leq c - 1\}$ and $A_2 = \{x; x \geq c\}$ defined by a cutoff score c. The cutoff score c_o, which corresponds to the minimax rule c_o, will be referred to as the minimax passing score. There are two degenerate cases which correspond to c = 0 and c = n + 1. When c = 0, A_1 is empty, and hence the examinee is declared a master regardless of his test score. On the other hand, A_2 is empty if c = n + 1. For this situation, mastery is always denied.

It follows that the minimax passing score may be found by minimizing the function $M(c) = \max \{L_1(c), L_2(c)\}$ where

$$L_1(c) = \sup_{\theta<\theta_o} C_f(\theta) \sum_{x=c}^{n} \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{4}$$

and

$$L_2(c) = \sup_{\theta\geq\theta_o} C_s(\theta) \sum_{x=o}^{c-1} \binom{n}{x} \theta^x (1-\theta)^{n-x}. \tag{5}$$

The following section will provide the detailed computations for the case of constant losses.

4. THE BINOMIAL ERROR MODEL WITH 0-1 REFERRAL SUCCESS AND CONSTANT LOSSES

Let ε_1 and ε_2 be two suitably chosen nonnegative constants such that $0 < \theta_o - \varepsilon_1 \leq \theta_o + \varepsilon_2 < 1$. Without loss of generality, the case of constant losses may be specified as follows:

$$C_f(\theta) = \begin{cases} 1 & \text{if} \quad \theta < \theta_o - \varepsilon_1 \\ 0 & \text{if} \quad \theta_o - \varepsilon_1 \leq \theta < \theta_o, \end{cases}$$

and

$$C_s(\theta) = \begin{cases} Q & \text{if} \quad \theta_o + \varepsilon_2 \leq \theta \\ 0 & \text{if} \quad \theta_o \leq \theta < \theta_o + \varepsilon_2. \end{cases}$$

Thus the region $\theta\varepsilon\left[\theta_o - \varepsilon_1, \theta_o + \varepsilon_2\right]$ is an indifference zone. For an examinee with a true ability within this region, it does not matter whether action a_1 or a_2 is taken. It may be noted that the constant Q is the ratio of the loss caused by a false negative decision to that incurred by a false positive decision (i.e., $Q = C_s(\theta) \div C_f(\theta)$).

It can be verified that the functions $L_1(c)$ and $L_2(c)$ as detailed in (4) and (5) are given as

$$L_1(c) = \sum_{x=c}^{n} \binom{n}{x} (\theta_o-\varepsilon_1)^x (1-\theta_o+\varepsilon_1)^{n-x} \tag{6}$$

and

$$L_2(c) = Q \sum_{x=o}^{c-1} \binom{n}{x} (\theta_o+\varepsilon_2)^x (1-\theta_o-\varepsilon_2)^{n-x}. \tag{7}$$

For the general case where ε_1 and ε_2 are not zero, the search for the minimax passing score c_o may be accomplished by computing the value of $M(c) = \max\{L_1(c), L_2(c)\}$ for each value $c = 0, 1, 2, \ldots, n+1$, and then selecting the value c_o at which $M(c)$ is the smallest.

Numerical Example

Assume $n = 5$, $\theta_o = .80$, $\varepsilon_1 = .10$, $\varepsilon_2 = .05$, and $Q = .80$. Table 1 reports the values of L_1, L_2, and M at the passing scores of 0, 1, 2, 3, 4, 5, and 6. Note that both 0 and 6 are degenerate passing scores. The minimax passing score is $c_o = 5$.

TABLE 1

Values of the Functions L_1, L_2, and M

	Passing Score						
Function	0	1	2	3	4	5	6
$L_1(c)$	1	.99757	.96922	.83692	.52822	.16807	0
$L_2(c)$	0	.00006	.00178	.02129	.13183	.44503	.80
$M(c)$	1	.99757	.96922	.83692	.52822	.44503	.80

The minimax passing score is $c_o = 5$. All computations were carried out with a table of cumulative binomial distributions.

The aforementioned discussion encompasses part of the presentation by Wilcox (1976) regarding the length and passing score of a mastery test. Table I of the Wilcox paper provides minimax passing scores for the following combinations: $n = 8\ (1)\ 20$, θ_o (Wilcox's π_o) = .70 (.05) .85, $\varepsilon_1 = \varepsilon_2$ (Wilcox's c) = .05, .10, and $Q = 1$. The maximum expected loss, $M(c_o)$, associated with the minimax passing score is obtained by subtracting from one the minimum probability of a correct decision as tabulated in Wilcox's Table I. For example, with $n = 10$, $\theta_o = .75$, $\varepsilon_1 = \varepsilon_2 = .05$, and $Q = 1$, the minimax passing score is $c_o = 8$. The corresponding maximum expected loss is $M(c_o) = 1 - .6172 = .3828$.

The remaining part of this paper will focus on the case $\varepsilon_1 = \varepsilon_2 = 0$. It follows from Equations (6) and (7) that

$$M(c) = \max\{L_1(c), Q\cdot(1-L_1(c))\}$$

where

$$L_1(c) = \sum_{x=c}^{n} \binom{n}{x} \theta_o^x (1-\theta_o)^{n-x} . \tag{8}$$

If the test score x were continuous, the minimax passing score c_o would be the one at which $L_1(c) = Q\cdot\left(1-L_1(c)\right)$. In other words, it would satisfy the equation

$$\sum_{x=c_o}^{n} \binom{n}{x} \theta_o^x (1-\theta_o)^{n-x} = \frac{Q}{1+Q}. \tag{9}$$

If this equation has an integer solution c_o, then c_o is the minimax passing score. Otherwise, let c_o' be the smallest integer such that

$$\sum_{x=c_o'}^{n} \binom{n}{x} \theta_o^x (1-\theta_o)^{n-x} < \frac{Q}{1+Q}. \tag{10}$$

The minimax passing score will be either c_o' or $c_o'-1$ (or possibly both), whichever minimizes the maximum expected loss M(c).

Numerical Example

Let n = 10, $\theta_o = .70$, and Q = .5. Then via a table of cumulative binomial distributions, it may be found that $c_o' = 9$. At the cutoff score 9, M(c) = 4253, and at the other cutoff score 8 ($=c_o'-1$), M(c) = .3828. Thus the minimax passing score is $c_o = 8$.

Now let I(p,q;t) denote the incomplete beta function as tabulated in Pearson (1934) and implemented via computer routines such as BDTR of the IBM Scientific Subroutine Package (1971) or MDBETA of the International Mathematical and Statistical Library (1977). Inequation (10) may now be written as

$$I(c_o', n-c_o' + 1; \theta_o) < \frac{Q}{1+Q}. \tag{11}$$

This inequality is reminiscent of the one defining the Bayes (or empirical Bayes) passing score for the beta-binomial model as presented in Huynh (1976, p. 70-72). In fact, let us impose on the true ability θ the prior beta density with parameters α and β. Then the Bayes (or empirical Bayes) passing score is the smallest integer c_1 at which

$$I(\alpha+c_1, n+\beta-c_1; \theta_o) \le \frac{Q}{1+Q}. \qquad (12)$$

It appears from (11) and (12) that the minimax passing score c_o and the Bayes passing score c_1 do not differ by more than one unit if $\beta = 1$ and if α is sufficiently small.

A special note is due for the case $Q = 1$, i.e., when the consequences associated with false positive decisions and false negative decisions are weighted equally. Equation (9) or Inequation (10) indicates that the minimax passing score c_o would be chosen such that, for an examinee with true ability θ_o, chances are about equal that he would be classified as a master or a nonmaster on the basis of the test score.

Finally, a normal approximation is available for reasonably large n and for θ_o not too close to 0 or 1. Let ξ be the 100/(1+Q) percentile of the unit normal distribution. The minimax passing score may be approximated by the quantity

$$c_o = n\theta_o + \xi\left(n\theta_o(1-\theta_o)\right)^{\frac{1}{2}}.$$

5. THE BINOMIAL ERROR MODEL WITH 0-1 REFERRAL SUCCESS AND POWER LOSSES CENTERING AROUND θ_o

Consider now the loss functions $C_f(\theta) = (\theta_o-\theta)^{p_1}$ for $\theta < \theta_o$ and $C_s(\theta) = Q(\theta-\theta_o)^{p_2}$ for $\theta \ge \theta_o$, where p_1, p_2, and Q are positive constants. Linear losses correspond to $p_1 = p_2 = 1$ and squared error losses are obtained by letting $p_1 = p_2 = 2$. At the cutoff score c, we have

$$L_1(c) = \sup_{\theta<\theta_o} (\theta_o-\theta)^{p_1} \sum_{x=c}^{n} \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

and

$$L_2(c) = \sup_{\theta\ge\theta_o} Q(\theta-\theta_o)^{p_2} \sum_{x=o}^{c-1} \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

For the special case $c = 0$, $L_1(c) = \theta_o^{p_1}$ and $L_2(c) = 0$, hence $M(c) = \theta_o^{p_1}$. On the other hand, when $c = n+1$, $L_1(c) = 0$ and $L_2(c) = Q(1-\theta_o)^{p_2}$, hence $M(c) = Q(1-\theta_o)^{p_2}$. For other situations where $1 \le c \le n$, it may be shown that there exist two values θ_1

and θ_2, $0 < \theta_1 < \theta_o < \theta_2 < 1$ such that at each cutoff c,

$$L_1(c) = (\theta_o - \theta_1)^{P_1} \sum_{x=c}^{n} \binom{n}{x} \theta_1^x (1-\theta_1)^{n-x} \tag{13}$$

and

$$L_2(c) = Q(\theta_2 - \theta_o)^{P_2} \sum_{x=o}^{c-1} \binom{n}{x} \theta_2^x (1-\theta_2)^{n-x}. \tag{14}$$

As in all previous discussions, $M(c) = \max \{L_1(c), L_2(c)\}$. The minimax passing score c_o is the one at which the maximum expected loss M(c) is minimized.

The determination of θ_1 and θ_2 at each cutoff score c may be carried out via numerical approximation procedures such as the Newton-Raphson algorithm for solving nonlinear equations.

5.1. Searching for $L_1(c)$

Consider now the function

$$Z_1(\theta) = \sum_{x=c}^{n} \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

The first derivative Z_1' of Z_1 with respect to θ is given as

$$Z_1'(\theta) = \sum_{x=c}^{n} \binom{n}{x} \left\{x\theta^{x-1}(1-\theta)^{n-x} - (n-x)\theta^x(1-\theta)^{n-x-1}\right\}.$$

Taking into account that

$$\binom{n}{x} x = n\binom{n-1}{x-1}$$

and

$$\binom{n}{x}(n-x) = n\binom{n-1}{x},$$

it follows that

$$Z_1'(\theta) = n\left\{\sum_{x=c}^{n} \binom{n-1}{x-1}\theta^{x-1}(1-\theta)^{n-x} - \sum_{x=c}^{n-1} \binom{n-1}{x}\theta^x(1-\theta)^{n-x-1}\right\}$$

or

$$Z_1'(\theta) = c\binom{n}{c}\theta^{c-1}(1-\theta)^{n-c}.$$

Now let

$$H_1(\theta) = (\theta_o - \theta)^{p_1} Z_1(\theta).$$

Then the value θ_o of θ which maximizes $H_1(\theta)$ satisfies the equation $H_1'(\theta_1) = 0$, where

$$H_1'(\theta) = -p_1(\theta_o-\theta)^{p_1-1} Z_1(\theta) + (\theta_o-\theta)^{p_1} Z_1'(\theta).$$

In other words, θ_1 satisfies the equation $D_1(\theta_1) = 0$, where

$$D_1(\theta) = -p_1 \sum_{x=c}^{n} \binom{n}{x} \theta^x (1-\theta)^{n-x} + c\binom{n}{c}(\theta_o-\theta)\theta^{c-1}(1-\theta)^{n-c} = 0. \quad (15)$$

To solve this equation via the Newton-Raphson algorithm, the derivative $D_1'(\theta)$ is needed. It is given as

$$D_1'(\theta) = c\binom{n}{c}\theta^{c-2}(1-\theta)^{n-c-1}G_1(\theta) \quad (16)$$

where

$$G_1(\theta) = -(p_1+1)\theta(1-\theta) + (\theta_o-\theta)\left(c-1-(n-1)\theta\right) \quad (17)$$

or

$$G_1(\theta) = (n+p_1)\theta^2 - \left(p_1+c+(n-1)\theta_o\right)\theta + (c-1)\theta_o. \quad (18)$$

Consider first the situation where $c > 1$. It may be seen from (17) that $G_1(0) = (c-1)\theta_o > 0$ and $G_1(\theta_o) = -(p_1+1)\theta_o(1-\theta_o) < 0$. Hence it may be seen that $G_1(\theta)$ vanishes at only one point, θ^* between 0 and θ_o. The value of θ^* is given as

$$\theta^* = \frac{p_1+c+(n-1)\theta_o - \{\left(p_1+c+(n-1)\theta_o\right)^2 - 4(n+p_1)(c-1)\theta_o\}^{1/2}}{2(n+p_1)}.$$

It follows that $D_1'(\theta)$ is positive when $0 < \theta < \theta^*$ and negative when $\theta^* < \theta < \theta_o$. In other words, $D_1(\theta)$ is increasing when $0 < \theta < \theta^*$, is decreasing when $\theta^* < \theta < \theta_o$, and reaches a maximum at $\theta = \theta^*$. Since $D_1(0) = 0$, $D_1(\theta_1) > 0$. On the other hand, $D_1(\theta_o) < 0$ as may be seen from (15). Hence $D_1(\theta) = 0$ at only θ_1 where $\theta^* < \theta_1 < \theta_o$. By entering $c = 1$ directly in Equation (15), it may also be argued that $D_1(\theta) = 0$ at only θ_1 somewhere between $\theta^* = 0$ and θ_o.

The above discussion indicates that the value θ_1 may be obtained via the Newton-Raphson iteration procedure with input data $D_1(\theta)$ and $D_1'(\theta)$ computed via (15), (16), and (17). The iteration process has been found to converge if the suitably chosen starting value for θ is somewhere between θ^* and θ_o.

5.2. Searching for $L_2(\theta)$

In the expression defining $L_2(c)$ at the beginning of this section, let $\xi_o = 1-\theta_o$, $\xi = 1-\theta$, $y = n-x$, and $d = n-c+1$. It then may be seen that

$$L_2(c) = Q \sup_{\xi \le \xi_o} (\xi_o-\xi)^{p_2} \sum_{y=d}^{n} \binom{n}{y} \xi^y (1-\xi)^{n-y}.$$

It follows that the search for θ_2, and hence $L_2(c)$, may be conducted in the same way as in the locating of θ_1.

6. A FRAMEWORK OF CORRECTION FOR GUESSING

Consider now the case where each test item has A alternatives, and let us assume that an examinee without knowledge on a given item will randomly choose one of the A alternatives as his response. Thus the framework of knowledge-or-random-guessing is used in the present section.

As in previous sections, let θ be the true proportion of items that an examinee has knowledge of and would respond correctly to if given. Since the examinee guesses randomly on the remaining items (which account for a proportion $1-\theta$), and since each item has A alternatives, the proportion of items that would be answered correctly by pure guessing is $(1-\theta)/A$. Thus an examinee with true ability θ will actually have a probability of $t = \theta+(1-\theta)/A$ to answer correctly each item of the pool of items from which the test is assembled. It may be noted that since $0 \le \theta \le 1$, $\frac{1}{A} \le t \le 1$.

Now let θ_o, p_1, and p_2 have the same meaning as in the beginning of Section 5, and let

$$t_o = \theta_o+(1-\theta_o)/A.$$

Then it may be seen that

$$\theta-\theta_o = \frac{A}{A-1}(t-t_o)$$

and hence

$$L_1(c) = \left(\frac{A}{A-1}\right)^{p_1} \sup_{\frac{1}{A} \leq t < t_o} (t_o - t)^{p_1} \sum_{x=c}^{n} \binom{n}{c} t^x (1-t)^{n-x}, \tag{19}$$

and

$$L_2(c) = Q\left(\frac{A}{A-1}\right)^{p_2} \sup_{t \geq t_o} (t - t_o)^{p_2} \sum_{x=o}^{c-1} \binom{n}{c} t (1-t)^{n-x}. \tag{20}$$

For the two degenerate cases $c = 0$ and $c = n+1$, the maximum expected loss $M(c)$ takes the values

$$M(0) = \left(\frac{A}{A-1}\right)^{p_1} \left(t_o - \frac{1}{A}\right)^{p_2}$$

and

$$M(n+1) = Q\left(\frac{A}{A-1}\right)^{p_2} (1-t_o)^{p_2}.$$

As for $1 \leq c \leq n$, the search for $L_2(c)$ of (20) may be conducted via the procedure described in Section 5.2. The value $L_1(c)$ from (19), with the constraint $\frac{1}{A} \leq t < t_o$, may be obtained by going through the steps described in Section 5.1 to obtain the maximum of the function

$$g(t) = (t_o - t)^{p_1} \sum_{x=c}^{n} \binom{n}{c} t^x (1-t)^{n-x}$$

under the constraint $t \leq t_o$ and the value t^* at which the maximum occurs. If $t^* > \frac{1}{A}$, then

$$L_1(c) = \left(\frac{A}{A-1}\right)^{p_1} g(t^*).$$

On the other hand, if $t^* \leq \frac{1}{A}$, then

$$L_1(c) = \left(\frac{A}{A-1}\right)^{p_1} g\left(\frac{1}{A}\right).$$

As in other cases, $M(c) = \max\ \{L_1(c), L_2(c)\}$ and the minimax passing score is the one at which $M(c)$ is the smallest.

<u>Numerical Example</u>

Let $n = 15$, $\theta_o = .60$, $A = 4$, $p_1 = p_2 = .5$, and $Q = .25$. The minimax passing score is 12. Without correction for guessing, the minimax passing score would be 11.

7. RELATIONSHIP BETWEEN MINIMAX PASSING SCORES AND OTHER PARAMETERS

Extensive computations as well as the examination of Appendix A reported in Section 8 reveal that, other things being the same, the minimax passing score is a nondecreasing function of n, θ_o, and p_2 and a nonincreasing function of A, p_1, and Q. These trends seem to be justified intuitively. For example, a low Q or a high p_2 will reduce the consequences incurred with a false negative error; hence, a higher passing score might be needed to dampen the overall expected loss associated with the decision problem. On the other hand, high values of p_1 will reduce the consequences of a false positive error, thus making a lower passing score tolerable. As for the number A of alternatives, a low value for A will provide opportunity for some extra probability of getting a correct answer beyond the true ability of the examinee. Thus it would be sensible to increase the passing score in order to offset this unwarranted benefit.

8. TABLES OF MINIMAX PASSING SCORES

The computations described in Sections 5 and 6 may be implemented where computer facilities are available. A FORTRAN IV routine will be described in the next section. In a number of instances, however, a passing score might be needed quickly. Appendix A presents a set of tables of passing scores for the case of no correction for guessing (Section 5) only.

All computations were carried out via the FORTRAN program described in Section 9. The tables are set up with the presumption that the false-negative consequences are less serious than those incurred by false positive errors. The parameter Q is set at .25, .50, .75, and 1.00. Sixteen combinations of p_1 and p_2 are used, namely those in which these parameters vary from .50 to 2.00 in steps of .50. The number of items is set at n = 3 (1) 20, and the criterion level at θ_o = .50 (.05) .90.

It is possible to get a passing score of n+1, especially when θ_o is large and/or Q is small. Such a mastery score indicates that nonmastery is always declared regardless of test score. This peculiarity is due to the discontinuous nature of the binomial probability density and produces the seeming paradox noted in the papers by Novick and Lewis (1974, p. 153-154) and by Wilcox (1976, p. 362, footnote) and in Section 10 of this report. In a practical sense, the peculiarity may be avoided by (i) not allowing θ_o to be unrealistically high, and (ii) not letting the loss associated with one type of error in decision (false positive or false negative) dominate that associated with the other type of error.

In a number of instances, it may be possible to deduce a passing score for nontabled entries by taking advantage of the relationships described in Section 7.

Example 1

Let $n = 10$, $p_1 = p_2 = .5$, and $Q = .75$. At $\theta_o = .70$ and $.75$, the passing score is 8. Hence for all θ between .70 and .75, it may be assumed that the passing score is also 8.

Example 2

Let $n = 10$, $p_1 = .5$, $\theta_o = .70$, and $Q = .25$. At both $p_2 = .5$ and 1.0, the passing score is 9. It may be assumed that the same passing score holds for any p_2 between the two given values.

9. COMPUTER PROGRAM

A FORTRAN IV routine for passing score computations based on Sections 5 and 6 is listed in Appendix B. The program requires two packaged subroutines, DRTNI from the Scientific Subroutine Package (1971) and MDBIN of the International Mathematical and Statistical Library (1977).

The main part of the program contains an attempt to solve Equation (15) iteratively at each c via the Newton-Raphson procedure for nonlinear equations, as implemented by DRTNI. A good starting value for θ is required for convergence; therefore, the following steps are built into the program.

1. First, the value θ^* of Section 5.1 is computed.

2. The interval (θ^*, θ_o) will then be divided into N equal intervals using (N-1) points. The value of $D_1(\theta)$ of (15) is computed at successive dividing points until two points, θ_a and θ_b, are found such that the product $D_1(\theta_a)D_1(\theta_b) < 0$.
3. Then the interval (θ_a, θ_b) will be subdivided in M equal intervals in order to search for two successive dividing points θ_t, θ_s such that $D_1(\theta_t)D_1(\theta_s) < 0$.
4. Finally, the starting value for DRTNI is set at $(\theta_t + \theta_s)/2$.

In the construction of the tables of Section 8, the following values were used: N = 20 and M = 50. The tolerance for θ was set at EPS = .0001. Subroutine DRTNI converged in all cases listed in the tables. For long tests along with θ_c very near 0 or 1, an M larger than 50 might be needed for convergence.

10. A SEEMING PARADOX

Consider the mastery decision defined by the parameters n = 3, $\theta_o = .8$, $p_1 = p_2 = .5$, and Q = .25. The nonrandomized minimax passing score is 3, at which the maximum expected loss M(c) is .218. Now let us suppose that the decision has been carried out on a continuous random variable Y independent of the ability θ of the examinee. Let c be any cutoff score. Then

$$L_1(c) = \sup_{\theta<\theta_o} (\theta_o-\theta)^{p_1} P(Y \geq c) = .89443\ P(Y \geq c)$$

and

$$L_2(c) = Q\cdot\sup_{\theta>\theta_o} (\theta-\theta_o)^{p_2} P(Y < c) = .11180(1-P(Y \geq c)).$$

It follows the maximum expected loss M(c) is minimized when $L_1(c) = L_2(c)$ at which $P(Y \geq c) = .111$, and M(c) = .100. Thus, as judged by the minimax principle, the decision rule of randomly assigning mastery status with an 11.1 percent probability and nonmastery status with an 88.9 percent probability is better than that based on the test score!

The apparent paradox is actually caused by the restriction of the decision problem to the class of <u>nonrandomized</u> classifications defined by the passing scores of 0, 1,..., n, n+1. A similar contradiction is also displayed in a paper by Wilcox (1976) in which the minimum probability of a correct decision is <u>not</u> an increasing function of the number of test items.

The paradox, however, may be resolved by a consideration of the entire class of <u>randomized decision</u> rules. It is well known (Ferguson, 1967, Section 2.8) that under fairly general conditions, there always exists a randomized decision rule which is as good as or better than a given nonrandomized decision rule. Randomized minimax decisions, unfortunately, seem harder to approach than nonrandomized decisions.

11. SUMMARY

In this report solutions are provided for the setting of passing scores within the context of nonrandomized decisions based on the binomial test score model. No assumption is required regarding the true ability distribution of the individual examinee or of the group of examinees under study. The model assumes that the test is formed by a random selection of items from a large (real or hypothetical) pool of items. In addition, it requires specification of the minimum true ability for mastery and of consequences incurred by misclassification errors. A scheme for correction-for-guessing within the minimax framework is also presented. Tables and descriptions of a computer program are also provided to facilitate the determination of passing scores.

BIBLIOGRAPHY

Ferguson, T. S. (1967). <u>Mathematical statistics: A decision-theoretic approach</u>. New York: Academic Press.

Fhanér, S. (1974). Item sampling and decision-making in achievement testing. <u>British Journal of Mathematical and Statistical Psychology 27</u>, 172-175.

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

Hambleton, R. K., Swaminathan, H., Algina, J. & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research 48, 1-47.

Huynh, H. (1976). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

Huynh, H. (1977). Two simple classes of mastery scores based on the beta-binomial model. Psychometrika 42, 601-608.

Huynh, H. & Perney, J. (1979). Determination of mastery scores when instructional units are linearly related. Educational and Psychological Measurement 39, 317 - 323.

IBM Application Program, System/360 (1971). Scientific Subroutine Package (360-CM-03X) Version III, programmer's manual. White Plains, New York: IBM Corporation Technical Publications Department.

IMSL Library 1 (1977). Houston: International Mathematical and Statistical Libraries.

Lindley, D. V. & Smith, A. F. M. (1972). Bayesian estimates for the linear model. Journal of the Royal Statistical Society (Series B) 34, 1-14.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Co.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testings: Views regarding mastery and standard-setting. Review of Educational Research 46, 133-158.

Novick, M. R., Isaacs, G. L., & DeKeyrel, D. F. (1977). Computer-assisted data analysis 1977, Manual for the Computer-Assisted Data Analysis (CADA) Monitor. Iowa City: Iowa Testing Program.

Novick, M. R. & Jackson, P. H. (1974). Statistical methods for educational and psychological research. New York: McGraw-Hill.

Novick, M. R. & Lewis, C. (1974). Prescribing test length for criterion-referenced 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, University of California.

Novick, M. R., Lewis, C. & Jackson, P. H. (1973). The estimation of proportions in m groups. Psychometrika 38, 19-45.

Pearson, K. (1934). Tables of the incomplete beta function. Cambridge: University Press.

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

van der Linden, W. J. & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. Applied Psychological Measurement 1, 593-599.

Wilcox, R. (1976). A note on the length and passing score of a mastery test. Journal of Educational Statistics 1, 359-364.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, principal investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance and comments of Joseph C. Saunders and Anthony J. Nitko are gratefully acknowledged.

APPENDIX A

Tables of Minimax Passing Scores
in the Binomial Error Model

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1=0.5$ and $p_2=0.5$

Q=0.25

n	$\theta_o(\%)=$ 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	3	3	4	4
4	3	4	4	4	4	4	4	5	5
5	4	4	4	5	5	5	5	5	6
6	4	5	5	5	6	6	6	6	7
7	5	5	6	6	6	7	7	7	7
8	6	6	6	7	7	7	8	8	8
9	6	7	7	7	8	8	9	9	9
10	7	7	8	8	9	9	10	10	10
11	7	8	8	9	9	10	10	11	11
12	8	8	9	10	10	11	11	12	12
13	8	9	10	10	11	11	12	13	13
14	9	10	10	11	12	12	13	13	14
15	9	10	11	12	12	13	14	14	15
16	10	11	12	12	13	14	15	15	16
17	10	11	12	13	14	15	15	16	17
18	11	12	13	14	15	15	16	17	18
19	12	12	13	14	15	16	17	18	19
20	12	13	14	15	16	17	18	19	20

Q=0.50

n	$\theta_o(\%)=$ 50	55	60	65	70	75	80	85	90
3	2	3	3	3	3	3	3	3	4
4	3	3	3	4	4	4	4	4	4
5	3	4	4	4	4	5	5	5	5
6	4	4	5	5	5	6	6	6	6
7	5	5	5	6	6	6	7	7	7
8	5	5	6	6	7	7	7	8	8
9	6	6	6	7	7	8	8	9	9
10	6	7	7	8	8	9	9	10	10
11	7	7	8	8	9	9	10	10	11
12	7	8	8	9	10	10	11	11	12
13	8	8	9	10	10	11	12	12	13
14	8	9	10	10	11	12	12	13	14
15	9	9	10	11	12	12	13	14	15
16	9	10	11	12	12	13	14	15	16
17	10	11	11	12	13	14	15	16	16
18	10	11	12	13	14	15	16	17	17
19	11	12	13	14	15	16	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=0.75

n	$\theta_o(\%)=$ 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	3
4	3	3	3	3	4	4	4	4	4
5	3	3	4	4	4	5	5	5	5
6	4	4	4	5	5	5	6	6	6
7	4	5	5	5	6	6	6	7	7
8	5	5	6	6	6	7	7	8	8
9	5	6	6	7	7	8	8	8	9
10	6	6	7	7	8	8	9	9	10
11	6	7	7	8	9	9	10	10	11
12	7	7	8	9	9	10	10	11	12
13	7	8	9	9	10	11	11	12	13
14	8	9	9	10	11	11	12	13	13
15	8	9	10	11	11	12	13	14	14
16	9	10	10	11	12	13	14	14	15
17	9	10	11	12	13	14	14	15	16
18	10	11	12	13	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=1.00

n	$\theta_o(\%)=$ 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	3
4	2	3	3	3	3	4	4	4	4
5	3	3	4	4	4	4	5	5	5
6	3	4	4	4	5	5	5	6	6
7	4	4	5	5	5	6	6	7	7
8	4	5	5	6	6	7	7	7	8
9	5	5	6	6	7	7	8	8	9
10	5	6	7	7	8	8	9	9	10
11	6	7	7	8	8	9	9	10	11
12	6	7	8	8	9	10	10	11	11
13	7	8	8	9	10	10	11	12	12
14	7	8	9	10	10	11	12	13	13
15	8	9	10	10	11	12	13	13	14
16	8	9	10	11	12	13	13	14	15
17	9	10	11	12	12	13	14	15	16
18	9	10	11	12	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	10	12	13	14	15	16	17	18	19

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1=0.5$ and $p_2=1.0$

Q=0.25

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	4	4	4	4
4	4	4	4	4	4	4	5	5	5
5	4	5	5	5	5	5	6	6	6
6	5	5	5	6	6	6	6	7	7
7	5	6	6	7	7	7	7	8	8
8	6	7	7	7	8	8	8	9	9
9	7	7	8	8	8	9	9	9	10
10	7	8	8	9	9	10	10	10	11
11	8	8	9	10	10	10	11	11	12
12	8	9	10	10	11	11	12	12	13
13	9	10	10	11	12	12	13	13	14
14	10	10	11	12	12	13	14	14	14
15	10	11	12	12	13	14	14	15	15
16	11	12	12	13	14	15	15	16	16
17	11	12	13	14	15	15	16	17	17
18	12	13	14	15	15	16	17	18	18
19	13	14	14	15	16	17	18	19	19
20	13	14	15	16	17	18	19	20	20

Q=0.50

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	3	4	4	4
4	3	4	4	4	4	4	4	5	5
5	4	4	4	5	5	5	5	6	6
6	5	5	5	5	6	6	6	6	7
7	5	5	6	6	7	7	7	7	8
8	6	6	6	7	7	8	8	8	9
9	6	7	7	8	8	8	9	9	10
10	7	7	8	8	9	9	10	10	10
11	7	8	9	9	10	10	11	11	11
12	8	9	9	10	10	11	11	12	12
13	9	9	10	11	11	12	12	13	13
14	9	10	11	11	12	13	13	14	14
15	10	10	11	12	13	13	14	15	15
16	10	11	12	13	13	14	15	16	16
17	11	12	13	13	14	15	16	16	17
18	11	12	13	14	15	16	17	17	18
19	12	13	14	15	16	17	17	18	19
20	12	13	14	15	16	17	18	19	20

Q=0.75

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	3	3	4	4
4	3	3	4	4	4	4	4	4	5
5	4	4	4	4	5	5	5	5	6
6	4	5	5	5	6	6	6	6	7
7	5	5	6	6	6	7	7	7	7
8	5	6	6	7	7	7	8	8	8
9	6	6	7	7	8	8	9	9	9
10	7	7	8	8	9	9	10	10	10
11	7	8	8	9	9	10	10	11	11
12	8	8	9	10	10	11	11	12	12
13	8	9	10	10	11	11	12	13	13
14	9	10	10	11	12	12	13	14	14
15	9	10	11	12	12	13	14	14	15
16	10	11	12	12	13	14	15	15	16
17	10	11	12	13	14	15	15	16	17
18	11	12	13	14	15	15	16	17	18
19	12	13	13	14	15	16	17	18	19
20	12	13	14	15	16	17	18	19	20

Q=1.00

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	3	3	3	3	3	3	3	4
4	3	3	3	4	4	4	4	4	5
5	4	4	4	4	5	5	5	5	6
6	4	4	5	5	5	6	6	6	6
7	5	5	5	6	6	6	7	7	7
8	5	6	6	6	7	7	8	8	8
9	6	6	7	7	8	8	9	9	9
10	6	7	7	8	8	9	9	10	10
11	7	7	8	9	9	10	10	11	11
12	7	8	9	9	10	10	11	12	12
13	8	9	9	10	11	11	12	13	13
14	9	9	10	11	11	12	13	13	14
15	9	10	11	11	12	13	14	14	15
16	10	10	11	12	13	14	14	15	16
17	10	11	12	13	14	14	15	16	17
18	11	12	13	13	14	15	16	17	18
19	11	12	13	14	15	16	17	18	19
20	12	13	14	15	16	17	18	19	20

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1=0.5$ and $p_2=1.5$

Q=0.25

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	4	4	4	4	4	4
4	4	4	4	4	5	5	5	5	5
5	5	5	5	5	5	6	6	6	6
6	5	5	6	6	6	6	7	7	7
7	6	6	7	7	7	7	8	8	8
8	6	7	7	8	8	8	9	9	9
9	7	8	8	8	9	9	9	10	10
10	8	8	9	9	10	10	10	11	11
11	8	9	9	10	10	11	11	12	12
12	9	10	10	11	11	12	12	13	13
13	10	10	11	11	12	13	13	13	14
14	10	11	12	12	13	13	14	14	15
15	11	12	12	13	14	14	15	15	16
16	11	12	13	14	14	15	16	16	17
17	12	13	14	15	15	16	17	17	18
18	13	14	14	15	16	17	18	18	19
19	13	14	15	16	17	18	18	19	20
20	14	15	16	17	18	19	19	20	21

Q=0.50

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	4	4	4	4
4	4	4	4	4	4	4	5	5	5
5	4	4	5	5	5	5	6	6	6
6	5	5	5	6	6	6	6	7	7
7	5	6	6	7	7	7	7	8	8
8	6	7	7	7	8	8	8	9	9
9	7	7	8	8	8	9	9	10	10
10	7	8	8	9	9	10	10	10	11
11	8	9	9	10	10	11	11	11	12
12	9	9	10	10	11	11	12	12	13
13	9	10	10	11	12	12	13	13	14
14	10	10	11	12	12	13	14	14	15
15	10	11	12	13	13	14	15	15	16
16	11	12	13	13	14	15	15	16	17
17	12	12	13	14	15	16	16	17	18
18	12	13	14	15	16	16	17	18	18
19	13	14	15	16	16	17	18	19	19
20	13	14	15	16	17	18	19	20	20

Q=0.75

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	3	4	4	4
4	3	4	4	4	4	4	5	5	5
5	4	4	5	5	5	5	5	6	6
6	5	5	5	6	6	6	6	7	7
7	5	6	6	6	7	7	7	8	8
8	6	6	7	7	7	8	8	8	9
9	6	7	7	8	8	9	9	9	10
10	7	8	8	9	9	10	10	10	11
11	8	8	9	9	10	10	11	11	12
12	8	9	10	10	11	11	12	12	13
13	9	10	10	11	11	12	13	13	14
14	9	10	11	12	12	13	14	14	15
15	10	11	12	12	13	14	14	15	15
16	11	11	12	13	14	15	15	16	16
17	11	12	13	14	15	15	16	17	17
18	12	13	14	15	15	16	17	18	18
19	12	13	14	15	16	17	18	19	19
20	13	14	15	16	17	18	19	20	20

Q=1.00

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	3	4	4	4
4	3	3	4	4	4	4	4	5	5
5	4	4	4	5	5	5	5	6	6
6	5	5	5	5	6	6	6	7	7
7	5	5	6	6	7	7	7	7	8
8	6	6	7	7	7	8	8	8	9
9	6	7	7	8	8	9	9	9	10
10	7	7	8	8	9	9	10	10	11
11	8	8	9	9	10	10	11	11	12
12	8	9	9	10	11	11	12	12	13
13	9	9	10	11	11	12	12	13	13
14	9	10	11	11	12	13	13	14	14
15	10	11	11	12	13	14	14	15	15
16	10	11	12	13	14	14	15	16	16
17	11	12	13	14	14	15	16	17	17
18	12	12	13	14	15	16	17	18	18
19	12	13	14	15	16	17	18	19	19
20	13	14	15	16	17	18	19	19	20

Table of Minimax Mastery Scores in the Binomial Error Model with p_1=0.5 and p_2=2.0

Q=0.25

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	4	4	4	4	4	4
4	4	4	4	4	5	5	5	5	5
5	5	5	5	5	6	6	6	6	6
6	5	6	6	6	6	7	7	7	7
7	6	6	7	7	7	8	8	8	8
8	7	7	8	8	8	8	9	9	9
9	7	8	8	9	9	9	10	10	10
10	8	9	9	10	10	10	11	11	11
11	9	9	10	10	11	11	11	12	12
12	9	10	11	11	12	12	12	13	13
13	10	11	11	12	12	13	13	14	14
14	11	11	12	13	13	14	14	15	15
15	11	12	13	13	14	15	15	16	16
16	12	13	13	14	15	16	16	17	17
17	13	13	14	15	16	16	17	18	18
18	13	14	15	16	17	17	18	18	19
19	14	15	16	17	17	18	19	19	20
20	14	15	16	17	18	19	20	20	21

Q=0.50

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	4	4	4	4	4
4	4	4	4	4	4	5	5	5	5
5	4	5	5	5	5	6	6	6	6
6	5	5	6	6	6	6	7	7	7
7	6	6	6	7	7	7	8	8	8
8	6	7	7	8	8	8	9	9	9
9	7	8	8	8	9	9	9	10	10
10	8	8	9	9	10	10	10	11	11
11	8	9	9	10	10	11	11	12	12
12	9	10	10	11	11	12	12	13	13
13	10	10	11	12	12	13	13	14	14
14	10	11	12	12	13	14	14	14	15
15	11	12	12	13	14	14	15	15	16
16	11	12	13	14	15	15	16	16	17
17	12	13	14	15	15	16	17	17	18
18	13	14	15	15	16	17	18	18	19
19	13	14	15	16	17	18	19	19	20
20	14	15	16	17	18	19	19	20	21

Q=0.75

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	4	4	4	4	4
4	4	4	4	4	4	5	5	5	5
5	4	5	5	5	5	5	6	6	6
6	5	5	6	6	6	6	7	7	7
7	6	6	6	7	7	7	7	8	8
8	6	7	7	7	8	8	8	9	9
9	7	7	8	8	9	9	9	10	10
10	8	8	9	9	9	10	10	11	11
11	8	9	9	10	10	11	11	12	12
12	9	9	10	11	11	12	12	12	13
13	9	10	11	11	12	12	13	13	14
14	10	11	11	12	13	13	14	14	15
15	11	11	12	13	14	14	15	15	16
16	11	12	13	14	14	15	16	16	17
17	12	13	14	14	15	16	17	17	18
18	12	13	14	15	16	17	17	18	19
19	13	14	15	16	17	18	18	19	20
20	14	15	16	17	18	18	19	20	21

Q=1.00

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	4	4	4	4
4	3	4	4	4	4	4	5	5	5
5	4	4	5	5	5	5	6	6	6
6	5	5	5	6	6	6	7	7	7
7	5	6	6	7	7	7	7	8	8
8	6	7	7	7	8	8	8	9	9
9	7	7	8	8	9	9	9	10	10
10	7	8	8	9	9	10	10	11	11
11	8	9	9	10	10	11	11	11	12
12	9	9	10	10	11	12	12	12	13
13	9	10	11	11	12	12	13	13	14
14	10	11	11	12	13	13	14	14	15
15	10	11	12	13	13	14	15	15	16
16	11	12	13	13	14	15	16	16	17
17	12	12	13	14	15	16	16	17	18
18	12	13	14	15	16	17	17	18	19
19	13	14	15	16	17	17	18	19	20
20	13	14	15	16	17	18	19	20	21

Table of Minimax Mastery Scores in the Binomial Error Model with p_1=1.0 and p_2=0.5

Q=0.25

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	3	3	3	3	3	3	3
4	3	3	3	3	4	4	4	4	4
5	3	4	4	4	4	5	5	5	5
6	4	4	4	5	5	5	6	6	6
7	4	5	5	5	6	6	6	7	7
8	5	5	6	6	6	7	7	8	8
9	5	6	6	7	7	8	8	8	9
10	6	6	7	7	8	8	9	9	10
11	6	7	7	8	8	9	10	10	11
12	7	7	8	9	9	10	10	11	12
13	7	8	9	9	10	10	11	12	12
14	8	8	9	10	11	11	12	13	13
15	8	9	10	10	11	12	13	13	14
16	9	[illegible]	10	11	12	13	14	14	15
17	9	10	11	12	13	13	14	15	16
18	10	11	11	12	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=0.50

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	3	3	3	3	3
4	2	3	3	3	3	4	4	4	4
5	3	3	3	4	4	4	5	5	5
6	3	4	4	4	5	5	5	6	6
7	4	4	5	5	5	6	6	6	7
8	4	5	5	6	6	6	7	7	8
9	5	5	6	6	7	7	8	8	9
10	5	6	6	7	7	8	8	9	9
11	6	6	7	7	8	9	9	10	10
12	6	7	7	8	9	9	10	11	11
13	7	7	8	9	9	10	11	11	12
14	7	8	9	9	10	11	11	12	13
15	8	8	9	10	11	11	12	13	14
16	8	9	10	10	11	12	13	14	15
17	8	9	10	11	12	13	14	15	16
18	9	10	11	12	13	14	15	15	16
19	9	10	11	12	13	14	15	16	17
20	10	11	12	13	14	15	16	17	18

Q=0.75

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	2	3	3	3	3
4	2	2	3	3	3	3	4	4	4
5	3	3	3	3	4	4	4	5	5
6	3	3	4	4	4	5	5	5	6
7	4	4	4	5	5	5	6	6	7
8	4	4	5	5	6	6	7	7	7
9	4	5	5	6	6	7	7	8	8
10	5	5	6	6	7	8	8	9	9
11	5	6	6	7	8	8	9	9	10
12	6	6	7	8	8	9	10	10	11
13	6	7	8	8	9	10	10	11	12
14	7	7	8	9	10	10	11	12	13
15	7	8	9	10	10	11	12	13	14
16	8	8	9	10	11	12	13	14	14
17	8	9	10	11	12	13	13	14	15
18	9	9	10	11	12	13	14	15	16
19	9	10	11	12	13	14	15	16	17
20	9	10	12	13	14	15	16	17	18

Q=1.00

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	2	3	3	3	3
4	2	2	2	3	3	3	3	4	4
5	2	3	3	3	4	4	4	4	5
6	3	3	4	4	4	5	5	5	6
7	3	4	4	4	5	5	6	6	6
8	4	4	5	5	6	6	6	7	7
9	4	5	5	6	6	7	7	8	8
10	5	5	6	6	7	7	8	8	9
11	5	6	6	7	7	8	9	9	10
12	6	6	7	7	8	9	9	10	11
13	6	7	7	8	9	9	10	11	12
14	6	7	8	9	9	10	11	12	13
15	7	8	8	9	10	11	12	13	13
16	7	8	9	10	11	12	12	13	14
17	8	9	10	10	11	12	13	14	15
18	8	9	10	11	12	13	14	15	16
19	9	10	11	12	13	14	15	16	17
20	9	10	11	12	13	14	15	17	18

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1=1.0$ and $p_2=1.0$

Q=0.25

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	3	3	4	4
4	3	3	4	4	4	4	4	4	5
5	4	4	4	5	5	5	5	5	6
6	4	5	5	5	6	6	6	6	7
7	5	5	6	6	6	7	7	7	7
8	5	6	6	7	7	7	8	8	8
9	6	6	7	7	8	8	9	9	9
10	6	7	7	8	8	9	9	10	10
11	7	8	8	9	9	10	10	11	11
12	8	8	9	9	10	11	11	12	12
13	8	9	9	10	11	11	12	13	13
14	9	9	10	11	11	12	13	13	14
15	9	10	11	11	12	13	14	14	15
16	10	10	11	12	13	14	14	15	16
17	10	11	12	13	14	14	15	16	17
18	11	12	13	13	14	15	16	17	18
19	11	12	13	14	15	16	17	18	19
20	12	13	14	15	16	17	18	19	20

Q=0.50

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	2	2	3	3	3	3	3	3	4
4	3	3	3	4	4	4	4	4	4
5	3	4	4	4	4	5	5	5	5
6	4	4	5	5	5	5	6	6	6
7	4	5	5	6	6	6	7	7	7
8	5	5	6	6	7	7	7	8	8
9	5	6	6	7	7	8	8	9	9
10	6	7	7	8	8	9	9	10	10
11	6	7	8	8	9	9	10	10	11
12	7	8	8	9	10	10	11	11	12
13	8	8	9	10	10	11	12	12	13
14	8	9	10	10	11	12	12	13	14
15	9	9	10	11	12	12	13	14	15
16	9	10	11	12	12	13	14	15	16
17	10	10	11	12	13	14	15	16	16
18	10	11	12	13	14	15	16	16	17
19	11	12	13	14	15	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=0.75

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	3
4	3	3	3	3	4	4	4	4	4
5	3	3	4	4	4	5	5	5	5
6	4	4	4	5	5	5	6	6	6
7	4	5	5	5	6	6	6	7	7
8	5	5	6	6	6	7	7	8	8
9	5	6	6	7	7	8	8	9	9
10	6	6	7	7	8	8	9	9	10
11	6	7	7	8	9	9	10	10	11
12	7	7	8	9	9	10	10	11	12
13	7	8	9	9	10	11	11	12	13
14	8	8	9	10	11	11	12	13	14
15	8	9	10	11	11	12	13	14	14
16	9	10	10	11	12	13	14	15	15
17	9	10	11	12	13	14	15	15	16
18	10	11	12	13	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=1.00

n	θ_0(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	3
4	2	3	3	3	3	4	4	4	4
5	3	3	4	4	4	4	5	5	5
6	3	4	4	5	5	5	6	6	6
7	4	4	5	5	6	6	6	7	7
8	4	5	5	6	6	7	7	8	8
9	5	5	6	6	7	7	8	8	9
10	5	6	7	7	8	8	9	9	10
11	6	7	7	8	8	9	10	10	11
12	6	7	8	8	9	10	10	11	12
13	7	8	8	9	10	10	11	12	12
14	7	8	9	10	10	11	12	13	13
15	8	9	10	10	11	12	13	13	14
16	8	9	10	11	12	13	14	14	15
17	9	10	11	12	13	13	14	15	16
18	9	10	11	12	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	10	12	13	14	15	16	17	18	19

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1=1.0$ and $p_2=1.5$

Q=0.25

n	$\theta_o(\%)=$ 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	4	4	4	4
4	3	4	4	4	4	4	5	5	5
5	4	4	5	5	5	5	5	6	6
6	5	5	5	6	6	6	6	7	7
7	5	6	6	6	7	7	7	7	8
8	6	6	7	7	7	8	8	8	9
9	6	7	7	8	8	9	9	9	10
10	7	7	8	8	9	9	10	10	11
11	8	8	9	9	10	10	11	11	12
12	8	9	9	10	11	11	12	12	12
13	9	9	10	11	11	12	12	13	13
14	9	10	11	11	12	13	13	14	14
15	10	11	11	12	13	14	14	15	15
16	10	11	12	13	14	14	15	16	16
17	11	12	13	14	14	15	16	17	17
18	12	12	13	14	15	16	17	18	18
19	12	13	14	15	16	17	18	18	19
20	13	14	15	16	17	18	18	19	20

Q=0.50

n	$\theta_o(\%)=$ 50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	3	3	4	4
4	3	3	4	4	4	4	4	5	5
5	4	4	4	5	5	5	5	5	6
6	4	5	5	5	6	6	6	6	7
7	5	5	6	6	6	7	7	7	8
8	5	6	6	7	7	7	8	8	9
9	6	6	7	7	8	8	9	9	9
10	7	7	8	8	9	9	10	10	10
11	7	8	8	9	9	10	10	11	11
12	8	8	9	10	10	11	11	12	12
13	8	9	10	10	11	12	12	13	13
14	9	9	10	11	12	12	13	14	14
15	9	10	11	12	12	13	14	15	15
16	10	11	12	12	13	14	15	15	16
17	10	11	12	13	14	15	16	16	17
18	11	12	13	14	15	15	16	17	18
19	12	12	13	14	15	16	17	18	19
20	12	13	14	15	16	17	18	19	20

Q=0.75

n	$\theta_o(\%)=$ 50	55	60	65	70	75	80	85	90
3	2	3	3	3	3	3	3	4	4
4	3	3	3	4	4	4	4	4	5
5	4	4	4	4	5	5	5	5	6
6	4	4	5	5	5	6	6	6	7
7	5	5	5	6	6	7	7	7	7
8	5	6	6	6	7	7	8	8	8
9	6	6	7	7	8	8	9	9	9
10	6	7	7	8	8	9	9	10	10
11	7	7	8	9	9	10	10	11	11
12	7	8	9	9	10	10	11	12	12
13	8	9	9	10	11	11	12	13	13
14	8	9	10	11	11	12	13	13	14
15	9	10	11	11	12	13	14	14	15
16	10	10	11	12	13	14	14	15	16
17	10	11	12	13	14	14	15	16	17
18	11	12	13	13	14	15	16	17	18
19	11	12	13	14	15	16	17	18	19
20	12	13	14	15	16	17	18	19	20

Q=1.00

n	$\theta_o(\%)=$ 50	55	60	65	70	75	80	85	90
3	2	2	3	3	3	3	3	3	4
4	3	3	3	4	4	4	4	4	5
5	3	4	4	4	5	5	5	5	6
6	4	4	5	5	5	6	6	6	6
7	4	5	5	6	6	6	7	7	7
8	5	5	6	6	7	7	8	8	8
9	6	6	7	7	8	8	8	9	9
10	6	7	7	8	8	9	9	10	10
11	7	7	8	8	9	10	10	11	11
12	7	8	8	9	10	10	11	12	12
13	8	8	9	10	10	11	12	12	13
14	8	9	10	10	11	12	13	13	14
15	9	10	10	11	12	13	13	14	15
16	9	10	11	12	13	13	14	15	16
17	10	11	12	13	13	14	15	16	17
18	10	11	12	13	14	15	16	17	18
19	11	12	13	14	15	16	17	18	19
20	11	13	14	15	16	17	18	19	20

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1=1.0$ and $p_2=2.0$

θ_0 (%) =

n	50	55	60	65	70	75	80	85	90
					Q=0.25				
3	3	3	3	3	4	4	4	4	4
4	4	4	4	4	4	5	5	5	5
5	4	5	5	5	5	5	6	6	6
6	5	5	6	6	6	6	7	7	7
7	6	6	6	7	7	7	7	8	8
8	6	7	7	7	8	8	8	9	9
9	7	7	8	8	9	9	9	10	10
10	7	8	8	9	9	10	10	11	11
11	8	9	9	10	10	11	11	11	12
12	9	9	10	10	11	11	12	12	13
13	9	10	10	11	12	12	13	13	14
14	10	10	11	12	13	13	14	14	15
15	10	11	12	13	13	14	15	15	16
16	11	12	13	13	14	15	16	16	17
17	12	12	13	14	15	16	16	17	18
18	12	13	14	15	16	16	17	18	19
19	13	14	15	16	16	17	18	19	19
20	13	14	15	16	17	18	19	20	20

θ_0 (%) =

n	50	55	60	65	70	75	80	85	90
					Q=0.50				
3	3	3	3	3	3	4	4	4	4
4	3	4	4	4	4	4	5	5	5
5	4	4	5	5	5	5	5	6	6
6	5	5	5	6	6	6	6	7	7
7	5	6	6	6	7	7	7	8	8
8	6	6	7	7	7	8	8	8	9
9	6	7	7	8	8	9	9	9	10
10	7	8	8	9	9	10	10	10	11
11	8	8	9	9	10	10	11	11	12
12	8	9	9	10	11	11	12	12	13
13	9	9	10	11	11	12	13	13	14
14	9	10	11	11	12	13	13	14	15
15	10	11	11	12	13	14	14	15	15
16	10	11	12	13	14	14	15	16	16
17	11	12	13	14	14	15	16	17	17
18	12	13	13	14	15	16	17	18	18
19	12	13	14	15	16	17	18	19	19
20	13	14	15	16	17	18	19	20	20

θ_0 (%) =

n	50	55	60	65	70	75	80	85	90
					Q=0.75				
3	3	3	3	3	3	3	4	4	4
4	3	3	4	4	4	4	4	5	5
5	4	4	4	5	5	5	5	6	6
6	4	5	5	5	6	6	6	7	7
7	5	5	6	6	7	7	7	7	8
8	6	6	6	7	7	8	8	8	9
9	6	7	7	8	8	9	9	9	10
10	7	7	8	8	9	9	10	10	11
11	7	8	8	9	10	10	11	11	12
12	8	9	9	10	10	11	12	12	12
13	8	9	10	11	11	12	12	13	13
14	9	10	11	11	12	13	13	14	14
15	10	10	11	12	13	13	14	15	15
16	10	11	12	13	13	14	15	16	16
17	11	12	13	13	14	15	16	17	17
18	11	12	13	14	15	16	17	18	18
19	12	13	14	15	16	17	18	18	19
20	12	13	15	16	17	17	18	19	20

θ_0 (%) =

n	50	55	60	65	70	75	80	85	90
					Q=1.00				
3	2	3	3	3	3	3	4	4	4
4	3	3	4	4	4	4	4	5	5
5	4	4	4	5	5	5	5	6	6
6	4	5	5	5	6	6	6	6	7
7	5	5	6	6	6	7	7	7	8
8	5	6	6	7	7	8	8	8	9
9	6	6	7	7	8	8	9	9	10
10	7	7	8	8	9	9	10	10	11
11	7	8	8	9	9	10	11	11	11
12	8	8	9	10	10	11	11	12	12
13	8	9	10	10	11	12	12	13	13
14	9	10	10	11	12	12	13	14	14
15	9	10	11	12	13	13	14	15	15
16	10	11	12	12	13	14	15	16	16
17	11	11	12	13	14	15	16	16	17
18	11	12	13	14	15	16	17	17	18
19	12	13	14	15	16	17	17	18	19
20	12	13	14	15	16	17	18	19	20

Table of Minimax Mastery Scores in the Binomial Error Model
with $p_1=1.5$ and $p_2=0.5$

Q=0.25

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	3	3	3	3	3
4	2	3	3	3	3	4	4	4	4
5	3	3	3	4	4	4	4	5	5
6	3	4	4	4	5	5	5	6	6
7	4	4	4	5	5	6	6	6	7
8	4	5	5	5	6	6	7	7	8
9	5	5	6	6	6	7	7	8	8
10	5	6	6	7	7	8	8	9	9
11	5	6	7	7	8	8	9	10	10
12	6	7	7	8	8	9	10	10	11
13	6	7	8	8	9	10	10	11	12
14	7	8	8	9	10	10	11	12	13
15	7	8	9	10	10	11	12	13	14
16	8	9	9	10	11	12	13	14	14
17	8	9	10	11	12	13	13	14	15
18	9	10	10	11	12	13	14	15	16
19	9	10	11	12	13	14	15	16	17
20	10	11	12	13	14	15	16	17	18

Q=0.50

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	2	3	3	3	3
4	2	2	3	3	3	3	3	4	4
5	2	3	3	3	4	4	4	4	5
6	3	3	4	4	4	5	5	5	6
7	3	4	4	4	5	5	6	6	6
8	4	4	5	5	5	6	6	7	7
9	4	5	5	6	6	7	7	8	8
10	5	5	6	6	7	7	8	8	9
11	5	6	6	7	7	8	9	9	10
12	5	6	7	7	8	9	9	10	11
13	6	7	7	8	9	9	10	11	11
14	6	7	8	8	9	10	11	12	12
15	7	7	8	9	10	11	[illegible]	[illegible]	13
16	7	8	9	10	10	11	12	[illegible]	14
17	8	8	9	10	11	12	13	14	15
18	8	9	10	11	12	13	14	15	16
19	8	9	10	11	12	13	14	16	17
20	9	10	11	12	13	14	15	16	17

Q=0.75

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	1	2	2	2	2	2	3	3	3
4	2	2	2	3	3	3	3	4	4
5	2	3	3	3	3	4	4	4	5
6	3	3	3	4	4	4	5	5	5
7	3	3	4	4	5	5	5	6	6
8	3	4	4	5	5	6	6	7	7
9	4	4	5	5	6	6	7	7	8
10	4	5	5	6	6	7	8	8	9
11	5	5	6	6	7	8	8	9	10
12	5	6	6	7	8	8	9	10	10
13	6	6	7	8	8	9	10	10	11
14	6	7	7	8	9	10	10	11	12
15	6	7	8	9	10	10	11	12	13
16	7	8	8	9	10	11	12	13	14
17	7	8	9	10	11	12	13	14	15
18	8	9	10	10	11	12	13	14	15
19	8	9	10	11	12	13	14	15	16
20	9	10	11	12	13	14	15	16	17

Q=1.00

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	1	2	2	2	2	2	2	3	3
4	2	2	2	2	3	3	3	3	4
5	2	2	3	3	3	4	4	4	5
6	2	3	3	4	4	4	5	5	5
7	3	3	4	4	4	5	5	6	6
8	3	4	4	5	5	5	6	6	7
9	4	4	5	5	6	6	7	7	8
10	4	5	5	6	6	7	7	8	9
11	4	5	6	6	7	7	8	9	9
12	5	6	6	7	7	8	9	10	10
13	5	6	7	7	8	9	10	10	11
14	6	6	7	8	9	9	10	11	12
15	6	7	8	9	9	10	11	12	13
16	7	7	8	9	10	11	12	13	14
17	7	8	9	10	11	12	12	13	14
18	7	8	9	10	11	12	13	14	15
19	8	9	10	11	12	13	14	15	16
20	8	9	10	11	12	14	15	16	17

Table of Minimax Mastery Scores in the Binomial Error Model with p_1=1.5 and p_2=1.0

Q=0.25

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	3	3	3	3	3	3	4
4	3	3	3	3	4	4	4	4	4
5	3	4	4	4	4	5	5	5	5
6	4	4	4	5	5	5	6	6	6
7	4	5	5	5	6	6	7	7	7
8	5	5	6	6	6	7	7	8	8
9	5	6	6	7	7	8	8	9	9
10	6	6	7	7	8	8	9	9	10
11	6	7	7	8	9	9	10	10	11
12	7	7	8	9	9	10	11	11	12
13	7	8	9	9	10	11	11	12	13
14	8	8	9	10	11	11	12	13	13
15	8	9	10	11	11	12	13	14	14
16	9	10	10	11	12	13	14	14	15
17	9	10	11	12	13	14	14	15	16
18	10	11	12	13	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=0.50

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	3
4	2	3	3	3	3	4	4	4	4
5	3	3	4	4	4	4	5	5	5
6	3	4	4	4	5	5	5	6	6
7	4	4	5	5	5	6	6	7	7
8	4	5	5	6	6	7	7	7	8
9	5	5	6	6	7	7	8	8	9
10	5	6	6	7	7	8	9	9	10
11	6	6	7	8	8	9	9	10	11
12	6	7	8	8	9	9	10	11	11
13	7	7	8	9	10	10	11	12	12
14	7	8	9	9	10	11	12	12	13
15	8	9	9	10	11	12	12	13	14
16	8	9	10	11	12	12	13	14	15
17	9	10	10	11	12	13	14	15	16
18	9	10	11	12	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	10	11	12	13	14	15	16	17	18

Q=0.75

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	3	3	3	3	3
4	2	3	3	3	3	4	4	4	4
5	3	3	3	4	4	4	5	5	5
6	3	4	4	4	5	5	5	6	6
7	4	4	4	5	5	6	6	6	7
8	4	5	5	5	6	6	7	7	8
9	5	5	6	6	7	7	8	8	9
10	5	6	6	7	7	8	8	9	9
11	6	6	7	7	8	9	9	10	10
12	6	7	7	8	9	9	10	11	11
13	6	7	8	9	9	10	11	11	12
14	7	8	8	9	10	11	11	12	13
15	7	8	9	10	11	11	12	13	14
16	8	9	10	10	11	12	13	14	15
17	8	9	10	11	12	13	14	15	16
18	9	10	11	12	13	14	15	16	16
19	9	10	11	12	13	14	15	16	17
20	10	11	12	13	14	15	16	17	18

Q=1.00

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	3	3	3	3	3
4	2	2	3	3	3	3	4	4	4
5	3	3	3	4	4	4	4	5	5
6	3	3	4	4	4	5	5	6	6
7	4	4	4	5	5	6	6	6	7
8	4	4	5	5	6	6	7	7	8
9	4	5	5	6	6	7	7	8	8
10	5	5	6	7	7	8	8	9	9
11	5	6	7	7	8	8	9	10	10
12	6	6	7	8	8	9	10	10	11
13	6	7	8	8	9	10	10	11	12
14	7	7	8	9	10	10	11	12	13
15	7	8	9	10	10	11	12	13	14
16	8	9	9	10	11	12	13	14	15
17	8	9	10	11	12	13	14	15	15
18	9	10	10	11	12	13	14	15	16
19	9	10	11	12	13	14	15	16	17
20	10	11	12	13	14	15	16	17	18

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1=1.5$ and $p_2=1.5$

Q=0.25

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	3	3	3	3	3	3	4	4
4	3	3	4	4	4	4	4	5	5
5	4	4	4	4	5	5	5	5	6
6	4	4	5	5	5	6	6	6	7
7	5	5	5	6	6	7	7	7	7
8	5	6	6	7	7	7	8	8	8
9	6	6	7	7	8	8	9	9	9
10	6	7	7	[illegible]	8	9	9	10	10
11	7	7	8	[illegible]	9	10	10	11	11
12	7	8	9	9	10	10	11	12	12
13	8	9	9	10	11	11	12	13	13
14	8	9	10	11	11	12	13	13	14
15	9	10	11	11	12	13	14	14	15
16	10	10	11	12	13	14	14	15	16
17	10	11	12	13	14	14	15	16	17
18	11	11	12	13	14	15	16	17	18
19	11	12	13	14	15	16	17	18	19
20	12	13	14	15	16	17	18	19	20

Q=0.50

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	3	3	3	3	3	3	4
4	3	3	3	4	4	4	4	4	5
5	3	4	4	4	4	5	5	5	5
6	4	4	5	5	5	6	6	6	6
7	4	5	5	6	6	6	7	7	7
8	5	5	6	6	7	7	7	8	8
9	5	6	6	7	7	8	8	9	9
10	6	6	7	8	8	9	9	10	10
11	6	7	8	8	9	9	10	10	11
12	7	8	8	9	10	10	11	11	12
13	7	8	9	10	10	11	12	12	13
14	8	9	9	10	11	12	12	13	14
15	8	9	10	11	12	12	13	14	15
16	9	10	11	12	12	13	14	15	16
17	10	10	11	12	13	14	15	16	16
18	10	11	12	13	14	15	16	17	17
19	11	12	13	14	14	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=0.75

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	4
4	3	3	3	3	4	4	4	4	4
5	[illegible]	3	4	4	4	5	5	5	5
6	4	4	4	5	5	5	6	6	6
7	4	5	5	5	6	6	[illegible]	7	7
8	5	5	6	6	6	7	7	8	8
9	5	6	6	7	7	8	8	9	9
10	6	6	7	7	8	8	9	9	10
11	6	7	7	8	9	9	10	10	11
12	7	7	8	9	9	10	11	11	12
13	7	8	9	9	10	11	11	12	13
14	8	8	9	10	11	11	12	13	14
15	8	9	10	11	11	12	13	14	15
16	9	10	10	11	12	13	14	15	15
17	9	10	11	12	13	14	15	15	16
18	10	11	12	13	14	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=1.00

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	4
4	2	3	3	3	3	4	4	4	4
5	3	3	4	4	4	4	5	5	5
6	3	4	4	5	5	5	6	6	6
7	4	4	5	5	6	6	6	7	7
8	4	5	5	6	6	7	7	8	8
9	5	5	6	6	7	7	8	8	9
10	5	6	7	7	8	8	9	9	10
11	6	7	7	8	8	9	10	10	11
12	6	7	8	8	9	10	10	11	12
13	7	8	8	9	10	10	11	12	13
14	7	8	9	10	10	11	12	13	13
15	8	9	10	10	11	12	13	14	14
16	8	9	10	11	12	13	14	14	15
17	9	10	11	12	13	13	14	15	16
18	9	10	11	12	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	10	12	13	14	15	16	17	18	19

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1=1.5$ and $p_2=2.0$

$\theta_o(\%)=$ — Q=0.25

n	50	55	60	65	70	75	80	85	90
3	3	3	3	3	3	4	4	4	4
4	3	4	4	4	4	4	5	5	5
5	4	4	4	5	5	5	5	6	6
6	4	5	5	5	6	6	6	7	7
7	5	5	6	6	7	7	7	7	8
8	6	6	6	7	7	8	8	8	9
9	6	7	7	8	8	9	9	9	10
10	7	7	8	8	9	9	10	10	11
11	7	8	8	9	10	10	11	11	12
12	8	9	9	10	10	11	12	12	12
13	8	9	10	10	11	12	12	13	13
14	9	10	10	11	12	13	13	14	14
15	10	10	11	12	13	13	14	15	15
16	10	11	12	13	13	14	15	16	16
17	11	12	12	13	14	15	16	17	17
18	11	12	13	14	15	16	17	17	18
19	12	13	14	15	16	17	17	18	19
20	12	13	14	15	16	17	18	19	20

$\theta_o(\%)=$ — Q=0.50

n	50	55	60	65	70	75	80	85	90
3	2	3	3	3	3	3	4	4	4
4	3	3	4	4	4	4	4	5	5
5	4	4	4	4	5	5	5	5	6
6	4	4	5	5	6	6	6	6	7
7	5	5	6	6	6	7	7	7	8
8	5	6	6	7	7	7	8	8	9
9	6	6	7	7	8	8	9	9	9
10	6	7	7	8	9	9	10	10	10
11	7	8	8	9	9	10	10	11	11
12	7	8	9	9	10	11	11	12	12
13	8	9	9	10	11	11	12	13	13
14	9	9	10	11	12	12	13	14	14
15	9	10	11	11	12	13	14	14	15
16	10	10	11	12	13	14	15	15	16
17	10	11	12	13	14	15	15	16	17
18	11	12	13	14	14	15	16	17	18
19	11	12	13	14	15	16	17	18	19
20	12	13	14	15	16	17	18	19	20

$\theta_o(\%)=$ — Q=0.75

n	50	55	60	65	70	75	80	85	90
3	2	3	3	3	3	3	3	4	4
4	3	3	3	4	4	4	4	4	5
5	3	4	4	4	5	5	5	5	6
6	4	4	5	5	5	6	6	6	7
7	5	5	5	6	6	6	7	7	8
8	5	6	6	6	7	7	8	8	8
9	6	6	7	7	8	8	9	9	9
10	6	7	7	8	8	9	9	10	10
11	7	7	8	8	9	10	10	11	11
12	7	8	9	9	10	10	11	12	12
13	8	8	9	10	11	11	12	13	13
14	8	9	10	11	11	12	13	13	14
15	9	10	10	11	12	13	14	14	15
16	9	10	11	12	13	14	14	15	16
17	10	11	12	13	13	14	15	16	17
18	10	11	12	13	14	15	16	17	18
19	11	12	13	14	15	16	17	18	19
20	12	13	14	15	16	17	18	19	20

$\theta_o(\%)=$ — Q=1.00

n	50	55	60	65	70	75	80	85	90
3	2	2	3	3	3	3	3	4	4
4	3	3	3	4	4	4	4	4	5
5	3	[illegible]	4	4	5	5	5	5	6
6	4	4	5	5	5	6	6	6	7
7	4	5	5	6	6	6	7	7	7
8	5	5	6	6	7	7	8	8	8
9	5	6	6	7	7	8	8	9	9
10	6	7	7	8	8	9	9	10	10
11	7	7	8	8	9	10	10	11	11
12	7	8	8	9	10	10	11	12	12
13	8	8	9	10	10	11	12	12	13
14	8	9	10	10	11	12	13	13	14
15	9	9	10	11	12	13	13	14	15
16	9	10	11	12	13	13	14	15	16
17	10	11	12	12	13	14	15	16	17
18	10	11	12	13	14	15	16	17	18
19	11	12	13	14	15	16	17	18	19
20	11	12	13	14	15	17	18	19	20

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1 = 2.0$ and $p_2 = 0.5$

θ_o(%)=

Q=0.25

n	50	55	60	65	70	75	80	85	90
3	2	2	2	2	2	3	3	3	3
4	2	2	3	3	3	3	4	4	4
5	2	3	3	3	4	4	4	5	5
6	3	3	4	4	4	5	5	5	6
7	3	4	4	4	5	5	6	6	6
8	4	4	5	5	5	6	6	7	7
9	4	5	5	6	6	7	7	8	8
10	4	5	6	6	7	7	8	8	9
11	5	5	6	7	7	8	8	9	10
12	5	6	7	7	8	9	9	10	11
13	6	6	7	8	9	9	10	11	11
14	6	7	8	8	9	10	11	11	12
15	7	7	8	9	10	11	11	12	13
16	7	8	9	10	10	11	12	13	14
17	7	8	9	10	11	12	13	14	15
18	8	9	10	11	12	13	14	15	16
19	8	9	10	11	12	13	14	15	16
20	9	10	11	12	13	14	15	16	17

θ_o(%)=

Q=0.50

n	50	55	60	65	70	75	80	85	90
3	1	2	2	2	2	2	3	3	3
4	2	2	2	3	3	3	3	4	4
5	2	2	3	3	3	4	4	4	5
6	3	3	3	4	4	4	5	5	5
7	3	3	4	4	4	5	5	6	6
8	3	4	4	5	5	6	6	6	7
9	4	4	5	5	6	6	7	7	8
10	4	5	5	6	6	7	7	8	9
11	4	5	6	6	7	7	8	9	9
12	5	5	6	7	7	8	9	9	10
13	5	6	7	7	8	9	10	10	11
14	6	6	7	8	9	9	10	11	12
15	6	7	8	8	9	10	11	12	13
16	6	7	8	9	10	11	12	13	14
17	7	8	9	10	10	11	12	13	14
18	7	8	9	10	11	12	13	14	15
19	8	9	10	11	12	13	14	15	16
20	8	9	10	11	12	13	15	16	17

θ_o(%)=

Q=0.75

n	50	55	60	65	70	75	80	85	90
3	1	1	2	2	2	2	2	3	3
4	2	2	2	2	3	3	3	3	4
5	2	2	3	3	3	3	4	4	4
6	2	3	3	3	4	4	4	5	5
7	3	3	3	4	4	5	5	6	6
8	3	3	4	4	5	5	6	6	7
9	3	4	4	5	5	6	6	7	8
10	4	4	5	5	6	7	7	8	8
11	4	5	5	6	7	7	8	9	9
12	5	5	6	7	7	8	9	9	10
13	5	6	6	7	8	9	9	10	11
14	5	6	7	8	8	9	10	11	12
15	6	7	7	8	9	10	11	12	12
16	6	7	8	9	10	10	11	12	13
17	7	7	8	9	10	11	12	13	14
18	7	8	9	10	11	12	13	14	15
19	7	8	9	10	11	12	14	15	16
20	8	9	10	11	12	13	14	15	17

θ_o(%)=

Q=1.00

n	50	55	60	65	70	75	80	85	90
3	1	1	2	2	2	2	2	3	3
4	2	2	2	2	2	3	3	3	4
5	2	2	2	3	3	3	4	4	4
6	2	3	3	3	4	4	4	5	5
7	3	3	3	4	4	5	5	5	6
8	3	3	4	4	5	5	6	6	7
9	3	4	4	5	5	6	6	7	7
10	4	4	5	5	6	6	7	8	8
11	4	5	5	6	6	7	8	8	9
12	4	5	6	6	7	8	8	9	10
13	5	5	6	7	8	8	9	10	11
14	5	6	7	7	8	9	10	11	12
15	6	6	7	8	9	10	10	11	12
16	6	7	8	9	9	10	11	12	13
17	6	7	8	9	10	11	12	13	14
18	7	8	9	10	11	12	13	14	15
19	7	8	9	10	11	12	13	14	16
20	8	9	10	11	12	13	14	15	16

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1 = 2.0$ and $p_2 = 1.0$

Q=0.25

n	$\theta_0(\%)$= 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	3
4	2	3	3	3	3	4	4	4	4
5	3	3	4	4	4	4	5	5	5
6	3	4	4	4	5	5	5	6	6
7	4	4	5	5	5	6	6	7	7
8	4	5	5	6	6	7	7	7	8
9	5	5	6	6	7	7	8	8	9
10	5	6	6	7	7	8	8	9	10
11	6	6	7	7	8	9	9	10	10
12	6	7	7	8	9	9	10	11	11
13	7	7	8	9	9	10	11	11	12
14	7	8	9	9	10	11	12	12	13
15	8	8	9	10	11	12	12	13	14
16	8	9	10	11	11	12	13	14	15
17	8	9	10	11	12	13	14	15	16
18	9	10	11	12	13	14	15	16	17
19	9	10	11	12	13	14	15	16	17
20	10	11	12	13	14	15	16	17	18

Q=0.50

n	$\theta_0(\%)$= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	3	3	3	3	3
4	2	2	3	3	3	3	4	4	4
5	3	3	3	4	4	4	4	5	5
6	3	3	4	4	4	5	5	6	6
7	4	4	4	5	5	6	6	6	7
8	4	4	5	5	6	6	7	7	8
9	4	5	5	6	6	7	7	8	8
10	5	5	6	6	7	8	8	9	9
11	5	6	6	7	8	8	9	10	10
12	6	6	7	8	8	9	10	10	11
13	6	7	8	8	9	10	10	11	12
14	7	7	8	9	10	10	11	12	13
15	7	8	9	9	10	11	12	13	14
16	7	8	9	10	11	12	13	14	14
17	8	9	10	11	12	12	13	14	15
18	8	9	10	11	12	13	14	15	16
19	9	10	11	12	13	14	15	16	17
20	9	10	11	12	14	15	16	17	18

Q=0.75

n	$\theta_0(\%)$= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	2	3	3	3	3
4	2	2	3	3	3	3	4	4	4
5	2	3	3	3	4	4	4	5	5
6	3	3	4	4	4	5	5	5	6
7	3	4	4	4	5	5	6	6	7
8	4	4	5	5	6	6	6	7	7
9	4	5	5	6	6	7	7	8	8
10	5	5	6	6	7	7	8	9	9
11	5	6	6	7	7	8	9	9	10
12	5	6	7	7	8	9	9	10	11
13	6	7	7	8	9	9	10	11	12
14	6	7	8	9	9	10	11	12	13
15	7	8	8	9	10	11	12	13	13
16	7	8	9	10	11	12	12	13	14
17	8	9	9	10	11	12	13	14	15
18	8	9	10	11	12	13	14	15	16
19	9	10	11	12	13	14	15	16	17
20	9	10	11	12	13	14	15	17	18

Q=1.00

n	$\theta_0(\%)$= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	2	3	3	3	3
4	2	2	2	3	3	3	3	4	4
5	2	3	3	3	4	4	4	4	5
6	3	3	3	4	4	5	5	5	6
7	3	4	4	4	5	5	6	6	6
8	4	4	4	5	5	6	6	7	7
9	4	4	5	6	6	7	7	8	8
10	4	5	6	6	7	7	8	8	9
11	5	5	6	7	7	8	9	9	10
12	5	6	7	7	8	9	9	10	11
13	6	6	7	8	9	9	10	11	12
14	6	7	8	8	9	10	11	12	12
15	7	7	8	9	10	11	11	12	13
16	7	8	9	10	10	11	12	13	14
17	7	8	9	10	11	12	13	14	15
18	8	9	10	11	12	13	14	15	16
19	8	9	10	11	12	13	14	16	17
20	9	10	11	12	13	14	15	16	18

Table of Minimax Mastery Scores in the Binomial Error Model with p_1=2.0 and p_2=1.5

Q=0.25

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	3	3	3	3	3	4	4
4	3	3	3	4	4	4	4	4	5
5	3	4	4		4	5	5	5	5
6	4	4	4	5	5	5	6	6	6
7	4	5	5	5	6	6	7	7	7
8	5	5	6	6	7	7	7	8	8
9	5	6	6	7	7	8	8	9	9
10	6	6	7	7	8	9	9	10	10
11	6	7	7	8	9	9	10	10	11
12	7	7	8	9	9	10	11	11	12
13	7	8	9	9	10	11	11	12	13
14	8	9	9	10	11	11	12	13	14
15	8	9	10	11	11	12	13	14	15
16	9	10	10	11	12	13	14	15	15
17	9	10	11	12	13	14	15	15	16
18	10	11	12	13	14	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=0.50

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	4
4	3	3	3	3	4	4	4	4	4
5	3	3	4	4	4	4	5	5	5
6	3	4	4	5	5	5	6	6	6
7	4	4	5	5	6	6	6	7	7
8	4	5	5	6	6	7	7	8	8
9	5	5	6	6	7	7	8	8	9
10	5	6	7	7	8	8	9	9	10
11	6	6	7	8	8	9	10	10	11
12	6	7	8	8	9	10	10	11	12
13	7	8	8	9	10	10	11	12	12
14	7	8	9	10	10	11	12	13	13
15	8	9	9	10	11	12	13	13	14
16	8	9	10	11	12	13	13	14	15
17	9	10	11	12	12	13	14	15	16
18	9	10	11	12	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	10	11	12	13	14	16	17	18	19

Q=0.75

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	3
4	2	3	3	3	3	4	4	4	4
5	3	3	3	4	4	4	5	5	5
6	3	4	4	4	5	5	5	6	6
7	4	4	5	5	5	6	6	7	7
8	4	5	5	6	6	7	7	7	8
9	5	5	6	6	7	7	8	8	9
10	5	6	6	7	7	8	9	9	10
11	6	6	7	7	8	9	9	10	11
12	6	7	7	8	9	9	10	11	11
13	7	7	8	9	9	10	11	12	12
14	7	8	9	9	10	11	12	12	13
15	8	8	9	10	11	12	12	13	14
16	8	9	10	11	11	12	13	14	15
17	8	9	10	11	12	13	14	15	16
18	9	10	11	12	13	14	15	16	17
19	9	10	11	12	13	15	16	17	18
20	10	11	12	13	14	15	16	17	18

Q=1.00

n	θ_o(%)= 50	55	60	65	70	75	80	85	90
3	2	2	2	2	3	3	3	3	3
4	2	2	3	3	3	4	4	4	4
5	3	3	3	4	4	4	5	5	5
6	3	4	4	4	5	5	5	6	6
7	4	4	4	5	5	6	6	7	7
8	4	5	5	5	6	6	7	7	8
9	5	5	6	6	7	7	8	8	9
10	5	6	6	7	7	8	8	9	10
11	5	6	7	7	8	9	9	10	10
12	6	7	7	8	9	9	10	11	11
13	6	7	8	9	9	10	11	11	12
14	7	8	8	9	10	11	11	12	13
15	7	8	9	10	11	11	12	13	14
16	8	9	10	10	11	12	13	14	15
17	8	9	10	11	12	13	14	15	16
18	9	10	11	12	13	14	15	16	17
19	9	10	11	12	13	14	15	16	17
20	10	11	12	13	14	15	16	17	18

Table of Minimax Mastery Scores in the Binomial Error Model with $p_1=2.0$ and $p_2=2.0$

Q=0.25 (θ_o(%)=)

n	50	55	60	65	70	75	80	85	90
3	2	3	3	3	3	3	4	4	4
4	3	3	4	4	4	4	4	5	5
5	4	4	4	4	5	5	5	5	6
6	4	4	5	5	5	6	6	6	7
7	5	5	5	6	6	7	7	7	8
8	5	6	6	7	7	7	8	8	8
9	6	6	7	7	8	8	9	9	9
10	6	7	7	8	8	9	9	10	10
11	7	7	8	9	9	10	10	11	11
12	7	8	9	9	10	10	11	12	12
13	8	9	9	10	11	11	12	13	13
14	8	9	10	11	11	12	13	13	14
15	9	10	10	11	12	13	14	14	15
16	9	10	11	12	13	14	14	15	16
17	10	11	12	13	13	14	15	16	17
18	10	11	12	13	14	15	16	17	18
19	11	12	13	14	15	16	17	18	19
20	11	13	14	15	16	17	18	19	20

Q=0.50 (θ_o(%)=)

n	50	55	60	65	70	75	80	85	90
3	2	2	3	3	3	3	3	4	4
4	3	3	3	4	4	4	4	4	5
5	3	4	4	4	4	5	5	5	6
6	4	4	5	5	5	6	6	6	6
7	4	5	5	6	6	6	7	7	7
8	5	5	6	6	7	7	8	8	8
9	5	6	6	7	7	8	8	9	9
10	6	6	7	8	8	9	9	10	10
11	6	7	8	8	9	9	10	11	11
12	7	8	8	9	10	10	11	11	12
13	7	8	9	10	10	11	12	12	13
14	8	9	9	10	11	12	12	13	14
15	8	9	10	11	12	12	13	14	15
16	9	10	11	12	12	13	14	15	16
17	9	10	11	12	13	14	15	16	17
18	10	11	12	13	14	15	16	17	17
19	10	11	13	14	15	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=0.75 (θ_o(%)=)

n	50	55	60	65	70	75	80	85	90
3	2	2	3	3	3	3	3	3	4
4	3	3	3	3	4	4	4	4	5
5	3	3	4	4	4	5	5	[illegible]	5
6	4	4	4	5	5	5	6	6	6
7	4	5	5	5	6	6	7	7	7
8	5	5	6	6	6	7	7	8	8
9	5	6	6	7	7	8	8	9	9
10	6	6	7	7	8	8	9	10	10
11	6	7	7	8	9	9	10	10	11
12	7	7	8	9	9	10	11	11	12
13	7	8	9	9	10	11	11	12	13
14	8	8	9	10	11	11	12	13	14
15	8	9	10	11	11	12	13	14	15
16	9	10	10	11	12	13	14	15	16
17	9	10	11	12	13	14	15	16	16
18	10	11	12	13	14	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	11	12	13	14	15	16	17	18	19

Q=1.00 (θ_o(%)=)

n	50	55	60	65	70	75	80	85	90
3	2	2	2	3	3	3	3	3	4
4	2	3	3	3	4	4	4	4	5
5	3	3	4	4	4	5	5	5	5
6	3	4	4	5	5	5	6	6	6
7	4	4	5	5	6	6	6	7	7
8	4	5	5	6	6	7	7	8	8
9	5	6	6	7	7	8	8	9	9
10	5	6	7	7	8	8	9	9	10
11	6	7	7	8	8	9	10	10	11
12	6	7	8	8	9	10	10	11	12
13	7	8	8	9	10	11	11	12	13
14	7	8	9	10	11	11	12	13	14
15	8	9	10	10	11	12	13	14	15
16	8	9	10	11	12	13	14	15	15
17	9	10	11	12	13	14	14	15	16
18	9	10	11	12	13	14	15	16	17
19	10	11	12	13	14	15	16	17	18
20	10	12	13	14	15	16	17	18	19

APPENDIX B

SUBROUTINE MIMAX

This subroutine computes the minimax passing (mastery) score for the binomial error model in mastery testing.

Disclaimer: The computer program hereafter isted has been written with care and tesced extensively under a variety of conditions. The author, however, makes no warranty as to its accuracy and functioning, nor shall the fact of its distribution imply such warranty.

MINIMAX PASSING SCORES

```
      SUBROUTINE MINMAX(N,TA,IA,P1,P2,Q,IZ)
C
C***********************************************************************
C
C     THIS SUBROUTINE COMPUTES THE MINIMAX PASSING (MASTERY) SCORE FOR
C     THE BINOMIAL ERROR MODEL IN MASTERY TESTING.
C
C     INPUT DATA ARE:
C          N ..... NUMBER OF TEST ITEMS
C          TA .... CRITERION LEVEL (THETA ZERO)
C          IA .... NUMBER OF OPTIONS (ALTERNATIVES)  FOR  EACH  MULTIPLE-
C                  CHOICE ITEM.  THIS INFORMATION IS NEEDED IF CORRECTION
C                  FOR GUESSING IS TO BE PERFORMED.  IF NO CORRECTION FOR
C                  GUESSING IS REQUIRED, SET IA = 0.
C          P1 .... EXPONENT FOR  FALSE POSITIVE ERROR LOSS
C          P2 .... EXPONENT FOR  FALSE NEGATIVE ERROR LOSS
C          Q ..... WEIGHTING CONSTANT FOR FALSE NEGATIVE ERROR LOSS
C
C     OUTPUT DATA IS
C          IZ .... MINIMAX PASSING (MASTERY) SCORE
C
C     SUBROUTINES REQUIRED:
C          DRTNI FROM SSP (NEWTON-RALPHSON ITERATION PROCESS)
C          MDBIN FROM IMSL (BINOMIAL PROBABILITY)
C
C***********************************************************************
C
      COMMON NKEEP,IC,R,TT,KODE,IOPT
      DOUBLE PRECISION FL1,FL2,FMAX,FMAX1
C
      WRITE (6.200) N,TA,IA,P1.P2,Q
  200 FORMAT('1'.T4,'NUMBER OF ITEMS .',I4/
     1            T4.'CRITERION LEVEL .',F10.5/
     2            T4,'NUMBER OF OPTIONS',I4/
     3            T4,'P1 ..... .......',F10.5/
     4            T4,'P2 .............',F10.5/
     5            T4,'LOSS RATIO Q ....',F10.5)
      DMAX=AMIN1(1.,Q)
      NKEEP=N
      DD=IA *1./(IA-1)
      IF(IA.EQ.0) DD=1.
      X1=DD**P1
      X2=DD**P2
      TZ=TA
      IF(IA.NE.0) TZ=TA*(1.-1./IA)+1./IA
      IC1=0
      FMAX1=1.D50
C
      DO 10 ID=1,N
C
      IC=ID
      R=P1
      TT=TZ
      IOPT=IA
      CALL LMAX(FL1)
      FL1=FL1*X1
      R=P2
      TT=1.-TZ
      IC=N-ID+1
      IOPT=-1
      CALL LMAX(FL2)
      FL2=FL2*Q
      FL2=FL2*X2
      FMAX=DMAX1(FL1.FL2)
      IF(FMAX.GE.FMAX1) GOTO 10
      IC1=ID
```

```
      FMAX1=FMAX
   10 CONTINUE
C
      AMAX=TZ**P1
      AMAX=AMAX*X1
      B=Q*(1.-TZ)**P2
      B=B*X2
      IX=0
      IF(AMAX.LE.B) GOTO 13
      IX=N+1
      AMAX=B
   13 IZ=IC1
      IF(AMAX.LT.FMAX1) IZ=IX
C
      WRITE(6.220) IZ
  220 FORMAT('0',2X,'MINIMAX PASSING'/3X,'SCORE ...........',I4)
      RETURN
      END
C
      SUBROUTINE LMAX(FL)
      COMMON N,IC,P,TZ,KODE,IA
      DOUBLE PRECISION T,F,DERF,TS,FL,T1,F1,DERF1
      EXTERNAL FCT
      XX=0.
      IF(IA.GT.0)XX=1.0/IA
      EPS=.0001
      IEND=200
      KODE=0
      NN=20
      MM=50
      H=P+IC+(N-1)*TZ
      T1=(H-SQRT(H*H-4*(N +P)*(IC-1)*TZ))/(2*(N+P))
      IF(T1.LE.0.D0) T1=1.D-20
      DD=(TZ-T1)/NN
      TS=T1
      CALL FCT(T1,F1,DERF1)
      DO 5 I=1,NN
      T=T1+I*DD
      CALL FCT(T,F,DERF)
      IF(F*F1.LE.0.0) GOTO 10
      TS=T
      F1=F
    5 CONTINUE
   10 DD=(T-TS)/MM
      CALL FCT(TS,F1,DERF1)
      T1=TS
      DO 15 I=1,MM
      T=T1+I*DD
      CALL FCT(T,F,DERF)
      IF(F1*F.LE.0.) GOTO 20
      TS=T
      F1=F
   15 CONTINUE
   20 TS=(TS+T)/2.0
      DD=T-TS
      IF(DD.LE.EPS) GOTO 25
      KODE=1
      CALL DRTNI(T,F,DERF,FCT.TS,EPS,IEND,IER)
      IF(IER.NE.0) WRITE(6,200) IER
  200 FORMAT('0','ERROR IN THE SSP SUBROUTINE DRTNI',I4)
   25 IF(IA.GT.0.AND.T.LT.XX)T=XX
      S=T
      CALL MDBIN(IC-1,N,S,D,PK,IER)
      IF(IER.NE.0) WRITE(6,210) IER
  210 FORMAT('0','ERROR IN THE IMSL SUBROUTINE MDBIN',I4)
```

```
      FL=(TZ-T)**P*(1.-D)
      RETURN
      END
C
      SUBROUTINE FCT(T,F,DERF)
      COMMON N,IC,P,TZ,KODE
      EXTERNAL BI
      INTEGER BI
      DOUBLE PRECISION T,F,DERF,G
      S=T
      LL=BI(N,IC)
      F=IC*LL*(TZ-T)*T**(IC-1)*(1.D0-T)**(N-IC)
      CALL MDBIN(IC-1,N,S,D,PK,IER)
      F=-P*(1.D0-D)+F
      IF(KODE.EQ.0) RETURN
      DERF=0
      IF(IC.EQ.N) GOTO 10
      G=(1.D0-T)**(N-IC-1)
      IF(IC.EQ.1) GOTO 5
      DERF=(IC-1)*TZ*T**(IC-2)*G
    5 DERF=((N+P)*T**IC-(P+IC+(N-1)*TZ) *T**(IC-1))*G+DERF
      DERF=DERF*IC*LL
      RETURN
   10 DERF=N*T**(N-2)*(-(N+P)*T+(N-1)*TZ)
      RETURN
      END
C
      FUNCTION BI(N,M)
      INTEGER BI
      BI=1
      IF(M*(N-M).EQ.0) RETURN
      MM=N-M
      IF(MM.GT.M)MM=M
      DO 15 J=1,MM
   15 BI=BI*(N-J+1)/J
      END
//LKED.SYSLIB DD
// DD  DSN=ACAD.IMSL.DP.SUBLIB,DISP=SHR
// DD  DSN=ACAD.IMSL.SP.SUBLIB,DISP=SHR
// DD  DSN=SSP.SUBLIB,DISP=SHR
```

2

BAYESIAN AND EMPIRICAL BAYES APPROACHES TO SETTING PASSING SCORES ON MASTERY TESTS

Huynh Huynh
Joseph C. Saunders

University of South Carolina

Presented at the symposium "Psychometric approaches to domain-referenced testing" sponsored jointly by the American Educational Research Association and the National Council on Measurement in Education at their annual meetings in San Francisco, April 8-12, 1979.

ABSTRACT

The Bayesian approach to setting passing scores as proposed by Swaminathan, Hambleton, and Algina is compared with the empirical Bayes approach to the same problem that is derived from Huynh's decision-theoretic framework. Comparisons are based on simulated data which follow an approximate beta-binomial distribution and on real test data sampled from a statewide testing program. It is found that the two procedures lead to setting identical or almost identical passing scores as long as the test score distribution is reasonably symmetric or when the minimum mastery level or criterion level is high. Larger discrepancies tend to occur when this level is low, especially when the distribution of test scores is concentrated at a few extreme scores or when the frequencies are irregular. However, in terms of mastery/nonmastery decisions, the two procedures result in the same classifications in practically all situations. However, the empirical Bayes procedure may be used for tests of any length, while the Bayesian procedure is recommended only for tests of 8 or more items. Additionally, the empirical

This paper has been distributed separately as RM 79-2, April, 1979.

Bayes procedure can be generalized and applied to more complex testing situations with less difficulty than the Bayesian procedure.

1. INTRODUCTION

Among the many decision-theoretic approaches to setting passing scores (or standards) for mastery tests, there are at least two methods which rely on test data collected from a group of examinees. The Bayesian procedure, as presented in Swaminathan, Hambleton, and Algina (1975), assumes that prior knowledge regarding the examinees is exchangeable (Novick, Lewis, & Jackson, 1973) and can be quantified in some appropriate manner. On the other hand, the empirical Bayes approach, as formulated in Huynh (1976a), uses only the true ability distribution of the examinees and makes no assumption regarding prior knowledge about the examinees. Both procedures use test data collected from a group of examinees and establish passing scores for mastery tests by minimizing certain loss functions. The purpose of this paper is to present a comparison of the two sets of standards (passing scores) formulated under a variety of conditions which can be expected to be encountered in mastery testing or in minimum competency testing. The comparison will be made first on the basis of approximate beta-binomial test scores. Further comparisons will be made using the Comprehensive Tests of Basic Skills (CTBS, 1973) data collected in the 1978 South Carolina Statewide Testing Program.

2. AN OVERVIEW OF THE BAYESIAN AND EMPIRICAL BAYES APPROACHES

Overall Framework

The Bayesian framework as presented by Swaminathan et al. and the special empirical Bayes procedure described in Huynh (1976a, p. 70-73) start with a typical four-corner setup used in decision theory. (See Figure I, p. 78, for the basic elements of this setup.) Let θ (π in the notation of Swaminathan et al.) be the true score (or true ability) of an examinee and x be the observed test score as obtained from an n-item test. For the binomial error model adopted in both standard setting approaches, θ is the proportion of items in a real or hypothetical item pool that an examinee answers correctly. Let a person be called a master if that person's true score θ is such that $\theta \geq \theta_o$ and a nonmaster if $\theta < \theta_o$. Here, θ_o is a given constant which defines the lower boundary of the mastery level or the criterion level. Since a person's true score cannot be observed directly, decisions about whether to call the person a master must be based on an observed test score. What remains to be determined is the cutoff score c that will be in some sense optimal.

On the basis of the test score x, a person is called a master if $x \geq c$ and a nonmaster if $x < c$. A correct decision is made whenever either (a) $\theta \geq \theta_o$ <u>and</u> $x \geq c$, or (b) $\theta < \theta_o$ <u>and</u> $x < c$. Otherwise, either a false positive error ($\theta < \theta_o$ <u>and</u> $x \geq c$) or a false negative error ($\theta \geq \theta_o$ <u>and</u> $x < c$) is encountered.

In the case where the loss associated with each error is constant, generality is not diminished if we let the loss incurred by a false positive error be equal to 1 and that associated with a false negative error be equal to Q. Here, Q expresses the <u>ratio</u> of the false negative error loss to the false positive error loss. (In the notation of Swaminathan et al., $Q = \ell_{21}/\ell_{12}$.)

Bayesian Approach

Now let an n-item test be given to m examinees. In the Bayesian procedure as implemented by Swaminathan <u>et al</u>., the prior information regarding the examinees is assumed to be <u>exchangeable</u> (i.e., prior knowledge regarding one examinee can be interchanged with that associated with another examinee without causing any disturbance in the decision problem). The model requires knowledge (prior belief) of the distribution of the variance of true scores for the group. (In point of fact, an arcsine transformation of θ is used.) This prior distribution is taken to be the inverse chi-square distribution with parameter λ and degrees of freedom ν. A recommended choice of ν is 8 (Novick, <u>et al</u>., 1973).

To assess λ, let t be the number of test items which would need to be administered to a typical examinee in order to obtain as much information about that examinee's θ as we already have. Then, $\lambda = 3/(2t+1)$. Wang (1973) has tables to facilitate computation in this procedure. In the setup of the Wang tables, λ/ν is chosen as .01, .02, .03, .04, and .05. These ratios correspond to the t values of 18.25, 8.875, 5.75, 4.1875, and 3.25. Given the prior information as revealed through λ and ν and the test data of m subjects, it is possible via the Wang tables to compute the two expected losses: $\Pr(\theta < \theta_o \mid \text{test data})$ and $Q \cdot \Pr(\theta \geq \theta_o \mid \text{test data})$, at each test score. A Bayesian passing score is then the smallest score at which the first expected loss is smaller than the second one. More details may be found in Swaminathan et al. (1975) and in Novick et al. (1973).

Empirical Bayes Approach

The empirical Bayes solution assumes that the m examinees constitute a random sample from a population for which the true ability θ follows a known distributional form such as the beta density with parameters α and β (Keats & Lord, 1962, page 68). Sample test data are used to obtain the estimates $\hat{\alpha}$ and $\hat{\beta}$, and the results are used to compute the probability of a false positive decision $\Pr(\theta < \theta_o, x \geq c)$ and of a false negative decision $Q \cdot \Pr(\theta \geq \theta_o, x < c)$ at a given cutoff score c. The optimum passing score (henceforth referred to simply as the passing score) will be the value of c at which the average loss, $\Pr(\theta < \theta_o, x \geq c) + Q \cdot \Pr(\theta \geq \theta_o, x < c)$, is the smallest.

The procedure is implemented as follows. Let $\bar{x}$ and s be the mean and standard deviation of the test scores, and let the Kuder-Richardson reliability coefficient be defined as

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{x}(n-\bar{x})}{ns^2}\right).$$

Then

$$\hat{\alpha} = (-1 + 1/\hat{\alpha}_{21})\bar{x}$$

and

$\hat{\beta} = -\hat{\alpha} + n/\hat{\alpha}_{21} - n.$

For test scores with insufficient variability, $\hat{\alpha}_{21}$ may be negative. If this occurs simply replace $\hat{\alpha}_{21}$ by the smallest positive reliability estimate which happens to be available. Let I denote the incomplete beta function as tabulated in Pearson (1934) and implemented via computer programs such as the IBM Scientific Subroutine Package (1971) or the IMSL (1977). Then the passing score is the smallest integer c, at which

$$I(\hat{\alpha}+c, n+\hat{\beta}-c; \theta_o) \leq Q/(1+Q). \tag{1}$$

A normal approximation is available if there is a sufficiently large number of items and if θ_o is not near 0 or 1. Let ξ denote the 100/(1+Q) percentile of the unit normal distribution. Then the test passing score is nearly equal to

$$c = (n+\hat{\alpha}+\hat{\beta}-1)\theta_o + \xi\left[(n+\hat{\alpha}+\hat{\beta}-1)\theta_o(1-\theta_o)\right]^{\frac{1}{2}} - \hat{\alpha} + .5. \tag{2}$$

The data presented in Huynh (1976b) indicate that the passing score computed from Equation (2) does not differ appreciably from the one deduced from Inequation (1) when the test consists of 20 items and when θ_o is within the range from .50 to .80.

3. A COMPARISON OF BAYESIAN AND EMPIRICAL BAYES PASSING SCORES FOR APPROXIMATE BETA-BINOMIAL TEST DATA

The passing score obtained via the empirical Bayes approach, as revealed by Inequation (1), is based on test score data that follow a beta-binomial distribution. It may be of interest to compare the Bayesian approach to setting a passing score with the empirical Bayes approach, using test data which follow closely a beta-binomial form.

Both the present comparison and the one detailed in the next section are based on tests with ten items. In these comparisons, the criterion or minimum mastery level is set at θ_o = .60, .70, and .80. The loss ratio is chosen to be Q = .25, .50, 1.00, and 2.00. (A loss ratio smaller than one indicates that a false positive error is less serious than a false negative error.) To compute a passing score via the Bayesian approach, it is necessary to specify the ratio λ/ν or, equivalently, the quantity t as described in Section 2. It may be recalled that t may be interpreted as the number of "test items" which are believed to be as informative as the prior belief about the examinees. In practical situations involving standard setting, it seems unreasonable to let the prior belief ν carry as much weight as the objective test data. In other words, it is unlikely that t is too close to n. Thus for the comparisons based on 10-item tests reported in this section and in Section 4 as well as the comparisons based on 20-item tests described in Section 5, the t-values are chosen to be 8.875 (λ/ν = .02), 5.75 (λ/ν = .03), 4.1875 (λ/ν = .04), and 3.25 (λ/ν = .05).

The first five test score frequency distributions (labeled A1 through A5 in Table 1) serve as the data base for the comparison of the passing scores computed by the two procedures using test score distributions that are approximately beta-binomial. Each is deliberately chosen (i) to yield an s_g^2 value (variance of the arcsine-square-root transformation of the test scores) conforming as closely as possible to the tabulated s_g^2 values of the Wang tables (so that no interpolation would be necessary) and (ii) to reflect several degrees of skewness and variability thought to be typical of mastery testing situations. (Also in Table 1, and explained below, are distributions of actual test scores from the South Carolina Statewide Testing Program.) It may be noted that in Table 1, the quantity D(%) represents the maximum percent difference between the observed and beta-binomial-fitted cumulative frequencies. A small D-value indicates a good fit.

Table 2 reports the Bayesian passing scores and the corresponding empirical Bayes passing scores *(in italics)* for several combinations of θ_o, Q, and t. The data indicate that for the situations under consideration, the Bayesian and empirical Bayes passing scores are identical, or nearly so, as long as the test score distribution is reasonably symmetrical (Cases A2, A4, and A5). For highly skewed distributions (Cases A1 and A3) the two passing

TABLE 1

Frequency Distributions of Test Scores Used in Comparisons of Passing Scores

Data Set	Source/ Subtest	m*	D(%)†	S.D.	Skew-ness	0	1	2	3	4	5	6	7	8	9	10
	Approximate Beta-Binomial															
A1	Fictitious	40	3.1	1.36	-0.61						1	3	6	8	11	11
A2	Fictitious	80	1.0	1.87	-0.31			1	3	6	10	13	16	15	11	5
A3	Fictitious	40	1.2	1.01	-1.51							1	2	4	10	23
A4	Fictitious	40	1.6	2.01	-0.02		1	3	5	6	7	7	5	4	2	0
A5	Fictitious	40	1.0	2.15	0.12	1	3	5	6	7	6	5	4	2	1	0
	Comprehensive Tests of Basic Skills															
B1	Mathematics concepts and application	20	6.7	1.28	-0.63							2	1	6	4	7
B2	Mathematics computations	20	9.2	1.45	-0.24							3	4	3	4	6
B3	Spelling	20	6.1	1.76	-1.04					2	0	1	2	6	4	5
B4	Social studies	40	6.2	2.11	0.27		1	4	5	9	5	5	6	3	1	1
B5	Language expression	40	8.2	1.86	-0.53			1	1	5	3	4	11	10	3	2
B6	Reading	40	4.1	1.22	-2.12						1	1	2	3	3	30
B7	Science	60	5.6	1.74	-0.22				2	6	10	8	14	8	12	0
B8	Reading vocabulary	60	3.2	1.56	-1.75				1	0	3	1	5	5	16	29
B9	Reading vocabulary	80	2.7	1.68	-1.49				2	1	2	5	6	11	23	30
B10	Spelling	80	2.1	1.50	-1.44				1	0	2	4	7	12	16	38

Columns 0–10: Frequency at score of

*m = total number of scores in the distribution.

†D(%) represents the maximum percent difference between the observed and beta-binomial-fitted cumulative frequencies. All are not significant at the ten percent level of significance.

scores rarely differ by more than one unit when the criterion level θ_o is relatively high (.70 or .80) and when λ/ν is such that t is not too close to n, say when λ/ν is at least .03. Large discrepancies, however, may occur at a low criterion level such as .60 or when t is close to n.

TABLE 2

Bayesian and Empirical Bayes Passing Scores for Five Approximate Beta-Binomial Test Score Distributions

Data Set	θ_o	Bayesian (at λ/ν = .02,.03,.04,.05) and empirical Bayes *(in italics)* at Q = .25	Q = .50	Q = 1.00	Q = 2.00
A1	.60	4, 5, 6, 6, *4*	3, 4, 5, 5, *2*	2, 3, 4, 4, *1*	1, 2, 3, 3, *0*
	.70	7, 8, 8, 8, *6*	6, 7, 7, 7, *5*	5, 5, 6, 6, *4*	4, 4, 5, 5, *3*
	.80	10,10,10,10, *9*	9, 9, 9, 9, *8*	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*
A2	.60	7, 8, 8, 8, *7*	6, 7, 7, 7, *6*	5, 6, 6, 6, *5*	4, 4, 5, 5, *4*
	.70	10,10, 9, 9, *9*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	7, 7, 7, 7, *7*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	9, 9, 9, 9, *9*
A3	.60	1, 3, 4, 4, *3*	1, 2, 3, 3, *2*	0, 1, 2, 2, *1*	0, 1, 1, 2, *0*
	.70	4, 5, 6, 6, *6*	3, 4, 5, 5, *5*	2, 3, 4, 4, *4*	1, 2, 3, 3, *3*
	.80	8, 8, 9, 9, *8*	7, 7, 8, 8, *7*	5, 6, 7, 7, *6*	4, 5, 6, 6, *5*
A4	.60	9, 9, 9, 9, *9*	9, 8, 8, 8, *8*	8, 7, 7, 7, *8*	7, 6, 6, 6, *6*
	.70	10,10,10,10,*10*	10, [illegible],10,10,*10*	10, 9, 9, 9,*10*	9, 9, 8, 8, *9*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*
A5	.60	10,10, [illegible], 9,*10*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	7, 7, 7, 7, *7*
	.70	10,10,10,10,*10*	10,10,10,10,*10*	10,10, 9, 9,*10*	9, 9, 9, 9, *9*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*

4. A COMPARISON OF BAYESIAN AND EMPIRICAL BAYES PASSING SCORES FOR CTBS TEST DATA

This phase of the study is based on a 10% systematic sample of the entire third grade CTBS-Level C data file compiled during the 1978 South Carolina Statewide Testing Program. To obtain the frequency distributions labeled as B1 to B10 (in Tables 1 and 3), the following procedure was used. First, ten 10-item subtests were assembled by random selection of items from each CTBS subtest. Next, for each 10-item subtest, a frequency distribution was constructed for each school district which had at least 20 students in the systematic sample, and the corresponding s_g^2 value was obtained. (The s_g^2 values were distributed as follows: .10 to .50 (32%), .51 to .75 (38%), .76 to 1.00 (20%), and more than 1.00 (10%). Large s_g^2 values tended to associate with subtests dealing with reading comprehension (sentences or paragraphs), language expression, and language mechanics.) Third, among the frequency distributions with s_g^2 values included between .01 and .05, ten were finally selected and altered slightly so that the total number of examinees (m) was exactly 20, 40, 60, or 80.

Table 3 lists the Bayesian and empirical Bayes passing scores under a variety of conditions. As in the previous section, the data

TABLE 3

Bayesian and Empirical Bayes Passing Scores
for Ten CTBS Test Score Distributions

Data Set	θ_o	Bayesian (at λ/ν = .02,.03,.04,.05) and empirical Bayes *(in italics)* at Q = .25	Q = .50	Q = 1.00	Q = 2.00
B1	.60	5, 5, 6, 6, *3*	4, 4, 5, 5, *2*	3, 3, 4, 4, *1*	2, 2, 3, 3, *0*
	.70	7, 7, 8, 8, *6*	6, 6, 7, 7, *5*	5, 5, 6, 6, *4*	4, 4, 5, 5, *3*
	.80	10,10,10,10, *9*	9, 9, 9, 9, *8*	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*
B2	.60	6, 6, 6, 6, *5*	5, 5, 5, 5, *4*	4, 4, 4, 5, *2*	3, 3, 3, 4, *1*
	.70	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*	6, 6, 6, 6, *5*	5, 5, 5, 6, *4*
	.80	10,10,10,10, *9*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	7, 7, 8, 8, *7*
B3	.60	6, 6, 7, 7, *6*	5, 5, 6, 6, *6*	4, 4, 5, 5, *5*	3, 4, 4, 4, *4*
	.70	8, 8, 8, 8, *8*	7, 7, 8, 8, *7*	6, 7, 7, 7, *6*	5, 6, 6, 6, *6*
	.80	10,10,10,10,*10*	9, 9, 9, 9, *9*	9, 9, 9, 9, *8*	8, 8, 8, 8,
B4	.60	9, 9, 9, 9, *9*	9, 8, 8, 8, *8*	8, 8, 7, 7, *7*	7, 7, 6, 6, *7*
	.70	10,10,10,10,*10*	10,10,10,10,*10*	10, 9, 9, 9, *9*	9, 9, 8, 8, *9*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*
B5	.60	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*	6, 6, 6, 6, *5*	4, 5, 5, 5, *4*
	.70	10,10, 9, 9,*10*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	7, 7, 7, 7, *7*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	9, 9, 9, 9, *9*
B6	.60	2, 3, 4, 5, *6*	1, 2, 3, 4, *6*	1, 2, 2, 3, *5*	0, 1, 1, 2, *4*
	.70	5, 5, 6, 7, *8*	3, 4, 5, 6, *7*	2, 3, 4, 5, *6*	2, 2, 3, 4, *6*
	.80	8, 8, 9, 9, *9*	7, 7, 8, 8, *8*	6, 6, 7, 7, *8*	4, 5, 6, 6, *7*
B7	.60	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*	5, 6, 6, 6, *5*	4, 5, 5, 5, *4*
	.70	10,10,10,10, *9*	9, 9, 9, 9, *9*	8, 8, 8 8, *8*	7, 7, 7, 7, *7*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	10,10, 9, 9,*10*
B8	.60	3, 4, 5, 6, *6*	2, 3, 4, 5, *6*	2, 2, 3, 4, *5*	1, 2, 2, 3, *4*
	.70	6, 7, 7, 8, *8*	5, 6, 6, 7, *7*	4, 5, 5, 6, *6*	3, 4, 4, 5, *6*
	.80	9, 9, 9, 9, *9*	8, 8, 9, 9, *8*	7, 7, 8, 8, *8*	6, 6, 7, 7, *7*
B9	.60	4, 5, 5, 6, *6*	3, 4, 4, 5, *6*	2, 3, 3, 4, *5*	1, 2, 3, 3, *4*
	.70	7, 7, 8, 8, *8*	4, 6, 7, 7, *7*	4, 5, 6, 6, *6*	3, 4, 5, 5, *6*
	.80	9,10,10,10, *9*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	6, 7, 7, 7, *7*
B10	.60	3, 4, 5, 6, *6*	2, 3, 4, 5, *5*	1, 2, 3, 4, *5*	1, 1, 2, 3, *4*
	.70	6, 7, 7, 8, *8*	5, 6, 6, 7, *7*	4, 4, 5, 6, *6*	3, 3, 4, 5, *5*
	.80	9, 9, 9, 9, *9*	8, 8, 9, 9, *8*	7, 7, 8, 8, *8*	6, 6, 7, 7, *7*

show that the two sets of passing scores are the same, or nearly so, as long as the test score distribution is reasonably symmetric (see cases B4, B5, and B7). Discrepancies in these situations are rarely larger than one unit. For most other situations, the difference between the two values for a passing score is seldom larger than one unit when the criterion θ_o is .70 or .80 and when λ/ν is at least .03. The same magnitude of difference, one unit, also tends to hold at θ_o = .60 unless the test scores pile up at extreme values (Case B6) or unless the frequencies are fairly irregular (Case B1).

5. ADDITIONAL DATA FOR MODERATELY SKEWED DISTRIBUTIONS

Additional comparisons were made for ten 20-item tests with distributions having skewness ranging from -1.109 to .117 (see Table 4). These tests were assembled in the same way as the 10-item tests described in Section 4. As in the previous sections, the criterion level θ_o was set at .60, .70, and .80, and the loss ratio Q at .25, .50, 1.00, and 2.00. The prior knowledge about the examinees was assumed to be equivalent to a number of items, t, of 8.875 (λ/ν = .02), 5.75 (λ/ν = .03), 4.1875 (λ/ν = .04), and 3.25 (λ/ν = .05). For all the 480 combinations under consideration, the

TABLE 4

Frequency Distribution of Scores on Ten CTBS Subtests Mentioned in Section 5

Subtest	Frequency at score of 5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Reading vocabulary							1	1	5	3	4	7	4	8	3	4
Spelling								1	1	2	3	2	3	8	12	8
Science		1	1	1	3	3	4	3	1	9	4	5	2	1	1	1
Social studies	2	0	2	0	3	1	2	2	6	9	1	4	4	1	3	0
Social studies		1	2	5	3	3	1	6	5	4	2	2	5	0	0	1
Reading vocabulary			2	0	0	2	1	4	4	3	3	4	8	3	4	2
Mathematics concepts and application		1	0	0	1	2	3	2	3	4	0	7	7	2	6	2
Reading vocabulary								1	2	3	2	5	5	6	9	7
Social studies	1	3	1	1	1	0	2	5	3	6	3	5	4	4	1	0
Science	1	1	4	2	2	2	4	2	4	2	3	4	3	5	0	1

absolute value of the discrepancies between the two computed passing scores are distributed as follows: 0 (35%), 1 (37%), 2 (15%), 3 (5%), and 4 or more (8%). Hence in about three-fourths of all situations, the Bayesian and empirical Bayes passing scores do not differ from each other by more than one unit.

6. AGREEMENT OF MASTERY/NONMASTERY DECISIONS

As noted in Section 4, there are situations (such as some cases associated with the A1, B1, and B6 data sets) where the passing scores obtained from the two methods differ appreciably. This may seem disheartening. However, the procedures provide mastery/nonmastery classifications which are in high agreement for most cases under consideration. For Data Set A1 with θ_o = .60 and .70, for example, the combined proportions of students identically classified in either the mastery or nonmastery category by the Bayesian procedure (with λ/μ = .05) and by the empirical Bayes procedure are 88%, 95%, 99%, and 100% for Q = .25, .50, 1.00, and 2.00 respectively. Over the fifteen data sets of Table 1 and with the same values for λ/ν and Q, the proportions of identical classifications reach 94%, 96%, 98, and 97% respectively. As for the data of Table 4, these proportions stand at 98%, 98%, 98%, and 97%.

Though the overall agreement for classifications is high for the data considered in this study, some individual cases may show less agreement than others. These cases include situations such as A2 with θ_o = .60, Q = .25, and λ/ν = .05 where the Bayesian passing score of 8 and the empirical Bayes passing score of 7 are located near the center of the test score distribution. The shift of only one unit in test score in this case actually causes 10 students out of a total of 80 to be classified differently by the two procedures. Visible disagreement between the classifications defined by the Bayesian and empirical Bayes procedures may occur in situations where scores with high frequencies of occurrence are selected as the passing scores. If this is the case, the proportion of students classified in the mastery (or nonmastery) category is not likely to be close to either 0% or 100%. In other situations where most students are declared masters (Data Set A1 with $\theta_o = .60$, $\lambda/\nu = .05$, and $Q = 2.00$) or nonmasters (Data Set A5 with $\theta_o = .70$, $\lambda/\nu = .05$, and $Q = 1.00$), the agreement in classifications is almost perfect.

7. DISCUSSION AND CONCLUSION

The results described in previous sections may be summarized as follows: (i) Bayesian passing scores and those computed via the empirical Bayes procedure are identical or almost identical as long as the test score frequency distribution is reasonably symmetric or when the criterion level θ_o is sufficiently high (.70 or .80); (ii) large discrepancies in passing scores may occur at criterion levels .60 (or below), especially when the test scores pile up at a few extreme values or when the frequency distribution is irregular; (iii) however, mastery/nonmastery decisions derived from the two procedures are most often identical. Overall, the combined proportion of students similarly classified by both procedures is about 97%.

All in all, there is little difference between the Bayesian approach as described by Swaminathan et al. and the Huynh empirical Bayes procedure described here, either in terms of the resulting passing scores or in terms of the mastery/nonmastery categorization.

It should be pointed out that the procedure by Swaminathan et al. relies on a normal arcsine-square-root transformation of the test data and is therefore considered adequate only when the test has at least 8 items. In addition, the scheme requires the evaluation of certain posterior probabilities. This may be done via the MARPRO computer program (mentioned in Wang, 1973) or via the Wang tables. To the chagrin of the writers, many frequency distributions such as those derived from the CTBS test data of the South Carolina Statewide Testing Program have s_g^2 values much larger than the upper bound of .05 allowed in the above-mentioned tables. In addition, the constraint of having at least 8 items seems to be quite severe in many practical situations involving objective-referenced testing. Such tests frequently have 5 or fewer items per objective.

The empirical Bayes approach in its simplest form, as presented in Huynh (1976a), requires that the test scores follow a beta-binomial distribution. There are indications (Keats & Lord, 1962; Duncan, 1974; Huynh & Saunders, 1979; also see Table 1) that the model adequately fits many test score distributions. Moreover, it is known (Subkoviak, 1978; Huynh & Saunders, 197ʳ) that the model is useful in the estimation of the reliability of mastery classification based on one test administration. In addition, using the empirical Bayes approach, passing scores may be computed for tests of <u>any</u> length and can be approximated quickly via Equation (2).

It may be noted that the Bayesian and empirical Bayes procedures discussed in this paper deal with the setting of passing scores for a particular test. Both procedures assume the availability of a minimum mastery or criterion level θ_o and the availability of other information such as Q, the ratio of the loss incurred by a false positive decision to that incurred by a false negative one. In the context of testing for instructional purposes, θ_o may be based on the judgment of a curriculum specialist or a knowledgeable teacher and Q may be assessed via the time losses encountered by a misdecision (Huynh, 1976a). The issue is much more involved for end-of-program certification, such as high school graduation (minimum competency) testing programs legislated in several states. The reader is referred to Jaeger (1976) and Shepard (1976) for insight regarding some of these issues.

The empirical Bayes approach with the availability of a predetermined criterion level, however, is <u>only</u> the simplest form of the general framework of mastery evaluation as approached by Huynh (1976a). The essential component of this model is an <u>external task</u> (real or hypothetical) that examinees are supposed to perform once they are granted mastery of the objectives or content upon which a test is based. Such an external task may be identified in the context of instruction, especially when instructional units are sequenced in some logical order. If this requirement is fulfilled, the specification of θ_o is no longer necessary. Some suggestions for solutions along this line have been presented elsewhere (Huynh, 1976a, p. 73-75; Huynh, 1977; Huynh & Perney, 1979). To the knowledge of the writers, the Bayesian approach as presented by Swaminathan et al. has not been generalized to situations other than those involving constant losses and when a criterion level is available. Although such a generalization may be made, the numerical analysis would be more involved than can be expected from the empirical Bayes approach.

As indicated previously, both standard setting procedures studied in this paper are based on group data and therefore are appropriate to the extent that minimization of loss is considered for the entire group of examinees. This may be the case for minimum competency testing where resources for remedial instruction are limited. Procedures relating to standard setting in the absence of group data are available (see, for example, Huynh, 1978).

In conclusion, the empirical Bayes approach yields mastery/nonmastery decisions identical in most cases to those based on the Bayesian approach. In addition, the former approach is simpler in terms of computations, is applicable to any test length, and has been generalized to more complex testing situations.

BIBLIOGRAPHY

Comprehensive Tests of Basic Skills, Level C (1973). Monterey, California: CTB/McGraw-Hill.

Duncan, G. T. (1974). An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association 69, 50-57.

Huynh, H. (1976a). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

Huynh, H. (1976b). On mastery scores and efficiency of criterion-referenced tests when losses are partially known. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 19-23.

Huynh, H. (1977). Two simple classes of mastery scores based on the beta-binomial model. Psychometrika 42, 601-608.

Huynh, H. (1978). A nonrandomized minimax solution for mastery scores in the binomial error model. Research Memorandum 78-2, Publication Series in Mastery Testing. University of South Carolina College of Education.

Huynh, H. & Perney, J. C. (1979). Determination of mastery scores when instructional units are linearly related. Educational and Psychological Measurement 39, 317-325.

Huynh, H. & Saunders, J. C. (1979). Accuracy of two procedures for estimating reliability of mastery tests. Research Memorandum 79-1, Publication Series in Mastery Testing. University of South Carolina College of Education. Also presented at the annual conference of the Eastern Education Research Association, Kiawah Island, South Carolina, February 22-24, 1979.

IBM Application Program, System/360 (1971). Scientific subroutines package (360-CM-03X) Version III, Programmer's manual. White Plains, New York: IBM Corporation Technical Publication Department.

IMSL Library 1 (1977). Houston: International Mathematical and Statistical Libraries.

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. Florida Journal of Educational Research 18, 22-27.

Keats, J. A. & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika 27, 59-72.

Novick, M. R., Lewis, C. & Jackson, P. H. (1973). The estimation of proportions in m groups. Psychometrika 38, 19-45.

Pearson, K. (1934). Tables of the Incomplete Beta Function. Cambridge: University Press.

Shepard, L. A. (1976). Setting standards and living with them. Florida Journal of Education Research 18, 23-32.

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

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

Wang, M. M. (1973). Tables of constants for the posterior marginal estimates of proportions in m groups. ACT Technical Bulletin No. 14. Iowa City, Iowa: The American College Testing Program.

FIGURE I

Four Categories of Decisions
Based on Observed Test Scores

Observed Score (X) / True Score (θ)	Observed Nonmastery	Observed Mastery
	c	
True Mastery	Nonmastery-Mastery (false negative decision)	Mastery-Mastery (accurate decision)
θ_c		
True Nonmastery	Nonmastery-Nonmastery (accurate decision)	Mastery-Nonmastery (false positive decision)

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not neccessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance of Anthony J. Nitko is gratefully acknowledged.

3

A CLASS OF PASSING SCORES BASED ON THE BIVARIATE NORMAL MODEL

Huynh Huynh

University of South Carolina

Proceedings of the 1979 meeting of the American Statistical Association (Social Statistics Section).

ABSTRACT

This study touches some aspects of the determination of passing (cutoff, mastery) scores on the basis of the bivariate normal test model. The loss ratio associated with classification errors is assumed to be constant, and the referral success function is assumed to belong to the normal ogive family. Alternately the model also provides a fairly simple way to assess the loss consequences associated with each passing score. Such information is deemed useful to the test user who may wish to examine these consequences before making a final choice of passing score.

1. INTRODUCTION

A general framework for setting passing (cutoff, mastery) scores in binary classification (or mastery testing) has been provided recently (Huynh, 1976). Applications of the procedure to test data distributed as the beta-binomial model have also been presented (Huynh, 1976, 1977). The framework assumes that the true

This paper has been distributed separately as RM 79-4, April, 1979.

ability of a population of subjects may be described by a random variable θ with probability density function $p(\theta)$. If only _one_ subject is involved, then $p(\theta)$ describes the prior information regarding this subject's ability. A test is administered to the subject and the resulting test score is denoted as x. The test score is then compared to a passing (or cutoff) score equal to a constant c. If x is equal to or greater than c, the subject passes (or is declared a "master"). If x is less than c, the subject does not pass (or is declared a "nonmaster"). The problem is to determine a value of c which is optimum in some sense.

The model, as proposed, postulates the availability of a referral task which the subjects are expected to be able to perform if they are classified as having mastered the competencies underlying the test scores. Performance on the referral task is categorized as success or failure. The probability of a successful performance on the task by a subject with true ability θ is defined via a nondecreasing function $s(\theta)$, the referral task. Each referral task corresponds to a unique function $s(\theta)$. Conversely, from a purely mathematical point of view, any nondecreasing function $s(\theta)$ may be conceptualized as a referral task.

The referral task, thus, may be real or hypothetical. For example, if an integer addition unit is to be followed by lessons on integer multiplication, then performance on a multiplication test may serve as a referral task for a test tapping the ability to add integers. Othe illustrations of _real_ referral tasks may also be found in situations where the sequence of instructional units forms a linear hierarchy. In a number of situations, a referral task can be _conceptualized_. For example, in minimum competency testing programs legislated in several states, a consensus on what constitutes a minimum level of performance for mastery may serve as a basis for a referral task. To be specific, let us agree that in order to qualify for mastery, an examinee must have a true ability of at least θ_o. Then the nondecreasing function $s(\theta)$ which takes the value of 0 if $\theta < \theta_o$ and the value of 1 for $\theta \geq \theta_o$ _mathematically_ defines the referral task for this case. The special 0-1 form for $s(\theta)$ has been considered by a number of writers including Hambleton and Novick (1973).

Now let $C_f(\theta)$ represent the opportunity loss incurred by granting mastery status to a subject who will eventually fail in performing the referral task (a false positive error). Likewise, let $C_s(\theta)$ be the loss associated with the denial of mastery to a subject whose performance on the task would be deemed successful (a false negative error). Under these conditions, reasonable choice for an optimum passing score would be the score c_o at which the average loss across all subjects in the population (or the Bayes risk in the case of only one subject) is smallest. Details regarding the computation of c_o may be found in Huynh (1976).

When test scores may be assumed to follow a beta-binomial model and when the referral success function $s(\theta)$ is of the 0-1, linear, or cubic form, closed-form solutions exist for c_o (Huynh, 1976, 1977). As is well known, the binomial error model is appropriate when each examinee is given an independent sample of items (Lord and Novick, 1968, chap. 23). There are indications that several test score distributions might fit the beta-binomial framework even if examinees in each distribution respond to the same set of items.

There are models other than the beta-binomial framework which could be used to represent test data. For example, many frequency distributions obtained from standardized tests are known to follow closely a normal distribution. Models using a bivariate normal distribution for the true score θ and the observed score x are not uncommon in educational measurement and Bayesian statistical literature. Moreover, as an implication of the Central Limit Theorem. the beta-binomial distribution will resemble a bivariate normal distribution when the number of test items is sufficiently large.

The purpose of this paper is to provide the computation for the optimum passing score (mastery score) for the bivariate normal test score model with constant losses and 0-1 or normal ogive $s(\theta)$.

Since normal test scores form a continuous scale, the optimum passing score c_o satisfies the equation

$$\int_\Omega \{(C_s(\theta) + C_f(\theta))s(\theta) - C_f(\theta)\}p(\theta|c_o)d\theta = 0. \tag{1}$$

In the above expression, Ω represents the sample space of θ. For the sake of completeness, a procedure will also be proposed for approximating the referral success function $s(\theta)$.

2. PASSING SCORE COMPUTATION FOR THE BIVARIATE NORMAL MODEL WITH CONSTANT LOSSES AND NORMAL OGIVE REFERRAL SUCCESS

Without any loss of generality, let $C_f(\theta) = 1$ and $C_s(\theta) = Q$. Here Q expresses the ratio of the loss incurred by a false negative error to that associated with a false positive error. Now let the referral success be defined as

$$s(\theta) = F_N\left(\frac{\theta-\theta_o}{\sigma}\right) \tag{2}$$

where θ_o and σ are two constants and $F_N(\cdot)$ denotes the cumulative distribution function of a unit normal random variable. In addition, let x be in its standardized form (with zero mean and unit variance). With ρ as the test reliability, the mean and variance of θ are respectively 0 and ρ, and the correlation between x and θ is $\sqrt{\rho}$.

It is now assumed that the vector (θ,x) follows a bivariate normal distribution. It may be then verified that the conditional density $p(\theta|c_o)$ is given as a normal density with mean ρc_o and variance $\rho(1-\rho)$. Equation (1) now becomes

$$\int_{-\infty}^{+\infty}\left[(Q+1)F_N\left(\frac{\theta-\theta_o}{\rho}\right) - 1\right]p(\theta|c_o)d\theta = 0$$

or

$$\int_{-\infty}^{+\infty} F_N\left(\frac{\theta-\theta_o}{\sigma}\right)p(\theta|c_o) = \frac{1}{1+Q}. \tag{3}$$

The integral in Equation (3) may be written as

$$A = \frac{1}{2\pi\sigma\sqrt{\rho-\rho^2}}\int_{-\infty}^{+\infty}\left\{\int_{-\infty}^{\theta}\exp\left[-\frac{(t-\theta_o)^2}{2\sigma^2}\right]dt\right\}\exp\left[-\frac{(\theta-\rho c_o)^2}{2(\rho-\rho^2)}\right]d\theta .$$

This integral may be viewed as the probability of the joint event $\{-\infty < \theta < \infty,\ t < \theta\}$ associated with two independent random variables t and θ. The random variable t has mean θ_o and variance σ^2; the second random variable θ has mean ρc_o and variance $\rho - \rho^2$. Now the difference $t - \theta$ follows a normal distribution with mean $\theta_o - \rho c_o$ and variance $\rho - \rho^2 + \sigma^2$. Since the mentioned joint event is equivalent to the condition $t - \theta < 0$, it follows that the value of A is

$F_N\left[(\rho c_o - \theta_o)/(\rho - \rho^2 + \sigma^2)^{\frac{1}{2}}\right]$. Let ξ be the 100/(1+Q) percentile of the unit normal distribution, e.g. $F_N(\xi) = 1/(1+Q)$. Then c_o is given as

$$c_o = \frac{\theta_o + \xi\sqrt{\rho-\rho^2+\sigma^2}}{\rho}. \tag{4}$$

If the test scores have mean μ_x and a standard deviation σ_x, then the test cutoff score is given as $C_o = \mu_x + c_o \cdot \sigma_x$.

The following remarks may be made about Equation (4). First by letting $\sigma^2 = 0$, the normal ogive $s(\theta)$ will degenerate to a 0-1 form with the jump occurring at θ_o. Thus if true nonmastery status is defined by $\theta < \theta_o$ and true mastery by $\theta \geq \theta_o$, then the cutoff score is $c_o = \theta_o/\rho + \xi\sqrt{1-\rho}$. Next, when misdecisions are weighted equally in terms of losses (i.e., when Q = 1), c_o and θ_o relate to each other via the equation $\theta_o = \rho c_o$. This expression is reminiscent of the Kelly formula which defines the regression of true score on test score (Lord and Novick, 1968, p. 65). Finally, when the relationship between the ability θ and the referral task is fuzzy, i.e., when σ^2 is large, the cutoff score c_o will shoot sharply above the "central value" θ_o/ρ if Q < 1 and will locate appreciably below this central value if Q > 1.

It may be noted that the unstandardized passing score C_o may be written as

$$C_o = \mu_x + \frac{\theta_o\sigma_x}{\rho} + \xi\sqrt{(1-\rho)\sigma_x^2 + \sigma^2\sigma_x^2/\rho^2}.$$

Let σ_e^2 be the squared standard error of measurement. Then $\sigma_e^2 = (1-\rho)\sigma_x^2$ and

$$C_o = \mu_x + \frac{\theta_o \sigma_x}{\rho} + \xi\sqrt{\sigma_e^2 + \sigma^2\sigma_x^2/\rho^2}. \tag{5}$$

Numerical Example 1

Let $\mu_x = 100$, $\sigma_x = 15$, $\rho = .90$, $\theta_o = 1$, $\sigma = .5$, and $Q = .5$. Then $\xi = .432$, and $c_o = 1.391$. The raw (unstandardized) cutoff score is found to be $C_o = 120.86$.

. ESTIMATION PROCEDURE FOR NORMAL OGIVE REFERRAL SUCCESS

Now let $g(x,1)$ be the proportion of subjects who have a test score of x and succeed in performing the referral task. Then from Equation (13) of Huynh (1976, p. 74), it may be seen that

$$g(x,1) = \int_{-\infty}^{+\infty} h(x,\theta)s(\theta)d\theta$$

where $h(x,\theta)$ is the bivariate normal density of x and θ. It follows that

$$g(x,1) = f_N(x)\int_{-\infty}^{+\infty} F_N\left(\frac{\theta-\theta_o}{\sigma}\right)p(\theta|x)d\theta$$

where $f_N(.)$ is the unit normal density. Hence from the derivations in the middle part of the previous section,

$$\frac{g(x,1)}{f_N(x)} = F_N\left(\frac{\rho x-\rho_o}{\sqrt{\rho-\rho^2+\sigma^2}}\right).$$

The ratio $p(x) = g(x,1)/f_N(x)$ represents the (conditional) proportion of students who, at the test score x, will succeed in performing the referral task. Now let

$$\alpha = \rho/(\rho-\rho^2+\sigma^2)^{1/2} \tag{6}$$

and

$$\beta = -\theta_o/(\rho-\rho^2+\sigma^2)^{1/2},$$

then

$$p(x) = F_N(\alpha x+\beta).$$

If $\xi(x)$ denotes the 100p(x) percentile of the unit normal distribution, then

$$\xi(x) = \alpha x + \beta \ . \tag{7}$$

Now let $\hat{p}(x)$, $\hat{\xi}(x)$ be the observed values of p(x) and $\xi(x)$. Let w(x) be a suitably chosen weight function at the score x. Then via the least squares technique, the estimates for α and β are given as

$$\hat{\alpha} = s(\hat{\xi})\cdot r(x,\hat{\xi}) \tag{8}$$

and

$$\hat{\beta} = \ddot{\xi}_{.} \ , \tag{9}$$

where $\hat{\xi}_{.}$ and $s(\hat{\xi})$ are the mean and standard deviation of the $\hat{\xi}(x)$ values, and $r(x,\hat{\xi})$ is the correlation between the x and $\hat{\xi}(x)$ values, each pair being weighted by w(x). The computation, of course, is carried out only over the x values at which the sample values $\hat{p}(x)$ are available. The reader may recall that the test scores x are in standardized form.

It may be noted that p(x) is an increasing function of x. Hence it seems reasonable to require that the sample value p(x) be a nondecreasing function of x. This may be done by applying the Pool-Adjacent-Violator algorithm (Barlow, Bartholomew, Bremner, and Brunk, 1972, p. 13) using w(x) as the weight function. In addition, since all p(x) values must be included strictly between 0 and 1, the algorithm must be conducted such that the adjusted values $\hat{p}(x)$ conform to this requirement. (See Table 1 for an illustration.)

As in any least square procedure, the weight function w(x) may be chosen in a variety of ways. It appears to the author that the number of subjects at each test score might serve as a reasonable choice for this function.

Once the estimates $\hat{\alpha}$ and $\hat{\beta}$ have been determined, the estimates for θ_o and σ^2 may be derived from Equations (5) and (6). These are

$$\hat{\theta}_o = -\rho\hat{\beta}/\hat{\alpha} \tag{10}$$

and

$$\hat{\sigma}^2 = \rho^2/\hat{\alpha}^2 - \rho + \rho^2 . \tag{11}$$

In the case where Equation (11) yields a negative value, a reasonable choice for σ^2 would be 0.

Numerical Example 2

Table 1 presents the basic data for this example. The test reliability is taken to be $\rho = .90$. The summary data are $\hat{\xi}_. =$ $-.2280$, $s(\hat{\xi}) = .8668$, and $r(x,\hat{\xi}) = .9723$. It follows that $\hat{\alpha} =$ $.8427$ and $\hat{\beta} = -.2280$, hence $\hat{\theta}_o = .244$ and $\hat{\sigma}^2 = 1.050$.

4. ASSESSING THE CONSEQUENCES OF SELECTING A MASTERY SCORE

Section 2 provides the computation of mastery scores when the loss ratio Q is known. In a number of applications, however, the test user may not be willing to specify in advance a value for Q. Instead the user may wish to look at the consequences associated with each cutoff score before making a final choice. Such a practice is not uncommon in real testing situations. Both Jaeger (1976) and Shepard (1976) have advocated an iterative process for setting cutoff scores in testing programs such as high school graduation or minimum competency testing.

As in Section 2, let $F_N(.)$ denote the cumulative distribution function of the unit normal variable. Given the loss ratio Q, the mastery score c_o is given by the equation

$$F_N\left\{(\rho c_o-\theta_o)/(\rho-\rho^2+\sigma^2)^{\frac{1}{2}}\right\} = \frac{1}{1+Q}.$$

Alternately the selection of c_o as the cutoff score would indicate that the weights (or losses) accorded to a false negative error and to a false positive error are in the ratio of Q to 1 where

$$Q = 1/F_N\left\{(\rho c_o-\theta_o)/(\rho-\rho^2+\sigma^2)^{\frac{1}{2}}\right\} - 1.$$

Q will degenerate to 0 when c_o goes to $+\infty$ (i.e., when all subjects are denied mastery) and to ∞ when c_o goes to $-\infty$ (i.e., when mastery is granted regardless of test score).

5. SUMMARY AND CONCLUSION

This study touches some aspects of the determination of passing scores on the basis of the bivariate normal test model. The

TABLE 1

Basic Data for Numerical Example 2

	Raw Test Score									
	1	2	3	4	5	6	7	8	9	10
Frequency of examinees	1	4	10	21	16	23	21	16	8	5
Frequency of referral-successful examinees	0	0	1	3	4	8	15	10	7	5
Unadjusted $\hat{p}(x)$	0	0	.100	.143	.250	.348	.714	.625	.875	1
Pool-Adjacent-Violator-Adjusted $\hat{p}(x)$	.067	.067	.067	.143	.250	.348	.676	.676	.923	.923
$\hat{\xi}(x)$	-1.450	-1.450	-1.450	-1.067	-.675	-.391	.457	.457	1.426	1.426

loss ratio associated with classification errors is assumed to be constant, and the referral success function is assumed to be in the normal ogive family. Alternately, the model also provides a fairly simple way to assess the loss consequences associated with each mastery score. Such information is deemed useful to the test user who may wish to examine these consequences before making a final choice of cutoff score.

It should be mentioned that the paper deals with group test data for a population of examinees. Thus the various results would be useful to the extent that loss consequences are considered jointly for the entire population. A procedure for setting passing scores on tests in the absence of group data is discussed elsewhere (Huynh, 1978; also in press).

BIBLIOGRAPHY

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. (1972). Statistical inference under order restrictions. New York: John Wiley & Sons.

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

Huynh, H. (1976). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

Huynh, H. (1977). Two simple classes of mastery scores based on the beta-binomial model. Psychometrika 42, 601-608.

Huynh, H. (1978). A nonrandomized minimax solution for mastery scores in the binomial error model. Research Memorandum 78-2, Publication Series in Mastery Testing. University of South Carolina College of Education.

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. Florida Journal of Educational Research 18, 22-27.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company.

Shepard, L. A. (1976). Setting standards and living with them. Florida Journal of Educational Research 18, 23-32.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not neccessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance of Joseph C. Saunders and Anthony J. Nitko are gratefully acknowledged.

4

AN EMPIRICAL BAYES APPROACH TO DECISIONS BASED ON MULTIVARIATE TEST DATA

Huynh Huynh

University of South Carolina

Presented at the annual meeting of the Psychometric Society, Iowa City, Iowa, May 28-30, 1980.

ABSTRACT

A general framework for making mastery/nonmastery decisions based on multivariate test data is described in this study. Over all, mastery is granted (or denied) if the posterior expected loss associated with such action is smaller than the one incurred by the denial (or grant) of mastery. An explicit form for the cutting contour which separates mastery and nonmastery states in the test score space is given for multivariate test scores which follow a normal distribution with a constant loss ratio. For the case involving multiple cutting scores in the true ability space, the test score cutting contour will resemble the boundary defined by multiple test cutting scores when the test reliabilities are reasonably close to unity. For tests with low reliabilities, decisions may very well be based simply on a suitably chosen composite score

1. INTRODUCTION

Application of mental measurement to selection or certification problems often involves the use of more than one test score. For

---

This paper has been distributed separately as RM 79-7, December, 1979.

example, the selection of students for an advanced program in some subject area may be based on several traits (variables), such as prior achievement, aptitude, interest, etc. Ideally, selection should be based on the subject's true measures on these traits; in reality, however, decisions are typically based on observed test scores which are contaminated with errors of measurement. Thus, misclassifications are bound to occur, and rules for decisions based on test data are typically formulated in such a way as to minimize the risks incurred by misclassification.

Decision problems based on one variable have been considered at length in the literature. Statistical issues involved in establishing a single cutoff (cutting, passing, or mastery) score are described in detail in a number of sources including Swaminathan, Hambleton, and Algina (1975); Huynh (1976, 1977, 1979, 1980); Wilcox (1976); and van der Linden and Mellenbergh (1977). Huynh (1979, 1980) also provides an explicit relationship among test cutting score, losses incurred by misclassification, and errors of measurement. In general, within the minimax or empirical Bayes decision framework, it is found that errors in measurement will reduce the test cutting score when a false negative error is more serious than a false positive error. Conversely, the test cutting score will increase when a false negative error is less serious than a false positive error.

The effect of errors of measurement in selection situations involving multiple true cutting scores has been considered by Lord (1962). The selection framework used involves the regression line expressing the amount of "desirability" assigned to different examinees as a function of the observed test scores. Using the multivariate normal distribution to describe the true and observed scores, Lord was able to plot the contour line in the observed test score plane which separates the subjects deemed acceptable (masters) from those judged as unacceptable (nonmasters). Lord's paper, however, does not appear to come naturally from decision theory as formulated by Wald (1950) or as prescribed in Ferguson (1967).

The purpose of this paper is twofold. First it will describe a general empirical Bayes solution to the "plotting" of a cutting contour in selection situations involving multiple test scores. Second, it will explore the influence of the loss ratio on the cutting contour and will reexamine the distortion caused by errors of measurement (Lord, 1962), using an empirical Bayes decision-theoretic framework. Examples based on the multivariate normal distribution with constant losses for misdecisions are provided to illuminate various points or procedures put forward in the paper.

2. EMPIRICAL BAYES APPROACH TO CUTTING CONTOUR

Now let the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$ denote the true scores (measures) of an individual subject on k traits (or selection variables). Let Ω represent the region in the true score space where a subject must be located in order to qualify for the true state of mastery. Thus a subject is defined as a true master if $\theta \in \Omega$. Let Ω^c be the complement of Ω. Then a subject is declared a true nonmaster when $\theta \in \Omega^c$.

Now let the vector $x = (x_1, x_2, \ldots, x_k)'$ represent the observed test scores of the subject. On the basis of x and other prior information regarding θ, a decision may be made concerning the subject: either to grant mastery (action a_1) or to deny mastery (action a_2). When $\theta \in \Omega$, the best course of action is a_1, and no loss will be encountered. Similarly, action a_2 is best when $\theta \in \Omega^c$. For other situations, classification errors occur. To be specific, the choice of action a_2 when $\theta \in \Omega$ constitutes a false negative error, whereas the selection of a_1 when $\theta \in \Omega^c$ produces a false positive error.

Let $C_s(\theta)$ be the loss associated with a false negative error and $C_f(\theta)$ be the loss encountered by a false positive error. Let $p(\theta|x)$ be the posterior probability density of θ given that the test score vector x has been observed. Given x, the posterior expected loss encountered in taking action a_1 is given by the integral $R(a_1|x) = \int_{\Omega^c} C_f(\theta)p(\theta|x)d\theta$.

Similarly, the posterior loss associated with the choice of action a_2 is $R(a_2|x) = \int_\Omega C_s(\theta)p(\theta|x)d\theta$.

It follows from Bayes (or empirical Bayes) decision theory as expressed, for example, in Ferguson (1967) that, in the test score space generated by the test score vector x, the cutting contour S separating the two actions a_1 (granting mastery) and a_2 (denying mastery) is defined by the equality $R(a_1|x) = R(a_2|x)$. In other words, the line (or surface) S consists of all points x at which

$$\int_\Omega C_s(\theta)p(\theta|x)d\theta = \int_{\Omega^c} C_f(\theta)p(\theta|x)d\theta. \quad (1)$$

The following section explores in detail the implications of Equation (1) for the case involving constant losses and multiple true cutting scores.

3. CUTTING CONTOUR FOR CONSTANT LOSSES AND MULTIPLE TRUE CUTTING SCORES

Let losses be constant and expressed as $C_f(\theta) = 1$ and $C_s(\theta) = Q$ in the region where they do not vanish. In other words, Q is the ratio of the false negative loss to the false positive loss. In addition, let Ω be the "upper right" corner defined by the true cutting scores $\theta_1^*, \theta_2^*, \ldots, \theta_k^*$. In other words,

$$\Omega = \{\theta; \theta_1^* \le \theta_1, \theta_2^* \le \theta_2, \ldots, \theta_k^* \le \theta_k\}.$$

With constant losses Equation (1) may now be written as

$$Q\int_\Omega p(\theta|x)d\theta = \int_{\Omega^c} p(\theta|x)d\theta.$$

Since $\Omega \cup \Omega^c$ spans the entire space for θ, it follows that

$$\int_\Omega p(\theta|x)d\theta + \int_{\Omega^c} p(\theta|x)d\theta = 1.$$

With this relationship, Equation (1) becomes

$$\int_\Omega p(\theta|x)d\theta = \frac{1}{1+Q}, \quad (2)$$

which may be written, using the given multiple true cutting scores, as

$$Pr(\theta_1^* \le \theta_1, \theta_2^* \le \theta_2, \ldots, \theta_k^* \le \theta_k|x) = \frac{1}{1+Q} \quad (3)$$

The line consisting of the points of coordinate x which satisfy Equation (2) or (3) defines the boundary between granting and denying mastery in the test score space. This boundary line will be referred to as a cutting contour.

4. CUTTING CONTOUR IN MULTIVARIATE NORMAL TEST SCORES

For illustrative purposes, let it be assumed that the true score vector θ for a population of subjects follows a multivariate normal distribution with mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_k)'$ and with covariance matrix $\Sigma_\theta = (\sigma_{ij})$. In the terminology of empirical Bayes statistics, this statement is equivalent to the requirement that the prior distribution of the true score vector θ be the same for all subjects in the population under study. This common prior distribution may be estimated from historical test score data or by procedures which are consistent with classical measurement theory and practice.

The difference vector $e = x - \theta$ represents the errors of measurement. It will be assumed that the k components of e are normally and independently distributed, each with a mean of zero and a variance of ε_{ii}, $i = 1,2,\ldots,k$, free of θ. In addition, it will be assumed that the two vectors e and θ are stochastically independent. To simplify the notation, let Σ_e be the diagonal matrix with elements ε_{ii}.

It follows from classical measurement theory and from known properties of multivariate normal distributions that the joint distribution of x and θ is multivariate normal with a mean vector of μ for both x and θ and with a covariance matrix defined as

$$\left(\begin{array}{c|c} \Sigma_x & \Sigma_\theta \\ \hline \Sigma_\theta & \Sigma_\theta \end{array}\right)$$

where $\Sigma_x = \Sigma_\theta + \Sigma_\varepsilon$. Hence the posterior distribution of θ given the test score x is multivariate normal with mean vector $\xi(x) = (\xi_1, \xi_2, \ldots, \xi_k)' = \mu + (x-\mu)'\Sigma_\theta\Sigma_x^{-1}$ and with covariance matrix $\Lambda = (\lambda_{ij}) = \Sigma_\theta - \Sigma_\theta\Sigma_x^{-1}\Sigma_\theta$. The vector $\xi(x)$ is a function of the test score vector x. On the other hand, the matrix Λ is free of x.

Now let us consider the standardized variables $y_1, y_2, \ldots, y_k$ where

$$y_i = (\theta_i - \xi_i(x))/\sqrt{\lambda_{ii}}, \quad i = 1,2,\ldots,k.$$

Each of these variables has zero mean and unit variance. Let Γ be the correlation matrix associated with Λ (i.e., Γ is the covariance matrix of the y_i variables). In addition, let

$$y_i^* = \left(\theta_i^* - \xi_i(x)\right)/\sqrt{\lambda_{ii}},\ i = 1,2,\ldots,k. \tag{4}$$

Then the cutting contour separating the two actions a_1 and a_2 in the test score space is defined by the equality

$$\Pr(y_1^* \leq y_1, y_2^* \leq y_2, \ldots, y_k^* \leq y_k) = \frac{1}{1+Q} \tag{5}$$

where the random vector $y = (y_1, y_2, \ldots, y_k)'$ follows a multivariate normal distribution with zero means, unit variances, and correlation matrix Γ free of x.

Consider now the set γ consisting of the points with coordinates $(y_1^*, y_2^*, \ldots, y_k^*)$ which satisfy Equation (5). Tihansky (1970) refers to this set as an equidistributional contour and provides ways to construct contours of this type for bivariate normal distributions. The contour γ depends only on Γ which does not involve the observed test score vector x. Once it has been constructed, the cutting contour C in the test score space may be plotted via the system of linear equations represented by

$$\mu + (x-\mu)'\Sigma_\theta \Sigma_x^{-1} = \xi, \tag{6}$$

where

$$\xi_i = \theta_i^* - y_i^* \sqrt{\lambda_{ii}},\ i = 1,2,\ldots,k.$$

Where computer facilities are available, equidistributional contours may be drawn via the Newton-Raphson iteration process for nonlinear equations. For example, let $(y_1, y_2)'$ follow a standardized bivariate normal distribution with correlation ρ. Let α be any number between 0 and 1, and u be such that $\Pr(u \leq y_1) < \alpha$. We will search for the value v at which $G(v) = 0$, where

$$G(v) = \Pr(y_1 \geq u, y_2 \geq v) - \alpha,$$

$$= \Pr(y_1 \leq -u, y_2 \leq -v) - \alpha. \tag{7}$$

The derivative of G(v) with respect to v is given as

$$G'(v) = -(2\pi)^{-\frac{1}{2}} \exp\left(-\frac{v^2}{2}\right) P(y_1 \leq -u \mid y_2 = -v). \tag{8}$$

Here the conditional distribution of y_1 given $y_2 = -v$ is a normal distribution with a mean of $-\rho v$ and a standard deviation of $(1-\rho^2)^{1/2}$. Hence

$$G'(v) = -(2\pi)^{-1/2}\exp\left(-\frac{v^2}{2}\right)P\left(Z \leq \frac{-u+\rho v}{(1-\rho^2)^{1/2}}\right) \tag{9}$$

where Z is the standardized normal variable. The values of G(v) and G'(v) may be obtained via computer programs such as MDBNOR (IMSL, 1977) and the Fortran IV library function ERFC. Both G(v) and G'(v) are needed in the Newton-Raphson iteration process. This procedure has been found to converge when u is not too close to the upper bound u_o at which $P(u_o \leq y_1) = \alpha$. (It may be noted that the bivariate equidistributional contour has two asymptotes defined as $u = u_o$ and $v = u_o$. Thus small variations in a u value near u_o will tend to associate with substantial changes in the v values; because of this, the iteration process may fail. However, since $P(y_1 \geq u, y_2 \geq v) = P(y_1 \geq v, y_2 \geq u)$, the contour is symmetric with respect to the first diagonal in the (u,v)-plane. Thus it is necessary to iterate the v value for each u sufficiently smaller than the upper bound u_o, and then to resort to symmetry to complete the drawing of the contour.)

The drawing of an equidistributional contour for any k-variate normal distribution may be accomplished in the same way via the Newton-Raphson iteration process previously described. The details are straightforward and therefore are not presented here. Multivariate normal probabilities of the form $P(y_1^* \leq y_1, y_2^* \leq y_2, \ldots, y_k^* \leq y_k)$ may be evaluated via computer programs such as the one described in Milton (1972).

It may be noted that the contour γ does not depend on the two vectors θ^* and μ. In addition, in the transformation from γ to C as defined by (6), these two vectors act only to indicate the new location of the transformed curve. It follows that the shape of the cutting contour C does not depend on either the vector μ or the vector θ^*.

5. AN ILLUSTRATION OF CUTTING CONTOUR

Consider now a selection based on two variables defined by the true scores θ_1 and θ_2, and by the observed test data x_1 and x_2. It will be assumed, as in Lord (1962), that both x_1 and x_2 are in their standardized form and have a common reliability coefficient of .90. In addition, let the correlation between x_1 and x_2 be .60. It follows that the matrices Σ_x and Σ_θ are defined as

$$\Sigma_x = \begin{pmatrix} 1.00 & .60 \\ .60 & 1.00 \end{pmatrix}$$

and

$$\Sigma_\theta = \begin{pmatrix} .90 & .60 \\ .60 & .90 \end{pmatrix}.$$

With

$$\Sigma_x^{-1} = \frac{1}{.64}\begin{pmatrix} 1.00 & -.60 \\ -.60 & 1.00 \end{pmatrix},$$

it follows that

$$\Sigma_\theta \Sigma_x^{-1} = \frac{1}{.64}\begin{pmatrix} .54 & .06 \\ .06 & .54 \end{pmatrix} = \begin{pmatrix} .84375 & .09375 \\ .09375 & .84375 \end{pmatrix}$$

and

$$\Lambda = \begin{pmatrix} .90 & .60 \\ .60 & .90 \end{pmatrix} - \frac{1}{.64}\begin{pmatrix} .522 & .378 \\ .378 & .522 \end{pmatrix} = \begin{pmatrix} .084375 & .009375 \\ .009375 & .084375 \end{pmatrix}.$$

Thus the posterior distribution of $\theta = (\theta_1, \theta_2)'$ given the test data $x = (x_1, x_2)'$ is bivariate normal with mean vector $\xi(x) = (\xi_1, \xi_2)'$ where $\xi_1 = .84375x_1 + .09375x_2$ and $\xi_2 = .09375x_1 + .84375x_2$. The posterior standard deviations are $(.084375)^{1/2} = .29047$ for both θ_1 and θ_2, and the posterior correlation between θ_1 and θ_2 is $.009375/.084375 = .11111$.

It may then be deduced from the equations represented by (4) that

$$y_1^* = \left(\theta_1^* - (.84375x_1 + .09375x_2)\right)/.29047$$

and

$$y_2^* = \left(\theta_2^* - (.09375x_1 + .84375x_2)\right)/.29047.$$

To draw the (x_1,x_2) contour line, let us suppose that $\theta_1^* = \theta_2^* = 0$. The two equations represented by (6) can be written as

$$.84375x_1 + .09375x_2 = -.29047y_1^*$$

$$.09375x_1 + .84375x_2 = -.29047y_2^*$$

or equivalently

$$x_1 = -.34857y_1^* + .03873y_2^*$$

$$x_2 = .03873y_1^* - .34857y_2^* .$$

In the above equations, the point at coordinate (y_1^*,y_2^*) belongs to the equidistributional contour line defined by $P(y_1^* \leq y_1, y_2^* \leq y_2) = 1/(1+Q)$, where $(y_1,y_2)'$ follows a standardized bivariate normal distribution with correlation .11111. It may be recalled that Q is the ratio of the false negative error loss to the false positive error loss.

For purposes of illustration, the steps previously described were implemented in drawing the cutting contours associated with the loss ratios Q = 1/3, 1, and 4. These contours are depicted in Figure I.

6. EFFECT OF LOSS RATIO ON CUTTING CONTOUR

In Figure I, the upper right region bounded by each cutting contour consists of the test score points at which mastery is granted. It may be observed that the mastery region expands as the loss ratio Q increases. This conclusion is to be expected. If the consequences due to a false negative error become more serious (i.e., Q increases), then the classification (or selection) procedure should be so designed as to reduce the probability of this error. Thus the size of the nonmastery set must be reduced, and as a consequence, it becomes more likely that mastery will be granted.

In general, let the set $A^*(Q_1)$ consist of all points $y^* = (y_1^*,y_2^*,\ldots,y_k^*)$ for which

$$P(y_1^* \leq y_1, y_2^* \leq y_2, \ldots, y_k^* \leq y_k) > 1/(1+Q_1) \qquad (10)$$

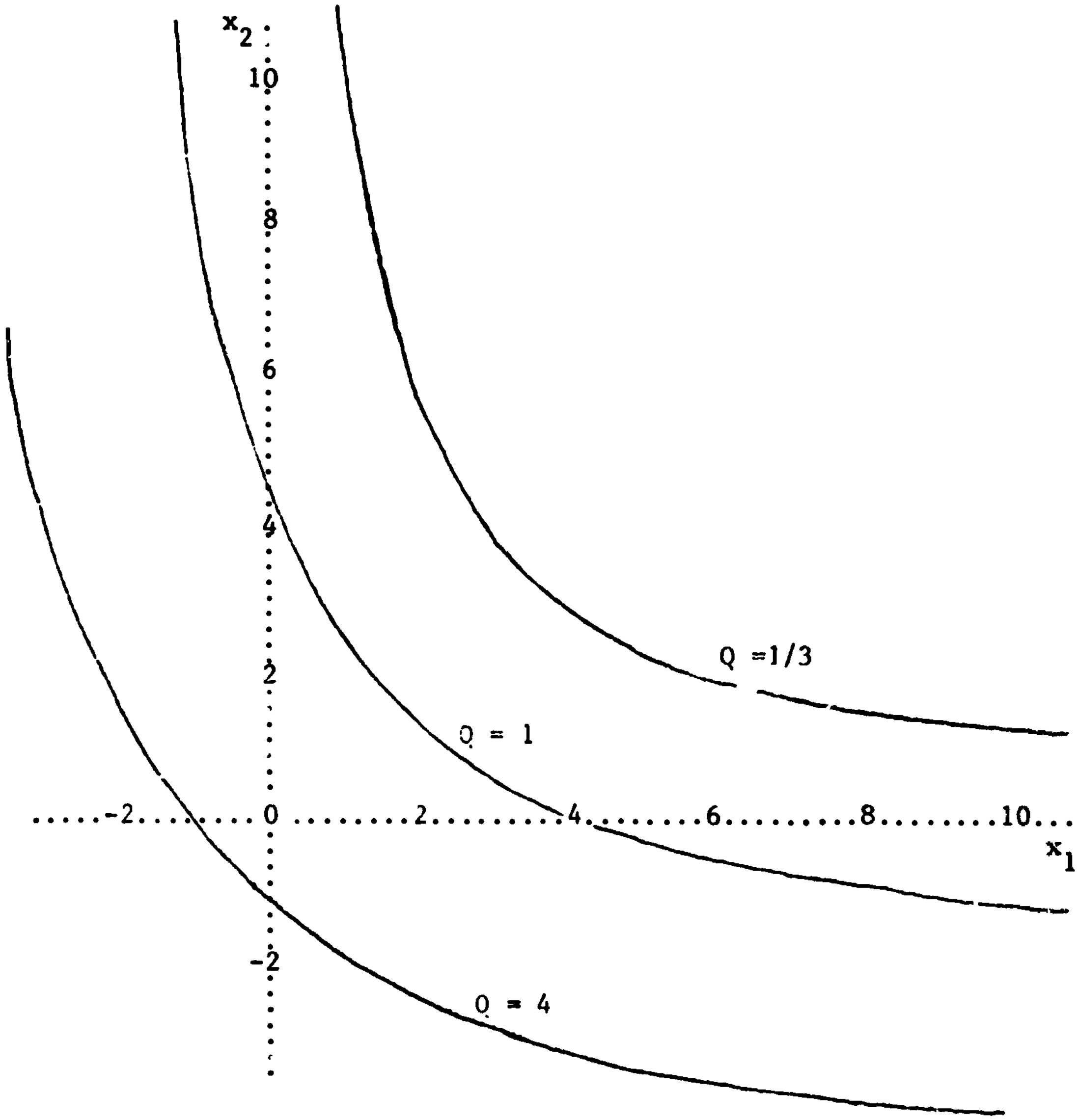

FIGURE I
Multivariate Cutting Contour
for three Q Values

and let $A(Q_1)$ be the corresponding region in the test score space. It may be verified that in $A(Q_1)$ the expected posterior losses associated with the two actions a_1 (granting mastery) and a_2 (denying mastery) satisfy the inequality $R(a_1|x) < R(a_2|x)$. Thus the set $A(Q_1)$ consists of test score points at which the subject is declared a master. Now let Q_2 be a second loss ratio such that $Q_1 < Q_2$. This is equivalent to $1/(1+Q_1) > 1/(1+Q_2)$. Let $A(Q_2)$ have the same meaning as above. Then any test score points which belong to $A(Q_1)$ must also belong to $A(Q_2)$. In other words, the inequality $Q_1 < Q_2$ implies that $A(Q_1) \subset A(Q_2)$. Thus, as the loss ratio Q increases, the mastery region in the test score space will expand. By the same line of reasoning, when Q decreases, the mastery region will be reduced in size.

7. EFFECT OF ERRORS OF MEASUREMENT ON CUTTING CONTOUR

To illustrate the effect of errors of measurement on the cutting contour in the test score space, let it be assumed as in the previous section that the test scores x_1 and x_2 are in their standardized forms and have a correlation of .60. In addition, let it be assumed that x_1 and x_2 are equally reliable with common reliability coefficient ρ, and that $\theta_1^* = \theta_2^* = 0$.

It follows from the equations represented by (6) that

$$\begin{aligned} 1.25(\rho-.36)x_1 + .75(1-\rho)x_2 &= (\rho^2-1.36\rho+.36)^{\frac{1}{2}}y_1^* \\ .75(1-\rho)x_1 + 1.25(\rho-.36)x_2 &= (\rho^2-1.36\rho+.36)^{\frac{1}{2}}y_2^* . \end{aligned} \tag{11}$$

In these expressions, the point (y_1^*, y_2^*) belongs to an appropriate equidistributional contour associated with the standardized bivariate normal distribution with correlation $\delta = .6(1-\rho)/(\rho-.36)$.

It may be deduced from the positive semidefiniteness of the covariance matrix of (θ_1, θ_2) that the common reliability ρ must be between .60 and 1.00. As a function of ρ, the posterior correlation δ is a decreasing function, assuming the value of 1.00 when $\rho = .60$ and having the limit of 0 when ρ tends to 1.00.

When ρ approaches the upper limit 1.00, the posterior distribution of (θ_1,θ_2) will degenerate at the point (x_1,x_2). (It may be noted that when $\rho = 1$, the posterior covariance matrix Λ as defined in Section 4, i.e., $\Sigma_\theta - \Sigma_\theta\Sigma_x^{-1}\Sigma_\theta$, will vanish.) Given the test score vector $x = (x_1,x_2)'$, formally, the posterior expected loss for taking action a_1, $R(a_1|x)$, is equal to 0 when $x \in \Omega$ and 1 when $x \in \Omega^c$. Similarly, $R(a_2|x)$ is equal to Q when $x \in \Omega$ and 0 when $x \in \Omega^c$. Thus, mastery is granted when $x_1 \geq 0$ and $x_2 \geq 0$. When either $x_1 < 0$ or $x_2 < 0$ (or both), mastery is denied. In summary, when ρ tends to unity, the cutting contour line in the test score space will approach the cutting contour line defined in the true score space.

Consider now the other limiting situation where ρ tends to .60 and δ goes to 1.00. The entire bivariate probability of $(x_1,x_2)'$ is now concentrated on the diagonal $x_1 = x_2$. Let y_o be the point at which $P(y_o \leq y_1) = 1/(1+Q)$ where y_1, as previously defined, is a standardized normal variable. The equidistributional contour line is now comprised of the two half lines defined by (i) $y_1^* = y_o$ and $y_2^* \leq y_o$, and (ii) $y_2^* = y_o$ and $y_1^* \leq y_o$. Both half lines start at the point (y_o,y_o) and extend to $-\infty$, one vertically and the other horizontally.

The equations (11) now become

$$x_1 + x_2 = -.32y_1^*$$

$$x_1 + x_2 = -.32y_2^* .$$

It follows that the cutting contour in the observed test score space is the straight line defined by the equation $x_1 + x_2 = -.32y_o$. The decision regarding granting or denying mastery in this case is actually based on the composite score $x_1 + x_2$ although <u>separate</u> cutting scores have been set in the true score space!

For purposes of illustration, cutting contours are drawn for the reliability coefficients of ρ = .95, .80, and .65, and with the loss ratio Q = 1. The contours are shown in Figure II.

MULTIVARIATE CUTTING CONTOUR

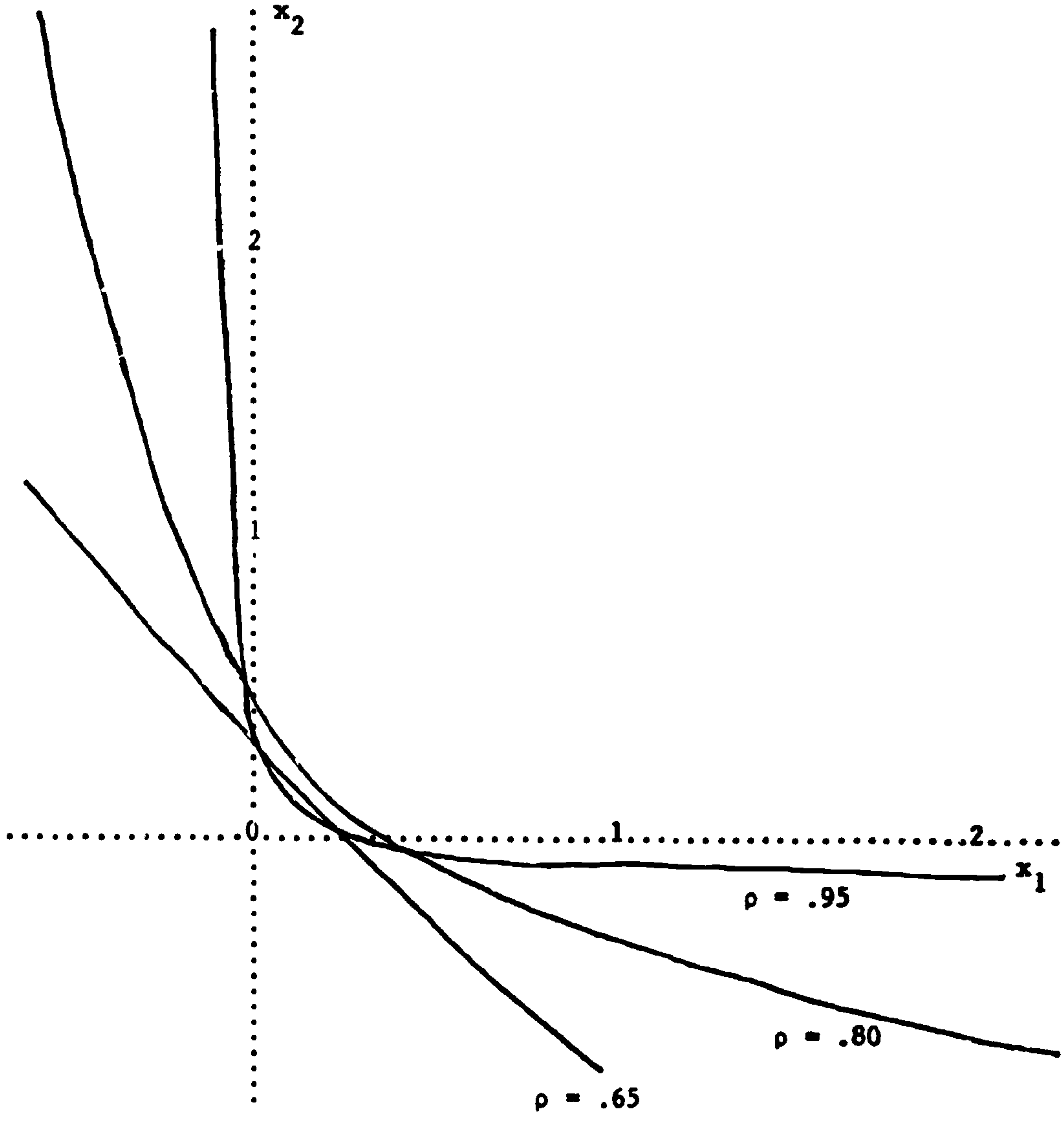

FIGURE I1

Multivariate Cutting Contours
for Three ρ Values

8. SUMMARY

A general framework for making mastery/nonmastery decisions based on multivariate test data is described in this study. Over all, mastery is granted (or denied) if the posterior expected loss associated with such action is smaller than the one incurred by the denial (or grant) of mastery. An explicit form for the cutting contour which separates mastery and nonmastery states in the test score space is given for multivariate test scores which follow a normal distribution with a constant loss ratio.

For the case involving multiple cutting scores in the true ability space, the test score cutting contour will resemble the boundary defined by multiple test cutting scores when the test reliabilities are reasonably close to unity. For tests with low reliabilities, decisions may very well be based simply on a suitably chosen composite score.

BIBLIOGRAPHY

Ferguson, T. S. (1976). Mathematical statistics: A decision-theoretic approach. New York: Academic Press.

Huynh, H. (1976). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

Huynh, H. (1977). Two simple classes of mastery scores based on the beta-binomial model. Psychometrika 42, 601-608.

Huynh, H. (1979). A class of passing scores based on the bivariate normal model. Proceedings of the American Statistical Association (Social Statistics Section).

Huynh, H. (1980). A nonrandomized minimax solution for passing scores in the binomial error model. Psychometrika 45, 167-182.

IMSL Library 1 (1977). Houston: International Mathematical and Statistical Libraries.

Lord, F. M. (1962). Cutting scores and errors of measurement. Psychometrika 27, 19-30.

Milton, R. C. (1972). Computer evaluation of the multivariate normal integral. Technometrics 14, 881-889.

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

Tihansky, D. P. (1970). Property of the bivariate normal cumulative distribution, Report P4400. Santa Monica, California: Rand Corporation.

van der Linden, W. J. & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. Applied Psychological Measurement 1, 593-599.

Wald, A. (1950). Statistical decision functions. New York: John Wiley & sons.

Wilcox, R. (1976). A note on the length and passing score of a mastery test. Journal of Educational Statistics 1, 359-364.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, principal investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no official endorsement should be inferred. The assistance of Joseph C. Saunders is gratefully acknowledged.

5

A COMPARISON OF TWO WAYS OF SETTING PASSING SCORES BASED ON THE NEDELSKY PROCEDURE

Joseph C. Saunders
Joseph P. Ryan
Huynh Huynh

University of South Carolina

Presented at the annual conference of the Eastern Educational Research Association, Norfolk, Virginia, March 5-8, 1980. Applied Psychological Measurement *(in press).*

ABSTRACT

Two versions of the Nedelsky procedure for setting minimum passing scores are compared. Two groups of judges, one using each version, set passing scores for a classroom test. Comparisons of the resulting sets of passing scores are made on the basis of (1) the raw distributions of passing scores, (2) the consistency of pass-fail decisions between the two versions, and (3) the consistency of pass-fail decisions between each version and the passing score established by the test designer. The two versions of the procedure are found to produce essentially equivalent results. In addition, a significant relationship is observed between the passing score set by a judge and that judge's level of achievement in the content area of the test.

This paper has been distributed separately as RM 80-1, March, 1980.

1. INTRODUCTION

Passing scores are needed in a broad variety of situations, including (a) entrance examinations, (b) tests for advancement of students from unit to unit in individually prescribed instructional programs, (c) minimum competency testing, and (d) certification or licensing examinations. Though writers such as Glass (1978) charge that passing scores for minimum competency testing are usually selected arbitrarily and frequently used unwisely, others (Hambleton, 1978; Shepard, 1976) have documented the need for cutoff scores in such areas as objectives-based programs and individualized instruction. This paper presumes the practical necessity of passing scores and explores ways in which they can be established more objectively.

Procedures for Setting Passing Scores

Various procedures for setting passing scores or "standards" have been developed (see Meskauskas, 1976). Most can be placed into one of three broad categories: (a) comparisons with the performance of others, (b) considerations of the consequences of misclassification, and (c) examinations of item content. Standard-setting procedures in the first two categories generally require actual student response data or assume a theoretical, statistical distribution of such data; content-based methods use judgements of content experts. Content-based methods frequently are used with tests when student performance data are not available.

Methods for determining passing scores by analyzing test content require a judge or group of judges to estimate the probable score of a hypothetical examinee responding at the level of minimum acceptable performance. Three of the best-known content-based procedures are those proposed by Angoff (1971), Ebel (1972), and Nedelsky (1954). In using the Angoff method, each judge estimates, the probability that the "minimally acceptable person" would respond correctly to each item; the passing score is determined by summing the estimated item probabilities (Angoff, 1971; Zieky and Livingston, 1977). In the Ebel procedure, judges sort items into categories of "relevance" and "difficulty." Each judge then estimates the proportion of correct answers in each category expected of a "minimally qualified" examinee. The passing score is the weighted sum of these proportions, with the weight for each category being the number of items it contains (Ebel, 1972). The Nedelsky method is restricted to multiple-choice tests. Every response option is considered by each judge, who decides which options could be rejected as incorrect by an examinee performing at the minimum passing level. The probability that someone at this level would respond correctly to the item is taken to be the reciprocal of the number of remaining options (i.e., one divided by the number of options that the minimally performing examinee should not be able to reject). The passing score is the sum of these reciprocals for all items. (In the original formulation, Nedelsky (1954) offers further refinements, such as, estimating the standard deviation of the chance distribution of scores and using it in conjunction with setting the passing score. These refinements are not considered in this paper.) In all cases, the passing score can be expressed as a fraction or percentage of the total number of items.

Comparisons of the Application of the Methods

The methods discussed above, though operationally quite different, have strong logical similarities. It might seem that they could be expected to produce equivalent passing scores. Research reported in the literature indicates that this equivalence is not always observed. In a study comparing the Ebel and Nedelsky procedures, Andrew and Hecht (1976) found that the two standard-setting methods produced significantly different passing scores. Perhaps an even more important consideration was that 45 percent of the examinees being tested were classified differently by the two passing scores (Glass, 1978). In research utilizing the Nedelsky and Angoff procedures, Brennan and Lockwood (1979) also reported a substantial difference in the resulting passing scores.

When several judges are used, the variation among judges' individual passing scores also can become an issue. A certain degree of variation might be expected. It is usually suggested that the different passing scores be reconciled either by averaging the scores or by requiring judges to reach a consensus passing score. Andrew and Hecht (1976) found that passing scores obtained by consensus and by averaging did not differ significantly. In at least one reported case, however, the amount of variation among passing scores set by a group of judges using the Nedelsky procedure was substantial, and the procedure was rejected as unfeasible (Meskauskas and Webster, 1975). The averaging process treats the variation in passing scores as random or "error" variation. It might be, however, that differences in passing scores are related systematically to characteristics of the judges. If passing scores are to be useful, they should not depend too much on the characteristics of a particular judge or group of judges. Such characteristics, once identified, possibly could be controlled to prevent them from exerting an undue influence on the standard-setting process. One characteristic which intuitively might be expected to show such a relationship is the judge's own level of achievement in the relevant area.

Focus of this Paper

This paper deals only with the Nedelsky procedure. Two versions of the procedure appear to be in use. In the first version, judges must classify response options into two categories: (a) those which should be rejected as incorrect by the minimally performing examinee, and (b) those which should not. In the alternative version, a third category, "undecided," also is used when the judge is unable to classify the response option as one that either should or should not be rejected. Decisions between the two versions seem to be based on the preferences of the judges, rather than any theoretical consideration (e.g., Paiva and Vu, 1979; Smilansky and Guerin, 1976). Nedelsky (1954) discussed the use of the alternative procedure; he apparently felt the two versions were equivalent.

The purpose of this paper is twofold. First, a comparison is made between the two versions of the Nedelsky procedure. Second, the relationship between the achievement levels of judges and the passing scores they set will be assessed.

2. METHOD

Subjects

In order to compare the two versions of the Nedelsky procedure, subjects acting as judges were divided into two groups. Group A used the two-category version of the procedure to set passing scores on an achievement test, while Group B used the three-category version. The results were compared using the distributions of passing scores, as well as the consistency of decisions based upon the scores. Also, to determine the relationship between judges' achievement and passing score, the correlation between measures of the two variables was calculated.

Data for the study were obtained from students in an introductory course in educational research and measurement. The course was conducted via videotape at a number of regional campuses of a large state university. All subjects were graduate students; many were experienced teachers.

Instrument

The instrument for which passing scores were set, and by which judges' achievement levels were determined, was the course midterm examination, a 40-item, four-option, multiple-choice test, constructed by the course instructor (the second author). The test covered such topics as the nature of the research process, observation and measurement, sampling, and item analysis. The exam has been revised over several years to reach a high degree of content validity, and in its most recent administration showed an internal consistency (KR20) reliability index of .82. Thus, scores on the test are taken to be valid and reliable measures of achievement.

Treatment Groups

All students enrolled in the course wrote the midterm examination as a regular course requirement. The exams routinely were graded and returned to the students for discussion in class. The students then were asked to participate in an exercise involving the use of the Nedelsky procedure to determine a passing score for the test. While participation in the exercise was voluntary, more than 95% of the students chose to participate. Of the 148 students agreeing to participate, 30 were deleted from the study due to failure to follow instructions, missing identification codes, or missing achievement data, leaving 118 students as the sample used in the experiment. Subjects were assigned randomly to groups, stratified by course section to control for possible differences among regional campuses. Then they were given copies of the test, along with detailed instructions on the Nedelsky procedure. Instructions for the two groups differed only with respect to the version of the procedure used.

Definition of Minimum Competence

Minimum acceptable performance was defined for the subjects as the lowest level of performance on the test for which a grade of "B" would be awarded. This level was chosen as appropriate, since one of the requirements of the subjects' degree programs is that a "B" average be maintained. For each incorrect response option on the test, the subjects were instructed to respond to the question "Should the student performing at the minimum acceptable level (as defined above) be able to reject this option as incorrect?" Spaces were provided for that purpose beside each option. For the two-category version (Group A) of the procedure, the possible responses were "yes" and "no." The three-category version (Group B) also allowed "undecided" as a possible choice. In order to minimize any possible confounding effect produced by the subjects' knowledge of previously existing course standards, the subjects were not required to calculate their resulting Nedelsky passing scores; this was done by the authors. Each subject responded individually; no attempt was made to determine consensus passing scores.

Comparison Procedures

The frequency distributions of passing scores produced by the two groups were compared using the Kolmogorov-Smirnov two-sample test, a broad test sensitive to any difference in the two distributions. The distributions of passing scores are given in Table 1. All passing scores were rounded upward to the nearest whole number, that is, the number of correctly-answered items necessary for an examinee to be classified as passing. Decision consistency was assessed via comparisons of the proportions of students writing the exam who were classified similarly by the two versions. Both the mean and median passing scores for each group were used in the comparisons. The results are shown in Table 2. Also, decisions based on the groups' passing scores were compared with those based on the standard established by the course instructor, as shown in Table 3. Finally, to assess the relationship between judges' achievement and passing score, the Pearson product-moment correlation coefficient was determined for the subjects' examination grades and their Nedelsky passing scores. For this calculation, the two groups were combined.

TABLE 1

Distributions of Passing Scores from Two Versions of the Nedelsky Procedure

Passing Score	Frequency Group A	Frequency Group B	Passing Score	Frequency Group A	Frequency Group B
13	0	1	26	2	4
14	1	0	27	1	0
15	0	0	28	5	2
16	2	1	29	4	4
17	0	1	30	0	1
18	1	0	31	3	5
19	0	0	32	5	3
20	3	1	33	2	3
21	1	0	34	6	10
22	1	0	35	6	5
23	2	2	36	3	2
24	2	4	37	3	5
25	1	2	38	5	3

	N	MEAN	MEDIAN	S.D.
Group A	59	29.88	31.17	6.38
Group B	59	30.51	31.37	5.79

Kolmogorov-Smirnov D = .170 ($p = .36$)

3. RESULTS

The overall passing score distributions for the two groups, displayed in Table 1, showed no significant difference ($p = .36$). As can be seen in Table 2, the two forms also produced highly consistent classification decisions. If the mean passing score for each group is used as a standard, only 7 of 185 students taking the test would have been classified differently, a percentage of agreement of 96%. The exact median passing scores from the two groups are 31.17 and 31.37, respectively. Rounding upward, both these values become 32. Thus, use of the median passing score produced the surprising result of complete agreement in classification.

The fact that the two versions produce passing scores yielding consistent decisions does not, in itself, mean that the scores are useful in practice. But further comparisons of decisions based on the Nedelsky passing scores with those based on standards previously established by the course instructor (32 correct answers for a grade of B) also show a high degree of agreement (Table 3). Using the group mean passing score as the standard, 11 of 185 students were classified differently by Group A (the two-category version) and the course instructor's pre-set standard (percentage agreement = 94%). For Group B (the three-category versions), this percentage was 98% (7 students classified differently). The group medians, rounded up to 32, coincide exactly with the course instructor's standard. Here again, use of the group medians produced complete agreement.

As was noted previously, subjects in both groups were combined to consider the relationship between judges' achievement and passing score. Such a relationship, if it exists, might be expected to hold across methods; in any event, the demonstrated equivalence of the two forms suggests the reasonableness of combining the two groups. The linear correlation between achievement and passing score for the subjects of the study was .30 ($p = .001$). Thus achievement in the subject matter area accounted for 9% of the observed variation in passing scores.

4. DISCUSSION

From the results of this study, the two- and three-category versions of the Nedelsky procedure yield equivalent results. The finding holds both in terms of the empirical distributions of passing scores, and of consistency in classification decisions. Additionally, there was a close correspondence both in distributions of passing scores and in classification decisions between passing scores set by the subjects and the pre-set standard established by the course instructor.

TABLE 2

Decision Consistency of Passing Scores
Two Versions of the Nedelsky Procedure

Case I: Using the mean of several judges.

		Group A fail	Group A pass	
Group B	fail	44	7	51
	pass	0	134	134
		44	141	185

Proportion of consistent decisions $= \frac{134 + 44}{185} = .96$

Case II: Using the median of several judges.

		Group A fail	Group A pass	
Group B	fail	55	0	55
	pass	0	134	134
		55	134	185

Proportion of consistent decisions $= \frac{134 + 55}{185} = 1.00$

While either the mean or median of several judges' passing scores could be used to set the final passing standards the median, rather than the mean, might be more appropriate. The median's resistance to the influence of extreme scores would seem to reduce some of the effect of variability in passing scores from a group of judges.

Some variation was observed in the scores from both groups of judges. The slightly smaller standard deviation of passing scores from Group B, using the three-category version of the procedure, might be a point in favor of the use of that version. The significant positive correlation between judges' achievement and passing score indicates that at least a small portion of the observed variation in passing scores was related systematically to a characteristic of the judges. Other relevant characteristics might be identified which also relate systematically to judges' passing scores. Knowledge of these characteristics and their relationship to passing scores could lead to their elimination, control, or utilization in the standard-setting process. This knowledge would make the setting of passing scores on the basis of expert judgement a more objective process.

In conclusion, this study has shown that the two versions of the Nedelsky procedure considered here produce equivalent passing scores. Also, it was shown that the passing scores set by different judges were related positively to the judges' own achievement. It should be noted that the study involved the setting of passing scores for a single test, using as judges students who took the test but who were not responsible for constructing it. Further, such judges are not likely to have the broad knowledge of other students, of how such tested content fits into the total curriculum, and of the subject-matter itself which, say, faculty members might have. It is an open question whether faculty members would tend to show the same pattern of consistenc, in applying the two Nedelsky methods. Thus the observed results must be seen as suggestive rather than conclusive. However, given the

TABLE 3

Decision Consistency of Course Instructor's Standard with Passing Scores from Two Versions of the Nedelsky Procedure

Case I: Using the mean of several judges.

Instructor's Pre-set Standard		Group A fail	Group A pass		Group B fail	Group B pass	
	fail	44	11	55	51	4	55
	pass	0	130	130	0	130	130
		44	141	185	51	134	185

Proportions of consistent decisions =

$$\frac{130 + 44}{185} = .94 \qquad \frac{130 + 51}{185} = .98$$

Case II: Using the median of several judges.

Instructor's Pre-set Standard		Group A fail	Group A pass		Group B fail	Group B pass	
	fail	55	0	55	55	0	55
	pass	0	130	130	0	130	130
		55	130	185	55	130	185

Proportions of consistent decisions =

$$\frac{130 + 55}{185} = 1.00 \qquad \frac{130 + 55}{185} = 1.00$$

results of this study, a choice between the two versions justifiably could be made on practical grounds, such as the preference of the judges.

BIBLIOGRAPHY

Andrew, B. J. & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement 45, 4-9.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education.

Brennan, R. L. & Lockwood, R. E. (1979). A comparison of two cutting scores using generalizability theory. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement 15, 237-261.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced test in instructional settings. Journal of Educational Measurement 15, 277-290.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research 46, 133-158.

Meskauskas, J. A. & Webster, G. W. (1975). The American Board of Internal Medicine recertification examination process and results. Annals of Internal Medicine 82, 577-581.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement 14, 3-19.

Paiva, R. E. A. & Vu, N. V. (1979). Standards for acceptable level of performance in an objectives-based medical curriculum: A case study. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Shepard, L. A. (1976). Setting standards and living with them. Florida Journal of Educational Research 18, 28-32.

Smilansky, J. & Guerin, R. O. (1976). Minimal acceptable performance levels for criterion-referenced multiple-choice examinations and their validation. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Zieky, M. J. & Livingston, S. A. (1977). Manual for setting standards on the basic skills assessment tests. Princeton, N. J.: Educational Testing Service.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no official endorsement should be inferred. The comments of Anthony J. Nitko and Elizabeth M. Haran are gratefully acknowledged.

120

PART TWO

Assessing the Consequences of Selecting a Passing Score

6

BUDGETARY CONSIDERATION IN SETTING PASSING SCORES

Huynh Huynh

University of South Carolina

Presented as part of the symposium "setting standards: Theory and practice" sponsored jointly by the American Educational Research Association and the National Council on Measurement in Education at their annual meetings in San Francisco, April 8-12, 1979.

ABSTRACT

A general model along with four illustrations is presented for the consideration of budgetary constraints in the setting of passing scores in instructional programs involving remedial action for poor test performers. Budgetary constraints normally put an upper limit on any choice of passing score. Given relevant information, this limit may be determined. Alternately, ways to assess the budgetary consequences associated with a given passing score are provided. Such information would be useful in any final decision regarding the passing score.

1. INTRODUCTION

In many instructional programs, such as Individually Prescribed Instruction (Glaser, 1968) or others of a similar nature (Atkinson, 1968; Flanagan, 1967), testing is conducted at the end of every instructional unit to provide feedback to the student and/or teacher in order that appropriate action can be taken. If a student's test score is high, it may be reasonable to grant that student mastery

This paper has been distributed separately as RM 79-3, April, 1979.

of the current unit and to allow him to proceed to a subsequent unit. On the other hand, a low score may indicate that the student might benefit from some remedial action. This is also the case for certification testing such as high school graduation or for minimum competency testing as legislated in several states. Funds are usually allocated for remediation for students whose scores are too low to warrant mastery of the competencies under consideration.

The statistical issues relating to granting or denying mastery status have been approached by several writers, including Huynh (1976, 1977, 1978). Most proposed schemes are by and large quota-free, i.e., the mastery/nonmastery decision process considered by the writers does not take into account the budgetary consequences associated with the denial of mastery status. If funds provided for remediation are limited, then a constraint will have to be imposed on the number of students declared as failures (nonmasters).

The purpose of this paper is to demonstrate how budgetary restrictions may be taken into account in the process of setting passing (mastery) scores or performance standards. Alternately, the presentation provides ways to assess the budgetary consequences associated with an arbitrary passing score. Section 2 describes the overall framework. Illustrations based on the beta-binomial and normal-normal test score models will be provided in subsequent sections.

2. OVERALL FRAMEWORK

It is now assumed that the true ability of a population of subjects may be described by a random variable θ which ranges in the sample space Ω. For the beta-binomial model, θ is the proportion of items that a subject answers correctly in an item pool and Ω is the interval from 0 to 1. For the normal test score model, θ is the traditional true score (Lord & Novick, 1968) and Ω is the entire real line. Let the probability density function (pdf) of θ be $p(\theta)$.

BUDGETARY CONSIDERATION

Let x be the score obtained from the administration of an n-item test and let $f(x)$ and $f(x|\theta)$ denote its marginal and conditional probability density functions with respect to θ.

It shall be assumed that all subjects with test scores smaller than a passing (mastery) score c will be denied mastery for the instructional objectives covered by the test and that these subjects will be provided with appropriate remedial learning activities. The remediation is assumed to be so devised that its conclusion will coincide with the mastery status which was previously denied the student. The cost of remediation will be assumed to be a non-increasing function of θ and will be denoted as $\delta(\theta)$. Thus, remediation will cost less for more able students than it will for less able ones.

Consider now a subject with true ability θ. The probability that this person will be declared in need of remediation is given as the sum $\Sigma f(x|\theta)$ or the integral $\int f(x|\theta)dx$, with $x < c$. For the purposes of this section, the summation notation will be used. It follows that the (conditional) expected remediation cost for this subject is

$$\sum_{x<c} f(x|\theta)\delta(\theta).$$

Hence the (unconditional or marginal) expected remediation cost for a subject drawn randomly from the population is

$$\gamma(c) = \int_\Omega \sum_{x<c} f(x|\theta)\delta(\theta)p(\theta)d\theta. \tag{1}$$

This function is nondecreasing with respect to its argument c. Its lowest limit is zero (when all subjects are granted mastery status) and its maximum value, $\gamma_{max} = \int_\Omega \delta(\theta)p(\theta)d\theta$, is reached when remediation is provided to all subjects regardless of their test scores.

Let us suppose, furthermore, that testing is to be conducted for a total of m subjects and the total cost of possible remediation cannot exceed the value B. If the passing score c is selected, then the total expected remediation cost will be $m\gamma(c)$. Hence any choice for c must satisfy the budgetary constraint $m\gamma(c) \leq B$. If $\gamma_{max} \leq B$, any cutoff score will be acceptable. However, if $B < \gamma_{max}$, then the passing score c must be less than or equal to c_1, where c_1 is the highest score satisfying the inequality

$$\gamma(c_1) \leq B/m. \tag{2}$$

For discrete test scores, such as those of the binomial error model, Inequation (2) may be solved by computing the values of $\gamma(c)$ one by one, starting with c as the smallest test score, and stopping when the value c_1 is reached. For continuous test data, numerical procedures for solving the nonlinear equation $\gamma(c_1) = B/m$ might be needed.

3. THE BETA-BINOMIAL MODEL WITH CONSTANT COSTS

Consider now the beta-binomial model as defined by the following pdf's:

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x = 0,1,\ldots,n$$

and

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)},\ 0 < \theta < 1.$$

The two parameters α and β may be estimated from sample data via one of several estimation techniques such as the moment procedure or the maximum likelihood procedure. Let $\bar{x}$ and s be the sample test score mean and standard deviation. In addition, let $\hat{\alpha}_{21}$ be the KR21 reliability coefficient as defined by

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{x}(n-\bar{x})}{ns^2}\right). \tag{3}$$

(In the case of a negative $\hat{\alpha}_{21}$, simply replace the value computed from Equation (3) by any positive reliability estimate.) The moment estimates for α and β are given as

$$\hat{\alpha} = (-1 + 1/\hat{\alpha}_{21})\bar{x} \tag{4}$$

and

$$\hat{\beta} = -\hat{\alpha} + n/\hat{\alpha}_{21} - n. \tag{5}$$

We will now focus on the simple case where a single true passing score (or criterion level) θ_o, separating true masters from true nonmasters, has been specified. Let the remediation cost be constant and equal to γ_o for a true nonmaster and γ_1 for a true master. Thus the cost function is of the form

$$\delta(\theta) = \begin{cases} \gamma_o & \text{if } \theta < \theta_o \\ \gamma_1 & \text{if } \theta \geq \theta_o . \end{cases}$$

The nonincreasing nature of $\delta(\theta)$ is satisfied whenever $\gamma_o > \gamma_1$.

The expected remediation cost per student as shown by Equation (1) is now given as

$$\gamma(c) = \frac{1}{B(\alpha,\beta)} \sum_{x=o}^{c-1} \binom{n}{x} \left[\gamma_1 \int_{\theta_o}^{1} \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1} d\theta \right.$$

$$\left. + \gamma_o \int_o^{\theta_o} \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1} d\theta \right]$$

or

$$\gamma(c) = \frac{1}{B(\alpha,\beta)} \sum_{x=o}^{c-1} \binom{n}{x} \left[\gamma_1 B(\alpha+x, n+\beta-x) \right.$$

$$\left. + (\gamma_o - \gamma_1) \int_o^{\theta_o} \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1} d\theta \right] .$$

It may be noted that the marginal beta-binomial pdf of x is given as

$$f(x) = \binom{n}{x} B(\alpha+x, n+\beta-x)/B(\alpha,\beta) \tag{6}$$

and that the incomplete beta function $I(\alpha+x, n+\beta-x; \theta_o)$ is defined as

$$I(\alpha+x, n+\beta-x; \theta_o) = \int_o^{\theta_o} \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1} d\theta / B(\alpha+x, n+\beta-x) .$$

It follows that

$$\gamma(c) = \sum_{x=o}^{c-1} f(x)\left(\gamma_1 + (\gamma_o - \gamma_1) I(\alpha+x, n+\beta-x; \theta_o)\right) . \tag{7}$$

The values of f(x) may be computed via the following inductive formulae:

$$f(0) = \prod_{i=1}^{n} \frac{n+\beta-i}{n+\alpha+\beta-1} \tag{8}$$

and

$$f(x+1) = f(x) \cdot \frac{(n-x)(\alpha+x)}{(x+1)(n+\beta-x-1)}, \quad x = 0,1,\ldots,n-1 . \tag{9}$$

The following recurrence formula, on the other hand, will quicken the evaluation of the incomplete beta functions:

$$I(\alpha+x+1, n+\beta-x-1; \theta_o) = -\frac{\theta^{\alpha+x}(1-\theta)^{n+\beta-x-1}}{(\alpha+x)B(\alpha+x, n+\beta-x)} + I(\alpha+x, n+\beta-x; \theta_o). \tag{10}$$

Finally, as in Section 2, let B be the maximum funds allocated for possible remediation involving a group of m subjects. Then the passing score cannot exceed the highest integer c_1 at which $\gamma(c_1) \leq B/m$.

Numerical Example 1

A maximum sum of B = \$4000 has been allocated for remediation in an instructional program with m = 100 students. Thus B/m = \$40. For the program under study, assume that $\theta_o = .60$ and the remediation costs are γ_o = \$150 for each student with true ability $\theta < .60$ and γ_1 = \$50 for students with $\theta \geq .60$. Now suppose a 5-item test is administered and the test scores yield the estimates $\hat{\alpha} = 3$ and $\hat{\beta} = 2$. At the passing scores c = 1, 2, 3, 4, and 5, the expected remediation costs $\gamma(c)$ are \$7.02, \$19.06, \$31.83, \$41.25, and \$47.19, respectively. Since $\gamma(c_1) \leq$ \$40, it follows that $c_1 = 3$. The budget constraint imposes an upper limit of 3 on the passing score. If 3 is used, the expected cost of remediation amounts to \$3183. If the next higher passing score, 4, were used, the expected remediation cost would be \$4125, over the maximum budgeted sum of \$4000.

4. THE BETA-BINOMIAL MODEL WITH LINEAR COSTS

Let us suppose now that the cost function may be written as

$$\delta(\theta) = (\gamma_o - \gamma_1)(1-\theta) + \gamma_1, \tag{11}$$

in which $\gamma_1 < \gamma_o$. Thus the cost is a linear function of θ. It is equal to γ_o when $\theta = 0$ and γ_1 when $\theta = 1$.

Under the beta-binomial model as described in the first paragraph of Section 3, the expected cost per student is given as

$$\gamma(c) = \frac{1}{B(\alpha,\beta)} \sum_{x=o}^{c-1} \binom{n}{x} \left[(\gamma_o-\gamma_1) \int_o^1 \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x+1-1} d\theta \right.$$

$$\left. + \gamma_1 \int_o^1 \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1} d\theta\right]$$

$$= \frac{1}{B(\alpha,\beta)} \sum_{x=o}^{c-1} \binom{n}{x} \left[(\gamma_o-\gamma_1)B(\alpha+x,n+\beta-x+1) + \gamma_1 B(\alpha+x,n+\beta-x)\right].$$

By noting that

$$B(\alpha+x,n+\beta-x+1) = \frac{n+\beta-x}{n+\alpha+\beta} B(\alpha+x,n+\beta-x)$$

it may be deduced that

$$\gamma(c) = \sum_{x=o}^{c-1} f(x) \frac{(\gamma_o-\gamma_1)(n+\beta-x)}{n+\alpha+\beta} + \gamma_1$$

$$= \sum_{x=o}^{c-1} f(x) \quad \frac{\alpha_o(n+\beta-x) + \gamma_1(\alpha+x)}{n+\alpha+\beta}. \qquad (12)$$

As in the previous section, the values of $f(x)$ may be computed inductively via Equations (8) and (9).

Numerical Example 2

Consider the basic data of the first numerical example, namely B/m = \$40, γ_o = \$150, γ_1 = \$50, $\hat{\alpha} = 3$, $\hat{\beta} = 2$, and $n = 5$ items. At the passing scores of 1, 2, 3, 4, and 5, the expected remediation costs $\gamma(c)$ are \$5.71, \$18.81, \$37.86, \$59.29, and \$78.33. Hence the passing score cannot exceed 3, where the maximum value of the expected cost of remediation would amount to \$3786. Had a score of 4 been selected, the expected cost would have amounted to as much as \$5929.

To close this section, it should be mentioned that simple expressions for $\gamma(c)$ such as the one of Equation (12) may be worked out for all cost functions $\delta(\theta)$ which can be represented as integral polynomials of θ.

5. THE BIVARIATE NORMAL MODEL WITH CONSTANT COSTS

Now consider the case where the true score θ and the observed score x are jointly distributed according to a bivariate normal distribution. Without any loss of generality, it may be assumed that x is in its standardized form with zero mean and unit variance. Let ρ be the reliability of the test for the normal population of subjects under consideration. The true score θ has a mean of zero, a standard deviation of $\sqrt{\rho}$, and a correlation of $\sqrt{\rho}$ with the test score x. The joint pdf of x and θ is

$$f(x,\theta) = \frac{1}{2\pi\sqrt{\rho(1-\rho)}} \exp\left(-\frac{1}{2(1-\rho)}\left(x^2 - 2x\theta + \frac{\theta^2}{\rho}\right)\right). \quad (13)$$

As in Section 3, it will be assumed that the cost function $\delta(\theta)$ is constant, taking the values of γ_o for $\theta < \theta_o$ and the value of γ_1 for $\theta \geq \theta_o$. It follows from Equation (1) that at any passing score c, the remediation cost for a subject drawn randomly from the population is expected to be

$$\gamma(c) = \gamma_o \int_{-\infty}^{c} \int_{-\infty}^{\theta_o} f(x,\theta)d\theta dx + \gamma_1 \int_{-\infty}^{c} \int_{\theta_o}^{\infty} f(x,\theta)d\theta dx$$

$$= \gamma_1 \Pr(x \leq c) + (\gamma_o - \gamma_1) \int_{-\infty}^{c} \int_{-\infty}^{\theta_o} f(x,\theta)d\theta dx. \quad (14)$$

The maximum passing score c_1 satisfies the equation $\gamma(c_1) = B/m$. This value of c_1 exists as long as $B < \gamma_{max}$ where

$$\gamma_{max} = \gamma_o \Pr(\theta < \theta_o) + \gamma_1 \Pr(\theta \geq \theta_o).$$

Solutions may be found via numerical procedures such as the Newton iterative solution for nonlinear equations. To apply this technique, it may be noted that the derivative of $\gamma(c)$ with respect to c is

$$\gamma'(c) = \gamma_1 f_N(c) + (\gamma_o - \gamma_1) \int_{-\infty}^{\theta_o} f(c,\theta)d\theta$$

where $f_N(.)$ denotes the pdf of x (the unit normal variable). In other words,

$$f_N(c) = \frac{1}{\sqrt{2\pi}} e^{-c^2/2}.$$

It may also be noted that

$$\int_{-\infty}^{\theta_o} f(c,\theta)d\theta = f_N(c) \cdot F_N\left(\frac{\theta_o-\rho c}{\sqrt{\rho-\rho^2}}\right)$$

where $F_N(.)$ is the (cumulative) distribution function of the unit normal variable.

In summary,

$$\gamma'(c) = f_N(c)\left(\gamma_1 + (\gamma_o-\gamma_1)F_N\left(\frac{\theta_o-\rho c}{\sqrt{\rho-\rho^2}}\right)\right). \tag{15}$$

Both $\gamma(c)$ and $\gamma'(c)$ may be evaluated via computer programs such as those described in the IMSL (1977). They may also be obtained by use of appropriate tables for the univariate and bivariate normal distributions.

Numerical Example 3

Let the parameters defining the problem be $\rho = .64$, $\theta_o = 1$, $\gamma_o = \$150$, $\gamma_1 = \$50$, and $B/m = \$40$. Numerical procedure yields the maximum standardized passing score $c_1 = -.475$. If the test scores have a mean of 50 and a standard deviation of 20, then the passing score cannot exceed 40.5.

6. THE BIVARIATE NORMAL MODEL WITH NORMAL-OGIVE COST

Now consider the case where the cost function $\delta(\theta)$ is of the form

$$\delta(\theta) = (\gamma_o-\gamma_1)\left(1 - F_N\left(\frac{\theta-\theta_o}{\sigma}\right)\right) + \gamma_1 \tag{16}$$

where, as before, $F_N(.)$ represents the distribution function of the unit normal variable. In the context of decision theory, expressions similar to those of Equation (16) have been proposed as utility functions (e.g., Lindley, 1976, and Novick and Lindley, 1978). As in the case of the beta-binomial model with linear costs, γ_o and γ_1 represent the remediation costs associated with the least able ($\theta = -\infty$) and the most able ($\theta = +\infty$) subjects. On the other hand, the parameter θ_o is the location at which the cost is

$(\gamma_o+\gamma_1)/2$ and $1/\sigma$ indicates the extent to which $\delta(\theta)$ decreases at this location.

The expected remediation cost $\gamma(c)$ may now be written as

$$\gamma(c) = \int_{-\infty}^{c} \int_{-\infty}^{+\infty} f(x,\theta)\,\delta(\theta)\,d\theta dx$$

$$= \gamma_o \Pr(x \le c) - (\gamma_o-\gamma_1) \int_{-\infty}^{c} \phi(x) f_N(x)\,dx \tag{17}$$

where

$$\phi(x) = \int_{-\infty}^{+\infty} f(\theta|x) F_N\left(\frac{\theta-\theta_o}{\sigma}\right) d\theta .$$

The conditional pdf $f(\theta|x)$ is given as

$$f(\theta|x) = \frac{1}{\sqrt{2\pi\rho(1-\rho)}} \exp\left(-\frac{(\theta-\rho x)^2}{2\rho(1-\rho)}\right).$$

It follows that

$$\phi(x) = \frac{1}{2\pi\sigma\sqrt{\rho(1-\rho)}} \int_{-\infty}^{+\infty} \{\exp\left(-\frac{(\theta-\rho x)^2}{2\rho(1-\rho)}\right) \int_{-\infty}^{\theta} \exp\left(-\frac{(t-\theta_o)^2}{2\sigma^2}\right) dt\} d\theta .$$

It should be noted that the expression

$$\frac{1}{2\pi\sigma\sqrt{\rho(1-\rho)}} \exp\left(-\frac{(\theta-\rho x)^2}{2\rho(1-\rho)} - \frac{(t-\theta_o)^2}{2\sigma^2}\right)$$

acts as the joint pdf of two <u>independent</u> normal random variables θ and t with means ρx and θ_o, and with variances $\rho(1-\rho)$ and σ^2.

Now let us introduce the new random variable $u = \theta - t$ for which the mean is $\rho x - \theta_o$ and the variance is $\rho - \rho^2 + \sigma^2$. Since the condition $t < \theta$ is equivalent to $u > 0$, it follows that $\phi(x)$ may be expressed simply as

$$\phi(x) = \int_o^{\infty} \int_{-\infty}^{\infty} g_{\theta u}(\theta,u)\,d\theta du,$$

where $g_{\theta u}(\theta,u)$ is the bivariate normal pdf of θ and u. Hence

$$\phi(x) = \Pr(u \ge 0) = 1 - \Pr(u < 0)$$

or

$$\phi(x) = 1 - F_N\left(\frac{\theta_o - \rho x}{\sqrt{\rho - \rho^2 + \sigma^2}}\right). \tag{18}$$

With this new expression for $\phi(x)$, the expected remediation cost as defined in Equation (17) may be written as

$$\gamma(c) = \gamma_1 \Pr(x < c) + (\gamma_o - \gamma_1) \int_{-\infty}^{c} F_N\left(\frac{\theta_o - \rho x}{\sqrt{\rho - \rho^2 + \sigma^2}}\right) f_N(x)dx. \tag{19}$$

The integral found in Equation (19) may be written as

$$Z(c) = \int_{-\infty}^{c} \int_{-\infty}^{h(x)} f_N(w) f_N(x) dwdx,$$

where $h(x) = (-\rho x + \theta_o)/\sqrt{\rho - \rho^2 + \sigma^2}$, and $f_N(.)$ is again the pdf of a unit normal variable. Let

$$v = w - h(x) = w + (\rho x - \theta_o)/\sqrt{\rho - \rho^2 + \sigma^2}.$$

Then x and v follow a joint bivariate normal pdf, $g_{xv}(x,v)$, with means, variances, and correlation given, respectively, as

$$\begin{aligned} \mu_x &= 0, \\ \mu_v &= -\theta_c/\sqrt{\rho - \rho^2 + \sigma^2}, \\ \sigma_x^2 &= 1, \\ \sigma_v^2 &= (\rho + \sigma^2)/(\rho - \rho^2 + \sigma^2), \end{aligned} \tag{20}$$

and

$$\rho_{xv} = \rho/\sqrt{\rho + \sigma^2}.$$

Hence the integral Z(c) takes a simpler form given as

$$Z(c) = \int_{-\infty}^{c} \int_{-\infty}^{o} g_{xv}(x,v) dvdx,$$

and the expected remediation cost $\gamma(c)$ may be written as

$$\gamma(c) = \gamma_1 \Pr(x < c) + (\gamma_o - \gamma_1) \int_{-\infty}^{c} \int_{-\infty}^{o} g_{xv}(x,v) dvdx. \tag{21}$$

The numerical values of $\gamma(c)$ may be computed via tables or computer programs dealing with the univariate and bivariate normal distributions.

Numerical procedures such as the Newton iteration process may be used to solve the equation $\gamma(c) = B/m$. The derivative of $\gamma(c)$ with respect to c, from Equation (19), is found to be

$$\gamma'(c) = f_N(c)\left\{\gamma_1 + (\gamma_o-\gamma_1)F_N\left(\frac{\theta_o-\rho c}{\sqrt{\rho-\rho^2+\sigma^2}}\right)\right\} \tag{22}$$

It may be noted that by taking $\sigma^2 = 0$, Equations (19) and (22) of this section will reduce to Equations (14) and (15) of Section 5. This is expected since the normal-ogive cost function $\delta(\theta)$ as defined in (16) will degenerate into the constant cost function of Section 5 when σ^2 tends to zero. Finally, the maximum expected remediation cost (per random subject) may be deduced from Equation (21) by letting $c = +\infty$. It is

$$\gamma_m = \gamma_1 + (\gamma_o-\gamma_1)F_N\left(\frac{\theta_o}{\sqrt{\rho+\sigma^2}}\right). \tag{23}$$

<u>Numerical Example 4</u>

Let the parameters of the problem be $\rho = .64$, $\theta_o = 1$, $\sigma = 2$, $\gamma_o = \$150$, $\gamma_1 = \$50$, and $B/m = \$40$. The Newton iteration procedure for solving the equation $\gamma(c_1) = B/m$ yields the solution $c_1 = -.362$. If the test scores have a mean of 50 and a standard deviation of 20, then the test passing score cannot exceed 42.76.

7. SOME CONCLUDING REMARKS

In this paper a general model along with four separate illustrations is provided for the consideration of budgetary constraints in the setting of passing scores in instructional programs involving remediation for subjects with poor test performance. The illustrations are not meant to be exhaustive. Budgetary constraints normally impose a limit on the number of students allowed to take remedial learning activities and, hence, restrict the range in which a choice for the passing score is to be made. The paper also provides ways to assess the budgetary requirement associated with each passing score. This information would be a factor in decisions regarding passing scores and budgets for remediation.

BIBLIOGRAPHY

Atkinson, R. C. (1968). Computer-based instruction in initial reading. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service.

Flanagan, J. C. (1967). Functional education for the seventies. Phi Delta Kappan 49, 27-32.

Glaser, R. (1968). Adapting the elementary school curriculum to individual performance. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service.

Huynh, H. (1976). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

Huynh, H. (1977). Two simple classes of mastery scores based on the beta-binomial model. Psychometrika 42, 601-608.

Huynh, H. (1978). A nonrandomized minimax solution for mastery scores in the binomial error model. Research Memorandum 78-2, Publication Series in Mastery Testing. University of South Carolina College of Education.

IMSL Library 1 (1977). Houston: International Mathematical and Statistical Libraries.

Lindley, D. V. (1976). A class of utility functions. Annals of Statistics 4, 1-10.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company.

Novick, M. R. & Lindley, D. V. (1978). The use of more realistic utility functions in educational applications. Journal of Educational Measurement 15, 181-191.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not neccessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance of Joseph C. Saunders and Anthony J. Nitko are gratefully acknowledged.

PART THREE

CONSISTENCY OF DECISIONS

7

COMPUTATION AND INFERENCE FOR TWO RELIABILITY INDICES IN MASTERY TESTING BASED ON THE BETA-BINOMIAL MODEL

Huynh Huynh

University of South Carolina

Presented at the 17th Annual Southeastern Invitational Conference on Measurement in Education, University of North Carolina at Greensboro, December 8, 1978. Journal of Educational Statistics, Fall, 1979.

ABSTRACT

In mastery testing the raw agreement index and the kappa index may be secured via one test administration when the test scores follow beta-binomial distributions. This paper reports tables and a computer program which facilitate the computation of those indices and of their standard errors of estimate. Illustrations are provided in the form of confidence intervals, hypothesis testing, and minimum sample sizes in reliability studies for mastery tests.

1. INTRODUCTION

As indicated by several writers including Carver (1970) and Hambleton and Novick (1973), one of the uses of criterion-referenced testing is to classify examinees in two or more achievement categories. In this context, referred to here as mastery testing, reliability would be most appropriately viewed as classification (or decision) consistency across repeated test administrations using the same form or two equivalent forms. Decision consistency

This paper has been distributed separately as RM 78-1, December, 1979.

may be quantified by the raw agreement index p which expresses the proportion of examinees classified in the same category by both testings. When the two test administrations yield equivalent (or exchangeable) test data, p is bounded from below by p_c, the proportion of consistent decisions which would be expected if no relationship existed between the two sets of data (Huynh, 1976, 1978). In other words, $p_c \leq p \leq 1$. In a number of instances, for example when decision consistency is to be compared for two testing situations involving different p_c values, it would be suitable to scale p so that it forms an index with a range from 0 to 1. The kappa coefficient (Cohen, 1960), as defined by $\kappa = (p-p_c)/(1-p_c)$, is such an index. This coefficient represents the extent of improvement in decision consistency which is reflected by the dependency between two equivalent sets of data.

The definitions of both p and kappa include the notion of repeated testings. However, there are at least two procedures by which p and kappa may be approximated via test data collected from one test administration (Huynh, 1976; Subkoviak, 1976). The Subkoviak procedure relies on the estimation of the true score for each individual examinee. When combined with the binomial or compound binomial error model, the estimated true score will yield a consistency index for each examinee. The average of this index over a population of examinees is the Subkoviak estimate for p.

The Huynh method, on the other hand, assumes that test scores on one form follow a beta-binomial model and test scores on both forms distribute jointly as a bivariate beta-binomial distribution. Both p and kappa (and other similar indices) may then be computed via the univariate and bivariate distributions. In a simulation study based on real test data, Subkoviak (1978) concluded that "all things considered, the Huynh approach seems worthy of recommendation. It is mathematically sound, requires only one testing, and provides reasonably accurate estimates, which appear to be slightly conservative for short tests" (p. 115).

This paper will consider only the Huynh procedure for the approximation of p and kappa. Section 2 will provide a review of the computation of p and kappa. Section 3 will present formulae for computing the asymptotic standard errors of their estimates. Section 4 will describe the arrangement of the tables regarding p and kappa and their standard errors. Section 5 describes the interpolation process for nontabulated entries. Some applications of the tables will be presented in Section 6. The last two sections deal with a computer program for the estimates and their standard errors.

2. COMPUTATIONS FOR p AND κ

Consider now the administration of an n-item test to a population of examinees with true ability distributed according to the beta density with parameters α and β. The frequency distribution of the observed test score x is given by the beta-binomial (or negative hypergeometric) density

$$f(x) = \binom{n}{x} B(\alpha + x, n + \beta - x)/B(\alpha,\beta). \tag{1}$$

In this formula as well as in all other subsequent ones, the notation B denotes the beta function. The density f(x) may be computed via any of the following inductive formulae

$$\begin{cases} f(0) = \prod_{i=1}^{n} \dfrac{n+\beta+i}{n+\alpha+\beta-i} \\ f(x+1) = f(x) \cdot \dfrac{(n-x)(\alpha+x)}{(x+1)(n+\beta-x-1)}, \quad x=0,1,\ldots,n-1; \end{cases} \tag{2}$$

or

$$\begin{cases} f(n) = \prod_{i=1}^{n} \dfrac{n+\alpha-i}{n+\alpha+\beta-i} \\ f(x-1) = f(x) \cdot \dfrac{x(n+\beta-x)}{(n-x+1)(\alpha+x-1)}, \quad x=1,\ldots,n. \end{cases} \tag{3}$$

The first recurrence scheme is more efficient for small test scores whereas the second set works better for large test scores.

Let x and y be the test scores obtained by administering two equivalent n-item tests to each examinee in the population. Under local independence with respect to true ability, x and y follow the bivariate beta-binomial (or negative hypergeometric) density

$$f(x,y) = \frac{\binom{n}{x}\binom{n}{y}}{B(\alpha,\beta)}\, B(\alpha+x+y, 2n+\beta-x-y).$$

This density is symmetric in the sense that $f(x,y) = f(y,x)$.

For values of x and y near 0, $f(x,y)$ may be evaluated inductively via the following formulae:

$$f(0,0) = \prod_{i=1}^{2n} \frac{2n+\beta-i}{2n+\alpha+\beta-i} = f(0) \cdot \prod_{i=1}^{n} \frac{2n+\beta-i}{2n+\alpha+\beta-i},$$

and

$$f(x+1,y) = f(x,y) \cdot \frac{(n-x)(\alpha+x+y)}{(x+1)(2n+\beta-x-y-1)}.$$

For values of x and y near n, it is more efficient to use the following formulae:

$$f(n,n) = \prod_{i=1}^{2n} \frac{2n+\alpha-i}{2n+\alpha+\beta-i} = f(n) \cdot \prod_{i=1}^{n} \frac{2n+\alpha-i}{2n+\alpha+\beta-i},$$

and

$$f(x-1,y) = f(x,y) \cdot \frac{x(2n+\beta-x-y)}{(n-x+1)(\alpha+x+y-1)}.$$

Consider now the case where it is desired to place examinees into k classifications or categories defined by k-1 cutoff scores denoted by the integers c_j, $j=1,2,\ldots,k-1$ with $0 < c_1 < \ldots < c_{k-1} < n$. The first category consists of all test scores between 0 and c_1-1 inclusive. For the second category, the test scores range between c_1 and c_2-1 inclusive, and so on. Finally, for the kth category, the test score limits are c_{k-1} and n. For binary classification, k=2 and the cutoff score c is traditionally referred to as a mastery or passing score. These two categories are represented as $\{x: 0 \le x \le c-1\}$ and $\{x: c \le x \le n\}$. For k classifications as defined above, the raw agreement index is expressed as

$$p = \sum_{j=1}^{k} \left(\sum_{x,y=c_{j-1}}^{c_j-1} f(x,y) \right).$$

Here $c_0 = 0$ and $c_k = n+1$. The lower limit for decision consistency is given as

$$p_c = \sum_{j=1}^{k} \left(\sum_{x=c_{j-1}}^{c_j-1} f(x) \right)^2 .$$

As previously mentioned, the kappa index is defined as $\kappa = (p-p_c)/(1-p_c)$.

The formulae become somewhat simpler for binary classifications. For the use of c near 0, let

$$p_o = \sum_{x=0}^{c-1} f(x)$$

and

$$p_{oo} = \sum_{x,y=0}^{c-1} f(x,y) .$$

Then

$$p = 1-2(p_o-p_{oo})$$

and

$$\kappa = (p_{oo}-p_o^2)/(p_c-p_o^2) .$$

On the other hand, for values of c near n, let

$$p_1 = \sum_{x=c}^{n} f(x)$$

and

$$p_{11} = \sum_{x,y=c}^{n} f(x,y) .$$

Then

$$p = 1-2(p_1-p_{11})$$

and

$$\kappa = (p_{11}-p_1^2)/(p_1-p_1^2) .$$

3. ASYMPTOTIC SAMPLING DISTRIBUTION OF THE ESTIMATES

The estimation for p and κ may be carried out by replacing α and β by their estimates in the appropriate formulae of Section 2. There are at least two ways to estimate α and β, namely the maximum likelihood (ML) principle and the moment method. Let $\bar{x}$ and s be the mean and standard deviation of the test scores of m examinees, and let the estimated KR21 reliability be defined as

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{x}(n-\bar{x})}{ns^2}\right).$$

The moment estimates of α and β are given as

$$\hat{\alpha} = (-1 + 1/\hat{\alpha}_{21})\bar{x}$$

and

$$\hat{\beta} = -\hat{\alpha} + n/\hat{\alpha}_{21} - n.$$

These estimates are positive (thus acceptable) only when $0 < \hat{\alpha}_{21} < 1$. When the test scores do not show sufficient variability, the computed value for $\hat{\alpha}_{21}$ may be zero or negative. If this happens, replace this computed value by the smallest positive estimate for test reliability which happens to be available.

Maximum likelihood estimations for α and β have been considered by Griffiths (1973). A fairly efficient algorithm has been provided by Huynh (1977). Starting with the moment estimates, the Newton-Raphson procedure as implemented by Huynh has been found to converge very quickly in practically all cases considered by the author. It has been found that the ML estimates, in most cases, do not differ appreciably from the moment estimates $\hat{\alpha}$ and $\hat{\beta}$, hence general sampling properties appropriate for the ML estimates would be applicable to $\hat{\alpha}$ and $\hat{\beta}$. For example, asymptotically, $\sqrt{m}(\hat{\alpha}-\alpha,\ \hat{\beta}-\beta)$ follows a bivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij}) = \| b_{pq} \|^{-1}$ where

$$b_{11} = \sum_{x=0}^{n} \left(\frac{\partial f(x)}{\partial \alpha}\right)^2 / f(x)\ ,$$

$$b_{12} = b_{21} = \sum_{x=0}^{n} \frac{\partial f(x)}{\partial \alpha} \cdot \frac{\partial f(x)}{\partial \beta} / f(x)$$

and

$$b_{22} = \sum_{x=0}^{n} \left(\frac{\partial f(x)}{\partial \beta}\right)^2 / f(x)\ .$$

Now let $p = p(\alpha,\beta)$ and $\kappa = \kappa(\alpha,\beta)$ be the functions of (α,β) defining the two reliability indices. By replacing α and β by $\hat{}$ and $\hat{\beta}$ respectively, the moment estimates $\hat{p}$ and $\hat{\kappa}$ may be obtained for p and κ. It may be noted that both p and κ are continuous with respect to (α,β). It follows from Rao (1973, p. 386-7), that as m goes to infinity, $\sqrt{m}(\hat{p}-p)$ and $\sqrt{m}(\hat{\kappa}-\kappa)$ converge to two normal distributions with zero means and with variances

$$V_p^2 = \sigma_{11}\left(\frac{\partial p}{\partial \alpha}\right)^2 + 2\sigma_{12}\,\frac{\partial p}{\partial \alpha}\cdot\frac{\partial p}{\partial \beta} + \sigma_{22}\left(\frac{\partial p}{\partial \beta}\right)^2$$

and

$$V^2 = \sigma_{11}\left(\frac{\partial \kappa}{\partial \alpha}\right)^2 + 2\sigma_{12}\,\frac{\partial \kappa}{\partial \alpha}\cdot\frac{\partial \kappa}{\partial \beta} + \sigma_{22}\left(\frac{\partial \kappa}{\partial \beta}\right)^2,$$

respectively. Thus, it may be said that p has an approximate normal distribution with mean p and standard deviation (standard error) of $\sigma_\infty(\hat{p}) = V_p/\sqrt{m}$ when m is sufficiently large. An estimated standard error for $\hat{p}$, namely $s_\infty(\hat{p})$, may be obtained by replacing α and β by their estimated values $\hat{\alpha}$ and $\hat{\beta}$. The discussion also holds for $\hat{\kappa}$. Thus $\hat{\kappa}$ has an approximate normal distribution with mean κ and standard error $\sigma_\infty(\hat{\kappa}) = V_\kappa/\sqrt{m}$. The estimated standard error $s_\infty(\hat{\kappa})$ may be obtained in the same way as $s_\infty(\hat{p})$.

4. TABLES FOR $\hat{p}$, V_p, $\hat{\kappa}$, AND V_κ FOR SHORT TESTS

Appendix A presents tables which facilitate the computations for the reliability indices p and κ and their standard errors for the case of tests having 5 to 10 items. All computations were carried out via the IBM 370/168 system at the University of South Carolina, using the double precision mode.

Input data to the tables are (1) number of test items, n, (2) mastery or passing score, c, (3) test mean, $\bar{x}$, and (4) the KR21 reliability estimate, $\hat{\alpha}_{21}$. It may be noted that if $\tilde{\alpha}$ and $\tilde{\beta}$ are any estimates of the parameters α and β other than the moment estimates, then the entries for test mean and KR21 are simply $n\tilde{\alpha}/(\tilde{\alpha}+\tilde{\beta})$ and $n/(n+\tilde{\alpha}+\tilde{\beta})$, respectively.

For each entry of $(n,c,\bar{x},\hat{\alpha}_{21})$, four values may be read out. They are $\hat{p}$, V_p, $\hat{\kappa}$, and V_κ respectively. Both V_p and V_κ are enclosed in parentheses.

The tables are constructed for $n = 5\ (1)\ 10$ and $\hat{\alpha}_{21} =$ $.10\ (.10)\ .90$. For each n, the mastery score c is set equal to $n_o, n_o+1, \ldots, n-1, n$ where n_o is the smallest integer such that $n_o \geq n/2$ and with $\bar{x} = n$ times a decimal which ranges from .10 to .90 in steps of .10. To read the values of $\hat{p}$, V_p, $\hat{\kappa}$, and V_κ for a mastery score of $c < n_o$, simply enter the tables with a mastery score of $n-c+1$ and a test mean of $n-\bar{x}$.

Numerical Example 1

Let $n = 10$, $\bar{x} = 6$, $\hat{\alpha}_{21} = .50$, and $c = 7$. Then $\hat{p} = .680$, $V_p = .278$, $\hat{\kappa} = .347$, and $V_\kappa = .582$. If the data are obtained from a random sample of $m = 36$ examinees, then the estimated standard errors are $s_\infty(\hat{p}) = .278/6 = .046$ for $\hat{p}$ and $s_\infty(\hat{\kappa}) = .582/6 = .097$ for $\hat{\kappa}$.

Numerical Example 2

Let $n = 8$, $\bar{x} = 6.4$, $\hat{\alpha}_{21} = .30$, and $c = 3$. Here $n_o = 4$. The values of $\hat{p}$, V_p, $\hat{\kappa}$, and V_κ may be obtained by using the entry $n = 8$, $\bar{x} = 8-6.4 = 1.6$, $\hat{\alpha}_{21} = .30$, and $c = 8-3+1 = 6$. The results are $\hat{p} = .988$, $V_p = .075$, $\hat{\kappa} = .050$, and $V_\kappa = .448$. With $m = 25$, for example, the estimated standard errors are $s_\infty(\hat{p}) = .015$ and $s_\infty(\hat{\kappa}) = .090$.

5. INTERPOLATION

As revealed through the tables, $\hat{p}$, V_p, $\hat{\kappa}$, and V_κ are not monotonically increasing or decreasing functions of $\bar{x}$ at each $\hat{\alpha}_{21}$, or of $\hat{\alpha}_{21}$ at each $\bar{x}$. Hence interpolation should not be carried out indiscriminately. However, in situations where $\hat{\alpha}_{21}$, $\bar{x}/n$, and c/n are not too extreme, for example when all these quantities are between .20 and .80, the monotonicity property usually holds. If so, bivariate linear interpolation may be safely carried out to approximate the values of $\hat{p}$, V_p, $\hat{\kappa}$, and V_κ.

Suppose $\hat{\alpha}_{21}$ and $\bar{x}$ represent the computed values of KR21 and the test mean. In general, let $f(\hat{\alpha}_{21},\bar{x})$ be any one of the quantities $\hat{p}$, V_p, $\hat{\kappa}$, or V_κ that are needed but not found in the tables. Let u_1 and u_2 (where $u_1 \leq \hat{\alpha}_{21} \leq u_2$) be the two tabulated values closest to the computed $\hat{\alpha}_{21}$-value. Also, let v_1 and v_2 (where $v_1 \leq \bar{x} \leq v_2$) be the two tabulated values closest to the computed $\bar{x}$-value. Define the following:

$$r = \frac{(\hat{\alpha}_{21}-u_1)}{(u_2-u_1)}$$

and

$$s = \frac{(\bar{x}-v_1)}{(v_2-v_1)}.$$

Then the linearly interpolated value for $f(\hat{\alpha}_{21},\bar{x})$ is given as

$$f(u,v) = (1-r)(1-s)f(u_1,v_1) + r(1-s)f(u_2,v_1) + s(1-r)f(u_1,v_2) + rsf(u_2,v_2)$$

(see Abramowitz & Stegun, 1968, Formula 25.2.66).

Numerical Example

Let $n = 10$, $\hat{\alpha}_{21} = .56$ $(=u)$, and $\bar{x} = 4.77$ $(=v)$. Here $u_1 = .50$, $u_2 = .60$, $r = .60$, $v_1 = 4.00$, $v_2 = 5.00$, and $s = .77$. At the mastery or passing score $c = 7$, it may be found that the $\hat{p}$-values are $f(u_1,v_1) = .839$, $f(u_2,v_1) = .836$, $f(u_1,v_2) = .742$, and $f(u_2,v_2) = .761$. Hence the linearly interpolated value for $\hat{p}$ at $\hat{\alpha}_{21} = .56$ and $\bar{x} = 4.77$ is given as $.40 \times .23 \times .839 + .60 \times .23 \times .836 + .77 \times .40 \times .742 + .60 \times .77 \times .761 = .773$. In the same way, other linearly interpolated values are $V_{\hat{p}} = .205$, $\hat{\kappa} = .365$, and $V_{\hat{\kappa}} = .574$. The exact values for $\hat{p}$, $V_{\hat{p}}$, $\hat{\kappa}$, and $V_{\hat{\kappa}}$ computed directly from the formulae of Section 3 are .771, .201, .364, and .574, respectively.

6. APPLICATIONS

Besides easing the computations for $\hat{p}$, $\hat{\kappa}$, and their tandard errors in the case of short tests, the tables may be used to establish confidence intervals for p and κ, to test the equality of two or several independent $\hat{p}$ or $\hat{\kappa}$'s, and to answer questions regarding sample size in reliability studies for mastery tests.

6.1. Inference for One Sample

Let a 5-item test be administered to 100 students and let the summary test data be $\bar{x} = 3.500$ and $\hat{\alpha}_{21} = .400$. At the mastery score $c = 4$, the tables yield the values $\hat{p} = .650$, $V_p = .386$, $\hat{\kappa} = .293$, and $V_\kappa = .760$. The estimated standard errors are $s_\infty(\hat{p}) = .386/10 = .039$ and $s_\infty(\hat{\kappa}) = .763/10 = .076$. The 90% confidence intervals are $.650 \pm 1.645 \times .039$ or $(.581, .714)$ for the parameter p, and $.293 \pm 1.645 \times .076$ or $(.168, .418)$ for the parameter κ.

Hypothesis testing may also be conducted for the one-sample case. To test the null hypothesis that p is equal to a specified value p_H against an appropriate alternative, simply compare the Student-like ratio $t_p = (\hat{p}-p_H)/s_\infty(\hat{p})$ with suitably chosen critical value(s) read from the unit normal distribution. For κ, use the ratio $t_\kappa = (\hat{\kappa}-\kappa_H)/s_\infty(\hat{\kappa})$. With the data provided in this section, the null hypothesis $p_H = .50$ corresponds to the Student-like ratio $t_p = (.650-.500)/.039 = 3.846$. The null hypothesis $\kappa_H = .350$ is associated with the ratio $t_\kappa = (.293-.350)/.076 = -.75$. If the alternatives are two-sided and if the level of significance is 10% (at which the critical values are $\pm$ 1.645), the null hypothesis for p_H is rejected, whereas the one for κ_H is accepted.

6.2. Inference for Two Independent Samples

Any inference for the case of two independent samples may be carried out by noting that the standard error of $\hat{p}_1-\hat{p}_2$, where $\hat{p}_1$ and $\hat{p}_2$ are two independent sample p-values, is

$$s_\infty(\hat{p}_1-\hat{p}_2) = \left[s^2_\infty(\hat{p}_1) + s^2_\infty(\hat{p}_2)\right]^{1/2}.$$

For two independent $\hat{\kappa}_1$ and $\hat{\kappa}_2$, the standard error of $\hat{\kappa}_1-\hat{\kappa}_2$ is given as

$$s_\infty(\hat{\kappa}_1-\hat{\kappa}_2) = \left[s^2_\infty(\hat{\kappa}_1) + s^2_\infty(\hat{\kappa}_2)\right]^{1/2}.$$

For example, let the data for the first sample be $n = 5$, $c = 4$, $\bar{x} = 4.000$, $\hat{\alpha}_{21} = .600$, and $m = 100$. It follows that $\hat{p}_1 = .785$, $s_\infty(\hat{p}_1) = .0289$, $\hat{\kappa}_1 = .464$, and $s_\infty(\hat{\kappa}_1) = .0675$. For the second sample, chosen independently from the first one, let $n = 8$, $c = 6$,

$\bar{x} = 4.8$ $\hat{\alpha}_{21} = .300$, and $m = 64$. It may be verified that $\hat{p}_2 = .633$, $s_\infty(\hat{p}_2) = .0398$, $\hat{\kappa}_2 = .196$, and $s_\infty(\hat{\kappa}_2) = .093$. It follows that

$$s_\infty(\hat{p}_1-\hat{p}_2) = .049$$

and

$$s_\infty(\hat{\kappa}_1-\hat{\kappa}_2) = .115.$$

These standard errors will allow the formulation of confidence intervals for the parameters p_1-p_2 and $\kappa_1-\kappa_2$. For example, at the 90% confidence level, the confidence intervals are $(.785-.633) \pm 1.645 \times .049$ or $(.071,.233)$ for p_1-p_2, and $(.464-.196) \pm 1.645 \times .115$ or $(.079,.457)$ for $\kappa_1-\kappa_2$. Student-like ratios may also be computed to test the equality hypothesis for p_1 and p_2, and for κ_1 and κ_2. For $p_1 = p_2$, the mentioned ratio is $t_{p_1-p_2} = (.785-.633)/.048 = 3.167$ and for $\kappa_1 = \kappa_2$, the corresponding ratio is $(.464-.196)/.115 = 2.330$. With two-sided alternatives and with a level of significance of 10% (at which the critical values are $\pm$ 1.645), both equality hypotheses are rejected.

6.3. Testing Equality of Several Independent p or κ's

The mechanism by which equality of several p (or κ) values is to be tested is similar to the one by which several independent correlations are compared (Rao, 1973, page 434). Let $\hat{p}_i$ and $s_\infty(\hat{p}_i)$, $i = 1,2,\ldots,I$, be the estimated raw agreement index and its standard error associated with the i-th sample. Let $u_i = 1/s_\infty^2(\hat{p}_i)$ be the reciprocal of the error variance, and let

$$T_1 = \sum_{i=1}^{I} u_i\hat{p}_i,$$

$$T_2 = \sum_{i=1}^{I} u_i\hat{p}_i^2,$$

and

$$B = \sum_{i=1}^{I} u_i.$$

Then the statistic for testing homogeneity of the p-values is

$$H = T_2 - (T_1^2/B),$$

which can be used as χ^2 with I-1 degrees of freedom. Table 1 presents the data and various computations for the statistic H. With the value H = 1.738 and I-1 = 3 degrees of freedom (at which the 5% critical value is 7.815), it may be concluded that the four independent $\hat{p}$ values do not differ significantly from each other at the 5% level of significance.

TABLE 1

An Illustration of Homogeneity Testing for p

n	c	m	$\bar{x}$	$\hat{\alpha}_{21}$	V_p	$s_\infty(\hat{p})$	u_i	$\hat{p}_i$	$u_i\hat{p}_i$	$u_i\hat{p}_i^2$
5	4	64	3.0	.60	.269	.033625	884.454	.730	645.652	471.326
8	7	25	4.8	.40	.239	.047800	437.667	.776	339.630	263.553
10	6	100	5.0	.70	.206	.020600	2356.490	.765	1802.715	1379.077
9	6	49	6.3	.50	.267	.038143	687.337	.721	495.570	357.306
						Total	4365.948		3283.567	2471.262

Summary data: $B = 4365.948$

$T_1 = 3283.567$

$T_2 = 2471.262$

Test statistic: H = 1.738 with df = 4-1 = 3

6.4. Sample Size Determination

In some reliability studies for mastery tests, it may be necessary to determine in advance the minimum number of examinees needed to achieve a given degree of accuracy. For example, if a standard error $s_\infty(\hat{p})$ of no more than 100γ% of the parameter p is acceptable, then how many examinees should be tested? The question, of course, may not have an answer unless there are some indications about the mean and variability of the test scores. In a number of situations involving an n-item test with a options for each item, it may not be unreasonable to assume that the test mean is about halfway between the chance score n/a and the maximum score n and that the standard deviation s is about one-fourth of the difference between these two scores. In other words, the "guessed-at" values for $\bar{x}$, s, and $\hat{\alpha}_{21}$ are given as

$$\bar{x} = (n + n/a)/2,$$

$$s = (n - n/a)/4,$$

and

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{x}(n-\bar{x})}{ns^2}\right).$$

By entering these values of $\bar{x}$ and $\hat{\alpha}_{21}$, along with n and c, those of $\hat{p}$ and $V_p = \sqrt{m}\, s_\infty(\hat{p})$ may be deduced. Then m may be approximated by noting that the ratio of $V_p/\sqrt{m}$ to $\hat{p}$ cannot exceed γ. In other words, the minimum number of examinees is $(V_p/(\gamma\hat{p}))^2$.

As in illustration, let n = 8, a = 5, c = 5, and $\gamma = 0.05$. Then $\bar{x} = 4.8$, s = 1.6, and $\hat{\alpha}_{21} = .29$. From the tables, it may be found that approximately $\hat{p} = .615$ and $V_p = .369$. The minimum number of examinees is 144. If γ is .10, then only 36 examinees would be needed.

7. COMPUTER PROGRAM

Appendix B lists a FORTRAN IV program which computes the values of $\hat{p}$, $s_\infty(\hat{p})$, $\hat{\kappa}$, and $s_\infty(\hat{\kappa})$ for situations with k classifications. The input data are to be keypunched on three cards detailed as follows.

First Card

This contains the title of the problem, keypunched anywhere between columns 1 and 80.

Second Card

This provides data on number of items (n), number of examinees (m), number of classifications (k), the test mean ($\bar{x}$), and the test standard deviations (s). These must be keypunched according to the format (3I5, 2F10.5).

Third Card

This contains the (k-1) cutoff scores, keypunched with the format (16I5). Thus reliability problems with 17 classifications

TABLE II

An Output of the Computer Program

```
ESTIMATES OF DECISION RELIABILITY
AND THEIR STANDARD ERRORS IN
MASTERY TESTING BASED ON THE BETA-
BINOMIAL MODEL
TITLE OF THIS JOB IS:
AN EXAMPLE OF RELIABILITY COMPUTATION

INPUT DATA ARE:

NUMBER OF ITEMS .. =    8
NUMBER OF SUBJECTS =   25
MEAN OF TEST SCORE ............. =      4.80000
STANDARD DEVIATION OF TEST SCORE =      2.22596
NUMBER OF CATEGORIES =    2
CUTOFF SCORE ....... =    5

OUTPUT DATA ARE:

ALPHA =     2.05710
BETA  =     1.37140
KR21  =     0.70000

RAW AGREEMENT INDEX P =  0.77095
STANDARD ERROR OF P.. =  0.04345

KAPPA INDEX ........... =  0.53165
STANDARD ERROR OF KAPPA =  0.08871


** NORMAL END FOR THIS JOB **
   PROGRAM WRITTEN BY HUYNH HUYNH
   COLLEGE OF EDUCATION
   UNIVERSITY OF SOUTH CAROLINA
   COLUMBIA, SOUTH CAROLINA 29208
   REVISED, DECEMBER 1979
```

may be implemented via this FORTRAN program.

The computer program starts with the computation of $\hat{\alpha}_{21}$. If $\hat{\alpha}_{21}$ is zero or negative, the following message will be printed:

NON-POSITIVE ESTIMATE KR21.

MOMENT ESTIMATES FOR ALPHA AND BETA DO NOT EXIST.

COMPUTATIONS DISCONTINUED FOR THIS CASE.

Otherwise, the estimates $\hat{\alpha}$ and $\hat{\beta}$ will be obtained. These, in turn, will be used as input in a subroutine which computes $\hat{p}$, $s_{\infty}(\hat{p})$, $\hat{\kappa}$, and $s_{\infty}(\hat{\kappa})$.

For example, let the input cards be as follows:

```
                      1    1    2    2    3    3
      Column :  1...5....0....5....0....5....0....5

  First Card :  AN EXAMPLE OF RELIABILITY COMPUTATION
 Second Card :        8   25    2   4.8        2.22596
  Third Card :        5
```

In other words, $n = 8$, $m = 25$, $k = 2$, $\bar{x} = 4.8$, $s = 2.22596$, $c = 5$. The output is printed in Table 2. It may be read that $\hat{p} = .77095$, $s_{\infty}(\hat{p}) = .04345$, $\hat{\kappa} = .53165$, and $s_{\infty}(\hat{\kappa}) = .08871$.

Several problems may be performed in one run by stacking the input cards together.

8. DISCLAIMER

The computer program presented in this report has been written with care and tested extensively under a variety of conditions using tests with 60 or fewer items. The author, however, makes no warranty as to its accuracy and functioning, nor shall the fact of its distribution imply such warranty.

BIBLIOGRAPHY

Abramowitz, M. & Stegun, I. A. (1968). Handbook of mathematical functions. Washington: National Bureau of Standards.

Carver, R. P. (1970). Special problems in measuring changes in psychometric devices. In Evaluative research: Strategies and methods. Pittsburgh: American Institute for Research.

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

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics 29, 637-648.

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

Huynh, H. (1977). Statistical inference for the kappa and kappamax reliability indices based on the beta-binomial model. Paper read at the Psychometric Society Meetings, The University of North Carolina at Chapel Hill, June 16-17.

Huynh, H. (1978). Reliability of multiple classifications. Psychometrika 43, 317-325.

Rao, C. R. (1973). Linear statistical inference and its applications (2nd ed.). New York: John Wiley & Sons.

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

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

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance and comments of Joseph C. Saunders and Anthony J. Nitko are gratefully acknowledged.

APPENDIX A

Tables of the Raw Agreement Index and Its Standard Error Times the Square Root of m, the Kappa Index and Its Standard Error Times the Square Root of m, When the Beta-Binomial Model is Assumed

(m = Number of Subjects)

Input data to the tables are (i) number of test items (n), (ii) mastery score (c), (iii) test mean ($\bar{x}$), and (iv) the KR21 reliability (α_{21}). (Note that if $\tilde{\alpha}$ and $\tilde{\beta}$ are any estimates of the parameters α and β other than the moment estimates, then the entries for test mean and KR21 are simply $n\tilde{\alpha}/(\tilde{\alpha}+\tilde{\beta})$ and $n/(n+\tilde{\alpha}+\tilde{\beta})$, respectively.)

For each entry of (n, c, $\bar{x}$, $\hat{\alpha}_{21}$), four values may be read out. They are $\hat{p}$, V_p, $\hat{\kappa}$, and V_κ, respectively. Both V_p and V_κ are enclosed in parentheses.

Example

Let n = 5, c = 3, $\bar{x}$ = 1.5, and $\hat{\alpha}_{21}$ = .400. The tables provide the values $\hat{p}$ = .755, V_p = .267, $\hat{\kappa}$ = .268, and V_κ = .784. With m = 100, for example, the estimated standard errors are $s(\hat{p})$ = .0267 and $s(\hat{\kappa})$ = .0784.

152

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 5
Mastery score C = 3

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.5	0.975	0.966	0.957	0.949	0.942	0.939	0.940	0.948	0.966
	(0.157)	(0.172)	(0.177)	(0.172)	(0.157)	(0.138)	(0.118)	(0.105)	(0.089)
	0.022	0.062	0.122	0.198	0.288	0.392	0.510	0.643	0.798
	(0.477)	(0.734)	(0.928)	(1.048)	(1.091)	(1.063)	(0.969)	(0.808)	(0.570)
1.0	0.879	0.869	0.862	0.858	0.858	0.864	0.877	0.901	0.938
	(0.297)	(0.276)	(0.252)	(0.226)	(0.202)	(0.180)	(0.162)	(0.146)	(0.119)
	0.042	0.096	0.162	0.239	0.325	0.421	0.529	0.652	0.800
	(0.706)	(0.808)	(0.858)	(0.863)	(0.831)	(0.769)	(0.680)	(0.563)	(0.405)
1.5	0.729	0.734	0.743	0.755	0.772	0.795	0.824	0.864	0.918
	(0.338)	(0.313)	(0.289)	(0.267)	(0.245)	(0.223)	(0.201)	(0.175)	(0.137)
	0.057	0.122	0.192	0.268	0.351	0.441	0.542	0.659	0.801
	(0.874)	(0.865)	(0.833)	(0.784)	(0.720)	(0.646)	(0.561)	(0.463)	(0.339)
2.0	0.591	0.617	0.645	0.675	0.709	0.746	0.789	0.840	0.906
	(0.431)	(0.397)	(0.365)	(0.332)	(0.299)	(0.266)	(0.232)	(0.195)	(0.147)
	0.067	0.137	0.209	0.285	0.365	0.453	0.550	0.662	0.802
	(0.973)	(0.898)	(0.821)	(0.744)	(0.666)	(0.587)	(0.505)	(0.417)	(0.309)
2.5	0.535	0.571	0.607	0.645	0.685	0.728	0.776	0.832	0.901
	(0.503)	(0.454)	(0.409)	(0.366)	(0.325)	(0.284)	(0.244)	(0.201)	(0.150)
	0.070	0.142	0.215	0.290	0.370	0.457	0.552	0.664	0.803
	(1.006)	(0.909)	(0.818)	(0.732)	(0.649)	(0.569)	(0.488)	(0.403)	(0.300)
3.0	0.591	0.617	0.645	0.675	0.709	0.746	0.789	0.840	0.906
	(0.431)	(0.397)	(0.365)	(0.332)	(0.299)	(0.266)	(0.232)	(0.195)	(0.147)
	0.067	0.137	0.209	0.285	0.365	0.453	0.550	0.662	0.802
	(0.973)	(0.898)	(0.821)	(0.744)	(0.666)	(0.587)	(0.505)	(0.417)	(0.309)
3.5	0.729	0.734	0.743	0.755	0.772	0.795	0.824	0.864	0.918
	(0.338)	(0.313)	(0.289)	(0.267)	(0.245)	(0.223)	(0.201)	(0.175)	(0.137)
	0.057	0.122	0.192	0.268	0.351	0.441	0.542	0.659	0.801
	(0.874)	(0.865)	(0.833)	(0.784)	(0.720)	(0.646)	(0.561)	(0.463)	(0.339)
4.0	0.879	0.869	0.862	0.858	0.858	0.864	0.877	0.901	0.938
	(0.297)	(0.276)	(0.252)	(0.226)	(0.202)	(0.180)	(0.162)	(0.146)	(0.119)
	0.042	0.096	0.162	0.239	0.325	0.421	0.529	0.652	0.800
	(0.706)	(0.808)	(0.858)	(0.863)	(0.831)	(0.769)	(0.680)	(0.563)	(0.405)
4.5	0.975	0.966	0.957	0.949	0.942	0.939	0.940	0.948	0.966
	(0.157)	(0.172)	(0.177)	(0.172)	(0.157)	(0.138)	(0.118)	(0.105)	(0.089)
	0.022	0.062	0.122	0.198	0.288	0.392	0.510	0.643	0.798
	(0.477)	(0.734)	(0.928)	(1.048)	(1.091)	(1.063)	(0.969)	(0.808)	(0.570)

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 5
Mastery score C = 4

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.5	0.998	0.996	0.992	0.987	0.981	0.974	0.968	0.964	0.971
	(0.028)	(0.045)	(0.064)	(0.084)	(0.101)	(0.108)	(0.102)	(0.083)	(0.068)
	0.005	0.021	0.055	0.111	0.192	0.297	0.427	0.583	0.768
	(0.142)	(0.355)	(0.611)	(0.855)	(1.041)	(1.136)	(1.118)	(0.971)	(0.682)
1.0	0.980	0.973	0.963	0.953	0.942	0.932	0.925	0.926	0.945
	(0.120)	(0.140)	(0.157)	(0.167)	(0.167)	(0.156)	(0.133)	(0.108)	(0.094)
	0.014	0.042	0.088	0.152	0.235	0.338	0.459	0.603	0.775
	(0.300)	(0.491)	(0.661)	(0.787)	(0.854)	(0.857)	(0.796)	(0.670)	(0.473)
1.5	0.928	0.916	0.903	0.891	0.882	0.876	0.876	0.889	0.923
	(0.242)	(0.243)	(0.237)	(0.223)	(0.202)	(0.175)	(0.148)	(0.127)	(0.114)
	0.027	0.067	0.123	0.192	0.276	0.374	0.487	0.620	0.782
	(0.483)	(0.620)	(0.715)	(0.764)	(0.767)	(0.727)	(0.650)	(0.537)	(0.384)
2.0	0.830	0.820	0.813	0.808	0.809	0.815	0.830	0.858	0.907
	(0.316)	(0.292)	(0.266)	(0.238)	(0.211)	(0.186)	(0.166)	(0.150)	(0.131)
	0.041	0.093	0.155	0.228	0.311	0.404	0.511	0.635	0.787
	(0.666)	(0.729)	(0.755)	(0.747)	(0.710)	(0.648)	(0.565)	(0.464)	(0.337)
2.5	0.697	0.701	0.709	0.721	0.738	0.761	0.793	0.836	0.899
	(0.323)	(0.299)	(0.277)	(0.256)	(0.237)	(0.218)	(0.199)	(0.178)	(0.146)
	0.055	0.116	0.184	0.258	0.339	0.429	0.530	0.647	0.792
	(0.827)	(0.817)	(0.785)	(0.737)	(0.674)	(0.600)	(0.517)	(0.424)	(0.313)
3.0	0.576	0.601	0.628	0.658	0.692	0.730	0.775	0.829	0.898
	(0.401)	(0.377)	(0.352)	(0.325)	(0.298)	(0.269)	(0.238)	(0.203)	(0.156)
	0.065	0.134	0.205	0.280	0.361	0.448	0.545	0.657	0.796
	(0.952)	(0.884)	(0.812)	(0.737)	(0.660)	(0.581)	(0.499)	(0.412)	(0.308)
3.5	0.538	0.574	0.612	0.650	0.691	0.735	0.784	0.839	0.908
	(0.521)	(0.473)	(0.429)	(0.386)	(0.345)	(0.304)	(0.262)	(0.216)	(0.159)
	0.071	0.144	0.217	0.293	0.374	0.460	0.555	0.664	0.800
	(1.027)	(0.932)	(0.844)	(0.760)	(0.678)	(0.598)	(0.516)	(0.429)	(0.323)
4.0	0.636	0.662	0.689	0.718	0.750	0.785	0.825	0.871	0.927
	(0.464)	(0.428)	(0.392)	(0.358)	(0.324)	(0.289)	(0.252)	(0.208)	(0.150)
	0.070	0.142	0.217	0.294	0.376	0.464	0.560	0.669	0.803
	(1.035)	(0.969)	(0.900)	(0.829)	(0.754)	(0.675)	(0.590)	(0.492)	(0.370)
4.5	0.845	0.844	0.847	0.853	0.864	0.879	0.899	0.925	0.958
	(0.317)	(0.291)	(0.267)	(0.247)	(0.231)	(0.214)	(0.195)	(0.167)	(0.121)
	0.057	0.124	0.198	0.279	0.365	0.458	0.559	0.671	0.805
	(0.952)	(1.028)	(1.052)	(1.036)	(0.988)	(0.913)	(0.810)	(0.677)	(0.502)

For the mastery score = 2 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 5
Mastery score C = 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.5	1.000	1.000	0.999	0.998	0.996	0.993	0.988	0.980	0.975
	(0.002)	(0.005)	(0.010)	(0.019)	(0.032)	(0.051)	(0.072)	(0.081)	(0.062)
	0.000	0.004	0.015	0.040	0.088	0.168	0.288	0.458	0.687
	(0.019)	(0.089)	(0.231)	(0.443)	(0.699)	(0.949)	(1.125)	(1.139)	(0.893)
1.0	0.999	0.997	0.995	0.992	0.986	0.978	0.966	0.954	0.950
	(0.015)	(0.024)	(0.037)	(0.055)	(0.077)	(0.100)	(0.116)	(0.111)	(0.080)
	0.002	0.010	0.028	0.062	0.119	0.205	0.326	0.488	0.702
	(0.059)	(0.158)	(0.303)	(0.476)	(0.649)	(0.787)	(0.853)	(0.807)	(0.613)
1.5	0.992	0.988	0.983	0.975	0.964	0.951	0.935	0.922	0.925
	(0.053)	(0.070)	(0.091)	(0.112)	(0.133)	(0.148)	(0.149)	(0.125)	(0.092)
	0.006	0.019	0.046	0.089	0.154	0.244	0.363	0.517	0.716
	(0.130)	(0.252)	(0.393)	(0.534)	(0.651)	(0.723)	(0.729)	(0.655)	(0.488)
2.0	0.973	0.965	0.954	0.942	0.927	0.911	0.895	0.887	0.904
	(0.127)	(0.147)	(0.165)	(0.180)	(0.188)	(0.184)	(0.164)	(0.127)	(0.105)
	0.012	0.034	0.070	0.122	0.192	0.284	0.400	0.545	0.729
	(0.236)	(0.364)	(0.487)	(0.591)	(0.660)	(0.682)	(0.651)	(0.562)	(0.416)
2.5	0.928	0.915	0.901	0.886	0.870	0.857	0.849	0.853	0.888
	(0.228)	(0.236)	(0.239)	(0.235)	(0.221)	(0.196)	(0.161)	(0.128)	(0.125)
	0.021	0.053	0.098	0.158	0.233	0.325	0.437	0.572	0.741
	(0.376)	(0.488)	(0.579)	(0.641)	(0.667)	(0.652)	(0.595)	(0.500)	(0.371)
3.0	0.843	0.830	0.817	0.806	0.799	0.796	0.803	0.826	0.880
	(0.311)	(0.296)	(0.275)	(0.248)	(0.218)	(0.185)	(0.158)	(0.148)	(0.151)
	0.033	0.076	0.131	0.197	0.275	0.366	0.477	0.597	0.753
	(0.544)	(0.620)	(0.668)	(0.686)	(0.673)	(0.629)	(0.557)	(0.461)	(0.347)
3.5	0.714	0.711	0.711	0.715	0.725	0.742	0.770	0.813	0.883
	(0.314)	(0.285)	(0.257)	(0.234)	(0.216)	(0.205)	(0.201)	(0.197)	(0.173)
	0.047	0.102	0.166	0.237	0.316	0.405	0.505	0.621	0.764
	(0.734)	(0.758)	(0.757)	(0.732)	(0.686)	(0.621)	(0.539)	(0.445)	(0.342)
4.0	0.576	0.597	0.621	0.649	0.683	0.722	0.769	0.827	0.901
	(0.349)	(0.346)	(0.343)	(0.337)	(0.328)	(0.313)	(0.291)	(0.256)	(0.196)
	0.063	0.130	0.201	0.277	0.357	0.443	0.537	0.643	0.775
	(0.945)	(0.910)	(0.861)	(0.799)	(0.727)	(0.646)	(0.558)	(0.464)	(0.366)
4.5	0.560	0.603	0.647	0.691	0.737	0.783	0.832	0.883	0.938
	(0.672)	(0.632)	(0.587)	(0.537)	(0.482)	(0.422)	(0.354)	(0.277)	(0.183)
	0.080	0.158	0.237	0.316	0.396	0.479	0.567	0.664	0.785
	(1.202)	(1.127)	(1.046)	(0.960)	(0.870)	(0.776)	(0.677)	(0.574)	(0.464)

For the mastery score = 1 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 6
Mastery score C = 3

Test Mean \ KR21=	.100	.200	.300	.400	.500	.600	.700	.800	.900
0.6	0.959	0.948	0.938	0.930	0.925	0.924	0.928	0.939	0.961
	(0.202)	(0.207)	(0.201)	(0.188)	(0.169)	(0.147)	(0.128)	(0.114)	(0.093)
	0.028	0.074	0.137	0.214	0.304	0.404	0.517	0.643	0.792
	(0.553)	(0.771)	(0.918)	(0.995)	(1.008)	(0.964)	(0.869)	(0.724)	(0.517)
1.2	0.815	0.811	0.811	0.814	0.822	0.836	0.857	0.887	0.931
	(0.320)	(0.293)	(0.267)	(0.242)	(0.220)	(0.199)	(0.179)	(0.157)	(0.123)
	0.051	0.111	0.180	0.256	0.340	0.431	0.533	0.650	0.793
	(0.793)	(0.837)	(0.842)	(0.816)	(0.766)	(0.697)	(0.611)	(0.506)	(0.368)
1.8	0.637	0.657	0.679	0.704	0.732	0.764	0.803	0.849	0.910
	(0.395)	(0.366)	(0.337)	(0.309)	(0.279)	(0.250)	(0.218)	(0.183)	(0.137)
	0.065	0.133	0.204	0.279	0.359	0.446	0.542	0.654	0.793
	(0.930)	(0.873)	(0.810)	(0.741)	(0.668)	(0.592)	(0.510)	(0.421)	(0.311)
2.4	0.538	0.573	0.609	0.646	0.685	0.727	0.774	0.829	0.898
	(0.487)	(0.440)	(0.396)	(0.354)	(0.314)	(0.274)	(0.235)	(0.193)	(0.143)
	0.069	0.140	0.212	0.286	0.365	0.450	0.544	0.654	0.792
	(0.973)	(0.880)	(0.793)	(0.710)	(0.629)	(0.550)	(0.470)	(0.387)	(0.287)
3.0	0.574	0.601	0.629	0.660	0.694	0.732	0.775	0.828	0.896
	(0.416)	(0.384)	(0.353)	(0.321)	(0.289)	(0.257)	(0.222)	(0.185)	(0.140)
	0.066	0.134	0.205	0.279	0.358	0.444	0.539	0.650	0.791
	(0.933)	(0.858)	(0.783)	(0.706)	(0.629)	(0.550)	(0.470)	(0.385)	(0.285)
3.6	0.708	0.713	0.721	0.734	0.750	0.773	0.803	0.844	0.903
	(0.328)	(0.304)	(0.281)	(0.258)	(0.236)	(0.214)	(0.191)	(0.166)	(0.132)
	0.055	0.117	0.185	0.259	0.340	0.428	0.528	0.643	0.788
	(0.820)	(0.807)	(0.774)	(0.724)	(0.660)	(0.586)	(0.503)	(0.411)	(0.300)
4.2	0.857	0.846	0.838	0.833	0.832	0.837	0.849	0.874	0.918
	(0.305)	(0.284)	(0.260)	(0.234)	(0.208)	(0.182)	(0.160)	(0.141)	(0.118)
	0.040	0.091	0.154	0.227	0.311	0.404	0.510	0.633	0.785
	(0.645)	(0.724)	(0.760)	(0.757)	(0.721)	(0.659)	(0.575)	(0.470)	(0.337)
4.8	0.957	0.946	0.934	0.923	0.913	0.906	0.905	0.913	0.940
	(0.192)	(0.203)	(0.206)	(0.200)	(0.185)	(0.163)	(0.137)	(0.115)	(0.099)
	0.022	0.061	0.115	0.185	0.271	0.371	0.486	0.619	0.780
	(0.429)	(0.603)	(0.731)	(0.804)	(0.822)	(0.788)	(0.708)	(0.585)	(0.413)
5.4	0.995	0.991	0.986	0.979	0.971	0.964	0.958	0.957	0.968
	(0.052)	(0.074)	(0.095)	(0.113)	(0.123)	(0.121)	(0.107)	(0.086)	(0.072)
	0.008	0.030	0.073	0.137	0.223	0.329	0.455	0.602	0.775
	(0.210)	(0.448)	(0.694)	(0.896)	(1.024)	(1.062)	(1.006)	(0.853)	(0.595)

For the Mastery score = 4 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its S.E.*SQRT(M), the Kappa Index and its S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 6
Mastery score C = 4

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.6	0.995	0.991	0.986	0.979	0.971	0.964	0.958	0.957	0.968
	(0.052)	(0.074)	(0.095)	(0.113)	(0.123)	(0.121)	(0.107)	(0.086)	(0.072)
	0.008	0.030	0.073	[illegible]	0.223	0.329	0.455	0.602	0.775
	(0.210)	(0.448)	(0.694)	(0.896)	(1.024)	(1.062)	(1.006)	(0.853)	(0.595)
1.2	0.957	0.946	0.934	0.923	0.913	0.906	0.905	[illegible]	0.940
	(0.192)	(0.203)	(0.206)	(0.200)	(0.185)	(0.163)	(0.137)	[illegible]	(0.099)
	0.022	0.061	0.115	0.185	0.271	0.371	0.486	0.619	0.780
	(0.429)	(0.603)	(0.731)	(0.804)	(0.822)	(0.788)	(0.708)	(0.585)	(0.413)
1.8	0.857	0.846	0.838	0.833	0.832	0.837	0.849	0.874	0.918
	(0.305)	(0.284)	(0.260)	(0.234)	(0.208)	(0.182)	(0.160)	(0.141)	(0.118)
	0.040	0.091	0.154	0.227	0.311	0.404	0.510	0.633	0.785
	(0.645)	(0.724)	(0.760)	(0.757)	(0.721)	(0.659)	(0.575)	(0.470)	(0.337)
2.4	0.708	0.713	0.721	0.734	0.750	0.773	0.803	0.844	0.903
	(0.328)	(0.304)	(0.281)	(0.258)	(0.236)	(0.214)	(0.191)	(0.166)	(0.132)
	0.055	0.117	0.185	0.259	0.340	0.428	0.528	0.643	0.788
	(0.820)	(0.807)	(0.774)	(0.724)	(0.660)	(0.586)	(0.503)	(0.411)	(0.300)
3.0	0.574	0.601	0.629	0.660	0.694	0.732	0.775	0.828	0.896
	(0.416)	(0.384)	(0.353)	(0.321)	(0.289)	(0.257)	(0.222)	(0.185)	(0.140)
	0.066	0.134	0.205	0.279	0.358	0.444	0.539	0.650	0.791
	(0.933)	(0.858)	(0.783)	(0.706)	(0.629)	(0.550)	(0.470)	(0.385)	(0.285)
3.6	0.538	0.573	0.609	0.646	0.685	0.727	0.774	0.829	0.898
	(0.487)	(0.440)	(0.396)	(0.354)	(0.314)	(0.274)	(0.235)	(0.193)	(0.143)
	0.069	0.140	0.212	0.286	0.365	0.450	0.544	0.654	0.792
	[illegible]	(0.880)	(0.793)	(0.710)	(0.629)	(0.550)	(0.470)	(0.387)	(0.287)
4.2	0.637	0.657	0.679	0.704	0.732	0.764	0.803	0.849	0.910
	(0.395)	(0.366)	(0.337)	(0.309)	(0.279)	(0.250)	(0.218)	(0.183)	(0.137)
	0.065	0.133	0.204	0.279	0.359	0.446	0.542	0.654	0.793
	(0.930)	(0.873)	(0.810)	(0.741)	(0.668)	(0.592)	(0.510)	(0.421)	(0.311)
4.8	0.815	0.811	0.811	0.814	0.822	0.836	0.857	0.887	0.931
	(0.320)	(0.293)	(0.267)	(0.242)	(0.220)	(0.199)	(0.179)	(0.157)	(0.123)
	0.051	0.111	0.180	0.256	0.340	0.431	0.533	0.650	0.793
	(0.793)	(0.837)	(0.842)	(0.816)	(0.766)	(0.697)	(0.611)	(0.506)	[illegible]
5.4	0.959	0.948	0.938	0.930	0.925	0.924	0.928	0.939	0.961
	(0.202)	(0.207)	(0.201)	(0.188)	(0.169)	(0.147)	(0.128)	(0.114)	(0.093)
	0.028	0.074	0.137	0.214	0.304	0.404	0.517	0.643	0.792
	(0.553)	(0.771)	(0.918)	(0.995)	(1.008)	(0.964)	(0.869)	(0.724)	(0.517)

For the mastery score = 3 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 6
Mastery score C = 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.6	1.000	0.999	0.998	0.996	0.992	0.986	0.979	0.972	0.973
	(0.006)	(0.013)	(0.024)	(0.039)	(0.058)	(0.0[illegible])	(0.088)	(0.081)	(0.059)
	0.001	0.009	0.029	0.069	0.137	0.235	0.366	0.532	0.737
	(0.048)	(0.175)	(0.381)	(0.631)	(0.871)	(1.045)	(1.101)	(1.001)	(0.714)
1.2	0.994	0.991	0.985	0.978	0.969	0.958	0.946	0.939	0.946
	(0.047)	(0.065)	(0.086)	(0.107)	(0.125)	(0.135)	(0.129)	(0.105)	(0 080)
	0.006	0.022	0.054	0.106	0.181	0.280	0.406	0.559	0.748
	(0.143)	(0.302)	(0.482)	(0.650)	(0.773)	(0.829)	(0.804)	(0.693)	(0.488)
1.8	0.971	0.962	0.951	0.938	0.925	[illegible].912	0.902	0.902	0.923
	(0.142)	(0.16[illegible])	(0.176)	(0.185)	(0.184)	(0.172)	(0.147)	(0.116)	(0.097)
	0.015	0.042	0.086	0.147	0.226	0.324	0.442	0.583	0.757
	(0.291)	(0.446)	(0.582)	(0.681)	(0.730)	(0.724)	(0.663)	(0.552)	(0.389)
2.4	0.909	0.895	0.882	0.869	0.859	0.852	0.853	0.866	0.905
	(0.261)	(0.258)	(0.249)	(0.233)	(0.211)	(0.182)	(0.152)	(0.128)	(0.114)
	0.028	0.068	0.121	0.188	0.269	0.364	0.474	0.604	0.766
	(0.472)	(0.584)	(0.661)	(0.698)	(0.694)	(0.651)	(0.575)	(0.469)	(0.335)
3.0	0.795	0.787	0.781	0.779	0.781	0.789	0.807	0.838	0.893
	(0.320)	(0.293)	(0.266)	(0.239)	(0.212)	(0.188)	(0.167)	(0.150)	(0.131)
	0.042	0.095	0.156	0.227	0.307	0.398	0.502	0.623	0.773
	(0.661)	(0.706)	(0.719)	(0.704)	(0.662)	(0.599)	(0.517)	(0.420)	(0.305)
3.6	0.649	0.659	0.673	0.690	0.712	0.739	0.775	0.823	0.890
	(0.321)	(0.301)	(0.282)	(0.264)	(0.246)	(0.227)	(0.206)	(0.181)	(0.146)
	0.057	0.119	0.187	0.260	0.339	0.426	0.524	0.638	0.780
	(0.831)	(0.805)	(0.763)	(0.708)	(0.642)	(0.568)	(0.486)	(0.397)	(0.294)
4.2	0.543	0.575	0.608	0.643	0.681	0.723	0.771	0.827	0.898
	(0.447)	(0.415)	(0.383)	(0.351)	(0.318)	(0.284)	(0.248)	(0.207)	(0.155)
	0.068	0.137	0.208	0.283	0.362	0.447	0.541	0.65[illegible]	0.786
	(0.959)	(0.880)	(0.802)	(0.724)	(0.647)	(0.569)	(0.488)	(0.403)	(0.303)
4.8	0.581	0.614	0.647	0.683	0.720	0.761	0.805	0.856	0.918
	(0.509)	(0.463)	(0.420)	(0.379)	(0.339)	(0.300)	(0.258)	(0.212)	(0.152)
	0.071	0.144	0.217	0.293	0.373	0.458	0.551	0.658	0.791
	(1.017)	(0.935)	(0.855)	(0.778)	(0.702)	(0.625)	(0.544)	(0.454)	(0.343)
5.4	0.798	0.803	0.811	0.823	0.839	0.859	0.883	0.914	0.952
	(0.344)	(0.318)	(0.295)	(0.274)	(0.255)	(0.234)	(0.210)	(0.177)	(0.126)
	0.062	0.130	0.204	0.283	0.367	0.457	0.554	0.663	0.795
	(0.967)	(0.996)	(0.990)	(0.957)	(0.903)	(0.829)	(0.736)	(0.617)	(0.462)

For the mastery score = 2 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 6
Mastery score C = 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.6	1.000	1.000	1.000	0.999	0.999	0.997	0.993	0.986	0.978
	(0.000)	(0.001)	(0.003)	(0.007)	(0.014)	(0.028)	(0.049)	(0.070)	(0.063)
	0.000	0.001	0.007	0.022	0.056	0.121	0.231	0.399	0.644
	(0.005)	(0.035)	(0.119)	(0.275)	(0.503)	(0.771)	(1.010)	(1.109)	(0.918)
1.2	1.000	0.999	0.998	0.997	0.994	0.988	0.979	0.965	0.953
	(0.004)	(0.008)	(0.015)	(0.026)	(0.042)	(0.065)	(0.091)	(0.105)	(0.081)
	0.001	0.004	0.014	0.038	0.082	0.156	0.270	0.434	0.663
	(0.022)	(0.078)	(0.182)	(0.332)	(0.509)	(0.680)	(0.797)	(0.801)	(0.628)
1.8	0.997	0.996	0.993	0.988	0.981	0.970	0.955	0.937	0.929
	(0.022)	(0.032)	(0.047)	(0.066)	(0.089)	(0.113)	(0.131)	(0.127)	(0.088)
	0.002	0.010	0.027	0.060	0.113	0.195	0.311	0.469	0.681
	(0.063)	(0.148)	(0.268)	(0.409)	(0.548)	(0.656)	(0.703)	(0.658)	(0.496)
2.4	0.988	0.983	0.976	0.967	0.954	0.939	0.920	0.903	0.905
	(0.068)	(0.086)	(0.106)	(0.128)	(0.148)	(0.162)	(0.161)	(0.135)	(0.094)
	0.006	0.021	0.047	0.089	0.151	0.238	0.353	0.503	0.698
	(0.137)	(0.245)	(0.368)	(0.488)	(0.586)	(0.643)	(0.641)	(0.567)	(0.418)
3.0	0.961	0.951	0.939	0.925	0.908	0.890	0.874	0.866	0.885
	(0.154)	(0.172)	(0.188)	(0.200)	(0.203)	(0.195)	(0.171)	(0.129)	(0.106)
	0.014	0.037	0.073	0.125	0.194	0.283	0.395	0.535	0.715
	(0.253)	(0.366)	(0.474)	(0.561)	(0.616)	(0.628)	(0.591)	(0.503)	(0.368)
3.6	0.898	0.884	0.869	0.854	0.839	0.827	0.822	0.831	0.873
	(0.263)	(0.265)	(0.260)	(0.248)	(0.227)	(0.196)	(0.159)	(0.130)	(0.131)
	0.024	0.059	0.106	0.166	0.240	0.330	0.437	0.567	0.730
	(0.410)	(0.505)	(0.579)	(0.625)	(0.637)	(0.613)	(0.552)	(0.458)	(0.338)
4.2	0.781	0.770	0.762	0.756	0.755	0.760	0.776	0.809	0.872
	(0.323)	(0.297)	(0.269)	(0.239)	(0.209)	(0.184)	(0.169)	(0.166)	(0.163)
	0.039	0.087	0.144	0.211	0.288	0.377	0.478	0.597	0.745
	(0.606)	(0.658)	(0.684)	(0.683)	(0.656)	(0.604)	(0.528)	(0.433)	(0.327)
4.8	0.620	0.630	0.644	0.662	0.687	0.718	0.759	0.814	0.889
	(0.297)	(0.285)	(0.277)	(0.272)	(0.268)	(0.264)	(0.254)	(0.235)	(0.190)
	0.056	0.118	0.185	0.258	0.337	0.423	0.517	0.625	0.758
	(0.836)	(0.825)	(0.797)	(0.751)	(0.691)	(0.618)	(0.534)	(0.441)	(0.343)
5.4	0.542	0.583	0.625	0.668	0.714	0.761	0.812	0.867	0.928
	(0.596)	(0.570)	(0.538)	(0.500)	(0.457)	(0.408)	(0.349)	(0.279)	(0.188)
	0.076	0.151	0.228	0.305	0.385	0.467	0.554	0.651	0.771
	(1.114)	(1.047)	(0.974)	(0.895)	(0.812)	(0.724)	(0.631)	(0.532)	(0.428)

For the mastery score = 1 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 7
Mastery score C = 4

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.7	0.990	0.985	0.978	0.970	0.961	0.953	0.949	0.951	0.964
	(0.081)	(0.104)	(0.123)	(0.136)	(0.139)	(0.131)	(0.113)	(0.091)	(0.076)
	0.011	0.039	0.087	0.156	0.244	0.349	0.471	0.610	0.775
	(0.274)	(0.516)	(0.738)	(0.901)	(0.986)	(0.992)	(0.919)	(0.772)	(0.541)
1.4	0.923	0.911	0.900	0.890	0.883	0.881	0.886	0.901	0.934
	(0.251)	(0.247)	(0.235)	(0.217)	(0.195)	(0.169)	(0.145)	(0.124)	(0.103)
	0.031	0.077	0.136	0.209	0.294	0.391	0.500	0.626	0.779
	(0.537)	(0.675)	(0.760)	(0.793)	(0.780)	(0.728)	(0.644)	(0.529)	(0.376)
2.1	0.775	0.772	0.774	0.779	0.788	0.804	0.826	0.860	0.911
	(0.323)	(0.296)	(0.270)	(0.245)	(0.221)	(0.199)	(0.176)	(0.152)	(0.121)
	0.050	0.109	0.176	0.250	0.331	0.420	0.521	0.637	0.782
	(0.758)	(0.779)	(0.768)	(0.733)	(0.678)	(0.607)	(0.524)	(0.428)	(0.309)
2.8	0.608	0.630	0.654	0.680	0.710	0.744	0.784	0.832	0.897
	(0.387)	(0.359)	(0.331)	(0.302)	(0.272)	(0.241)	(0.209)	(0.174)	(0.131)
	0.064	0.130	0.200	0.274	0.353	0.438	0.533	0.643	0.784
	(0.897)	(0.835)	(0.768)	(0.697)	(0.623)	(0.546)	(0.466)	(0.379)	(0.278)
3.5	0.534	0.569	0.604	0.641	0.680	0.722	0.768	0.823	0.892
	(0.472)	(0.426)	(0.383)	(0.342)	(0.303)	(0.263)	(0.224)	(0.182)	(0.134)
	0.068	0.138	0.209	0.282	0.360	0.443	0.537	0.645	0.784
	(0.945)	(0.853)	(0.767)	(0.685)	(0.605)	(0.527)	(0.448)	(0.365)	(0.269)
4.2	0.608	0.630	0.654	0.680	0.710	0.744	0.784	0.832	0.897
	(0.387)	(0.359)	(0.331)	(0.302)	(0.272)	(0.241)	(0.209)	(0.174)	(0.131)
	0.064	0.130	0.200	0.274	0.353	0.438	0.533	0.643	0.784
	(0.897)	(0.835)	(0.768)	(0.697)	(0.623)	(0.546)	(0.466)	(0.379)	(0.278)
4.9	0.775	0.772	0.774	0.779	0.788	0.804	0.826	0.860	0.911
	(0.323)	(0.296)	(0.270)	(0.245)	(0.221)	(0.199)	(0.176)	(0.152)	(0.121)
	0.050	0.109	0.176	0.250	0.331	0.420	0.521	0.637	0.782
	(0.758)	(0.779)	(0.768)	(0.733)	(0.678)	(0.607)	(0.524)	(0.428)	(0.309)
5.6	0.923	0.911	0.900	0.890	0.883	0.881	0.886	0.901	0.934
	(0.251)	(0.247)	(0.235)	(0.217)	(0.195)	(0.169)	(0.145)	(0.124)	(0.103)
	0.031	0.077	0.136	0.209	0.294	0.391	0.500	0.626	0.779
	(0.537)	(0.675)	(0.760)	(0.793)	(0.780)	(0.728)	(0.644)	(0.529)	(0.376)
6.3	0.990	0.985	0.978	0.970	0.961	0.953	0.949	0.951	0.964
	(0.031)	(0.104)	(0.123)	(0.136)	(0.139)	(0.131)	(0.113)	(0.091)	(0.076)
	0.011	0.039	0.087	0.156	0.244	0.349	0.471	0.610	0.775
	(0.274)	(0.516)	(0.738)	(0.901)	(0.986)	(0.992)	(0.919)	(0.772)	(0.541)

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 7
Mastery score C = 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.7	0.999	0.998	0.996	0.992	0.987	0.980	0.972	0.966	0.970
	(0.014)	(0.025)	(0.041)	(0.060)	(0.080)	(0.095)	(0.098)	(0.083)	(0.062)
	0.003	0.014	0.041	0.092	0.168	0.272	0.403	0.561	0.751
	(0.082)	(0.249)	(0.479)	(0.721)	(0.918)	(1.028)	(1.025)	(0.895)	(0.625)
1.4	0.986	0.979	0.971	0.961	0.949	0.938	0.929	0.926	0.942
	(0.093)	(0.115)	(0.135)	(0.150)	(0.156)	(0.152)	(0.133)	(0.105)	(0.084)
	0.011	0.035	0.077	0.138	0.220	0.321	0.443	0.586	0.760
	(0.237)	(0.417)	(0.588)	(0.719)	(0.791)	(0.796)	(0.736)	(0.613)	(0.427)
2.1	0.932	0.920	0.907	0.894	0.884	0.876	0.875	0.886	0.918
	(0.230)	(0.234)	(0.231)	(0.220)	(0.201)	(0.176)	(0.147)	(0.121)	(0.102)
	0.025	0.064	0.118	0.186	0.268	0.365	0.476	0.607	0.767
	(0.443)	(0.577)	(0.672)	(0.719)	(0.719)	(0.677)	(0.597)	(0.486)	(0.342)
2.8	0.815	0.807	0.801	0.798	0.799	0.807	0.823	0.851	0.901
	(0.316)	(0.291)	(0.265)	(0.238)	(0.212)	(0.186)	(0.163)	(0.142)	(0.118)
	0.042	0.095	0.157	0.228	0.309	0.400	0.503	0.623	0.774
	(0.653)	(0.705)	(0.721)	(0.706)	(0.663)	(0.598)	(0.515)	(0.416)	(0.297)
3.5	0.657	0.668	0.682	0.699	0.721	0.748	0.783	0.828	0.892
	(0.330)	(0.308)	(0.287)	(0.266)	(0.244)	(0.221)	(0.196)	(0.167)	(0.131)
	0.057	0.120	0.188	0.261	0.339	0.426	0.523	0.635	0.778
	(0.826)	(0.795)	(0.749)	(0.692)	(0.624)	(0.549)	(0.468)	(0.379)	(0.276)
4.2	0.544	0.575	0.609	0.643	0.681	0.722	0.767	0.822	0.892
	(0.444)	(0.407)	(0.370)	(0.334)	(0.299)	(0.263)	(0.225)	(0.186)	(0.138)
	0.067	0.136	0.206	0.280	0.357	0.441	0.535	0.644	0.782
	(0.932)	(0.848)	(0.768)	(0.689)	(0.611)	(0.533)	(0.454)	(0.370)	(0.274)
4.9	0.573	0.603	0.634	0.668	0.703	0.742	0.786	0.837	0.902
	(0.456)	(0.415)	(0.376)	(0.338)	(0.302)	(0.265)	(0.227)	(0.187)	(0.137)
	0.068	0.137	0.209	0.283	0.361	0.446	0.539	0.648	0.785
	(0.948)	(0.867)	(0.788)	(0.710)	(0.634)	(0.557)	(0.478)	(0.394)	(0.292)
5.6	0.749	0.754	0.762	0.773	0.789	0.811	0.838	0.874	0.924
	(0.339)	(0.313)	(0.288)	(0.264)	(0.241)	(0.218)	(0.194)	(0.166)	(0.126)
	0.057	0.121	0.191	0.267	0.348	0.437	0.535	0.647	0.786
	(0.851)	(0.849)	(0.823)	(0.777)	(0.717)	(0.646)	(0.563)	(0.466)	(0.343)
6.3	0.938	0.927	0.918	0.911	0.908	0.909	0.916	0.931	0.957
	(0.238)	(0.233)	(0.220)	(0.200)	(0.178)	(0.157)	(0.138)	(0.122)	(0.098)
	0.034	0.084	0.149	0.227	0.315	0.412	0.520	0.642	0.787
	(0.616)	(0.794)	(0.903)	(0.948)	(0.941)	(0.889)	(0.797)	(0.665)	(0.479)

For the mastery score = 3 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 7
Mastery score C = 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.7	1.000	1.000	0.999	0.998	0.997	0.993	0.987	0.979	0.975
	(0.001)	(0.003)	(0.008)	(0.016)	(0.030)	(0.04[illegible])	(0.069)	(0.077)	(0.057)
	0.000	0.003	0.015	0.042	0.096	0.183	0.311	0.484	0.706
	(0.015)	(0.081)	(0.226)	(0.446)	(0.703)	(0.934)	(1.064)	(1.021)	(0.747)
1.4	0.998	0.997	0.994	0.990	0.984	0.975	0.963	0.951	0.949
	(0.016)	(0.027)	(0.042)	(0.061)	(0.083)	(0.104)	(0.115)	(0.105)	(0.074)
	0.002	0.011	0.032	0.072	0.136	0.230	0.356	0.518	0.721
	(0.064)	(0.175)	(0.334)	(0.515)	(0.678)	(0.785)	(0.804)	(0.715)	(0.506)
2.1	0.989	0.983	0.976	0.966	0.954	0.940	0.925	0.916	0.925
	(0.072)	(0.092)	(0.113)	(0.133)	(0.149)	(0.155)	(0.146)	(0.117)	(0.086)
	0.008	0.025	0.058	0.109	0.182	0.278	0.399	0.548	0.734
	(0.166)	(0.305)	(0.455)	(0.588)	(0.680)	(0.712)	(0.676)	(0.571)	(0.399)
2.8	0.953	0.942	0.929	0.915	0.900	0.887	0.877	0.878	0.904
	(0.181)	(0.196)	(0.205)	(0.208)	(0.201)	(0.183)	(0.155)	(0.120)	(0.100)
	0.018	0.047	0.092	0.152	0.229	0.324	0.439	0.575	0.746
	(0.322)	(0.454)	(0.565)	(0.641)	(0.672)	(0.654)	(0.589)	(0.482)	(0.338)
3.5	0.869	0.856	0.843	0.832	0.824	0.8[illegible]1	0.826	0.844	0.890
	(0.292)	(0.231)	(0.264)	(0.241)	(0.214)	(0.134)	(0.155)	(0.132)	(0.117)
	0.032	0.075	0.130	0.196	0.275	0.367	0.474	0.599	0.756
	(0.513)	(0.599)	(0.652)	(0.670)	(0.653)	(0.604)	(0.526)	(0.424)	(0.302)
4.2	0.728	0.726	0.727	0.731	0.741	0.757	0.783	0.821	0.884
	(0.315)	(0.287)	(0.262)	(0.238)	(0.217)	(0.197)	(0.179)	(0.161)	(0.136)
	0.048	0.103	0.166	0.237	0.316	0.404	0.504	0.620	0.766
	(0.712)	(0.727)	(0.717)	(0.685)	(0.634)	(0.566)	(0.485)	(0.392)	(0.286)
4.9	0.578	0.600	0.625	0.[illegible]53	0.684	0.721	0.765	0.818	0.889
	(0.362)	(0.344)	(0.325)	(0.304)	(0.282)	(0.257)	(0.229)	(0.196)	(0.150)
	0.062	0.128	0.197	0.270	0.348	0.433	0.527	0.636	0.774
	(0.884)	(0.829)	(0.767)	(0.700)	(0.629)	(0.554)	(0.474)	(0.388)	(0.290)
5.6	0.548	0.584	0.621	0.659	0.699	0.742	0.789	0.843	0.909
	(0.513)	(0.467)	(0.423)	(0.382)	(0.341)	(0.301)	(0.259)	(0.213)	(0.153)
	0.071	0.142	0.215	0.289	0.368	0.451	0.543	0.649	0.781
	(0.990)	(0.904)	(0.822)	(0.744)	(0.668)	(0.592)	(0.513)	(0.427)	(0.323)
6.3	0.753	0.764	0.777	0.794	0.815	0.839	0.868	0.903	0.946
	(0.376)	(0.349)	(0.324)	(0.300)	(0.276)	(0.251)	(0.222)	(0.185)	(0.131)
	0.065	0.135	0.209	0.286	0.368	0.456	0.551	0.657	0.786
	(0.977)	(0.972)	(0.944)	(0.900)	(0.841)	(0.769)	(0.681)	(0.573)	(0.433)

For the mastery score = 2 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 7
Mastery score C = 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.7	1.000	1.000	1.000	1.000	1.000	0.999	0.996	0.990	0.981
	(0.000)	(0.000)	(0.001)	(0.002)	(0.006)	(0.01[illegible])	(0.031)	(0.057)	(0.064)
	0.000	0.000	0.003	0.012	0.036	0.088	0.184	0.347	0.604
	(0.001)	(0.014)	(0.060)	(0.168)	(0.356)	(0.616)	(0.893)	(1.068)	(0.940)
1.4	1.000	1.000	0.999	0.999	0.997	0.994	0.987	0.974	0.958
	(0.001)	(0.003)	(0.006)	(0.011)	(0.022)	(0.040)	(0.066)	(0.093)	(0.084)
	0.000	0.002	0.007	0.023	0.056	0.118	0.223	0.386	0.627
	(0.008)	(0.038)	(0.107)	(0.227)	(0.392)	(0.578)	(0.736)	(0.790)	(0.644)
2.1	0.999	0.998	0.997	0.994	0.990	0.982	0.969	0.950	0.934
	(0.009)	(0.014)	(0.023)	(0.036)	(0.055)	(0.080)	(0.108)	(0.122)	(0.091)
	0.001	0.005	0.016	0.040	0.083	0.155	0.265	0.425	0.649
	(0.030)	(0.085)	(0.179)	(0.307)	(0.453)	(0.568)	(0.672)	(0.660)	(0.508)
2.8	0.995	0.992	0.988	0.982	0.972	0.959	0.940	0.919	0.909
	(0.035)	(0.048)	(0.064)	(0.085)	(0.109)	(0.132)	(0.148)	(0.139)	(0.092)
	0.003	0.013	0.031	0.064	0.118	0.198	0.311	0.464	0.670
	(0.078)	(0.162)	(0.272)	(0.396)	(0.514)	(0.600)	(0.628)	(0.574)	(0.425)
3.5	0.979	0.972	0.963	0.951	0.936	0.918	0.898	0.880	0.886
	(0.098)	(0.117)	(0.137)	(0.157)	(0.173)	(0.181)	(0.172)	(0.138)	(0.096)
	0.009	0.025	0.054	0.098	0.160	0.246	0.358	0.502	0.690
	(0.168)	(0.271)	(0.382)	(0.486)	(0.566)	(0.604)	(0.589)	(0.510)	(0.370)
4.2	0.935	0.922	0.908	0.892	0.875	0.857	0.844	0.841	0.870
	(0.205)	(0.218)	(0.227)	(0.229)	(0.222)	(0.203)	(0.170)	(0.128)	(0.114)
	0.018	0.045	0.085	0.139	0.209	0.297	0.406	0.539	0.709
	(0.308)	(0.410)	(0.500)	(0.568)	(0.604)	(0.600)	(0.552)	(0.461)	(0.335)
4.9	0.835	0.821	0.808	0.796	0.787	0.783	0.788	0.810	0.865
	(0.309)	(0.295)	(0.275)	(0.250)	(0.219)	(0.186)	(0.157)	(0.144)	(0.148)
	0.032	0.073	0.124	0.187	0.262	0.350	0.453	0.575	0.728
	(0.501)	(0.572)	(0.620)	(0.641)	(0.632)	(0.593)	(0.524)	(0.430)	(0.318)
5.6	0.667	0.668	0.673	0.683	0.699	0.722	0.755	0.804	0.878
	(0.297)	(0.271)	(0.251)	(0.237)	(0.228)	(0.224)	(0.221)	(0.213)	(0.184)
	0.050	0.106	0.170	0.240	0.318	0.404	0.499	0.609	0.745
	(0.743)	(0.754)	(0.744)	(0.715)	(0.666)	(0.601)	(0.521)	(0.428)	(0.328)
6.3	0.536	0.573	0.611	0.653	0.697	0.744	0.796	0.853	0.919
	(0.517)	(0.504)	(0.485)	(0.459)	(0.428)	(0.389)	(0.341)	(0.278)	(0.193)
	0.072	0.145	0.219	0.295	0.374	0.456	0.543	0.641	0.761
	(1.043)	(0.985)	(0.920)	(0.848)	(0.770)	(0.687)	(0.599)	(0.504)	(0.402)

For the mastery score = 1 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 8
Mastery score C = 4

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	0.984	0.977	0.968	0.959	0.950	0.943	0.940	0.944	0.961
	(0.112)	(0.133)	(0.149)	(0.155)	(0.152)	(0.1[illegible])	(0.118)	(0.097)	(0.080)
	0.015	0.043	0.100	0.171	0.259	0.363	0.481	0.615	0.773
	(0.334)	(0.568)	(0.763)	(0.892)	(0.947)	(0.931)	(0.852)	(0.712)	(0.502)
1.6	0.881	0.871	0.862	0.856	0.854	0.858	0.869	0.890	0.928
	(0.290)	(0.273)	(0.251)	(0.227)	(0.202)	(0.177)	(0.154)	(0.133)	(0.107)
	0.039	0.090	0.153	0.227	0.311	0.404	0.509	0.629	0.776
	(0.627)	(0.724)	(0.770)	(0.773)	(0.741)	(0.680)	(0.595)	(0.488)	(0.350)
2.4	0.693	0.703	0.715	0.731	0.751	0.776	0.807	0.848	0.905
	(0.342)	(0.317)	(0.293)	(0.268)	(0.244)	(0.218)	(0.191)	(0.161)	(0.123)
	0.058	0.122	0.190	0.264	0.343	0.429	0.525	0.637	0.778
	(0.833)	(0.807)	(0.765)	(0.709)	(0.643)	(0.570)	(0.488)	(0.398)	(0.290)
3.2	0.549	0.581	0.615	0.649	0.686	0.726	0.771	0.824	0.892
	(0.451)	(0.409)	(0.369)	(0.331)	(0.293)	(0.256)	(0.217)	(0.177)	(0.130)
	0.067	0.136	0.206	0.279	0.356	0.439	0.532	0.640	0.778
	(0.923)	(0.838)	(0.756)	(0.677)	(0.600)	(0.522)	(0.444)	(0.360)	(0.264)
4.0	0.564	0.592	0.622	0.653	0.688	0.726	0.769	0.821	0.889
	(0.414)	(0.381)	(0.348)	(0.315)	(0.281)	(0.247)	(0.212)	(0.173)	(0.128)
	0.065	0.133	0.202	0.275	0.352	0.436	0.529	0.637	0.777
	(0.901)	(0.825)	(0.749)	(0.673)	(0.597)	(0.520)	(0.440)	(0.356)	(0.260)
4.8	0.714	0.717	0.724	0.735	0.751	0.771	0.799	0.833	0.896
	(0.324)	(0.299)	(0.275)	(0.252)	(0.229)	(0.206)	(0.181)	(0.154)	(0.120)
	0.054	0.114	0.180	0.253	0.332	0.419	0.516	0.630	0.774
	(0.777)	(0.769)	(0.739)	(0.691)	(0.630)	(0.557)	(0.474)	(0.382)	(0.275)
5.6	0.878	0.866	0.855	0.847	0.843	0.844	0.852	0.872	0.913
	(0.290)	(0.275)	(0.255)	(0.232)	(0.206)	(0.179)	(0.153)	(0.130)	(0.107)
	0.035	0.083	0.143	0.215	0.297	0.389	0.495	0.617	0.770
	(0.572)	(0.665)	(0.713)	(0.720)	(0.691)	(0.631)	(0.547)	(0.442)	(0.313)
6.4	0.971	0.962	0.951	0.939	0.928	0.918	0.912	0.915	0.937
	(0.147)	(0.165)	(0.177)	(0.181)	(0.176)	(0.161)	(0.137)	(0.109)	(0.088)
	0.017	0.049	0.098	0.164	0.248	0.348	0.464	0.600	0.764
	(0.330)	(0.507)	(0.652)	(0.745)	(0.778)	(0.753)	(0.678)	(0.557)	(0.388)
7.2	0.998	0.996	0.992	0.987	0.981	0.973	0.965	0.961	0.967
	(0.025)	(0.040)	(0.059)	(0.080)	(0.098)	(0.108)	(0.104)	(0.085)	(0.065)
	0.004	0.019	0.053	0.109	0.191	0.296	0.425	0.576	0.756
	(0.119)	(0.312)	(0.548)	(0.767)	(0.924)	(0.990)	(0.955)	(0.817)	(0.568)

For the mastery score = 5 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 8
Mastery score C = 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	0.998	0.996	0.992	0.987	0.981	0.973	0.965	0.961	0.967
	(0.025)	(0.040)	(0.059)	(0.080)	(0.098)	(0.108)	(0.104)	(0.085)	(0.065)
	0.004	0.019	0.053	0.109	0.191	0.296	0.425	0.576	0.756
	(0.119)	(0.312)	(0.548)	(0.767)	(0.924)	(0.990)	(0.955)	(0.817)	(0.568)
1.6	0.971	0.962	0.951	0.939	0.928	0.918	0.912	0.915	0.937
	(0.147)	(0.165)	(0.177)	(0.181)	(0.176)	(0.161)	(0.137)	(0.109)	(0.068)
	0.017	0.049	0.098	0.164	0.248	0.348	0.464	0.600	0.764
	(0.330)	(0.507)	(0.652)	(0.745)	(0.778)	(0.753)	(0.678)	(0.557)	(0.388)
2.4	0.878	0.866	0.855	0.847	0.843	0.844	0.852	0.872	0.913
	(0.290)	(0.275)	(0.255)	(0.232)	(0.206)	(0.179)	(0.153)	(0.130)	(0.107)
	0.035	0.083	0.143	[illegible]	0.297	0.389	0.495	0.617	0.770
	(0.572)	(0.665)	(0.713)	(0.720)	(0.691)	(0.631)	(0.547)	(0.442)	(0.313)
3.2	0.714	0.717	0.724	0.735	0.751	0.771	0.799	0.838	0.896
	(0.324)	(0.299)	(0.275)	(0.252)	(0.229)	(0.206)	(0.181)	(0.154)	(0.120)
	0.054	0.114	0.180	0.253	0.332	0.419	0.516	0.630	0.774
	(0.777)	(0.769)	(0.739)	(0.691)	(0.630)	(0.557)	(0.474)	(0.382)	(0.275)
4.0	0.564	0.592	0.622	0.653	0.688	0.726	0.769	0.821	0.889
	(0.414)	(0.381)	(0.348)	(0.315)	(0.281)	(0.247)	(0.212)	(0.173)	(0.128)
	0.065	0.133	0.202	0.275	0.352	0.436	0.529	0.637	0.777
	(0.901)	(0.825)	(0.749)	(0.673)	(0.597)	(0.520)	(0.440)	(0.356)	(0.260)
4.8	0.549	0.581	0.615	0.649	0.686	0.726	0.771	0.824	0.892
	(0.451)	(0.409)	(0.369)	(0.331)	(0.293)	(0.256)	(0.217)	(0.177)	(0.130)
	0.067	0.136	0.206	0.279	0.356	0.439	0.532	0.640	0.778
	(0.923)	(0.838)	(0.756)	(0.677)	(0.600)	(0.522)	(0.444)	(0.360)	(0.264)
5.6	0.693	0.703	0.715	0.731	0.751	0.776	0.807	0.848	0.905
	(0.342)	(0.317)	(0.293)	(0.268)	(0.244)	(0.218)	(0.191)	(0.161)	(0.123)
	0.058	0.122	0.190	0.264	0.343	0.429	0.525	0.637	0.778
	(0.833)	(0.807)	(0.765)	(0.709)	(0.643)	(0.570)	(0.488)	(0.398)	(0.290)
6.4	0.881	0.871	0.862	0.856	0.854	0.858	0.869	0.890	0.928
	(0.290)	(0.273)	(0.251)	(0.227)	(0.202)	(0.177)	(0.154)	(0.133)	(0.107)
	0.039	0.090	0.153	0.227	0.311	0.404	0.509	0.629	0.776
	(0.627)	(0.724)	(0.770)	(0.773)	(0.741)	(0.680)	(0.595)	(0.488)	(0.350)
7.2	0.984	0.977	0.968	0.959	0.950	0.943	0.940	0.944	0.961
	(0.112)	(0.133)	(0.149)	(0.155)	(0.152)	(0.139)	(0.118)	(0.097)	(0.080)
	0.015	0.048	0.100	0.171	0.259	0.363	0.481	0.615	0.773
	(0.334)	(0.568)	(0.763)	(0.892)	(0.947)	(0.931)	(0.852)	(0.712)	(0.502)

For the mastery score = 4 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 8
Mastery score C = 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	1.000	0.999	0.999	0.997	0.994	0.989	0.982	0.974	0.972
	(0.003)	(0.008)	(0.016)	(0.028)	(0.046)	(0.066)	(0.082)	(0.080)	(0.057)
	0.001	0.006	0.023	0.060	0.124	0.222	0.354	0.521	0.727
	(0.029)	(0.128)	(0.312)	(0.552)	(0.791)	(0.967)	(1.025)	(0.930)	(0.656)
1.6	0.996	0.992	0.988	0.981	0.972	0.96[illegible]	0.948	0.939	0.945
	(0.038)	(0.055)	(0.075)	(0.097)	(0.116)	(0.1[illegible]3)	(0.126)	(0.105)	(0.075)
	0.005	0.019	0.050	0.100	0.175	0.275	0.400	0.553	0.740
	(0.121)	(0.270)	(0.448)	(0.615)	(0.737)	(0.788)	(0.757)	(0.642)	(0.444)
2.4	0.970	0.960	0.949	0.936	0.923	0.910	0.900	0.899	0.920
	(0.143)	(0.162)	(0.176)	(0.184)	(0.183)	(0.171)	(0.147)	(0.115)	(0.090)
	0.015	0.043	0.087	0.148	0.227	0.325	0.442	0.580	0.750
	(0.286)	(0.438)	(0.572)	(0.664)	(0.705)	(0.690)	(0.622)	(0.507)	(0.350)
3.2	0.892	0.879	0.866	0.855	0.846	0.842	0.845	0.861	0.901
	(0.275)	(0.268)	(0.254)	(0.235)	(0.210)	(0.182)	(0.153)	(0.127)	(0.106)
	0.030	0.073	0.128	0.196	0.276	0.369	0.477	0.602	0.759
	(0.497)	(0.597)	(0.659)	(0.682)	(0.665)	(0.614)	(0.533)	(0.428)	(0.299)
4.0	0.747	0.744	0.745	0.749	0.758	0.772	0.796	0.831	0.889
	(0.317)	(0.290)	(0.265)	(0.240)	(0.217)	(0.195)	(0.173)	(0.150)	(0.121)
	0.048	0.103	0.167	0.238	0.317	0.405	0.504	0.620	0.767
	(0.706)	(0.723)	(0.713)	(0.679)	(0.627)	(0.557)	(0.475)	(0.381)	(0.272)
4.8	0.588	0.609	0.633	0.660	0.691	0.726	0.767	0.818	0.886
	(0.365)	(0.342)	(0.318)	(0.294)	(0.268)	(0.240)	(0.210)	(0.175)	(0.133)
	0.062	0.127	0.196	0.268	0.346	0.430	0.523	0.633	0.772
	(0.866)	(0.808)	(0.744)	(0.675)	(0.603)	(0.527)	(0.447)	(0.362)	(0.265)
5.6	0.540	0.574	0.610	0.646	0.685	0.727	0.773	0.827	0.895
	(0.476)	(0.430)	(0.388)	(0.347)	(0.308)	(0.269)	(0.229)	(0.187)	(0.137)
	0.069	0.138	0.209	0.282	0.359	0.442	0.534	0.641	0.777
	(0.940)	(0.852)	(0.769)	(0.689)	(0.612)	(0.536)	(0.458)	(0.375)	(0.278)
6.4	0.687	0.701	0.717	0.737	0.760	0.788	0.821	0.863	0.917
	(0.370)	(0.343)	(0.316)	(0.289)	(0.263)	(0.235)	(0.206)	(0.174)	(0.129)
	0.062	0.129	0.199	0.274	0.353	0.439	0.534	0.643	0.780
	(0.889)	(0.853)	(0.805)	(0.746)	(0.680)	(0.608)	(0.528)	(0.437)	(0.323)
7.2	0.915	0.904	0.896	0.891	0.890	0.894	0.904	0.923	0.952
	(0.267)	(0.253)	(0.233)	(0.211)	(0.183)	(0.166)	(0.148)	(0.130)	(0.102)
	0.039	0.093	0.159	0.237	0.323	0.418	0.522	0.640	0.781
	(0.668)	(0.809)	(0.886)	(0.908)	(0.887)	(0.830)	(0.741)	(0.619)	(0.450)

For the mastery score = 3 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 8
Mastery score C = 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	1.000	1.000	1.000	1.000	0.999	0.997	0.992	0.985	0.977
	(0.000)	(0.001)	(0.003)	(0.006)	(0.014)	(0.029)	(0.050)	(0.068)	(0.057)
	0.000	0.001	0.007	0.025	0.066	0.142	0.264	0.440	0.677
	(0.005)	(0.036)	(0.129)	(0.305)	(0.551)	(0.815)	(1.009)	(1.031)	(0.780)
1.6	1.000	0.999	0.998	0.996	0.992	0.985	0.975	0.961	0.953
	(0.005)	(0.010)	(0.019)	(0.031)	(0.050)	(0.073)	(0.096)	(0.102)	(0.073)
	0.001	0.005	0.018	0.048	0.101	0.187	0.311	0.478	0.695
	(0.027)	(0.097)	(0.222)	(0.394)	(0.577)	(0.726)	(0.792)	(0.734)	(0.527)
2.4	0.996	0.993	0.989	0.982	0.973	0.960	0.945	0.929	0.928
	(0.034)	(0.048)	(0.066)	(0.088)	(0.110)	(0.129)	(0.136)	(0.119)	(0.081)
	0.004	0.015	0.038	0.080	0.144	0.236	0.358	0.514	0.712
	(0.091)	(0.201)	(0.343)	(0.493)	(0.618)	(0.690)	(0.684)	(0.591)	(0.412)
3.2	0.977	0.969	0.959	0.947	0.932	0.916	0.900	0.891	0.905
	(0.112)	(0.133)	(0.152)	(0.168)	(0.177)	(0.175)	(0.157)	(0.122)	(0.090)
	0.011	0.032	0.068	0.121	0.193	0.287	0.404	0.547	0.727
	(0.212)	(0.342)	(0.470)	(0.576)	(0.641)	(0.652)	(0.604)	(0.499)	(0.345)
4.0	0.920	0.907	0.892	0.878	0.864	0.852	0.847	0.854	0.888
	(0.237)	(0.242)	(0.241)	(0.232)	(0.215)	(0.189)	(0.156)	(0.124)	(0.106)
	0.023	0.058	0.105	0.167	0.244	0.336	0.446	0.576	0.740
	(0.389)	(0.499)	(0.583)	(0.632)	(0.641)	(0.610)	(0.539)	(0.435)	(0.303)
4.8	0.798	0.788	0.781	0.776	0.775	0.780	0.795	0.824	0.879
	(0.317)	(0.293)	(0.266)	(0.239)	(0.211)	(0.185)	(0.162)	(0.144)	(0.126)
	0.039	0.088	0.146	0.214	0.292	0.381	0.483	0.602	0.752
	(0.599)	(0.650)	(0.671)	(0.663)	(0.628)	(0.570)	(0.490)	(0.394)	(0.282)
5.6	0.628	0.640	0.655	0.673	0.697	0.726	0.763	0.813	0.882
	(0.313)	(0.295)	(0.279)	(0.263)	(0.247)	(0.229)	(0.208)	(0.183)	(0.145)
	0.056	0.118	0.184	0.256	0.333	0.418	0.513	0.623	0.763
	(0.805)	(0.777)	(0.736)	(0.682)	(0.619)	(0.547)	(0.468)	(0.380)	(0.280)
6.4	0.535	0.570	0.606	0.644	0.685	0.728	0.776	0.832	0.901
	(0.482)	(0.444)	(0.406)	(0.369)	(0.332)	(0.295)	(0.256)	(0.211)	(0.154)
	0.069	0.139	0.210	0.284	0.361	0.444	0.535	0.640	0.772
	(0.956)	(0.874)	(0.795)	(0.719)	(0.645)	(0.570)	(0.493)	(0.4[illegible]8)	(0.309)
7.2	0.710	0.727	0.746	0.768	0.793	0.821	0.854	0.893	0.940
	(0.410)	(0.379)	(0.351)	(0.322)	(0.294)	(0.265)	(0.233)	(0.193)	(0.136)
	0.067	0.138	0.211	0.288	0.369	0.454	0.547	0.651	0.779
	(0.981)	(0.952)	(0.909)	(0.855)	(0.794)	(0.723)	(0.640)	(0.540)	(0.410)

For the mastery score = 2 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 8
Mastery score C = 6

Test Mean \ KR21=	.100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.994	0.984
	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)	(0.007)	(0.019)	(0.043)	(0.063)
	0.000	0.000	0.001	0.007	0.023	0.063	0.147	0.302	0.566
	(0.000)	(0.005)	(0.030)	(0.101)	(0.249)	(0.486)	(0.780)	(1.018)	(0.959)
1.6	1.000	1.000	1.000	1.000	0.999	0.997	0.992	0.982	0.963
	(0.000)	(0.001)	(0.002)	(0.005)	(0.011)	(0.023)	(0.046)	(0.078)	(0.086)
	0.000	0.001	0.004	0.014	0.038	0.089	0.184	0.343	0.593
	(0.003)	(0.018)	(0.062)	(0.152)	(0.297)	(0.485)	(0.671)	(0.772)	(0.660)
2.4	1.000	0.999	0.999	0.997	0.995	0.990	0.980	0.962	0.940
	(0.003)	(0.006)	(0.011)	(0.019)	(0.033)	(0.055)	(0.084)	(0.111)	(0.096)
	0.000	0.003	0.010	0.026	0.060	0.123	0.226	0.385	0.619
	(0.014)	(0.048)	(0.117)	(0.226)	(0.368)	(0.519)	(0.636)	(0.658)	(0.521)
3.2	0.998	0.996	0.994	0.990	0.983	0.973	0.956	0.933	0.914
	(0.017)	(0.025)	(0.037)	(0.054)	(0.076)	(0.103)	(0.128)	(0.137)	(0.096)
	0.002	0.007	0.020	0.046	0.091	0.164	0.273	0.427	0.644
	(0.044)	(0.105)	(0.198)	(0.317)	(0.444)	(0.554)	(0.611)	(0.581)	(0.435)
4.0	0.989	0.984	0.978	0.969	0.957	0.940	0.918	0.895	0.889
	(0.060)	(0.076)	(0.096)	(0.118)	(0.140)	(0.159)	(0.166)	(0.145)	(0.093)
	0.005	0.017	0.039	0.076	0.132	0.212	0.323	0.471	0.668
	(0.110)	(0.199)	(0.305)	(0.416)	(0.514)	(0.577)	(0.585)	(0.519)	(0.376)
4.8	0.959	0.949	0.936	0.922	0.904	0.884	0.865	0.853	0.869
	(0.152)	(0.170)	(0.187)	(0.200)	(0.206)	(0.200)	(0.177)	(0.134)	(0.102)
	0.013	0.035	0.068	0.116	0.181	0.267	0.376	0.513	0.691
	(0.230)	(0.331)	(0.429)	(0.513)	(0.570)	(0.586)	(0.554)	(0.468)	(0.335)
5.6	0.878	0.863	0.848	0.833	0.818	0.807	0.803	0.814	0.859
	(0.277)	(0.276)	(0.268)	(0.252)	(0.228)	(0.196)	(0.159)	(0.131)	(0.133)
	0.025	0.061	0.107	0.166	0.238	0.326	0.430	0.555	0.712
	(0.413)	(0.497)	(0.562)	(0.601)	(0.610)	(0.585)	(0.525)	(0.431)	(0.313)
6.4	0.713	0.708	0.706	0.708	0.715	0.730	0.756	0.798	0.869
	(0.310)	(0.281)	(0.253)	(0.228)	(0.209)	(0.197)	(0.194)	(0.193)	(0.176)
	0.044	0.096	0.156	0.224	0.300	0.386	0.483	0.594	0.733
	(0.661)	(0.691)	(0.699)	(0.684)	(0.647)	(0.590)	(0.513)	(0.420)	(0.317)
7.2	0.539	0.571	0.606	0.643	0.685	0.731	0.782	0.841	0.911
	(0.444)	(0.440)	(0.431)	(0.417)	(0.396)	(0.367)	(0.329)	(0.275)	(0.195)
	0.068	0.138	0.211	0.286	0.364	0.446	0.534	0.631	0.752
	(0.981)	(0.934)	(0.877)	(0.8[illegible])	(0.739)	(0.660)	(0.575)	(0.482)	(0.382)

For the mastery score = 1 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 9
Mastery score C = 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	0.996	0.993	0.988	0.982	0.974	0.966	0.958	0.955	0.964
	(0.039)	(0.058)	(0.079)	(0.099)	(0.114)	(0.119)	(0.110)	(0.088)	(0.068)
	0.006	0.025	0.063	0.124	0.208	0.314	0.440	0.585	0.758
	(0.159)	(0.367)	(0.597)	(0.791)	(0.914)	(0.949)	(0.896)	(0.758)	(0.528)
1.8	0.951	0.939	0.927	0.915	0.905	0.898	0.896	0.905	0.932
	(0.199)	(0.209)	(0.210)	(0.204)	(0.189)	(0.167)	(0.140)	(0.114)	(0.093)
	0.023	0.061	0.116	0.185	0.269	0.367	0.479	0.607	0.764
	(0.418)	(0.577)	(0.691)	(0.750)	(0.755)	(0.712)	(0.631)	(0.515)	(0.361)
2.7	0.812	0.805	0.801	0.801	0.805	0.814	0.831	0.860	0.907
	(0.315)	(0.290)	(0.264)	(0.238)	(0.213)	(0.188)	(0.164)	(0.139)	(0.110)
	0.045	0.099	0.163	0.235	0.316	0.405	0.506	0.622	0.769
	(0.676)	(0.722)	(0.730)	(0.708)	(0.660)	(0.593)	(0.509)	(0.411)	(0.293)
3.6	0.625	0.643	0.663	0.687	0.714	0.745	0.782	0.828	0.892
	(0.362)	(0.336)	(0.311)	(0.284)	(0.257)	(0.228)	(0.197)	(0.163)	(0.121)
	0.061	0.126	0.194	0.267	0.344	0.428	0.522	0.631	0.771
	(0.852)	(0.800)	(0.740)	(0.674)	(0.603)	(0.527)	(0.447)	(0.360)	(0.260)
4.5	0.534	0.568	0.603	0.639	0.677	0.718	0.764	0.817	0.886
	(0.457)	(0.412)	(0.370)	(0.331)	(0.292)	(0.253)	(0.214)	(0.172)	(0.125)
	0.067	0.136	0.205	0.278	0.354	0.436	0.527	0.634	0.772
	(0.913)	(0.824)	(0.741)	(0.661)	(0.583)	(0.506)	(0.428)	(0.345)	(0.251)
5.4	0.625	0.643	0.663	0.687	0.714	0.745	0.782	0.828	0.892
	(0.362)	(0.336)	(0.311)	(0.284)	(0.257)	(0.228)	(0.197)	(0.163)	(0.121)
	0.061	0.126	0.194	0.267	0.344	0.428	0.522	0.631	0.771
	(0.852)	(0.800)	(0.740)	(0.674)	(0.603)	(0.527)	(0.447)	(0.360)	(0.260)
6.3	0.812	0.805	0.801	0.801	0.805	0.814	0.831	0.860	0.907
	(0.315)	(0.290)	(0.264)	(0.238)	(0.213)	(0.188)	(0.164)	(0.139)	(0.110)
	0.045	0.099	0.163	0.235	0.316	0.405	0.506	0.622	0.769
	(0.676)	(0.722)	(0.730)	(0.708)	(0.660)	(0.593)	(0.509)	(0.411)	(0.293)
7.2	0.951	0.939	0.927	0.915	0.905	0.898	0.896	0.905	0.932
	(0.199)	(0.209)	(0.210)	(0.204)	(0.189)	(0.167)	(0.140)	(0.114)	(0.093)
	0.023	0.061	0.116	0.185	0.269	0.367	0.479	0.607	0.764
	(0.418)	(0.577)	(0.691)	(0.750)	(0.755)	(0.712)	(0.631)	(0.515)	(0.361)
8.1	0.996	0.993	0.988	0.982	0.974	0.966	0.958	0.955	0.964
	(0.039)	(0.058)	(0.079)	(0.099)	(0.114)	(0.119)	(0.110)	(0.088)	(0.068)
	0.006	0.025	0.063	0.124	0.208	0.314	0.440	0.585	0.758
	(0.159)	(0.367)	(0.597)	(0.791)	(0.914)	(0.949)	(0.896)	(0.758)	(0.528)

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 9
Mastery score C = 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	1.000	0.999	0.998	0.995	0.991	0.985	0.977	0.969	0.970
	(0.007)	(0.014)	(0.025)	(0.041)	(0.061)	(0.0[illegible])	(0.091)	(0.082)	(0.058)
	0.001	0.009	0.031	0.075	0.146	0.248	0.381	0.542	0.737
	(0.047)	(0.174)	(0.380)	(0.620)	(0.831)	(0.961)	(0.975)	(0.857)	(0.598)
1.8	0.990	0.985	0.978	0.968	0.957	0.945	0.934	0.929	0.940
	(0.070)	(0.091)	(0.112)	(0.131)	(0.143)	(0.145)	(0.132)	(0.105)	(0.079)
	0.008	0.029	0.067	0.125	0.205	0.306	0.428	0.572	0.748
	(0.186)	(0.356)	(0.530)	(0.671)	(0.754)	(0.766)	(0.708)	(0.587)	(0.404)
2.7	0.939	0.927	0.914	0.901	0.889	0.880	0.877	0.885	0.915
	(0.215)	(0.223)	(0.223)	(0.215)	(0.199)	(0.176)	(0.148)	(0.118)	(0.095)
	0.023	0.060	0.112	0.179	0.260	0.356	0.467	0.596	0.756
	(0.405)	(0.542)	(0.641)	(0.693)	(0.696)	(0.654)	(0.574)	(0.462)	(0.320)
3.6	0.811	0.802	0.796	0.794	0.796	0.804	0.819	0.847	0.896
	(0.314)	(0.290)	(0.264)	(0.238)	(0.212)	(0.186)	(0.162)	(0.137)	(0.110)
	0.042	0.094	0.156	0.227	0.307	0.396	0.497	0.615	0.763
	(0.640)	(0.688)	(0.702)	(0.684)	(0.640)	(0.574)	(0.490)	(0.392)	(0.276)
4.5	0.633	0.648	0.665	0.686	0.711	0.740	0.776	0.822	0.886
	(0.339)	(0.317)	(0.295)	(0.272)	(0.248)	(0.222)	(0.193)	(0.161)	(0.122)
	0.059	0.122	0.189	0.261	0.339	0.423	0.517	0.627	0.768
	(0.824)	(0.782)	(0.729)	(0.667)	(0.599)	(0.524)	(0.443)	(0.355)	(0.256)
5.4	0.534	0.568	0.603	0.639	0.677	0.718	0.764	0.818	0.887
	(0.455)	(0.412)	(0.371)	(0.332)	(0.293)	(0.255)	(0.216)	(0.175)	(0.128)
	0.067	0.135	0.205	0.278	0.354	0.436	0.527	0.634	0.772
	(0.913)	(0.826)	(0.743)	(0.664)	(0.587)	(0.510)	(0.432)	(0.349)	(0.255)
6.3	0.624	0.644	0.667	0.692	0.721	0.753	0.791	0.837	0.899
	(0.385)	(0.356)	(0.326)	(0.297)	(0.267)	(0.236)	(0.203)	(0.168)	(0.125)
	0.063	0.130	0.199	0.272	0.350	0.433	0.527	0.635	0.773
	(0.878)	(0.820)	(0.756)	(0.689)	(0.617)	(0.542)	(0.463)	(0.377)	(0.276)
7.2	0.834	0.827	0.822	0.822	0.826	0.836	0.852	0.879	0.923
	(0.311)	(0.286)	(0.261)	(0.236)	(0.211)	(0.187)	(0.164)	(0.141)	(0.111)
	0.045	0.102	0.167	0.241	0.323	0.413	0.514	0.630	0.773
	(0.700)	(0.756)	(0.771)	(0.752)	(0.707)	(0.640)	(0.557)	(0.457)	(0.330)
8.1	0.976	0.967	0.957	0.947	0.938	0.932	0.931	0.937	0.957
	(0.144)	(0.161)	(0.171)	(0.172)	(0.163)	(0.145)	(0.123)	(0.102)	(0.083)
	0.019	0.056	0.111	0.184	0.272	0.373	0.488	0.617	0.771
	(0.389)	(0.610)	(0.778)	(0.878)	(0.909)	(0.880)	(0.798)	(0.666)	(0.473)

For the mastery score = 4 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 9
Mastery score C = 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	1.000	1.000	1.000	0.999	0.997	0.994	0.989	0.980	0.975
	(0.001)	(0.002)	(0.005)	(0.012)	(0.024)	(0.042)	(0.063)	(0.074)	(0.056)
	0.000	0.003	0.012	0.038	0.091	0.179	0.309	0.483	0.704
	(0.010)	(0.062)	(0.193)	(0.405)	(0.659)	(0.886)	(1.007)	(0.956)	(0.688)
1.8	0.999	0.997	0.995	0.991	0.985	0.975	0.963	0.951	0.948
	(0.014)	(0.024)	(0.038)	(0.057)	(0.079)	(0.100)	(0.112)	(0.104)	(0.071)
	0.002	0.010	0.031	0.071	0.137	0.232	0.360	0.520	0.720
	(0.058)	(0.165)	(0.324)	(0.506)	(0.666)	(0.764)	(0.769)	(0.669)	(0.463)
2.7	0.987	0.981	0.973	0.963	0.951	0.936	0.922	0.913	0.923
	(0.078)	(0.098)	(0.119)	(0.139)	(0.152)	(0.1[illegible])	(0.145)	(0.116)	(0.083)
	0.008	0.027	0.062	0.115	0.190	0.287	0.407	0.553	0.733
	(0.175)	(0.318)	(0.468)	(0.596)	(0.676)	(0.694)	(0.644)	(0.530)	(0.351)
3.6	0.940	0.928	0.914	0.900	0.886	0.875	0.868	0.873	0.901
	(0.207)	(0.218)	(0.221)	(0.217)	(0.205)	(0.183)	(0.153)	(0.120)	(0.096)
	0.021	0.054	0.102	0.165	0.244	0.338	0.449	0.581	0.745
	(0.363)	(0.490)	(0.589)	(0.648)	(0.661)	(0.628)	(0.554)	(0.444)	(0.304)
4.5	0.824	0.814	0.805	0.799	0.797	0.800	0.812	0.837	0.887
	(0.311)	(0.289)	(0.265)	(0.239)	(0.211)	(0.184)	(0.159)	(0.136)	(0.112)
	0.038	0.087	0.145	0.214	0.293	0.382	0.484	0.604	0.755
	(0.585)	(0.644)	(0.671)	(0.665)	(0.630)	(0.570)	(0.488)	(0.388)	(0.272)
5.4	0.651	0.660	0.673	0.690	0.711	0.737	0.771	0.817	0.882
	(0.317)	(0.297)	(0.277)	(0.257)	(0.237)	(0.216)	(0.192)	(0.164)	(0.127)
	0.056	0.116	0.182	0.254	0.331	0.416	0.511	0.622	0.763
	(0.787)	(0.761)	(0.720)	(0.666)	(0.602)	(0.529)	(0.449)	(0.360)	(0.260)
6.3	0.535	0.569	0.603	0.639	0.677	0.718	0.765	0.819	0.889
	(0.448)	(0.409)	(0.372)	(0.336)	(0.300)	(0.263)	(0.226)	(0.185)	(0.136)
	0.067	0.135	0.205	0.277	0.354	0.436	0.528	0.634	0.770
	(0.914)	(0.831)	(0.752)	(0.675)	(0.599)	(0.523)	(0.446)	(0.364)	(0.268)
7.2	0.634	0.656	0.680	0.706	0.735	0.768	0.806	0.852	0.911
	(0.410)	(0.377)	(0.345)	(0.313)	(0.281)	(0.249)	(0.216)	(0.179)	(0.131)
	0.065	0.133	0.204	0.278	0.356	0.440	0.533	0.640	0.774
	(0.911)	(0.852)	(0.788)	(0.722)	(0.652)	(0.579)	(0.502)	(0.415)	(0.308)
8.1	0.888	0.879	0.873	0.871	0.873	0.880	0.893	0.915	0.948
	(0.288)	(0.267)	(0.244)	(0.220)	(0.197)	(0.176)	(0.157)	(0.137)	(0.106)
	0.043	0.100	0.168	0.245	0.329	0.422	0.524	0.638	0.777
	(0.712)	(0.820)	(0.870)	(0.874)	(0.842)	(0.782)	(0.696)	(0.583)	(0.427)

For the mastery score = 3 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 9
Mastery score C = 8

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	1.000	1.000	1.000	1.000	0.999	0.998	0.996	0.989	0.980
	(0.000)	(0.000)	(0.001)	(0.002)	(0.007)	(0.016)	(0.034)	(0.057)	(0.058)
	0.000	0.000	0.004	0.015	0.045	0.109	0.222	0.398	0.648
	(0.001)	(0.015)	(0.071)	(0.203)	(0.422)	(0.697)	(0.942)	(1.029)	(0.812)
1.8	1.000	1.000	0.999	0.998	0.996	0.992	0.984	0.970	0.957
	(0.002)	(0.004)	(0.008)	(0.016)	(0.029)	(0.049)	(0.075)	(0.094)	(0.074)
	0.000	0.002	0.010	0.031	0.074	0.150	0.270	0.440	0.670
	(0.011)	(0.051)	(0.143)	(0.292)	(0.479)	(0.659)	(0.770)	(0.749)	(0.548)
2.7	0.998	0.997	0.995	0.991	0.984	0.974	0.960	0.942	0.932
	(0.015)	(0.024)	(0.037)	(0.054)	(0.076)	(0.101)	(0.120)	(0.118)	(0.081)
	0.002	0.008	0.025	0.057	0.113	0.199	0.320	0.481	0.690
	(0.048)	(0.127)	(0.251)	(0.402)	(0.549)	(0.656)	(0.685)	(0.611)	(0.427)
3.6	0.989	0.984	0.977	0.967	0.955	0.939	0.921	0.905	0.908
	(0.065)	(0.084)	(0.105)	(0.126)	(0.145)	(0.157)	(0.153)	(0.127)	(0.085)
	0.006	0.021	0.049	0.094	0.161	0.252	0.370	0.519	0.708
	(0.135)	(0.250)	(0.381)	(0.507)	(0.601)	(0.642)	(0.616)	(0.517)	(0.354)
4.5	0.952	0.941	0.928	0.913	0.897	0.881	0.868	0.865	0.888
	(0.175)	(0.191)	(0.203)	(0.208)	(0.205)	(0.189)	(0.161)	(0.124)	(0.096)
	0.016	0.043	0.084	0.141	0.214	0.307	0.419	0.554	0.725
	(0.288)	(0.407)	(0.512)	(0.588)	(0.624)	(0.613)	(0.553)	(0.448)	(0.307)
5.4	0.855	0.842	0.829	0.818	0.809	0.806	0.810	0.829	0.876
	(0.297)	(0.285)	(0.267)	(0.244)	(0.216)	(0.186)	(0.156)	(0.132)	(0.116)
	0.032	0.074	0.127	0.192	0.269	0.358	0.462	0.585	0.740
	(0.497)	(0.574)	(0.622)	(0.639)	(0.623)	(0.575)	(0.500)	(0.400)	(0.280)
6.3	0.684	0.686	0.692	0.701	0.716	0.737	0.767	0.810	0.876
	(0.305)	(0.281)	(0.259)	(0.239)	(0.222)	(0.206)	(0.189)	(0.169)	(0.139)
	0.050	0.107	0.170	0.240	0.318	0.403	0.499	0.611	0.753
	(0.725)	(0.726)	(0.705)	(0.667)	(0.614)	(0.546)	(0.467)	(0.377)	(0.274)
7.2	0.539	0.570	0.603	0.639	0.677	0.719	0.767	0.823	0.894
	(0.432)	(0.404)	(0.375)	(0.346)	(0.316)	(0.283)	(0.248)	(0.207)	(0.153)
	0.066	0.134	0.204	0.277	0.354	0.436	0.527	0.632	0.764
	(0.917)	(0.845)	(0.773)	(0.701)	(0.628)	(0.555)	(0.478)	(0.395)	(0.297)
8.1	0.671	0.694	0.718	0.744	0.773	0.805	0.841	0.883	0.934
	(0.442)	(0.407)	(0.374)	(0.342)	(0.310)	(0.277)	(0.241)	(0.199)	(0.140)
	0.069	0.140	0.213	0.289	0.368	0.452	0.544	0.647	0.773
	(0.982)	(0.935)	(0.880)	(0.821)	(0.757)	(0.686)	(0.607)	(0.513)	(0.391)

For the mastery score = 2 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N = 9
Mastery score C = 9

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.996	0.987
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.003)	(0.011)	(0.032)	(0.060)
	0.000	0.000	0.001	0.004	0.015	0.045	0.117	0.263	0.530
	(0.000)	(0.002)	(0.015)	(0.060)	(0.172)	(0.380)	(0.675)	(0.962)	(0.972)
1.8	1.000	1.000	1.000	1.000	0.999	0.998	0.995	0.987	0.969
	(0.000)	(0.000)	(0.001)	(0.002)	(0.005)	(0.013)	(0.031)	(0.063)	(0.085)
	0.000	0.000	0.002	0.008	0.026	0.067	0.151	0.304	0.561
	(0.001)	(0.003)	(0.035)	(0.100)	(0.222)	(0.401)	(0.605)	(0.749)	(0.675)
2.7	1.000	1.000	0.999	0.999	0.997	0.994	0.987	0.971	0.946
	(0.001)	(0.003)	(0.005)	(0.010)	(0.019)	(0.036)	(0.063)	(0.097)	(0.099)
	0.000	0.001	0.006	0.017	0.044	0.097	0.192	0.343	0.590
	(0.006)	(0.026)	(0.075)	(0.164)	(0.295)	(0.452)	(0.595)	(0.653)	(0.535)
3.6	0.999	0.998	0.9[illegible]7	0.994	0.990	0.982	0.968	0.946	0.920
	(0.003)	(0.013)	(0.021)	(0.033)	(0.052)	(0.077)	(0.107)	(0.129)	(0.101)
	0.001	0.004	0.013	0.033	0.071	0.135	0.239	0.394	0.619
	(0.024)	(0.067)	(0.142)	(0.249)	(0.379)	(0.505)	(0.590)	(0.585)	(0.446)
4.5	0.994	0.991	0.987	0.981	0.971	0.956	0.936	0.910	0.893
	(0.035)	(0.048)	(0.064)	(0.085)	(0.109)	(0.134)	(0.153)	(0.147)	(0.095)
	0.003	0.012	0.028	0.059	0.108	0.183	0.291	0.441	0.646
	(0.072)	(0.143)	(0.240)	(0.352)	(0.462)	(0.547)	(0.578)	(0.526)	(0.383)
5.4	0.974	0.966	0.956	0.944	0.927	0.908	0.885	0.865	0.870
	(0.108)	(0.128)	(0.148)	(0.167)	(0.183)	(0.189)	(0.179)	(0.142)	(0.095)
	0.009	0.026	0.054	0.096	0.157	0.239	0.348	0.489	0.673
	(0.170)	(0.264)	(0.365)	(0.461)	(0.534)	(0.571)	(0.555)	(0.477)	(0.339)
6.3	0.910	0.896	0.881	0.864	0.847	0.831	0.819	0.821	0.856
	(0.237)	(0.246)	(0.249)	(0.244)	(0.230)	(0.204)	(0.167)	(0.127)	(0.120)
	0.020	0.051	0.092	0.147	0.216	0.30[illegible]	0.407	0.535	0.698
	(0.339)	(0.430)	(0.508)	(0.563)	(0.588)	(0.578)	(0.527)	(0.435)	(0.311)
7.2	0.757	0.747	0.739	0.735	0.735	0.742	0.760	0.795	0.862
	(0.319)	(0.292)	(0.263)	(0.233)	(0.206)	(0.184)	(0.174)	(0.174)	(0.167)
	0.039	0.086	0.143	0.208	0.283	0.369	0.467	0.580	0.722
	(0.539)	(0.635)	(0.658)	(0.656)	(0.631)	(0.582)	(0.509)	(0.417)	(0.309)
8.1	0.549	0.576	0.605	0.639	0.677	0.721	0.771	0.820	0.903
	(0.381)	(0.383)	(0.381)	(0.375)	(0.363)	(0.344)	(0.315)	(0.270)	(0.197)
	0.065	0.132	0.203	0.277	0.354	0.436	0.524	0.623	0.744
	(0.927)	(0.890)	(0.841)	(0.782)	(0.715)	(0.640)	(0.557)	(0.466)	(0.367)

For the mastery score = 1 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-Binomial Model
M = Number of subjects
Number of items N =10
Mastery score C = 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	0.994	0.98[illegible]	0.983	0.976	0.967	0.958	0.951	0.950	0.961
	(0.056)	(0.077)	(0.099)	(0.117)	(0.128)	(0.12[illegible])	(0.114)	(0.091)	(0.071)
	0.008	0.031	0.073	0.138	0.223	0.328	0.451	0.591	0.758
	(0.199)	(0.416)	(0.634)	(0.803)	(0.897)	(0.910)	(0.846)	(0.710)	(0.497)
2.0	0.924	0.912	0.900	0.889	0.882	0.878	0.881	0.895	0.927
	(0.245)	(0.243)	(0.234)	(0.218)	(0.196)	(0.171)	(0.145)	(0.120)	(0.096)
	0.029	0.073	0.132	0.203	0.286	0.381	0.489	0.612	0.764
	(0.499)	(0.632)	(0.714)	(0.745)	(0.730)	(0.676)	(0.592)	(0.482)	(0.341)
3.0	0.743	0.744	0.749	0.757	0.770	0.788	0.813	0.849	0.902
	(0.324)	(0.298)	(0.273)	(0.249)	(0.225)	(0.201)	(0.175)	(0.148)	(0.113)
	0.052	0.112	0.178	0.250	0.329	0.416	0.513	0.625	0.767
	(0.756)	(0.759)	(0.736)	(0.693)	(0.633)	(0.562)	(0.480)	(0.388)	(0.278)
4.0	0.563	0.592	0.623	0.655	0.689	0.727	0.770	0.821	0.887
	(0.421)	(0.384)	(0.349)	(0.313)	(0.278)	(0.243)	(0.206)	(0.16[illegible])	(0.122)
	0.066	0.133	0.202	0.274	0.350	0.432	0.523	0.630	0.768
	(0.889)	(0.811)	(0.735)	(0.660)	(0.585)	(0.508)	(0.430)	(0.346)	(0.250)
5.0	0.553	0.587	0.617	0.649	0.684	0.722	0.765	0.816	0.884
	(0.413)	(0.378)	(0.344)	(0.311)	(0.277)	(0.242)	(0.206)	(0.167)	(0.121)
	0.065	0.132	0.201	0.272	0.348	0.431	0.522	0.629	0.767
	(0.882)	(0.806)	(0.730)	(0.655)	(0.580)	(0.503)	(0.424)	(0.340)	(0.245)
6.0	0.722	0.724	0.730	0.739	0.753	0.772	0.798	0.835	0.892
	(0.320)	(0.295)	(0.271)	(0.248)	(0.224)	(0.201)	(0.175)	(0.147)	(0.113)
	0.052	0.111	0.176	0.248	0.326	0.412	0.509	0.621	0.764
	(0.747)	(0.744)	(0.718)	(0.673)	(0.614)	(0.541)	(0.459)	(0.367)	(0.260)
7.0	0.897	0.884	0.872	0.862	0.855	0.852	0.857	0.873	0.910
	(0.272)	(0.264)	(0.249)	(0.229)	(0.206)	(0.179)	(0.151)	(0.124)	(0.099)
	0.032	0.076	0.134	0.204	0.285	0.378	0.483	0.606	0.759
	(0.515)	(0.620)	(0.681)	(0.698)	(0.676)	(0.619)	(0.535)	(0.429)	(0.299)
8.0	0.981	0.[illegible]74	0.964	0.953	0.941	0.929	0.920	0.9[illegible]	0.936
	(0.109)	(0.130)	(0.149)	(0.161)	(0.164)	(0.156)	(0.136)	(0.108)	(0.082)
	0.012	0.039	0.083	0.146	0.228	0.329	0.447	0.584	0.751
	(0.256)	(0.432)	(0.590)	(0.701)	(0.751)	(0.737)	(0.665)	(0.544)	(0.375)
9.0	0.999	0.998	0.996	0.993	0.987	0.980	0.971	0.964	0.967
	(0.011)	(0.021)	(0.036)	(0.055)	(0.075)	(0.092)	(0.098)	(0.085)	(0.060)
	0.002	0.012	0.038	0.088	0.164	0.268	0.399	0.555	0.742
	(0.068)	(0.218)	(0.436)	(0.665)	(0.847)	(0.941)	(0.926)	(0.799)	(0.555)

For the mastery score = 6 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N =10
Mastery score C = 6

Test KR21= Mean	.100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	0.999	0.998	0.996	0.993	0.987	0.980	0.971	0.964	0.967
	(0.011)	(0.021)	(0.036)	(0.055)	(0.075)	(0.092)	(0.098)	(0.085)	(0.060)
	0.002	0.012	0.038	0.088	0.164	0.268	0.399	0.555	0.742
	(0.068)	(0.218)	(0.436)	(0.665)	(0.847)	(0.941)	(0.926)	(0.799)	(0.555)
2.0	0.981	0.974	0.964	0.953	0.941	0.929	0.920	0.919	0.936
	(0.109)	(0.130)	(0.149)	(0.161)	(0.164)	(0.156)	(0.136)	(0.108)	(0.082)
	0.012	0.039	0.083	0.146	0.228	0.329	0.447	0.584	0.751
	(0.256)	(0.432)	(0.590)	(0.701)	(0.751)	(0.737)	(0.665)	(0.544)	(0.375)
3.0	0.897	0.884	0.872	0.862	0.855	0.352	[illegible]857	0.873	0.910
	(0.272)	(0.264)	(0.249)	(0.229)	(0.206)	(0.179)	([illegible]51)	(0.124)	(0.099)
	0.032	0.076	0.134	0.204	0.285	0.378	0.483	0.606	0.759
	(0.515)	(0.620)	(0.681)	(0.698)	(0.676)	(0.619)	(0.535)	(0.429)	(0.299)
4.0	0.722	0.724	0.730	0.739	0.753	0.772	0.798	0.835	0.892
	(0.320)	(0.295)	(0.271)	(0.248)	(0.224)	(0.201)	(0.175)	(0.147)	(0.113)
	0.052	0.111	0.176	0.248	0.326	0.412	0.509	0.621	0.764
	(0.747)	(0.744)	(0.718)	(0.673)	(0.614)	(0.541)	(0.459)	(0.367)	(0.260)
5.0	0.558	0.587	0.617	0.649	0.684	0.722	0.765	0.816	0.884
	(0.413)	(0.378)	(0.344)	(0.311)	(0.277)	(0.242)	(0.206)	(0.167)	(0.121)
	0.065	0.132	0.201	0.272	0.348	0.431	0.522	0.629	0.767
	(0.832)	(0.806)	(0.730)	(0.655)	(0.580)	(0.503)	(0.424)	(0.340)	(0.245)
6.0	0.563	0.592	0.623	0.655	0.689	0.727	0.770	0.821	0.887
	(0.421)	(0.384)	(0.349)	(0.313)	(0.278)	(0.243)	(0.206)	(0.167)	(0.122)
	0.066	0.133	0.202	0.274	0.350	0.432	0.523	0.630	0.768
	(0.889)	(0.811)	(0.735)	(0.660)	(0.585)	(0.508)	(0.430)	(0.346)	(0.250)
7.0	0.743	0.744	0.749	0.757	0.770	0.788	0.813	0.849	0.902
	(0.324)	(0.298)	(0.273)	(0.249)	(0.225)	(0.201)	(0.175)	(0.148)	(0.113)
	0.052	0.112	0.173	0.250	0.329	0.416	0.513	0.625	0.767
	(0.756)	(0.759)	(0.736)	(0.693)	(0.633)	(0.562)	(0.480)	(0.388)	(0.278)
8.0	0.924	0.912	0.900	0.889	0.882	0.878	0.881	0.895	0.927
	(0.245)	(0.243)	(0.234)	(0.218)	(0.196)	(0.171)	(0.145)	(0.120)	(0.096)
	0.029	0.073	0.132	0.203	0.286	0.381	0.489	0.612	0.764
	(0.499)	(0.632)	(0.714)	(0.745)	(0.730)	(0.676)	(0.592)	(0.482)	(0.341)
9.0	0.994	0.989	0.983	0.976	0.967	0.958	0.951	0.950	0.961
	(0.056)	(0.077)	(0.099)	(0.117)	(0.128)	(0.127)	(0.114)	(0.091)	(0.071)
	0.008	0.031	0.073	0.138	0.223	0.328	0.451	0.591	0.758
	(0.199)	(0.416)	(0.634)	(0.803)	(0.897)	(0.910)	(0.846)	(0.710)	(0.497)

For the mastery score = 5 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-Binomial Model
M = Number of subjects
Number of items N =10
Mastery score C = 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	1.000	1.000	0.999	0.998	0.996	0.992	0.985	0.976	0.972
	(0.002)	(0.004)	(0.010)	(0.019)	(0.034)	(0.054)	(0.073)	(0.078)	(0.056)
	0.000	0.004	0.017	0.050	0.110	0.206	0.339	0.508	0.717
	(0.017)	(0.091)	(0.251)	(0.480)	(0.722)	(0.908)	(0.977)	(0.891)	(0.627)
2.0	0.997	0.994	0.990	0.984	0.976	0.964	0.952	0.941	0.944
	(0.030)	(0.044)	(0.063)	(0.085)	(0.105)	(0.121)	(0.123)	(0.105)	(0.073)
	0.004	0.016	0.044	0.092	0.166	0.265	0.391	0.544	0.731
	(0.098)	(0.235)	(0.410)	(0.581)	(0.709)	(0.764)	(0.732)	(0.616)	(0.421)
3.0	0.972	0.962	0.951	0.938	0.925	0.911	0.901	0.899	0.918
	(0.136)	(0.156)	(0.171)	(0.180)	(0.181)	(0.170)	(0.147)	(0.115)	(0.086)
	0.014	0.041	0.084	0.145	0.225	0.322	0.438	0.575	0.743
	(0.271)	(0.422)	(0.556)	(0.650)	(0.689)	(0.672)	(0.601)	(0.485)	(0.329)
4.0	0.883	0.870	0.858	0.847	0.839	0.836	0.841	0.857	0.897
	(0.281)	(0.271)	(0.256)	(0.235)	(0.210)	(0.182)	(0.153)	(0.126)	(0.101)
	0.032	0.075	0.131	0.199	0.279	0.371	0.476	0.599	0.753
	(0.506)	(0.599)	(0.654)	(0.670)	(0.648)	(0.593)	(0.510)	(0.405)	(0.279)
5.0	0.714	0.716	0.720	0.729	0.742	0.761	0.788	0.826	0.885
	(0.316)	(0.291)	(0.267)	(0.244)	(0.222)	(0.199)	(0.176)	(0.149)	(0.115)
	0.051	0.109	0.173	0.244	0.322	0.408	0.504	0.616	0.760
	(0.730)	(0.729)	(0.705)	(0.663)	(0.605)	(0.534)	(0.452)	(0.359)	(0.254)
6.0	0.555	0.583	0.613	0.645	0.680	0.718	0.762	0.814	0.883
	(0.405)	(0.373)	(0.341)	(0.310)	(0.278)	(0.244)	(0.209)	(0.170)	(0.125)
	0.065	0.131	0.200	0.271	0.347	0.430	0.521	0.628	0.765
	(0.378)	(0.804)	(0.730)	(0.656)	(0.582)	(0.506)	(0.427)	(0.343)	(0.248)
7.0	0.573	0.602	0.632	0.664	0.698	0.736	0.778	0.828	0.893
	(0.431)	(0.392)	(0.355)	(0.319)	(0.284)	(0.248)	(0.211)	(0.172)	(0.125)
	0.066	0.134	0.203	0.276	0.352	0.435	0.526	0.632	0.768
	(0.900)	(0.823)	(0.747)	(0.672)	(0.598)	(0.523)	(0.446)	(0.362)	(0.265)
8.0	0.783	0.781	0.783	0.789	0.799	0.815	0.837	0.869	0.917
	(0.323)	(0.296)	(0.271)	(0.246)	(0.222)	(0.198)	(0.174)	(0.148)	(0.114)
	0.051	0.111	0.178	0.252	0.332	0.420	0.517	0.629	0.770
	(0.758)	(0.779)	(0.768)	(0.732)	(0.677)	(0.609)	(0.527)	(0.433)	(0.315)
9.0	0.965	0.955	0.944	0.935	0.926	0.922	0.922	0.931	0.953
	(0.175)	(0.187)	(0.191)	(0.185)	(0.171)	(0.151)	(0.128)	(0.107)	(0.087)
	0.023	0.063	0.121	0.195	0.282	0.382	0.493	0.618	0.768
	(0.441)	(0.643)	(0.786)	(0.862)	(0.875)	(0.836)	(0.753)	(0.628)	(0.449)

For the mastery score = 4 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N =10
Mastery score C = 8

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	1.000	1.000	1.000	1.000	0.999	0.997	0.993	0.985	0.978
	(0.000)	(0.001)	(0.002)	(0.005)	(0.012)	(0.025)	(0.046)	(0.065)	(0.056)
	0.000	0.001	0.006	0.024	0.065	0.143	0.268	0.446	0.681
	(0.003)	(0.029)	(0.115)	(0.289)	(0.535)	(0.794)	(0.973)	(0.974)	(0.719)
2.0	1.000	0.999	0.998	0.996	0.992	0.985	0.975	0.961	0.952
	(0.005)	(0.010)	(0.018)	(0.031)	(0.050)	(0.073)	(0.094)	(0.100)	(0.071)
	0.001	0.005	0.019	0.049	0.105	0.194	0.321	0.488	0.700
	(0.026)	(0.096)	(0.226)	(0.402)	(0.586)	(0.725)	(0.771)	(0.693)	(0.482)
3.0	0.995	0.991	0.987	0.979	0.969	0.956	0.940	0.926	0.927
	(0.039)	(0.055)	(0.074)	(0.096)	(0.118)	(0.134)	(0.137)	(0.117)	(0.079)
	0.004	0.017	0.043	0.087	0.156	0.250	0.373	0.526	0.717
	(0.102)	(0.221)	(0.370)	(0.519)	(0.634)	(0.687)	(0.661)	(0.554)	(0.374)
4.0	0.968	0.958	0.947	0.933	0.918	0.903	0.890	0.885	0.904
	(0.141)	(0.160)	(0.176)	(0.186)	(0.188)	(0.178)	(0.155)	(0.120)	(0.089)
	0.014	0.039	0.079	0.136	0.212	0.307	0.422	0.560	0.731
	(0.255)	(0.389)	(0.512)	(0.604)	(0.648)	(0.638)	(0.574)	(0.462)	(0.311)
5.0	0.883	0.869	0.856	0.844	0.834	0.829	0.831	0.846	0.886
	(0.278)	(0.271)	(0.258)	(0.238)	(0.214)	(0.185)	(0.155)	(0.127)	(0.104)
	0.029	0.071	0.124	0.189	0.267	0.358	0.464	0.588	0.744
	(0.472)	(0.564)	(0.623)	(0.646)	(0.632)	(0.583)	(0.503)	(0.399)	(0.274)
6.0	0.718	0.717	0.720	0.726	0.737	0.754	0.780	0.818	0.879
	(0.312)	(0.286)	(0.262)	(0.239)	(0.217)	(0.197)	(0.175)	(0.152)	(0.121)
	0.049	0.104	0.167	0.237	0.315	0.400	0.497	0.610	0.754
	(0.701)	(0.709)	(0.693)	(0.657)	(0.604)	(0.536)	(0.454)	(0.362)	(0.257)
7.0	0.554	0.581	0.611	0.643	0.677	0.716	0.760	0.814	0.884
	(0.394)	(0.367)	(0.340)	(0.312)	(0.282)	(0.251)	(0.218)	(0.180)	(0.134)
	0.064	0.130	0.199	0.271	0.347	0.429	0.520	0.627	0.763
	(0.375)	(0.806)	(0.736)	(0.664)	(0.591)	(0.517)	(0.439)	(0.356)	(0.261)
8.0	0.591	0.619	0.649	0.680	0.714	0.751	0.793	0.842	0.905
	(0.445)	(0.405)	(0.368)	(0.331)	(0.295)	(0.259)	(0.223)	(0.183)	(0.133)
	0.067	0.136	0.206	0.280	0.357	0.439	0.530	0.636	0.769
	(0.921)	(0.847)	(0.774)	(0.702)	(0.630)	(0.557)	(0.482)	(0.399)	(0.296)
9.0	0.860	0.853	0.850	0.851	0.856	0.866	0.882	0.907	0.944
	(0.303)	(0.279)	(0.254)	(0.230)	(0.207)	(0.186)	(0.166)	(0.144)	(0.110)
	0.048	0.107	0.175	0.251	0.335	0.425	0.525	0.637	0.773
	(0.749)	(0.827)	(0.855)	(0.844)	(0.805)	(0.742)	(0.660)	(0.553)	(0.409)

For the mastery score = 3 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N =10
Mastery score C = 9

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	1.000	1.000	1.000	1.000	1.000	0.999	0.997	0.993	0.983
	(0.000)	(0.000)	(0.000)	(0.001)	(0.003)	(0.009)	(0.022)	(0.046)	(0.058)
	0.000	0.000	0.002	0.009	0.031	0.083	0.186	0.359	0.620
	(0.000)	(0.006)	(0.039)	(0.132)	(0.317)	(0.586)	(0.867)	(1.017)	(0.840)
2.0	1.000	1.000	1.000	0.999	0.998	0.995	0.989	0.977	0.962
	(0.001)	(0.001)	(0.003)	(0.007)	(0.016)	(0.031)	(0.056)	(0.083)	(0.076)
	0.000	0.001	0.006	0.020	0.054	0.120	0.233	0.404	0.646
	(0.004)	(0.027)	(0.090)	(0.212)	(0.389)	(0.587)	(0.737)	(0.757)	(0.570)
3.0	0.999	0.999	0.998	0.995	0.991	0.984	0.971	0.953	0.937
	(0.006)	(0.011)	(0.019)	(0.032)	(0.050)	(0.075)	(0.101)	(0.113)	(0.083)
	0.001	0.005	0.016	0.040	0.088	0.166	0.284	0.449	0.669
	(0.024)	(0.079)	(0.178)	(0.319)	(0.478)	(0.614)	(0.677)	(0.627)	(0.443)
4.0	0.995	0.992	0.987	0.980	0.970	0.956	0.938	0.918	0.912
	(0.036)	(0.050)	(0.068)	(0.090)	(0.113)	(0.133)	(0.143)	(0.129)	(0.084)
	0.004	0.014	0.035	0.073	0.133	0.220	0.338	0.492	0.691
	(0.084)	(0.178)	(0.302)	(0.437)	(0.554)	(0.625)	(0.623)	(0.535)	(0.365)
5.0	0.972	0.963	0.953	0.939	0.923	0.906	0.888	0.877	0.890
	(0.122)	(0.142)	(0.161)	(0.176)	(0.185)	(0.182)	(0.164)	(0.127)	(0.090)
	0.011	0.032	0.066	0.117	0.187	0.278	0.392	0.532	0.710
	(0.208)	(0.325)	(0.442)	(0.540)	(0.601)	(0.612)	(0.566)	(0.463)	(0.313)
6.0	0.898	0.884	0.870	0.855	0.842	0.831	0.827	0.836	0.874
	(0.259)	(0.260)	(0.254)	(0.241)	(0.219)	(0.191)	(0.158)	(0.126)	(0.108)
	0.025	0.061	0.109	0.170	0.245	0.335	0.442	0.568	0.728
	(0.405)	(0.501)	(0.572)	(0.612)	(0.615)	(0.581)	(0.511)	(0.408)	(0.282)
7.0	0.739	0.733	0.731	0.732	0.739	0.751	0.773	0.809	0.872
	(0.313)	(0.286)	(0.259)	(0.234)	(0.211)	(0.191)	(0.173)	(0.157)	(0.133)
	0.044	0.096	0.156	0.225	0.301	0.387	0.485	0.599	0.743
	(0.648)	(0.673)	(0.675)	(0.653)	(0.610)	(0.548)	(0.470)	(0.377)	(0.270)
8.0	0.555	0.581	0.609	0.641	0.675	0.714	0.760	0.815	0.888
	(0.377)	(0.359)	(0.339)	(0.317)	(0.294)	(0.268)	(0.238)	(0.202)	(0.152)
	0.063	0.129	0.198	0.269	0.346	0.428	0.519	0.624	0.757
	(0.874)	(0.815)	(0.752)	(0.686)	(0.617)	(0.545)	(0.469)	(0.386)	(0.288)
9.0	0.637	0.664	0.692	0.722	0.755	0.790	0.829	0.874	0.928
	(0.470)	(0.430)	(0.393)	(0.357)	(0.322)	(0.286)	(0.248)	(0.204)	(0.143)
	0.070	0.141	0.214	0.289	0.367	0.450	0.540	0.642	0.768
	(0.980)	(0.919)	(0.857)	(0.793)	(0.727)	(0.657)	(0.581)	(0.491)	(0.375)

For the mastery score = 2 enter N-xbar in the test mean column

Table of the Raw Agreement Index and its
S.E.*SQRT(M), the Kappa Index and its
S.E.*SQRT(M) in the Beta-binomial Model
M = Number of subjects
Number of items N =10
Mastery score C =10

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.997	0.989
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.006)	(0.023)	(0.055)
	0.000	0.000	0.000	0.002	0.009	0.032	0.093	0.229	0.497
	(0.000)	(0.001)	(0.007)	(0.036)	(0.118)	(0.294)	(0.579)	(0.901)	(0.981)
2.0	1.000	1.000	1.000	1.000	1.000	0.999	0.997	0.991	0.973
	(0.000)	(0.000)	(0.000)	(0.001)	(0.003)	(0.007)	(0.020)	(0.049)	(0.082)
	0.000	0.000	0.001	0.005	0.017	0.050	0.124	0.269	0.530
	(0.000)	(0.004)	(0.020)	(0.066)	(0.164)	(0.329)	(0.541)	(0.721)	(0.687)
3.0	1.000	1.000	1.000	0.999	0.999	0.997	0.991	0.978	0.952
	(0.000)	(0.001)	(0.002)	(0.005)	(0.011)	(0.023)	(0.046)	(0.082)	(0.100)
	0.000	0.001	0.003	0.011	0.032	0.076	0.162	0.314	0.563
	(0.003)	(0.014)	(0.047)	(0.117)	(0.234)	(0.390)	(0.551)	(0.642)	(0.548)
4.0	1.000	0.999	0.998	0.997	0.994	0.989	0.977	0.957	0.927
	(0.004)	(0.007)	(0.012)	(0.020)	(0.034)	(0.056)	(0.087)	(0.118)	(0.106)
	0.000	0.003	0.009	0.023	0.054	0.111	0.209	0.363	0.595
	(0.013)	(0.042)	(0.100)	(0.194)	(0.319)	(0.456)	(0.565)	(0.58[illegible]	(0.458)
5.0	0.997	0.995	0.992	0.988	0.980	0.969	0.950	0.923	0.899
	(0.021)	(0.030)	(0.042)	(0.060)	(0.082)	(0.109)	(0.136)	(0.145)	(0.101)
	0.002	0.008	0.020	0.045	0.088	0.157	0.262	0.413	0.626
	(0.046)	(0.102)	(0.187)	(0.295)	(0.412)	(0.513)	(0.567)	(0.535)	(0.393)
6.0	0.984	0.978	0.970	0.960	0.946	0.927	0.903	0.879	0.872
	(0.076)	(0.093)	(0.114)	(0.136)	(0.157)	(0.173)	(0.175)	(0.149)	(0.093)
	0.006	0.019	0.043	0.080	0.135	0.214	0.322	0.465	0.656
	(0.124)	(0.209)	(0.303)	(0.410)	(0.498)	(0.553)	(0.554)	(0.486)	(0.344)
7.0	0.935	0.922	0.908	0.891	0.873	0.853	0.836	0.829	0.854
	(0.197)	(0.212)	(0.223)	(0.228)	(0.224)	(0.208)	(0.175)	(0.130)	(0.109)
	0.016	0.042	0.079	0.129	0.195	0.280	0.386	0.517	0.685
	(0.277)	(0.371)	(0.457)	(0.526)	(0.566)	(0.570)	(0.529)	(0.441)	(0.312)
8.0	0.795	0.783	0.771	0.762	0.756	0.757	0.767	0.795	0.856
	(0.317)	(0.297)	(0.272)	(0.243)	(0.212)	(0.182)	(0.162)	(0.157)	(0.158)
	0.034	0.078	0.130	0.193	0.267	0.352	0.451	0.567	0.712
	(0.524)	(0.582)	(0.619)	(0.631)	(0.617)	(0.576)	(0.508)	(0.416)	(0.304)
9.0	0.564	0.585	0.610	0.639	0.673	0.714	0.762	0.821	0.896
	(0.333)	(0.335)	(0.337)	(0.336)	(0.331)	(0.320)	(0.299)	(0.263)	(0.198)
	0.061	0.126	0.195	0.268	0.345	0.427	0.516	0.615	0.737
	(0.877)	(0.850)	(0.810)	(0.758)	(0.696)	(0.624)	(0.544)	(0.454)	(0.354)

For the mastery score = 1 enter N-xbar in the test mean column

APPENDIX B

A Computer Program To Compute the Reliability Indices for Decision in Mastery Testing and Their Standard Errors of Estimate Based on the Beta-Binomial Model

Disclaimer: The computer program hereafter listed has been written with care and tested extensively under a variety of conditions using tests with 60 or fewer items. The author, however, makes no warranty as to its accuracy and functioning, nor shall the fact of its distribution imply such warranty.

```
C***********************************************************************
C
C       A COMPUTER PROGRAM TO COMPUTE THE RELIABILITY INDICES
C       FOR DECISION IN MASTERY TESTING AND THEIR STANDARD
C       ERRORS OF ESTIMATE BASED ON THE BETA-BINOMIAL MODEL.
C
C       INPUT DATA ARE:
C       FIRST CARD: TITLE CARD. ENTER ANYTHING YOU WANT.
C
C       SECOND CARD: MUST CONTAIN THE FOLLOWING INFORMATION
C
C       N......NUMBER OF ITEMS
C       M......NUMBER OF SUBJECTS OR EXAMINEES
C       K......NUMBER OF CLASSIFICATION CATEGORIES
C       XBAR...MEAN OF TEST SCORES
C       SD.....STANDARD DEVIATION OF TEST SCORES
C       FORMAT FOR SECOND CARD IS (3I5,2F10.5).
C
C       THIRD CARD: MUST CONTAIN THE (K-1) CUTOFF SCORES.
C       FORMAT IS (16I5).
C
C***********************************************************************
C
C       REMARK: THIS PROGRAM IS SET UP FOR TESTS WITH
C       UP TO 60 ITEMS. FOR LONGER TESTS USE THE FOLLOWING
C       DIMENSION MODIFICATIONS IN SUBROUTINE KAPPA.
C
C       LET N BE THE NUMBER OF TEST ITEMS.
C       THEN THE DIMENSION OF F(.), XA(.) XB(.) AND CF(.) IS N+1.
C
C
C       ALSO UP TO 17 CLASSIFICATION CATEGORIES CAN BE ACCOMMODATED.
C       FOR MORE CATEGORIES CHANGE L(17) TO L(K) IN THE MAIN
C       PROGRAM, K BEING THE NUMBER OF CATEGORIES.
C
C***********************************************************************
C
      DIMENSION TITLE(20),L(17)
      DOUBLE PRECISION A,B,F
1     READ(5,100,END=99) TITLE
100   FORMAT(20A4)
      WRITE(6,200) TITLE
200   FORMAT('1' /////T10,'ESTIMATES OF DECISION RELIABILITY'/
     *           T10,'AND THEIR STANDARD ERRORS IN'/
     *           T10,'MASTERY TESTING BASED ON THE BETA-'/
     *           T10,'BINOMIAL MODEL'/
     *           T10,'TITLE OF THIS JOB IS:'/
     *           T10,20A4/)
      READ(5,105)N,M,K,XBAR,SD
105   FORMAT(3I5,2F10.5)
```

```
      KM1=K-1
      READ(5,110) (L(I),I=1,KM1)
110   FORMAT(16I5)
      WRITE(6,205) N,M,XBAR,SD,K
205   FORMAT(T10,'INPUT DATA ARE:'//
     *       T10,'NUMBER OF ITEMS .. = ',I4/
     *       T10,'NUMBER OF SUBJECTS = ',I4/
     *       T10,'MEAN OF TEST SCORE ............ = ',F10.5/
     *       T10,'STANDARD DEVIATION OF TEST SCORE = ',F10.5/
     *       T10,'NUMBER OF CATEGORIES = ',I4)
      IF(K.EQ.2) WRITE(6,206) L(1)
206   FORMAT(T10,'CUTOFF SCORE ....... = ',I4)
      IF(K.GT.2) WRITE(6,207) (L(I),I=1,KM1)
207   FORMAT(T10,'CUTOFF SCORES ...... = ',I4,16I5)
      F=N/(N-1.)*(1.-XBAR*(N-XBAR)/(N*SD**2))
      IF(F.GT.0.) GOTO 5
      WRITE(6,210)
210   FORMAT(/T10,'NON-POSITIVE ESTIMATE KR21.'/
     *       T10,'MOMENT ESTIMATES FOR ALPHA AND BETA DO NOT EXIST.'/
     *       T10,'COMPUTATIONS DISCONTINUED FOR THIS CASE.')
      GOTO 1
5     A=(-1.+1./F)*XBAR
      B=-A+N/F-N
      CALL      KAPPA(N,A,B,K,L,M,XP,S    XK,SDK)
      WRITE(6,215) A,B,F,XP,SDP,XK,SDK
215   FORMAT(/T10,'OUTPUT DATA ARE:'//
     *       T10,'ALPHA = ',F10.5/
     *       T10,'BETA  = ',F10.5/
     *       T10,'KR21  = ',F10.5//
     *       T10,'RAW AGREEMENT INDEX P = ',F8.5/
     *       T10,'STANDARD ERROR OF P.. = ',F8.5//
     *       T10,'KAPPA INDEX ........... = ',F8.5/
     *       T10,'STANDARD ERROR OF KAPPA = ',F8.5)
      WRITE(6,220)
220   FORMAT('0',//,T7,'** NORMAL END FOR THIS JOB **'/
     *              T10,'PROGRAM WRITTEN BY HUYNH HUYNH'/
     *              T10,'COLLEGE OF EDUCATION'/
     *              T10,'UNIVERSITY OF SOUTH CAROLINA'/
     *              T10,'COLUMBIA, SOUTH CAROLINA 29208'/
     *              T10,'REVISED, DECEMBER 1979')
      GOTO 1
99    STOP
      END
      SUBROUTINE KAPPA(N,A,B,K,L,M,XP,SL  ,XK,SDK)
      DIMENSION F(61),CF(61),XA(61),XB(61),L(1)
      DOUBLE PRECISION A,B,F,CF,XA,XB,P,PC,A1,A2,A3,VA,VB,VAB,TWO,VKP,
     *                 VP,DPA,DPB,DPCA,DPCB,BFZ,DBFA,DBFB,DSA,DSB,SUMBF
      TWO=2.D0
C
      L(K)=N+1
```

```
C
      CALL        NEHY(N,A,B,F,CF)
      CALL        VARAB(N,A,B,VA,VB,VAB,M,F,XA,XB)
      CALL        ZERLAB(N,A,B,XA,XB,F)
C
      PC=CF(L(1))**2
      DPCA=TWO*CF(L(1))*XA(L(1))
      DPCB=TWO*CF(L(1))*XB(L(1))
C
      DO 5 I=2,K
      IM1=I-1
      A1=CF(L(I))-CF(L(IM1))
      PC=PC+A1*A1
      DPCA=DPCA+TWO*A1*(XA(L(I))-XA(L(IM1)))
    5 DPCB=DPCB+TWO*A1*(XB(L(I))-XB(L(IM1)))
C
      IF(K.GT.2) GOTO 9
C
C     OTHERWISE THERE ARE TWO CATEGORIES.
C
      ICUT=L(1)-1
      IF(2*L(1).LE.N) GOTO 6
      ICUT=N-L(1)
      CALL BF(N,0,ICUT,B,A,BFZ,DBFB,DBFA,DSB,DSA,SUMBF)
      A1=CF(L(2))-CF(L(1))
      P=1.D0-2.0*(A1-SUMBF)
      DPA=-2.D0*(XA(L(2))-XA(L(1))-DSA)
      DPB=-2.D0*(XB(L(2))-XB(L(1))-DSB)
      GOTO 15
C
    6 CALL BF(N,0,ICUT,A,B,BFZ,DBFA,DBFB,DSA,DSB,SUMBF)
      A1=CF(L(1))
      P=1.D0-2.D0*(A1-SUMBF)
      DPA=-2.D0*(XA(L(1))-DSA)
      DPB=-2.D0*(XB(L(1))-DSB)
      GOTO 15
C
    9 DPA=0.D0
      DPB=0.D0
      P=0.D0
C
      DO 10 I=1,K
      LL=0
      IF(I.GT.1) LL=L(I-1)
      LU=L(I)-1
      CALL BF(N,LL,LU,A,B,BFZ,DBFA,DBFB,DSA,DSB,SUMBF)
      P=P+SUMBF
      DPA=DPA+DSA
   10 DPB=DPB+DSB
C
```

```
   15 A1=1.D0-PC
      A2=1.D0-P
      A3=A1*A1
      DKA=(DPA*A1-DPCA*A2)/A3
      DKB=(DPB*A1-DPCB*A2)/A3
C
      VKP=VA*DKA**2+VB*DKB**2+2*VAB*DKA*DKB
      VP=VA*DPA**2+VB*DPB**2+2*VAB*DPA*DPB
      SDK=VKP**.5
      XP=P
      SDP=VP**.5
      XK=(P-PC)/A1
C
      RETURN
      END
      SUBROUTINE NEHY(N,A,B,F,CF)
      DIMENSION F(1),CF(1)
      DOUBLE PRECISION A,B,F,CF,Z1,Z2
      Z1=DFLOAT(N)+B
      Z2=Z1+A
      K=0
      F(1)=1.D0
      DO 5 I=1,N
 5    F(1)=F(1)*(Z1-DFLOAT(I))/(Z2-DFLOAT(I))
 10   KP1=K+1
      KP2=K+2
      F(KP2)=F(KP1)*DFLOAT(N-K)*(A+DFLOAT(K))/
     *                       (DFLOAT(KP1)*(Z1-DFLOAT(KP1)))
      K=K+1
      IF(K-N) 10,15,15
 15   CF(1)=F(1)
      DO 20 I=1,N
      IP1=I+1
 20   CF(IP1)=CF(I)+F(IP1)
 25   RETURN
      END
      SUBROUTINE BF(N,LL,LU,A,B,BFZ,DBFA,DBFB,DSA,DSB,SUMBF)
      DOUBLE PRECISION A,B,Z1,Z2,BFZ,SUMBF,AA,T,X,Y,DBFA,DBFB,DSA,
     *                 DSB,Z1M1,XA,XB,DN,AAHOLD,XAHOLD,XBHOLD,DLL
      N2=N+N
      IR=LU-LL+1
      DN=DFLOAT(N)
      Z1=DFLOAT(N2)+B
      Z1M1=Z1-1.D0
      Z2=Z1+A
      DLL=DFLOAT(LL)
C
      IF(LL.NE.0) GOTO 10
C
      AA=1.D0
```

```
      XA=0.D0
      XB=0.D0
C
      DO 5 I=1,N2
      T=DFLOAT(I)
      AA=AA*(Z1-T)/(Z2-T)
      XA=XA-1.D0/(Z2-T)
    5 XB=XB+1.D0/(Z1-T)
C
      XB=XB+XA
C
      GOTO 15
C
   10 X=DLL-1.D0
      Y=DLL-1.D0
      AA=BFZ*(DN-X)*(A+X+Y)/((X+1.D0)*(Z1M1-X-Y))
      XA=DBFA+1.D0/(A+X+Y)
      XB=DBFB-1.D0/(Z1M1-X-Y)
C
      X=LL
      AA=AA*(DN-Y)*(A+X+Y)/((Y+1.D0)*(Z1M1-X-Y))
      XA=XA+1.D0/(A+X+Y)
      XB=XB-1.D0/(Z1M1-X-Y)
C
   15 SUMBF=AA
      DSA=XA*AA
      DSB=XB*AA
C
      IF(IR.EQ.1) GOTO 90
C
      AAHOLD=AA
      XAHOLD=XA
      XBHOLD=XB
C
      DO 50 I=2,IR
      X=DLL+DFLOAT(I-2)
      Y=DLL
      AA=AAHOLD*(DN-X)*(A+X+Y)/((X+1.D0)*(Z1M1-X-Y))
      XA=XAHOLD+1.D0/(A+X+Y)
      XB=XBHOLD-1.D0/(Z1M1-X-Y)
C
      DSA=DSA+2.D0*XA*AA
      DSB=DSB+2.D0*XB*AA
      SUMBF=SUMBF+2.D0*AA
C
      AAHOLD=AA
      XAHOLD=XA
      XBHOLD=XB
C
      X=X+1.D0
```

```
      DO 50 J=2,I
      Y=DLL+DFLOAT(J)-2.D0
C
      AA=AA*(DN-Y)*(A+X+Y)/((Y+1.D0)*(Z1M1-X-Y))
      XA=XA+1.D0/(A+X+Y)
      XB=XB-1.D0/(Z1M1-X-Y)
C
      IF(I.EQ.J) GOTO 40
      SUMBF=SUMBF+2.D0*AA
      DSA=DSA+2.D0*XA*AA
      DSB=DSB+2.D0*XB*AA
      GOTO 50
C
   40 SUMBF=SUMBF+AA
      DSA=DSA+XA*AA
      DSB=DSB+XB*AA
   50 CONTINUE
C
   90 BFZ=AA
      DBFA=XA
      DBFB=XB
C
      RETURN
      END
      SUBROUTINE ZERLAB(N,A,B,XA,XB,F)
      DIMENSION XA(1),XB(1),F(1)
      DOUBLE PRECISION A,B,Z1,Z2,XA,XB,F,ONE
      ONE=1.D0
C
C
      XA(1)=0.D0
      XB(1)=0.D0
      Z1=DFLOAT(N)+B
      Z2=Z1+A
      NP1=N+1
      DO 5 I=1,N
      XA(1)=XA(1)-ONE/(Z2-DFLOAT(I))
    5 XB(1)=XB(1)+ONE/(Z1-DFLOAT(I))
      XB(1)=XB(1)+XA(1)
      DO 10 I=1,N
      IP1=I+1
      IX=I-1
      XA(IP1)=XA(I)+ONE/(A+DFLOAT(IX))
   10 XB(IP1)=XB(I)-ONE/(Z1-DFLOAT(I))
      XA(1)=XA(1)*F(1)
      XB(1)=XB(1)*F(1)
      DO 30 I=2,NP1
      IM1=I-1
      XA(I)=XA(IM1)+XA(I)*F(I)
   30 XB(I)=XB(IM1)+XB(I)*F(I)

C
      RETURN
      END
      SUBROUTINE VARAB(N,A,B,VA,VB,VAB,M,F,DA,DB)
      DIMENSION F(1),DA(1),DB(1)
      DOUBLE PRECISION A,B,DA,DB,F,B11,B12,B22,D,VA,VB,VAB
      CALL DERLAB (N,A,B,DA,DB)
      B11=0.
      B12=0.
```

```
      B22=0.D0
      NP1=N+1
      DO 15 I=1,NP1
      B11=B11+DA(I)*DA(I)*F(I)
      B12=B12+DA(I)*DB(I)*F(I)
15    B22=B22+DB(I)*DB(I)*F(I)
      B11=B11*M
      B12=B12*M
      B22=B22*M
      D=B11*B22-B12*B12
      VA=B22/D
      VB=B11/D
      VAB=-B12/D
      RETURN
      END
      SUBROUTINE DERLAB(N,A,B,DA,DB)
      DIMENSION DA(1),DB(1)
      DOUBLE PRECISION A,B,DA,DB,Z1,Z2
      DOUBLE PRECISION ONE
      ONE=1.D0
      DA(1)=0.D0
      DB(1)=0.D0
      Z1=DFLOAT(N)+B
      Z2=Z1+A
      NP1=N+1
C
      DO 5 I=1,N
      DA(1)=DA(1)-ONE/(Z2-DFLOAT(I))
5     DB(1)=DB(1)+ONE/(Z1-DFLOAT(I))
      DB(1)=DB(1)+DA(1)
C
      DO 10 I=1,N
      IP1=I+1
      IX=I-1
      DA(IP1)=DA(I)+ONE/(A+DFLOAT(IX))
10    DB(IP1)=DB(I)-ONE/(Z1-DFLOAT(I))
      RETURN
      END
```

8

ACCURACY OF TWO PROCEDURES FOR ESTIMATING RELIABILITY OF MASTERY TESTS

Huynh Huynh
Joseph C. Saunders

University of South Carolina

Presented at the annual conference of the Eastern Educational Research Association, Kiawah Island, South Carolina, February 22-24, 1979. A short version of this paper will appear in Journal of Educational Measurement (in press).

ABSTRACT

Single administration (beta-binomial) estimates for the raw agreement index p and the corrected-for-chance kappa index in mastery testing are compared with those based on repeated test administrations in terms of estimation bias and sampling variability. Across a variety of test score distributions, test lengths, and mastery (cutoff) scores. the beta-binomial estimates tend to underestimate the corresponding population values. The percent of bias is small (about 2.5%) and p and somewhat larger (about 10%) for kappa. Both beta-binomial estimates have standard errors about one-half the size of the standard errors of estimates based on repeated test administrations. Though the beta-binomial estimates presume equality of item difficulty, the data presented indicate that even gross departures from equality of item difficulty do not affect the amount of bias of the estimates.

This paper has been distributed separately as RM 79-1, February, 1979.

1. INTRODUCTION

In mastery testing reliability is often viewed as the consistency of mastery-nonmastery decisions across repeated test administrations (Huynh, 1976, 1978a; Subkoviak, 1976). Two reliability indices have been proposed and studied for mastery tests. They are the raw agreement index p and the corrected-for-chance kappa index (κ). The first index represents the proportion of examinees consistently classified in the same (mastery or nonmastery) category over two test administrations using the same form or two equivalent forms. It is assumed, of course, that the first testing does not induce any lasting change in the examinees. The second index, kappa, is defined as $\kappa = (p-p_c)/(1-p_c)$, where p_c is the proportion of consistent classification expected under complete random assignment. Thus kappa reflects the extent to which test scores will improve the consistency of decisions beyond the level expected by random classification. The relationship between kappa and other parameters such as cutoff score and classical test reliability may be found in Huynh (1978a).

The definitions of both p and kappa assume the feasibility of repeated test administrations. This may not be practical in many instances. Under some conditions, p and kappa may be approximated from a single test administration. There are at least two procedures to accomplish this, namely, those described in Huynh (1976) and Subkoviak (1976). The Huynh procedure assumes that the test scores are distributed as predicted by a univariate or bivariate beta-binomial model. On the other hand, the Subkoviak technique, in its simplest form, assumes that test scores are distributed as predicted by a binomial distribution and that the regression of true score on observed test score is linear.

Subkoviak (1978) has provided a comparison of these two procedures using simulations with fifty repetitions. The data reported in Table 2 of his paper clearly indicate that both procedures act almost identically in terms of estimation bias and standard error. This is an expected result. Linear regression of true score on observed score in the binomial error model automatically implies that the test score distribution under study must belong to the negative hypergeometric (beta-binomial) family (Lord & Novick, 1968, p. 516). Hence it appears that the conditions underlying the Subkoviak procedure are those of the beta-binomial distributio assumed in Huynh's paper (1976). For this reason and for inherent complexities in formulating inferential techniques associated with the Subkoviak procedure, this paper will be restricted to the beta-binomial model in the estimation of reliability for mastery tests.

The purpose of this paper is to compare the accuracy of two procedures for estimating reliability of decisions in mastery testing. One procedure is based on two test administrations; the other procedure relies on only one test administration and performs all computations assuming the appropriateness of the beta-binomial model for the test data under study. Sections 2, 3, and 4 deal with the asymptotic (large sample of examinees) nature of the estimates. Section 5 reports a simulation study for the case of small samples.

2. ASYMPTOTIC BIAS AND STANDARD ERRORS

Though the number of classification categories may be arbitrary, we will consider only the case of two categories, labeled mastery and nonmastery. The lowest score for which an examinee will be classified as a master will be referred to as the mastery (or passing) score in subsequent discussion.

First let us consider estimating p and κ by testing a sample of m examinees twice. Let p_{ij} be the proportion of examinees classified in the i-th category on the first testing and in the j-th category in the second testing. Here let i = 0 for a nonmaster and i = 1 for a master. Let the dot (.) bear the regular summation meaning. For example, the marginal proportion of masters on the first testing is $p_{1.} = p_{10} + p_{11}$.

The observed proportion* of consistent classifications in the sample at hand is $\hat{p}_R = p_{00} + p_{11}$ and the kappa index for this sample is

*The subscript R means repeated testings.

$$\hat{\kappa}_R = (\hat{p}_R - \hat{p}_c)/(1-\hat{p}_c) \quad (1)$$

where $p_c = p_{o.}p_{.o} + p_{1.}p_{.1}$. Under random sampling, $\hat{p}_R$ is an <u>efficient</u> statistic for the parameter p (Hogg & Craig, 1970, p. 372). In other words, $\hat{p}_R$ is <u>unbiased</u> and its standard error is equal to the Rao-Cramér lower bound. This standard error is $(p(1-p)/m)^{1/2}$. It may also be noted that $\hat{p}_R$ is also the maximum likelihood (ML) estimate of the population value of p and that $\hat{\kappa}_R$ is an ML estimate of the population value of κ. Its asymptotic (large sample) properties are well known. For example, $\hat{\kappa}_R$ follows an approximate normal distribution with mean κ and with a variance of

$$\frac{1}{m}\left\{\frac{p(1-p)}{(1-p_c)^2} + \frac{2(1-p)(2pp_c-a)}{(1-p_c)^3} + \frac{(1-p)^2(b-4p_c^2)}{(1-p_c)^4}\right\} \quad (2)$$

where

$$a = p_{oo}(p_{o.} + p_{.o}) + p_{11}(p_{1.} + p_{.1}) \quad (3)$$

and

$$b = \sum_{i,j} p_{ij}(p_{j.} + p_{.i})^2 \quad (4)$$

(Bishop <u>et al</u>., 1974, p. 396). In these formulae, all quantities listed are population values. When sample proportions are used in (2), the resulting value is an estimate for the variance of κ_R. Finally, since the asymptotic mean of $\hat{\kappa}_R$ is κ, $\hat{\kappa}_R$ is asymptotically an unbiased estimate for this parameter.

Consider now estimating p and κ from a single test administration. The estimates*, $\hat{p}_B$ and $\hat{\kappa}_B$, are described in detail in Huynh (1976); the asymptotic standard errors of both estimates may be obtained via the formulae, tables, or computer program described elsewhere (Huynh, 1978b). In the latter paper it is also shown that $\hat{p}_B$ and $\hat{\kappa}_B$ are asymptotically unbiased estimates of p and κ.

3. A COMPARISON OF THE ASYMPTOTIC STANDARD ERRORS OF ESTIMATE FOR BETA-BINOMIAL TEST DATA

Whether estimation is based on repeated or single testings, $\sqrt{m}$ times the standard error (S.E.) of the estimate is (or is

*The subscript B refers to the <u>b</u>eta-binomial model.

asymptotically) not a function of the sample size m. Thus m is not a significant factor in any comparison of the estimates as long as sufficiently large samples are to be considered. In this section and most subsequent ones, only the quantity $G = \sqrt{m} \times S.E.$ will be considered.

The comparisons described in this section are limited to test score distributions that follow the beta-binomial distribution. Strictly speaking, the procedure for estimating from a single administration (Huynh, 1976) is formulated only for this type of data.

The comparison was made for selected situations with $n = 5$, 10, 20, and 30 test items. The test mean (μ) and KR21 reliability (α_{21}) were chosen such that the resulting test score distribution would be one of the following types: (i) U-shaped with the higher-density mode at the upper end of the score range, (ii) symmetric, (iii) unimodal with a mode somewhere between μ and n, or (iv) J-shaped. The passing score c was chosen such that the ratio c/n would be 60, 70, or 80%. The G-values for $\hat{\kappa}_R$ were computed via Equations (2), (3), and (4) with the p_{ij} proportions generated by the bivariate beta-binomial model. The G-values for $\hat{p}_B$ and $\hat{\kappa}_B$ were obtained via the computer program described in Huynh (1978b).

Table 1 reports the obtained G-values when the two procedures for estimating p and κ are used. The G-values in the table clearly demonstrate that the standard error associated with the single administration (beta-binomial) procedure is uniformly smaller than that encountered with the procedure using two test administrations. Over the thirteen situations reported in Table 1, the standard errors for the single administration procedure average 59.3% of those from repeated administrations for the p index and 53.2% for the kappa index.

4. A COMPARISON OF THE ASYMPTOTIC BIAS AND STANDARD ERRORS OF ESTIMATE FOR CTBS TEST DATA

This phase of the study is motivated by the fact that real test data rarely conform exactly to a well-specified model such as

TABLE 1

G-Values for Beta-Binomial Test Data

							Index p			Kappa		
α	β	Shape	n	μ	σ	c	p	$G(\hat{p}_B)$	$G(\hat{p}_R)$	κ	$G(\hat{\kappa}_B)$	$G(\hat{\kappa}_R)$
5.0	3.0	Unimodal	5	3.125	1.301	3	.687	.320	.464	.270	.763	1.021
						4	.645	.350	.479	.273	.752	.970
2.0	.5	J-Shaped	5	4.000	1.309	3	.872	.168	.334	.492	.713	1.226
						4	.811	.265	.391	.526	.619	.953
.5	.2	U-Shaped	5	3.571	1.850	3	.907	.145	.291	.765	.379	.727
6.0	6.0	Symmetric	5	2.500	1.279	3	.605	.412	.489	.210	.823	.978
10.0	5.0	Unimodal	10	6.667	1.863	7	.644	.331	.479	.277	.663	.966
						8	.661	.280	.473	.262	.660	.966
8.0	2.0	Unimodal	10	8.000	1.706	7	.799	.222	.401	.332	.677	1.175
						8	.714	.295	.452	.357	.630	.984
4.5	.5	J-Shaped	10	9.000	1.500	7	.921	.135	.269	.454	.785	1.637
12.0	8.0	Unimodal	20	12.000	3.024	12	.678	.269	.467	.342	.550	.949
						14	.704	.235	.456	.326	.561	.998
12.0	3.0	Unimodal	20	16.000	2.646	12	.918	.169	.275	.304	.677	1.796
						14	.821	.192	.383	.370	.591	1.201
3.0	.5	J-Shaped	20	17.143	3.576	12	.940	.087	.237	.637	.478	1.369
16.0	14.0	Unimodal	30	16.000	3.801	20	.787	.212	.409	.290	.585	1.178
						24	.964	.123	.185	.142	.557	2.448
18.0	2.0	Unimodal	30	27.000	2.535	20	.982	.081	.133	.246	.775	3.716
						24	.888	.169	.315	.373	.650	1.496
19.5	.5	J-Shaped	30	29.250	1.319	24	.990	.062	.099	.273	1.105	5.038

the beta-binomial distribution. It is based on a portion of the Comprehensive Tests of Basic Skills (CTBS) test data collected in the 1978 South Carolina Statewide Testing Program. Table 2 describes the various tests artificially assembled from CTBS subtests or from the entire battery. For each test in the listing, two alternate (hopefully equivalent) forms were created by pairing items on the basis of content and/or difficulty and randomly assigning the items in each pair to the alternate forms. For reasons which will be obvious later on, a number of tests were deliberately constructed of items of similar difficulty.

The number of items (n) was set at 5, 10, 15, and 20. The number of students, selected by taking every tenth case from the entire South Carolina file, ranged from m = 1684 to 6035. For each test, the value D_{max} represents the maximum discrepancy between the observed relative cumulative frequency and the corresponding expected frequency from the beta-binomial model. A significance level (P-value) of more than .20 indicates that the test data follow closely the beta-binomial distribution. On the other hand, P-values of less than .05 or .01 reveal substantial departures from the theoretical distribution.

For each test described in Table 2, the population values p_R, $G(\hat{p}_R)$, κ_R, and $G(\hat{\kappa}_R)$ were computed using the bivariate frequency distribution generated by the alternate forms. The corresponding parameters p_B, $G(\hat{p}_B)$, κ_B, and $G(\hat{\kappa}_B)$ were obtained by imposing the beta-binomial model on each of the two alternate forms and averaging the two sets of results. Now both $\hat{p}_B$ and $\hat{\kappa}_B$ are asymptotic unbiased estimates of p_B and κ_B (Huynh, 1978b). Also, since $\hat{p}_R$ is an unbiased estimate of p_R, and $\hat{\kappa}_B$ is an asymptotically unbiased estimate of κ_R, only the asymptotic bias of $\hat{p}_B$ and $\hat{\kappa}_B$ in estimating p_R and κ_R was explored. Thus, it follows that the percent asymptotic bias for $\hat{p}_P$ and $\hat{\kappa}_B$ is $100(p_B - p_R)/p_R$ and $100(\kappa_B - \kappa_R)/\kappa_R$, respectively. A negative bias indicates underestimation whereas a positive bias documents an overestimation. (We focused on p_R and κ_R because test reliability is typically approached from the standpoint of equivalent forms.) All computations reported in this section were carried out as in the previous section.

TABLE 2

Description of the CTBS Data Used in Sections 4 and 5

Case	n	M	diff	D_{max} (%)	P-value	Grade	Description
5.1	5	1684	.056	1.80	>.20	3	Reading comprehension (paragraph)
5.2	5	1684	.107	0.68	>.20	3	Language expression
5.3	5	5543	.003	0.50	>.20	3	Total battery
10.1	10	1684	.060	2.24	>.20	3	Reading comprehension (sentences)
10.2	10	6035	.081	1.54	>.15	6	Reading vocabulary
10.3	10	5543	.007	2.02	<.05	3	Total battery
15.1	15	1684	.175	1.72	>.20	3	Science
15.4	15	1335	.022	3.85	<.05	6	Total battery
20.1	20	1684	.099	4.01	<.01	3	Mathematics
20.3	20	5543	.015	7.65	<.01	3	Total battery

Table 3 details the results of the various estimates for p_R and κ_R. The data indicate that the beta-binomial estimates ($\hat{p}_B$ and $\hat{\kappa}_B$) tend to underestimate the alternate-form population values. For the p index, the percent of bias ranges from -4.2 to 0.1 with an average of -2.3. A larger degree of bias, however, occurs in the estimation of kappa via $\hat{\kappa}_B$. The percent of bias for this estimate ranges from -17.5 to 0.9 with an average of about -7.8.

The larger bias of $\hat{\kappa}_B$ as compared with that of $\hat{p}_B$ is to be expected. With the factor 1 - p (which cannot exceed .50) in the denominator of Equation (1) defining kappa, the bias of $\hat{\kappa}_B$ is at least twice as large as that associated with $\hat{p}_B$. For situations in which a high proportion of examinees are to be classified either as masters or nonmasters, $1 - p_c$ is close to zero. As a consequence, the bias of $\hat{\kappa}_B$ will become more pronounced in those cases.

The beta-binomial model assumes that test items are equally difficult (Huynh, 1976). It would be natural to expect that the bias of the beta-binomial estimates would bear a positive (or direct) relationship with variation in item difficulty. This is not the case, however. The values of D_{max} in Table 2 clearly indicate that departures from the beta-binomial distribution show no resemblance to the standard deviation (σ_{diff}) of item difficulty.

TABLE 3

Percent Asymptotic Bias and G-Values for CTBS Test Data

			Index p			Kappa		
Case	n	Cutoff Score	% Bias	$G(\hat{p}_B)$	$G(\hat{p}_R)$	% Bias	$G(\hat{\kappa}_B)$	$G(\hat{\kappa}_R)$
5.1	5	3	-1.5	.174	.331	-7.1	.540	.403
		4	-3.5	.236	.350	-9.2	.485	.774
5.2	5	3	-2.6	.192	.348	-13.7	.664	1.064
		4	-4.7	.287	.391	-14.1	.593	.856
5.3	5	3	-2.8	.211	.364	-17.5	.734	1.148
		4	-3.4	.325	.429	-11.3	.667	.921
10.1	10	6	-2.9	.113	.256	-10.2	.329	.668
		8	-4.2	.147	.281	-9.7	.294	.604
10.2	10	6	-1.3	.136	.330	-5.3	.384	.832
		8	-3.6	.176	.347	-8.7	.345	.707
10.3	10	6	0.7	.136	.332	2.5	.537	1.165
		8	-1.2	.208	.385	-4.4	.441	.862
15.1	15	9	-2.6	.203	.403	-8.1	.407	.809
		13	-3.7	.164	.317	-7.6	.530	1.300
15.2	15	9	-1.9	.168	.393	-4.0	.351	.881
		13	-0.4	.141	.295	-7.1	.506	1.313
20.1	20	12	-2.7	.098	.241	-12.9	.412	1.040
		14	-2.8	.115	.292	-7.7	.353	.880
20.2	20	12	0.1	.132	.370	0.9	.267	.751
		14	-0.7	.121	.355	0.0	.283	.805

The same observation holds for the bias of $\hat{p}_B$ and $\hat{\kappa}_B$ as displayed in Table 3.

The G-values of Table 3 clearly show that the estimates based on the beta-binomial model have a smaller standard error of estimate than those based on alternate forms. Over all the situations considered, the standard error of $\hat{p}_B$ is about 50.4% of that of $\hat{p}_R$; the standard error of $\hat{\kappa}_B$ is about 50.2% of that of $\hat{\kappa}_R$. These results are consistent with those of Section 3.

5. A COMPARISON OF FINITE-SAMPLE BIAS AND STANDARD ERRORS OF ESTIMATE FOR CTBS TEST DATA

A simulation was conducted to study the sampling fluctuations of the estimates $\hat{p}_B$, $\hat{\kappa}_B$, and $\hat{\kappa}_R$ when sample sizes are of small or moderate size. This was done for samples of size m = 20, 40, and 60. For each test, one thousand replications were used to obtain the observed percent of bias and G-value for $\hat{\kappa}_R$. As for estimates based on the beta-binomial model, one thousand replications were simulated for each alternate form and the averages of the two sets of results were used to determine the bias and G-value for $\hat{p}_B$ and $\hat{\kappa}_B$.

Table 4 presents a summary of the results of simulation. The adequacy of the random number generator (more specifically, the IMSL (1977) subroutine GGUB) is documented by the near zero bias of $\hat{p}_R$ and the small fluctuation of the $G(\hat{p}_R)$ values for various sample sizes around the corresponding true values (enclosed in parentheses). The data reported in the table clearly show that, as in the case of large samples, the beta-binomial model tends to underestimate the parameters p_R and κ_R. The bias of $\hat{p}_B$ in estimating p_B averages -2.6%. For kappa, the bias of $\hat{\kappa}_B$ fluctuates around -11.0%. It is also interesting to note that the alternate form estimate, $\hat{\kappa}_R$, also tends to have a small negative bias.

TABLE 4

Percent Finite-Sample Bias and G-Values for CTBS Test Data

Case	n	Cutoff Score	m	$\hat{p}_B$ % Bias	$G(\hat{p}_B)$	$\hat{p}_R$ % Bias	$G(\hat{p}_R)$	$\hat{\kappa}_B$ % Bias	$G(\hat{\kappa}_B)$	$\hat{\kappa}_R$ % Bias	$G(\hat{\kappa}_R)$
5.1	5	3	20	-0.5	.186	-.4	.325	-8.6	.617	+1.5	1.005
			40	-0.1	.184	-.1	.335	-7.7	.569	-1.3	.936
			60	-1.1	188	-.1	.334	-7.4	.553	-0.3	.930
				(Exact	value	0	.331)				
10.1	10	7	20	-3.6	.141	-.1	.225	-11.9	.376	-1.3	.678
			40	-3.9	.146	.2	.269	-11.6	.327	-1.2	.644
			60	-4.0	.145	-.1	.268	-11.4	.304	-0.4	.625
				(Exact	value	0	.259)				
15.1	15	11	20	-3.4	.210	-.4	.395	-15.1	.543	-2.4	.949
			40	-3.8	.206	.3	.402	-13.4	.525	-2.2	.927
			60	-3.7	.203	-.2	.397	-13.0	.523	-0.1	.927
				(Exact	value	0	.392)				
20.1	20	14	20	-0.7	.141	-.2	.293	-12.7	.585	-5.0	1.017
			40	-2.6	.137	0	.306	-10.2	.519	-1.3	.961
			60	-2.6	.142	.2	.312	-9.2	.499	-2.2	.942
				(Exact	value	0	.292)				

The data in Table 4 show that the beta-binomial estimates have smaller sampling fluctuations than the alternate form estimates. For all situations reported in this table, the standard error of $\hat{p}_B$ is about 51.4% of that of $\hat{p}_R$; and the standard error of $\hat{\kappa}_B$ is about 56.9% of the standard error of $\hat{\kappa}_R$. These trends are very similar to those reported in the previous section.

6. DISCUSSION AND CONCLUSION

In this study the performance of a single administration estimate of reliability for mastery tests is compared with the behavior of the estimate based on two test administrations. The results clearly indicate that the single administration (beta-binomial) estimate for the raw agreement index p behaves very well. Not only does it show a negligible amount of negative bias, its sampling error is about half of that of the test-retest procedure. As for the kappa index, a moderate degree of negative bias (about ten percent) is displayed by the beta-binomial estimate. This estimate of kappa also has a standard error that is about one-half the corresponding value for the alternate form estimate. Though the beta-binomial estimates are originally derived for tests with items of equal difficulty, the data presented indicate that the bias of these estimates does not depend on the assumption of equal difficulty for test items. Our conclusion is that for testing situations involving tests like the CTBS (with items of a wide range of difficulty), the estimation for consistency of decisions in mastery tests may be safely carried out via one test administration with the beta-binomial model as a vehicle for computation.

BIBLIOGRAPHY

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, Massachusetts: The MIT Press.

Hogg, R. V. & Craig, A. T. (1970). An introduction to mathematical statistics (Third edition). New York: The MacMillan Company.

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

Huynh, H. (1978a). Reliability of multiple classification. Psychometrika 43, 317-325.

Huynh, H. (1978b). Computation and inference for two reliability indices in mastery testing based on the beta-binomial model. Research Memorandum 78-1, Publication Series in Mastery Testing. University of South Carolina College of Education.

IMSL Library 1 (1977). Houston: International Mathematical and Statistical Libraries.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company.

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

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

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not neccessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance and comments of Anthony J. Nitko are gratefully acknowledged.

9

AN APPROXIMATION TO THE TRUE ABILITY DISTRIBUTION IN THE BINOMIAL ERROR MODEL AND APPLICATIONS

Huynh Huynh

Garrett K. Mandeville

University of South Carolina

ABSTRACT

Assuming that the density p of the true ability θ in the binomial test score model is continuous in the closed interval [0,1], a Bernstein polynomial can be used to uniformly approximate p. Then via quadratic programming techniques, least-square estimates may be obtained for the coefficients defining the polynomial. The approximation, in turn will yield estimates for any indices based on the univariate and/or bivariate density function associated with the binomial test score model. Numerical illustrations are provided for the projection of decision reliability and proportion of success in mastery testing.

1. INTRODUCTION

The binomial error model (Lord and Novick, 1968) has been used extensively in analyses of mental test data. The model is deemed suitable in computer-assisted testing in which each examinee is

This paper has been distributed separately as RM 79-5, June, 1979.

given a random sample of items drawn from a large item universe. When the same test is given to all examinees, the binomial distribution implies that all items share the same difficulty level. There are indications (Keats and Lord, 1962; Duncan, 1974) that several test score distributions based on the same test fit the binomial (or more specifically the beta-binomial) model quite well, especially when similarity of item difficulty holds strictly or nearly. Let x denote the test score obtained from the administration of an n-item test to an examinee with true ability θ (the proportion of items in the universe that he/she knows, or the probability of answering each item correctly). Then the conditional density of x given θ is

$$f(x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \ x = 0,1,\ldots,n.$$

Let $p(\theta)$ be the density of the true ability for a population of examinees. The marginal density of x for this population is given as

$$f(x) = \binom{n}{x} \int_0^1 \theta^x (1 - \theta)^{n-x} p(\theta) \, d\theta.$$

As indicated in Lord and Novick (1968; Chapter 23), the knowledge of $f(x)$ implies the knowledge of the first n moments of the distribution of θ. Any distribution sharing these n moments will yield the same marginal density $f(x)$, hence the solution for $p(\theta)$ given $f(x)$ is not unique. We will seek an approximation for $p(\theta)$ via a polynomial and will show how such approximation is useful in the projection of decision reliability and proportion of successes in mastery testing.

2. A SOLUTION BASED ON THE BERNSTEIN POLYNOMIAL

We shall assume that $p(\theta)$ is continuous in the closed interval [0,1]. Then (Feller, 1966, p. 220) $p(\theta)$ can be uniformly approximated by a Bernstein polynomial of the form

$$B_m(\theta) = \sum_{k=0}^{m} z_k \binom{m}{k} \theta^k (1 - \theta)^{m-k}.$$

Thus given any arbitrarily small and positive ε, there exists an integer m and (m + 1) constants z_i such that $|B_m(\theta) - p(\theta)| < \varepsilon$ for all $\theta \in [0,1]$. We propose to use $B_m(\theta)$ to approximate $p(\theta)$. Procedures will be presented for the determination of the constants $m, z_0, z_1, \ldots, z_m$.

It may first be noted that the z_k constants must be non-negative and satisfy the constraint $\int_0^1 B_m(\theta)\, d\theta = 1$ in order for $B_m(\theta)$ to be a density. Hence

$$\sum_{k=0}^{m} z_k \binom{m}{k} \int_0^1 \theta^k (1 - \theta)^{m-k}\, d\theta = 1$$

or equivalently

$$\sum_{k=0}^{m} z_k = m + 1.$$

The Bernstein approximated value for the marginal density of x is now given as

$$f_B(x) = \binom{n}{x} \sum_{k=0}^{m} z_k \binom{m}{k} J(n + m; x + k)$$

where

$$J(n + m; x + k) = \int_0^1 \theta^{x+k} (1 - \theta)^{n+m-(x+k)} d\theta.$$

The J integrals may be computed inductively by noting that

$$J(p;0) = 1/(p + 1)$$

and

$$J(p;y + 1) = (y + 1)\, J(p;y)/(p - y).$$

Now let

$$c(k,x) = \binom{n}{x}\binom{m}{k} J(x + k)$$

and

$$\alpha(k,x) = c(k,x) - c(0,x).$$

Then the approximated marginal density of x becomes

$$f_B(x) = \sum_{k=1}^{m} \alpha(k,x)\, z_k + (m + 1)\, c(0,x)$$

where the z_k, k = 1, 2,..., m are nonnegative and sum up to no more than m + 1.

To determine the constants m, $z_1, z_2, \ldots, z_m$, we focus on the least-square criterion with the weight function w(x)

$$H(z_1, z_2, \ldots, z_m; m) = \sum_{x=0}^{n} w(x)\,[f_B(x) - f(x)]^2. \tag{1}$$

In other words, we will seek these constants in such a way that the H criterion is minimized. This may be done by first considering m as fixed and computing the z constants along with the minimum H_m of the criterion H. This process will be repeated many times starting with m = 0 [$p(\theta)$ and $f_B(x)$ are constant], 1, 2, etc. until an integer m can be located at which H_m is minimized. Following are the details for the algorithm.

<u>2.1 Minimizing H at Each Integer m.</u> Let

$\beta(x) = (m + 1)\, c(0,x) - f(x).$

Then (1) becomes

$$H = \sum_{x=0}^{n} [w(x) \sum_{k=1}^{m} \alpha(k,x) z_k + \beta(x)]^2. \tag{2}$$

At each given integer m, the nonnegative $z_1, z_2, \ldots, z_m$ may be obtained by minimizing H under the constraint $\Sigma z_k \leq m + 1$. Since H is continuous and the z's are located in a closed region, the solution for z <u>always</u> exists. To obtain such solution, standard routines for quadratic programming may be called upon. In this paper, Algorithm 431 (Ravindran, 1972) was used.

To enter into Algorithm 431, we note that the criterion H of (2) may be written as

$H = Z'DZ + 2BZ + C.$

In this formula, Z is the vector $(z_1, z_2, \ldots, z_m)'$, $D = (d_{kk'})$ is the matrix defined by

$$d_{kk} = \sum_{x=0}^{n} w(x)\,[\alpha(k,x)]^2$$

$$d_{kk'} = \sum_{x=0}^{n} w(x)\,\alpha(k,x)\,\alpha(k',x)$$

and $B = (b_k)$ is a vector with components

$$b_k = \sum_{x=0}^{n} w(x)\ \alpha(k,w)\ \beta(x).$$

The remaining quantity C is the constant

$$C = \sum_{x=0}^{n} w(x)\ [\beta(x)]^2.$$

2.2 Searching the Least Square Solution. We note that when $m = 0$, the minimum value H_0 of H is simply

$$H_0 = \sum_{x=0}^{n} w(x)\ [f(x) - \bar{f}]^2$$

where

$$\bar{f} = \Sigma w(x)\ f(x)/\Sigma w(x).$$

As for other m values, the minimum may be deduced from the quadratic programming. Thus the least square solution for the Bernstein polynomial may be obtained by computing H_0, H_1, H_2,... for several consecutive values of m, and locating the value of m at which H_m is the smallest. Since the criterion for minimization H is non-negative, all computations shall stop whenever $H_m = 0$. In other situations, a tolerance difference between H_m and H_{m-1} might have to be set up in order to end the approximation process.

3. NUMERICAL ILLUSTRATION

To illustrate the computational algorithm described in the previous section, three score frequency distributions based on $n = 10$ test items are used. For Data Set 1, almost all frequencies are concentrated at the upper end of the score range. Data Set 2 is slightly asymmetric and Data Set 3 has two modes, one near each end of the score range. Details regarding these data sets are presented in Table 1.

It appears from Table I that the goodness of fit via the

Bernstein polynomial improves when the degree of the polynomial increases. For unimodal distributions, the algorithm tends to put all the weights at only a few terms which correspond to some

TABLE 1

Observed and Fitted Frequency Distributions
for Three Data Sets

Test Score	Data Set 1		Data Set 2		Data Set 3	
	Observed	Fitted	Observed	Fitted	Observed	Fitted
0	0	.00	0	.06	4	6.09
1	0	.00	0	.37	10	10.28
2	0	.01	1	1.26	15	10.16
3	0	.07	3	3.07	2	8.68
4	1	.23	6	5.97	6	9.16
5	1	.69	10	9.66	10	12.64
6	3	1.82	13	13.28	20	17.49
7	5	4.42	16	15.47	25	20.22
8	8	9.93	15	14.88	15	17.89
9	15	20.96	11	11.00	10	10.89
10	47	41.91	5	4.97	4	3.50
Degree of the Bernstein polynomial:		10		10		24
Minimum H_m:		.0106		.0001		.0052
The positive z constants:						
		$z_{10} = 11.0000$		$z_7 = 9.891$[illegible]		$z_4 = 6.2830$
				$z_8 = 1.1088$		$z_5 = 1.3349$
						$z_{17} = 14.4010$
						$z_{18} = 3.9830$

consecutive z_i values. On the other hand, for a bimodal distribution such as Data Set 3, the algorithm puts the total weight on two blocks, each being formed by some consecutive z_i values.

4. PROJECTION OF DECISION RELIABILITY

Consider now two equivalent tests X and Y, each with n items. If the test score distributions are binomial, then the bivariate density is given as

$$f(x,y) = \binom{n}{x}\binom{n}{y} \int_0^1 \theta^{x+y} (1-\theta)^{2n-(x+y)}\, p(\theta)\, d\theta.$$

Let the density p be approximated from the data collected with one test as

$$B_m(\cdot) = \sum_{k=0}^{m} z_k \binom{m}{k} \theta^k (1-\theta)^{m-k}.$$

Then f(x,y) will be given by the expression

$$f_B(x,y) = \binom{n}{x}\binom{n}{y} \sum_{k=0}^{m} z_k \binom{m}{k} J(2n+m;\, x+y+k)$$

where the function J is defined as previously in Section 2. The expressions for $f_B(x)$ and $f_B(x,y)$ may now be used to project practically all agreement indices for decisions in mastery testing. Let the examinees now be classified in k categories A_i defined by $A_i = \{x; c_{i-1} \leq x < c_i\}$ where $c_0 = 0$ and $c_k = n + 1$. For binary classifications k = 2. In this case c_1 is usually referred to as <u>the cutoff (mastery) score</u>. The raw agreement index

$$P = \sum_{i=1}^{m} P\,[(X,Y) \in A_i \times A_i]$$

can be computed by the formula

$$P = \sum_{i=1}^{k} \Big[\sum_{c_{i-1} \leq x,y < c_i} f_B(x,y) \Big].$$

On the other hand, the corrected-for-chance kappa index is given as $\kappa = (P - P_c)/(1 - P_c)$ where

$$P_c = \sum_{i=1}^{k} \Big[\sum_{c_{i-1} \leq x < c_i} f_B(x) \Big]^2.$$

4.1 Numerical Example. Consider the case where $n = 5$, $m = 4$ and $z_0 = 1.0$, $z_1 = 1.5$, $z_2 = 2.0$, $z_3 = 0$ and $z_4 = .5$. The Bernstein polynomial generates the marginal frequency density of .20040, .21230, .20040, .16865, .12698 and .09127 at the test scores of 0, 1, 2, 3, 4, and 5. For the binary classifications with cutoff score 4, the raw agreement index is .8197 and the kappa index is .4716.

5. PROJECTION OF TEST SCORE DISTRIBUTIONS FOR LENGTHENED TESTS

There are situations in which a test needs to be lengthened in order to accomodate new conditions and data are available for the short version of the test. If the binomial model holds, then it is possible to project the test score distribution for a lengthened test, assuming that the ability distribution of the examinees remains unchanged. From the data for the short form, it may be possible to approximate the true ability distribution via the Bernstein polynomial

$$B_m(\theta) = \sum_{k=0}^{m} z_k \binom{m}{k} \theta^k (1 - \theta)^{m-k}.$$

For a lengthened test consisting of ℓ items, the projected density function for the test score is given as

$$f(x) = \binom{\ell}{x} \int_0^1 \theta^x (1 - \theta)^{\ell-x} p(\theta)\, d\theta$$

$$= \binom{\ell}{x} \sum_{k=0}^{m} z_k \binom{m}{k} J(\ell + m, x + k).$$

5.1 Numerical Example. Consider the case where the fitting via a 4th degree Bernstein polynomial ($m = 4$) yields the constants $z_0 = 1.0$, $z_1 = 1.5$, $z_2 = 2.0$, $z_3 = 0$ and $z_4 = .5$. For a test with $\ell = 10$ items, the projected density is .10406, .11372, .11888, .11905, .11422, .10489, .09207, .07726, .06244 and .05012 at the test scores of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.

BIBLIOGRAPHY

Duncan, G. T. (1974). An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association 69, 50-57.

Feller, W. (1966). An introduction to probability theory and its applications (Vol. 2). New York: John Wiley & Sons.

Keats, J. A. & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika 27, 59-72.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Co.

Ravindran, A. (1972). Algorithm 431: A computer routine for quadratic and linear programming problems. Communication of the Association for Computing Machinery 15, 818-820.

ACKNOWLEDGEMENT

This work was performed while Huynh Huynh was supported by Grant No. NIE-G-73-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE position or policy and no endorsement should be inferred.

10

ADEQUACY OF ASYMPTOTIC NORMAL THEORY IN ESTIMATING RELIABILITY FOR MASTERY TESTS BASED ON THE BETA-BINOMIAL MODEL

Huynh Huynh

University of South Carolina

ABSTRACT

Simulated data based on five test score distributions indicate that a slight modification of the asymptotic normal theory for the estimation of the p and kappa indices in mastery testing will provide results which are in close agreement with those based on small samples. The modification is achieved through the multiplication of the asymptotic standard errors of estimate by the constant $1+m^{3/4}$ where m is the sample size.

1. INTRODUCTION

A primary purpose of mastery testing is to classify examinees in several achievement (or ability) categories. Typically, there are two such categories, mastery and nonmastery. The reliability of mastery tests is often viewed as the consistency of the various classifications across two test administrations; this consistency may be quantified via the raw agreement index (p) or the kappa

This paper has been distributed separately as RM 80-2, July, 1980.

index (κ). The raw agreement index is simply the combined proportion of examinees classified consistently as masters or nonmasters (if there are only two categories) on the two test administrations. The kappa index, on the other hand, expresses the extent to which the test scores improve the consistency of decisions beyond what would be expected by chance. Details regarding the nature and use of these indices may be found in Swaminathan, Hambleton, and Algina (1974), Huynh (1976, 1978a), and Subkoviak (1976, 1980).

Although p and κ are defined in terms of repeated testing, practical considerations often necessitate their estimation on the basis of test data collected from a single test administration. This may be done, for example, via the beta-binomial model (Huynh, 1976, 1979). The data reported in Subkoviak (1978), and by Huynh and Saunders (in press) tend to indicate that the beta-binomial model yields reasonably accurate estimates for p and κ in situations involving educational tests such as the Scholastic Aptitude Test and the Comprehensive Tests of Basic Skills.

The beta-binomial model also provides a convenient way to study the asymptotic sampling characteristics of the estimates. Let $\hat{p}$ and $\hat{\kappa}$ denote the (moment or maximum likelihood) estimates for p and κ, and let m be the number of examinees. Then $\sqrt{m}\,(\hat{p} - p)$ and $\sqrt{m}\,(\hat{\kappa} - \kappa)$ follow asymptotically two normal distributions, each with a mean of zero and a standard deviation of $G(p)$ or $G(\kappa)$ (Huynh, 1978b, 1979). The constants $G(p)$ and $G(\kappa)$ depend only on the number of items (n), the mean (μ) and standard deviation (σ) of the test scores, and the cutoff score (c). They are not functions of the sample size m, and may be computed via formulae, tables, or computer program (Huynh, 1978b, 1979).

The asymptotic considerations just summarized indicate that the estimates $\hat{p}$ and $\hat{\kappa}$ follow approximately normal distributions with means of zero and standard deviations of $\sigma_\infty(\hat{p}) = G(p)/\sqrt{m}$ and $\sigma_\infty(\hat{\kappa}) = G(\kappa)/\sqrt{m}$ when the sample size m is sufficiently large.

ADEQUACY OF ASYMPTOTIC THEORY

The extent to which these "asymptotic" standard errors reveal adequately the corresponding values in small samples appears to be unknown. Further, if $s_\infty(\hat{p})$ and $s_\infty(\hat{\kappa})$ represent the asymptotic standard errors computed from the sample data, asymptotic theory holds that the sampling distributions of the two ratios, $z(\hat{p}) = (\hat{p} - p)/s_\infty(\hat{p})$ and $z(\hat{\kappa}) = (\hat{\kappa} - \kappa)/s_\infty(\hat{\kappa})$, are approximately normal distributions with zero means and unit variances. The degree with which this asymptotic normality is true for small samples has yet to be investigated.

The purpose of this paper is threefold. It will first assess the adequacy of using the asymptotic standard errors to approximate the actual values encountered in small samples. Then, it will look at the degree to which asymptotic normal distributions can be used to describe the actual sampling distributions of the ratios $z(\hat{p})$ and $z(\hat{\kappa})$ when small samples are used. Finally, the paper also suggests a slight adjustment to the results of the asymptotic theory so that they will resemble more closely the results associated with small or moderate samples.

2. PROCEDURES

Let $\sigma_m(\hat{p})$, and $\sigma_m(\hat{\kappa})$ be the actual standard errors associate with a sample of size m. The closeness of the asymptotic approximations to these actual standard errors, when small samples are employed, may be assessed by computing the relative errors of approximation: $\varepsilon(\hat{p}) = [\sigma_m(\hat{p}) - \sigma_\infty(\hat{p})]/\sigma_m(\hat{p})$ and $\varepsilon(\hat{\kappa}) = [\sigma_m(\hat{\kappa}) - \sigma_\infty(\hat{\kappa})]/\sigma_m(\hat{\kappa})$, respectively. Approximations are said to be good when the ratios, $\varepsilon(\hat{p})$ and $\varepsilon(\hat{\kappa})$, are close to zero. In most practical situations, a ratio falling between ±5% should probably be considered as evidence of acceptable approximation.

As stated in the introduction, the asymptotic standard errors $\left(\sigma_\infty(\hat{p}) \text{ and } \sigma_\infty(\hat{\kappa})\right)$ may be computed for a given test score distribution. Since no simple formulae appeared available for the computation of the small sample standard errors $\sigma_m(\hat{p})$ and $\sigma_m(\hat{\kappa})$, computer simulation with 5000 replications was used in order to estimate their values as well as the relative errors of approximation $\varepsilon(\hat{p})$ and $\varepsilon(\hat{\kappa})$.

Computer simulation with 5000 replications was also used to assess the adequacy of using the unit normal distribution to describe the sampling distributions of the ratios $z(\hat{p})$ and $z(\hat{\kappa})$. The proportions of the simulated z-ratios which fell within selected (two-sided) critical values were computed and compared with the corresponding values expected from a normal distribution. The extent to which the proportions from the computer simulated distributions resembled the corresponding normal distribution probabilities was used to assess the adequacy of the asymptotic normal distribution. For this study, (two-sided) critical values were selected so that the central portion of the unit normal distribution was covered corresponding to probabilities of 80%, 90%, 95%, and 99%.

Both the moment and maximum likelihood (ML) estimates were used in this study. Moment estimates exist when the sample reliability index, KR21, is positive. When this was not the case, it was then assumed (as in Wilcox, 1977) that the beta-binomial model degenerated to a binomial distribution with an estimated success probability of $\lambda = \bar{x}/n$ where $\bar{x}$ is the test mean. Under these conditions, the estimate for κ was taken as zero, and that for p was computed via the expression $\hat{p} = p_o^2 + (1 - p_o)^2$ where

$$p_o = \sum_{x=o}^{c} \binom{n}{x} \lambda^x (1-\lambda)^{n-x}.$$

In addition, following the intuitive reasoning that degenerate cases only represent extreme situations, both the $z(\hat{p})$ and $z(\hat{\kappa})$ ratios were taken as extremely large whenever the degenerate case occurred.

Although the moment estimates are considerably easier to compute than the corresponding ML estimates, ML estimates often have been considered better than the moment estimates. (The asymptotic sampling distributions of the moment and ML estimates are the same however.) Because of this, the comparisons previously described for the moment estimates were also made for ML estimates. The ML estimates were obtained via a Newton-Raphson iteration scheme described elsewhere (Huynh, 1977). In the rare instances where the ML iteration did not converge, the moment estimates were used.)

The data base for this study consisted of five beta-binomial distributions. Four tests consisting of $n = 5$, 10, 15, and 20 items each were assembled by random selection of items from the Comprehensive Tests of Basic Skills, Form S, Level 1, which had been used in the South Carolina 1978 Statewide Testing Program. The actual frequency distribution for each of these tests was altered slightly so that the resulting distribution would conform almost exactly to a (marginal) beta-binomial distribution. Another beta-binomial distribution, with $\alpha = 8.970$ and $\beta = 1.994$, was patterned after the one used in the Wilcox (1977) study. Details regarding these distributions and the selected cutoff scores c may be found in Table 1. For each case listed in this table, five thousand replications were simulated to estimate various standard errors and sampling distributions. The sample size m was selected to be 25, 50, 100, 200, and 400.

TABLE 1

Descriptions of the Five Tests used in the Simulation

Case	Source	n	Mean	SD	α	β	KR21	c
1	CTBS	5	3.7066	1.5445	1.2512	0.4367	.7476	3
2	CTBS	10	7.4702	2.9435	1.1285	0.3822	.8688	6
3	Wilcox	10	8.1814	1.6147	8.9703	1.9940	.4770	8
4	CTBS	15	8.8630	3.3588	3.3273	2.3039	.7271	9
5	CTBS	20	11.1811	5.1115	1.9115	1.5077	.8540	12

Preliminary simulations indicated that the asymptotic standard errors tended to underestimate the smaller sample standard errors, and that an adjustment via the multiplicative constant, $h = 1 + 1/m^{3/4}$, would substantially improve the adequacy of the results deduced from the asymptotic theory. Hence, adjusted asymptotic standard errors of the form $\sigma_\infty^* = \sigma_\infty(1 + 1/m^{3/4})$ and adjusted z ratios of the type $z^* = z/(1 + 1/m^{3/4})$ were also incorporated in the study.

3. RESULTS

Table 2 reports the relative errors of approximation, $\varepsilon(\hat{p})$ and $\varepsilon(\hat{\kappa})$, for the asymptotic standard errors of the moment and ML estimates. Values associated with the adjusted asymptotic standard errors are enclosed within parentheses. The table reveals the following points. (a) The unadjusted asymptotic standard errors for both p and κ are slightly closer to the finite-sample standard errors of the ML estimates than to those associated with the moment estimates. This result does not appear unexpected: Strictly speaking, asymptotic theory deals mainly with ML estimates which are asymptotically efficient (i.e., unbiased with minimum variance). The asymptotic results, however, may be applied to the less efficient moment estimates because these are asymptotically equivalent to the ML estimates. Hence, the asymptotic standard error should more accurately depict the sampling variability of the ML than those of the moment estimates. However, the difference in accuracy is minimal when sample sizes as small as 25 or 50 are used. (b) The unadjusted asymptotic standard errors underrepresent the corresponding finite-sample standard errors; the extent of underrepresentation is less for $\sigma_\infty(\hat{p})$ than for $\sigma_\infty(\hat{\kappa})$. As seen in the last four rows of Table 2, the absolute relative errors of approximation $\varepsilon(\hat{p})$ average 8.3, 4.9, 3.3, 2.9, and 3.0 percent for sample sizes of 25, 50, 100, 200, and 400, respectively. For $\hat{\kappa}$, these percentages are 13.8, 7.6, 4.6, 4.0, and 2.9%. (c) As mentioned in the last section, the multiplicative adjustment via the constant $1 + 1/m^{3/4}$ produced adjusted asymptotic standard errors σ_∞^* which were substantially closer to their finite-sample values σ_m. For these

TABLE 2

Relative approximation errors associated with the asymptotic standard errors and with the adjusted asymptotic standard errors[a]

Case	Index	Estimate	Relative approximation error (in percent) at m = 25	50	100	200	400
1	p	Moment	10.8(2.8)	6.2(1.2)	1.9(-1.2)	1.9(0.0)	1.4(0.3)
		ML	8.2(0.0)	3.6(-1.6)	0.3(-2.9)	0.3(-1.6)	-.2(-1.3)
	κ	Moment	13.1(5.3)	7.9(3.0)	2.3(-0.8)	2.6(0.7)	1.9(0.8)
		ML	11.8(3.9)	6.1(1.1)	0.9(-2.2)	1.1(-0.8)	0.3(-0.8)
2	p	Moment	7.8(-0.4)	5.7(0.7)	5.7(2.7)	5.9(4.1)	5.9(4.8)
		ML	4.4(-4.1)	1.5(-3.8)	1.3(-1.9)	0.3(-1.6)	0.2(-0.9)
	κ	Moment	20.4(13.3)	10.4(5.6)	6.2(3.2)	4.7(2.9)	3.6(2.5)
		ML	17.8(10.4)	7.6(2.7)	3.4(0.3)	1.4(-0.5)	0.0(-1.1)
3	p	Moment	6.0(-2.4)	4.0(1.1)	3.2(0.1)	2.9(1.1)	2.7(1.6)
		ML	7.0(-1.3)	3.7(-1.4)	1.8(-1.2)	1.0(-0.9)	0.3(-0.8)
	κ	Moment	6.7(-1.7)	6.8(1.8)	5.8(2.8)	4.8(3.0)	3.7(2.7)
		ML	6.0(-2.4)	5.7(0.6)	4.3(1.3)	2.5(0.6)	1.2(0.1)
4	p	Moment	8.8(0.0)	4.3(-0.8)	2.5(-0.6)	2.8(1.0)	2.4(1.3)
		ML	9.5(1.4)	4.2(-0.9)	2.0(-1.1)	2.1(0.2)	1.6(0.5)
	κ	Moment	14.9(7.2)	6.3(1.3)	4.3(1.3)	3.6(1.8)	2.6(1.5)
		ML	15.7(8.2)	6.2(1.2)	4.0(1.0)	3.1(1.2)	1.9(0.8)
5	p	Moment	7.9(-0.3)	4.3(-0.8)	3.2(0.1)	3.7(1.9)	2.4(1.3)
		ML	7.1(-1.3)	2.7(-2.5)	1.5(-1.7)	1.5(-0.4)	0.1(-1.0)
	κ	Moment	13.7(6.0)	6.6(1.6)	4.6(1.6)	4.4(2.6)	2.8(1.7)
		ML	13.3(5.6)	5.1(0.0)	2.9(-0.1)	2.2(0.3)	0.4(-0.7)
Average of absolute error							
	p	Moment	8.3(1.3)	4.9(0.9)	3.3(0.9)	2.9(1.6)	3.0(1.9)
		ML	7.2(1.6)	3.1(2.0)	1.4(1.8)	1.0(0.9)	0.5(0.9)
	κ	Moment	13.8(6.7)	7.6(2.7)	4.6(1.9)	4.0(2.2)	2.9(1.8)
		ML	12.9(6.1)	6.1(1.1)	3.1(1.0)	2.1(0.7)	0.8(0.7)

[a]Values in parentheses represent relative errors of approximation when the adjustment h is used.

adjusted asymptotic standard errors, the absolute relative errors of approximation of $\hat{p}$ average 1.3, 0.9, 0.9, 1.6 and 1.9 percent for m = 25, 50, 100, 200, and 400, respectively. As for $\hat{\kappa}$, these average absolute relative errors stand at 6.7, 2.7, 1.9, 2.2, and 1.8%. (d) As expected, the asymptotic standard errors resemble more closely those estimated for finite samples as the sample size m becomes larger. Sampling errors associated with the simulation probably account for the erratic variation behavior of the estimated finite-sample standard errors found at a few places in Table 2.

Table 3 reports the empirical percentages of simulated z and z^* values which fall around zero with a nominal normal probability of 80%, 90%, 95%, and 99% (The results are reported only for the moment estimates, which differ only slightly from those associated with the ML estimates.) Two major points may be inferred from the reported data. (a) The use of unadjusted asymptotic standard errors produces z ratios which show less concentration around 0 than that predicted from a unit normal distribution. This is consistent with the results previously reported regarding the under-approximation associated with the unadjusted asymptotic standard errors. This under approximation produces z ratios with a standard deviation slightly larger than one; hence the corresponding distribution for these z ratios would show less probability around the central value of zero than that of a unit normal distribution. (b) Adjustment via the factor $1 + 1/m^{3/4}$ results in adjusted z^* ratios which cluster around zero with (empirical) probabilities very close to the nominal values predicted from the asymptotic normal theory. The degree of similarity between the empirical and nominal probabilities is quite adequate even with samples of size m = 25. The empirical and nominal probabilities are, within sampling error, nearly identical when the sample size is larger, say when m is 50 or higher.

TABLE 3

Empirical percentages of unadjusted (and adjusted) $z(\hat{p})$ values which fall around zero with selected nominal probabilities

Case	Nominal Prob. (%)	Empirical percentage at m = 25	50	100	200	400
1	80	75.1(79.6)	77.0(79.4)	79.1(80.1)	78.8(79.7)	78.9(79.4)
	90	86.4(89.5)	87.0(88.8)	88.8(89.9)	89.2(89.8)	89.3(89.8)
	95	92.0(94.0)	92.9(94.6)	94.7(95.3)	94.3(94.8)	95.0(95.3)
	99	97.4(98.1)	98.1(98.7)	98.8(99.0)	98.7(98.9)	78.9(99.0)
2	80	74.7(78.6)	75.9(78.7)	76.3(78.0)	77.2(78.0)	77.0(77.5)
	90	85.4(88.5)	86.6(88.5)	87.2(88.6)	87.7(88.4)	87.8(88.3)
	95	91.3(93.1)	92.2(93.4)	92.9(93.6)	93.2(93.7)	93.8(94.1)
	99	96.2(97.3)	97.7(98.0)	98.0(98.2)	98.2(98.4)	98.3(98.3)
3	80	75.7(79.8)	78.1(80.6)	79.2(80.6)	78.8(79.6)	78.7(79.3)
	90	85.4(87.6)	89.0(90.6)	89.4(90.6)	89.2(89.7)	88.7(89.2)
	95	89.7(91.0)	93.5(94.7)	94.5(95.3)	94.6(95.0)	94.4(94.6)
	99	93.8(94.5)	97.8(98.2)	98.5(98.8)	98.7(98.8)	98.7(98.8)
4	80	77.4(81.3)	78.5(81.0)	78.6(80.0)	78.6(79.6)	79.3(79.9)
	90	87.9(90.7)	88.5(90.2)	89.2(90.0)	88.9(89.5)	89.1(89.5)
	95	93.3(95.4)	93.8(95.4)	94.1(94.9)	94.4(94.8)	94.4(94.6)
	99	98.0(98.8)	98.7(99.0)	98.5(98.8)	98.5(98.7)	98.7(98.7)
5	80	75.8(79.9)	78.0(80.1)	78.3(80.0)	78.7(79.6)	79.1(79.7)
	90	86.6(89.7)	88.2(89.9)	88.6(89.6)	88.5(89.1)	89.3(89.6)
	95	92.3(94.7)	93.7(94.7)	94.2(95.0)	93.7(94.3)	94.5(94.7)
	99	98.0(98.7)	98.3(98.8)	98.7(89.9)	98.5(98.6)	98.7(98.8)

4. SUMMARY AND CONCLUSION

The study indicates that the asymptotic normal theory for the estimation of p and κ via the estimates $\hat{p}$ and $\hat{\kappa}$ produces asymptotic standard errors which are slightly smaller than the actual standard errors associated with small samples. As a result, the sampling distribution of the z type ratios has fewer cases around zero than is predicted by a normal distribution. However, multiplication of the asymptotic standard errors by the constant $1 + 1/m^{3/4}$ results in adjusted asymptotic standard errors which show close agreement with the actual finite-sample standard errors, even with samples as small as 25 cases. In addition, the adjustment produces z ratios which follow very closely a normal distribution, at least with respect to the combined tail probabilities. This conclusion also holds for samples as small as 25 cases.

All in all, it appears that, with the multiplicative adjustment factor of $1 + 1/m^{3/4}$ imposed on the asymptotic standard errors, the asymptotic normal theory for the estimation of decision reliability in mastery testing (Huynh, 1978b, 1979) can be used safely with samples with as few as 25 cases. This conclusion, of course, is restricted to situations similar to these considered here.

BIBLIOGRAPHY

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

Huynh, H. (1977). Statistical inference for the kappa and kappamax realiabilty indices based on the beta-binomial model. Paper read at the Psychometric Society Meetings, The University of North Carolina at Chapel Hill, June 16-17.

Huynh, H. (1978a). Reliability of multiple classifications. Psychometrika 43, 317-325.

Huynh, H. (1978b). Computation and inference for two reliability indices in mastery testing based on the beta-binomial model. Research Memorandum 78-1, Publication Series in Mastery Testing. University of South Carolina College of Education.

Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics 4, 231-246.

Huynh, H. & Saunders, J. C. (in press). Accuracy of two procedures for estimating reliability of mastery tests. Journal of Educational Measurement.

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

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

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision theoretic formulation. Journal of Educational Measurement 1, 263-267.

Wilcox, R. (1977). Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics 2, 289-307.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance of Anthony J. Nitko and Joseph C. Saunders is gratefully acknowledged.

11

CONSIDERATIONS FOR SAMPLE SIZE IN RELIABILITY STUDIES FOR MASTERY TESTS

Joseph C. Saunders
Huynh Huynh

University of South Carolina

Presented at the annual conference of the Eastern Educational Research Association, Norfolk, Virginia, March 5-8, 1980.

ABSTRACT

In most reliability studies, the precision of a reliability estimate varies inversely with the number of examinees (sample size). Thus, to achieve a given level of accuracy, some minimum sample size is required. An approximation for this minimum size may be made if some reasonable assumptions regarding the mean and standard deviation of the test score distribution can be made. To facilitate the computations, tables are developed based on the Comprehensive Tests of Basic Skills. The tables may be used for tests ranging in length from five to thirty items, with percent cutoff scores of 60%, 70%, or 80%, and with examinee populations for which the test difficulty can be described as low, moderate, or high, and the test variability as low or moderate. The tables also reveal that for a given degree of accuracy, an estimate of kappa would require a considerably greater number of examinees than would an estimate of the raw agreement index.

This paper has been distributed separately as RM 80-3, March, 1980.

1. INTRODUCTION

In many applications of educational and psychological testing, an empirical demonstration of the reliability of the measuring instrument is desirable. Such demonstration is most meaningful when the estimate for the reliability has been obtained with a reasonable degree of accuracy. That is, the standard error of estimate must be within some acceptable limit. In most instances, the standard error is a decreasing function of the number of examinees (sample size) to be included in the reliability study. Thus, some minimum sample size is needed to achieve a given level of precision. The purpose of this paper is to illustrate how this sample size can be assessed in estimating the reliability of mastery tests.

The paper consists of three major parts. The first part presents an overview of the procedures for estimating two reliability indices for mastery tests by using data collected from one test administration. The use of the estimation process to determine the minimum sample size is illustrated in the second part. Finally, a set of tables is developed to facilitate the determination of the minimum sample size in reliability studies for mastery tests.

2. OVERVIEW OF SINGLE-ADMINISTRATION ESTIMATES FOR RELIABILITY

Mastery tests are commonly used to classify examinees into two achievement categories, usually referred to as mastery and nonmastery. The reliability of such tests is often viewed as the consistency of mastery-nonmastery decisions. It may be quantified via the raw agreement index (p) or the kappa index (κ). The p index is simply the combined proportion of examinees classified consistently as masters or nonmasters by two repeated testings using the same form or two equivalent forms of a mastery test. The kappa index, on the other hand, takes into account the level of decision consistency which would result from random category assignment. It expresses the extent to which the test scores improve the consistency of decisions beyond the chance level.

MINIMUM SAMPLE SIZE

Though both p and κ are defined in terms of repeated testings, there are many practical situations in which they may be estimated from the scores collected from a single test administration (Huynh, 1976). The estimation process assumes that the test scores conform to a beta-binomial (negative hypergeometric) model, and may be carried out via formulae, tables, and a computer program reported elsewhere (Huynh, 1978; 1979). The data reported by Subkoviak (1978) and by Huynh and Saunders (1979) tend to indicate that the beta-binomial model yields reasonably accurate estimates for p and κ in situations involving educational tests such as the Scholastic Aptitude Test and the Comprehensive Test of Basic Skills.

The beta-binomial model also provides asymptotic (large sample) standard errors for the estimates. Simulation studies indicate that the asymptotic standard errors tend to underestimate the actual standard errors when the sample size is small (Huynh, 1980). The degree of underestimation is not substantial when the sample has sixty or more examinees. Since the beta-binomial model will be used throughout the remaining part of this paper, a minimum sample size of sixty examinees will be assumed to hold uniformly for all cases under consideration.

3. ILLUSTRATIONS FOR SAMPLE SIZE DETERMINATION

The standard error (s.e.) of estimates for p and for κ are functions of sample size m. The quantity $G = \text{s.e.} \times \sqrt{m}$ is asymptotically (i.e., in large samples) a constant, however. This constant depends only on the number of items (n), the mean (μ) and standard deviation (σ) of the test scores, and the cutoff score (c). Given the availability of these parameters, the value of G may be determined via the tables or the computer program presented elsewhere (Huynh, 1978). Once G is determined, a minimum sample size m can be calculated which will restrict the standard error of estimate to whatever tolerable range is required.

Suppose, for example, that an estimate of κ is needed for a short (n = 6 items) test to be used with a particular population of students. Passing or mastery on the test is to be granted if an examinee attains a score of 5 or 6. Further, suppose that we want the standard error of this estimate to be smaller than 10% of κ, that is, s.e. $(\kappa) \leq .10\kappa$.

What sample size would be needed to obtain the specified degree of accuracy in the estimate? To answer this question using the above mentioned Huynh procedure, a preliminary knowledge of the test mean and standard deviation is needed. Suppose past data suggest that the students are generally well-prepared on the content of the test in question and can be expected to be fairly homogeneous in achievement. We might suppose that in the population the mean will be 5.0 and the standard deviation will be 1.2. Using these values, and the cutoff score of 5, a value of G can be read from the tables (or computed): $G(\kappa) = .7390$. If the population mean and standard deviation are as given, then, assuming the beta-binomial model, the population value of κ is .3778. These results are then used to estimate the sample size needed to bring the standard error of estimate with the desired limits (i.e. less than $.10\kappa$).

Since the standard error of estimate is approximately $G/\sqrt{m}$, the standard error must be such that

$$\frac{G(\kappa)}{\sqrt{m}} \leq .10\kappa$$

or, equivalently,

$$m \geq [G(\kappa)/.10\kappa]^2.$$

For this example, then,

$$m \geq [.7390/(.10)(.3778)]^2 = 382.62.$$

Thus, to have no more than 10% relative error requires that at lease 383 examinees be tested to estimate κ.

A similar computation can be made for s.e. $(p) \leq .10p$ when the above assumed population values hold. Thus, using the tables,

MINIMUM SAMPLE SIZE

$$G(p) = .3210,$$

$$p = .7532,$$

and

$$m \geq [G(p)/.10p]^2 = 18.16.$$

Because of the previously mentioned problems of underestimation in small samples, a sample size of at least sixty is recommended regardless of the above computation.

It might be disheartening to note that a much larger sample size is needed to keep the standard error of the κ estimate within the desired limits than is required when an estimate of p is used. However, the standard error for κ is much larger than that of p (Huynh, 1978). Thus, for the same relative size of errors of estimation, larger samples are needed to estimate κ than to estimate p. It could be argued that the same degree of accuracy of estimation is not required. If so, then a less accurate estimate of κ would allow a smaller sample size.

The above illustration presumes that the mean and standard deviation of the test scores can be projected prior to the real test administration. In a number of instances involving the use of standardized tests for a heterogeneous group of students, reasonable assumptions may be made, which will yield projected values for both μ and σ. For example, when an n-item multiple-choice is built to maximize the discrimination among individual examinees, it is not unreasonable to assume that the test mean is half way between the expected chance score and the maximum score n, and that the standard deviation is about one-sixth of the test score range from 0 to n. (If there are A options per item, the expected chance score is n/A.) In other words, it is not unreasonable to presume that

$$\mu = (n+n/A)/2$$

and

$$\sigma = n/6.$$

For example, consider a test consisting of 10 four-option items. Then A = 4, and the projected mean and standard deviation are

$\mu = 6.25$ and $\sigma = 1.66667$. Presuming a cutoff score of $c = 6$, it may be found that $p = .6140$, $G(p) = .3661$, $\kappa = .1118$, and $G(\kappa) = .8213$. If a relative error of 5% is acceptable for p, then a sample of at least $[.3661/(.05\times.6140)]^2 = 143$ students would be needed. On the other hand, a relative error of 25% for kappa would require $[.8213/(.25\times.1118)]^2 = 864$ students.

4. PRACTICAL CONSIDERATIONS IN SETTING SAMPLE SIZE IN BASIC SKILLS TESTING

Some general formulae are given for expressing the relationships among s.e., G, m, p, k, and the proportion of sampling error desired in an estimate. These general expressions will then be used in a series of simulations designed to explore their typical numerical values for real tests. Tables are developed to help the practitioner decide on the sample size needed to obtain estimates of p and κ for various degrees of precision.

General expressions

Since $G = \text{s.e.} \times \sqrt{m}$ is a constant for large samples, this expression forms the basis for the formulations in this section. In the previous section .10 and .05 were used as examples of desired degrees of precision for a sample estimate of p. In general, we will call this quantity γ, using γ_p and γ_κ to distinguish precisions desired for p and κ, respectively. Thus, the general expressions for minimum sample size are:

$$m \geq \left[\frac{G(p)}{\gamma_p p} \right]^2$$

and

$$m \geq \left[\frac{G(\kappa)}{\gamma_\kappa \kappa} \right]^2$$

A further simplification is to let $R(p) = [G(p)/p]^2$ and $R(\kappa) = [G(p)/\kappa]^2$. The above expressions for minimum sample size, m, become

$$m \geq R(p)/(\gamma_p)^2$$

and

$$m \geq R(\kappa)/(\gamma_\kappa)^2.$$

These expressions will allow minimum sample size to be determined from knowledge of two quantities, R and γ.

Determining typical values of $R(p)$ and $R(\kappa)$

In practical applications, the values $R(p)$ and $R(\kappa)$ depend on a test score distribution which is not yet available. So, as in the previous section, conjectures must be made regarding the mean and standard deviation of the test score in order to project the minimum sample size.

In this section, typical values for $R(p)$ and $R(\kappa)$ will be reported for practical testing situations involving the assessment of basic skills. Several combination of test length, difficulty, variability, and cutoff scores will be used. To arrive at the values of $R(p)$ and $R(\kappa)$ reported in Tables 1-5, the following series of steps was taken.

First, a series of subtests was developed, using items found in the Comprehensive Test of Basic Skills (CTBS), Form S, Level 1. The items composing each subtest were randomly selected from one of five CTBS content areas, to reflect a variety of subjects and skills. For each content area, subtests were constructed with 5, 10, 15, 20, 25, and 30 items, producing a total of 30 subtests.

Second, the administration of the subtests was simulated using actual student responses. Data for the simulation came from 5,543 students, comprising a systematic sample (every tenth case) of the third grade students tested using Level 1 of the CTBS by the 1978 South Carolina Statewide Testing Program. From the students' responses to each item in the CTBS, raw scores were generated for each student on all 30 subtests.

Third, values of the mean and standard deviation of raw scores on each test were obtained. District means and standard deviations were calculated for each school district with 40 or more students in the sample. For each of the 30 subtests, means and standard deviations were plotted in a bivariate scatter diagram. The scatter-plots were divided into areas representing different categories of test difficulty and variability. Then districts were selected with means and standard deviations considered to be typical of six categories of difficulty and variability. These six categories (tests of low, moderate, and high difficulty, with low and moderate variability) were chosen to represent types of test score distributions typically encountered in mastery testing.

Fourth, the typical values obtained in the previous step were used to determine R(p) and R(κ). For each of the 30 subtests, the computer program described elsewhere (Huynh, 1978) was used to obtain estimates of G(p), p, G(κ), and κ when the cutoff scores were equivalent to 60%, 70%, and 80%. These data were used to calculate R(p) and R(κ) in each case.

Finally, the values of R(p) and R(κ) obtained above were averaged over the five CTBS content areas and the resulting values were compiled in tabular form. Tables 1, 2, and 3 provide values of R(p) and R(κ) for percent cutoff scores of 60%, 70%, and 80%, respectively.

The data needed to enter the tables are: (1) test length (n), (2) an idea of test difficulty (high, moderate, or low), (3) test variability (low or moderate), and (4) percentage cutoff score (60%, 70%, or 80%). The minimum sample size needed is simply R/γ^2, that is, the value of R obtained from the tables divided by the square of the acceptable proportion of sampling error in the estimate.

Numerical example

Suppose a study is planned to assess the reliability of a twenty-item test (n = 20) using the kappa index when a cutoff score of 14 (c = 70%) is employed. The students for whom the test is

TABLE 1

Values of R for p and κ for Six Categories of Tests at the Percent Cutoff Score of 60%

Test Category (diff)	(var)		Number of Items 5	10	15	20	25	30
High	Low	(p)	0.219	0.075	0.050	0.031	0.023	0.018
		(κ)	5.349	1.623	0.666	0.391	0.307	0.209
High	Mod	(p)	0.164	0.061	0.036	0.025	0.018	0.014
		(κ)	2.589	0.908	0.327	0.280	0.209	0.139
Mod	Low	(p)	0.244	0.085	0.056	0.032	0.025	0.020
		(κ)	5.809	1.485	0.613	0.367	0.269	0.200
Mod	Mod	(p)	0.148	0.068	0.036	0.027	0.021	0.015
		(κ)	2.215	0.838	0.312	0.266	0.198	0.126
Low	Low	(p)	0.199	0.095	0.044	0.031	0.025	0.020
		(κ)	5.502	1.345	0.560	0.365	0.247	0.186
Low	Mod	(p)	0.142	0.068	0.032	0.024	0.020	0.016
		(κ)	2.371	0.770	0.298	0.249	0.176	0.128

intended are known to be a homogeneous group of relatively high ability. Thus, it might be expected that the test would be of low difficulty (i.e., easy), with low variability. Let us say that a fairly precise estimate of κ is desired, so γ_κ is set at .05. Entering Table 2, in the row corresponding to low difficulty and low variability, it if found that R(κ) for n = 20 items is .362. The minimum sample size needed to estimate kappa with 5% allowable error is then computed as $m = R(\kappa)/\gamma_\kappa^2 = .362/(.05)^2 = 144.8$. Thus, a sample of at least 145 students is necessary to achieve the desired degree of precision. If reliability is to be determined via the raw agreement index p, a similar procedure is followed using R(p) and γ_p. Again, at least 60 students should be used in the sample, even if it is found that $m < 60$.

TABLE 2

Values of R for p and κ for Six Categories of Tests at the Percent Cutoff Score of 70%

Test Category (diff)	(var)		Number of Items 5	10	15	20	25	30
High	Low	(p)	0.219	0.075	0.046	0.029	0.022	0.017
		(κ)	5.349	1.623	0.776	0.455	0.410	0.272
High	Mod	(p)	0.164	0.061	0.033	0.023	0.017	0.013
		(κ)	2.589	0.908	0.360	0.324	0.276	0.178
Mod	Low	(p)	0.244	0.085	0.053	0.031	0.023	0.019
		(κ)	5.809	1.485	0.646	0.396	0.322	0.242
Mod	Mod	(p)	0.148	0.068	0.035	0.026	0.019	0.014
		(κ)	2.215	0.838	0.321	0.289	0.237	0.149
Low	Low	(p)	0.199	0.095	0.050	0.031	0.024	0.019
		(κ)	5.502	1.345	0.512	0.362	0.265	0.203
Low	Mod	(p)	0.142	0.068	0.036	0.023	0.019	0.015
		(κ)	2.371	0.770	0.280	0.254	0.190	0.137

<u>Some observations on the tabled values</u>

<u>In every case R(κ) > R(p)</u>. This fact implies that the sample size necessary to estimate kappa will be larger than that needed to estimate p, for any fixed degree of precision, γ. As noted previously, practical limitations may require that larger proportions of error be tolerated when estimating kappa than when estimating p.

<u>R-values for the case of low variability are larger than those for moderate variability</u>. If there is doubt about the expected degree of variability, the value of R for the low variability case would produce the more conservative estimate of m.

<u>R decreases as the number of test items increases</u>. The relationship between R and n is not linear, however. Hence, linear interpolation would not be appropriate for determining R for non-

TABLE 3

Values of R and p and κ for Six Categories of Tests at the Percent Cutoff Score of 80%

Test Category (diff)	(var)		Number of Items 5	10	15	20	25	30
High	Low	(p)	0.132	0.063	0.032	0.021	0.018	0.013
		(κ)	7.076	2.805	1.494	1.055	0.887	0.660
High	Mod	(p)	0.098	0.045	0.024	0.018	0.015	0.011
		(κ)	3.510	1.678	0.608	0.717	0.568	0.404
Mod	Low	(p)	0.174	0.064	0.038	0.025	0.020	0.015
		(κ)	6.831	2.283	1.087	0.812	0.640	0.558
Mod	Mod	(p)	0.113	0.047	0.026	0.021	0.017	0.012
		(κ)	2.633	1.337	0.484	0.571	0.458	0.311
Low	Low	(p)	0.189	0.060	0.044	0.029	0.022	0.017
		(κ)	5.849	1.906	0.652	0.611	0.471	0.417
Low	Mod	(p)	0.122	0.046	0.029	0.023	0.018	0.014
		(κ)	2.675	1.113	0.348	0.430	0.325	0.248

tabled values of n. The value of R listed for the largest tabled n <u>less</u> than the actual number of items should yield a conservative estimate for m. For example, suppose the test considered in the numerical example above actually contained 22 items. The tabled value of R corresponding to n = 25 would produce an underestimate of m, and the resulting proportion of error in estimating kappa would exceed γ_κ. The R-value for n = 20 would overestimate m, and the observed proportion of error would then be less than γ_κ.

<u>The relationships between R and test difficulty or cutoff scores are more complex</u>. No simple trends can be observed in the tables. In many testing situations, the cutoff score typically ranges from 60% to 80% correct. For cutoff scores falling between the values in the tables, find R for both bracketing values and use the larger. Again, consider the situation in the numerical example above.

Suppose the cutoff score was 13 (65% correct). From Tables 1 and 2, the values of R corresponding to c = 60% and 70% are .365 and .362, respectively. The larger of these (corresponding to c = 60%) should provide a reasonable value for R.

4. CONCLUSIONS

In this paper, an approximation method has been presented for determining the minimum sample size necessary to achieve a specified degree of precision in estimating raw agreement (p) and kappa (κ) indices of reliability for mastery tests. The method uses the quantity R which can be calculated for known test score distributions. Tables of R have been constructed for test score distributions typically found in mastery testing, for a variety of test lengths and cutoff scores. In addition, suggestions have been made for obtaining reasonable estimates of R for situations not directly covered by the tables.

Of course, precision is only one of the factors that must be considered in any study. Feasibility, cost, and classroom management considerations also play important roles. However, knowledge of necessary sample sizes should facilitate and simplify the planning of reliability studies. The tables presented here should be particularly useful for tests involving the basic skills, and perhaps other tests of similar construction.

BIBLIOGRAPHY

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

Huynh, H. (1978). Computation and inference for two reliability indices in mastery testing based on the beta-binomial model. Research Memorandum 78-1, Publication Series in Mastery Testing. University of South Carolina College of Education.

Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics 4, 231-246.

Huynh, H. (1980). Adequacy of the asymptotic error in estimating reliability for mastery tests based on the beta-binomial model. Research Memorandum 80-2, Publication Series in Mastery Testing, University of South Carolina College of Education.

Huynh, H. & Saunders, J. C. (1979). Accuracy of two procedures for estimating reliability of mastery tests. Research Memorandum 79-1, Publication Series in Mastery Testing. University of South Carolina College of Education.

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

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare. Huynh Huynh, Principal Investigator. Points of view or opinions stated do not neccessarily reflect NIE positions or policy and no official endorsement should be inferred. The comments of Anthony J. Nitko and Elizabeth M. Haran are gratefully acknowledged.

PART FOUR

ACCURACY OF DECISIONS

12

STATISTICAL INFERENCE FOR FALSE POSITIVE AND FALSE NEGATIVE ERROR RATES IN MASTERY TESTING (COMPUTER PROGRAMS AND TABLES ADDED)

Huynh Huynh

University of South Carolina

Psychometrika, March 1980.

ABSTRACT

This paper describes an asymptotic inferential procedure for the estimates of the false positive and false negative error rates. Formulae and tables are described for the computations of the standard errors. A simulation study indicates that the asymptotic standard errors may be used even with samples of 25 cases as long as the Kuder-Richardson Formula 21 reliability is reasonably large. Otherwise, a large sample would be required.

1. INTRODUCTION

A primary purpose of mastery testing is to use test data in order to classify an examinee in one of several achievement (or ability) categories. Typically there are two such categories, mastery and nonmastery. For example, let θ be the true ability of a person. Then true nonmastery status is defined by the condition $\theta < \theta_o$ and true mastery by $\theta \geq \theta_o$, θ_o being a given constant often referred to as a criterion level. In the reality of testing,

---

This paper has been distributed separately as RM 79-6, July, 1979.

however, decisions are normally made on the basis of the observed test data. Let x be the test score and c an appropriately chosen passing (or mastery) score. Then nonmastery status is declared if $x < c$ and mastery status is granted if $x \geq c$. A correct decision on the basis of test data is made when $\theta < \theta_o$ and $x < c$ or when $\theta \geq \theta_o$ and $x \geq c$. The other two situations represent errors in classification: a false positive error is committed when $\theta < \theta_o$ and $x \geq c$; a false negative error is encountered when $\theta \geq \theta_o$ and $x < c$.

The likelihood (or rate) of false positive and false negative errors may be assessed via several schemes. For example, using the binomial error model and the notion of an indifference zone, it is possible to compute the maximum error rates in classification for an individual (Wilcox, 1976). On the other hand, the error rates for a group of examinees may be assessed if a reasonable form for the (group) distribution of θ is available. Such is the case of the beta-binomial model (Keats & Lord, 1962) explored by Huynh (1976a, 1976b, 1977a, 1978) and Wilcox (1977) in several technical problems regarding mastery testing.

The beta-binomial model requires that test items be exchangeable, i.e., they can replace each other without changing the distribution of test scores. Item exchangeability implies that the items are equally difficult. This condition can be considered only as approximately satisfied in most testing situations. However, there are indications (Keats & Lord, 1962; Duncan, 1974) that several test score distributions fit into the beta-binomial model adequately. There are more complex models (Lord, 1965, 1969) which take into account variation in item difficulty. However, as far as estimation of error rates is concerned, the data in Wilcox (1977) seem to suggest that the more complex models do not increase substantially the accuracy of the estimates.

The purpose of this paper is to describe an asymptotic inferential procedure for false positive and false negative error rates. The beta-binomial model is used as a vehicle for computation.

2. COMPUTATIONS FOR ERROR RATES

Let n be the number of test items randomly selected from an item pool, θ (true ability) be the true proportion of items in the total item pool that would be answered correctly by an examinee, and x be the examinee's observed test score. Then the conditional density of x is given as

$$f(x|\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}, \quad x = 0,1,\ldots,n.$$

Let the density p of θ be of the beta form with parameters α and β, i.e.,

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < \theta < 1.$$

Both α and β are positive constants. The joint density of (x,θ) is given as

$$g(x,\theta) = \frac{\binom{n}{x}}{B(\alpha,\beta)}\, \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1}.$$

With the criterion level θ_o and passing score c, the false positive error rate is given as

$$F_p = P(x \geq c, \theta < \theta_o)$$

$$= \frac{1}{B(\alpha,\beta)} \sum_{x=c}^{n} \binom{n}{x} \int_{o}^{\theta_o} \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1}d\theta.$$

Let

$$D(u,v;\theta_o) = \int_{o}^{\theta_o} t^{u-1}(1-t)^{v-1}dt.$$

Then

$$F_p = \frac{1}{B(\alpha,\beta)} \sum_{x=c}^{n} \binom{n}{x} D(\alpha+x, n+\beta-x; \theta_o).$$

As for the likelihood F_n of a false negative error, it may be noted that

$$F_n = P(x \leq c-1, \theta \geq \theta_o)$$

$$= \frac{1}{B(\alpha.\beta)} \sum_{x=o}^{c-1} \binom{n}{x} \int_{\theta_o}^{1} \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1}d\theta.$$

Let $\xi = 1-\theta$, $\xi_o = 1-\theta_o$, $y = n-x$, and $d = n-c+1$. Then it may be verified that

$$F_n = \frac{1}{B(\alpha,\beta)} \sum_{y=d}^{n} \binom{n}{y} \int_o^{\xi_o} \xi^{\beta+y-1}(1-\xi)^{n+\alpha-y-1} d\xi .$$

From this it may be seen that F_n may be computed in exactly the same way as F_p.

The computations of F_p can be carried out with some degree of efficiency by noting that

$$D(u+1,v-1;\theta_o) = \left(-\theta_o^u(1-\theta_o)^{v-1} + uD(u,v;\theta_o)\right)/(v-1)$$

and that

$$D(u,v;\theta_o) = B(u,v)\ I(u,v;\theta_o) .$$

In this formula, $I(u,v;\theta_o)$ denotes the incomplete beta function as tabulated in Pearson (1934) and implemented via the IBM subroutine BDTR (1970) or the IMSL subroutine MDBETA (1977).

3. ASYMPTOTIC STATISTICAL INFERENCE FOR ESTIMATES

Maximum likelihood estimation for α and β has been considered by several authors including Griffiths (1973). A fairly efficient computer routine is described in Huynh (1977b). The data generated by Huynh indicate that the maximum likelihood estimates $\tilde{\alpha}$ and $\tilde{\beta}$ and the moment estimates $\hat{\alpha}$ and $\hat{\beta}$ do not differ markedly from each other when the number m of examinees is reasonably large. Hence, for the numerical examples described in this paper, only $\hat{\alpha}$ and $\hat{\beta}$ shall be used. They are to be computed as follows. Let $\overline{x}$ and s be the mean and standard deviation of the test score, and let

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left(1 - \frac{\overline{x}(n-\overline{x})}{ns^2}\right)$$

be the estimated KR21 reliability. Then the moment estimates are

$$\hat{\alpha} = (-1 + 1/\hat{\alpha}_{21})\overline{x}$$

and

$$\hat{\beta} = -\hat{\alpha} + n/\hat{\alpha}_{21} - n .$$

The estimates are positive when $0 < \hat{\alpha}_{21} < 1$. (If the computed value for $\hat{\alpha}_{21}$ is zero or negative, replace it by the smallest positive estimate of reliability which happens to be available.)

For reasons previously mentioned, general sampling properties appropriate for the maximum likelihood estimates would be applicable to $\hat{\alpha}$ and $\hat{\beta}$. For example, $\sqrt{m}(\hat{\alpha}-\alpha, \hat{\beta}-\beta)$ follows asymptotically a bivariate normal distribution with zero mean and covariance matrix $\xi = (\sigma_{ij}) = ||b_{pq}||^{-1}$ where

$$b_{11} = \sum_{x=o}^{n} \left(\frac{\partial f(x)}{\partial \alpha}\right)^2 / f(x),$$

$$b_{12} = \sum_{x=o}^{n} \frac{\partial f(x)}{\partial \alpha} \cdot \frac{\partial f(x)}{\partial \beta} / f(x),$$

and

$$b_{22} = \sum_{x=o}^{n} \left(\frac{\partial f(x)}{\partial \beta}\right)^2 / f(x).$$

Now let $F_p = Z(\alpha,\beta)$ be the function of (α,β) defining the false positive error rate. Let $\hat{F}_p = Z(\hat{\alpha},\hat{\beta})$ be the estimate of F_p computed on the basis of $(\hat{\alpha},\hat{\beta})$. Then it may be deduced (Rao, 1973, p. 386-387) that $\sqrt{m}(\hat{F}_p - F_p)$ asymptotically follows a normal distribution with zero mean and with variance

$$V_{fp}^2 = \sigma_{11}\left(\frac{\partial F_p}{\partial \alpha}\right)^2 + 2\sigma_{12} \frac{\partial F_p}{\partial \alpha} \cdot \frac{\partial F_p}{\partial \beta} + \sigma_{22}\left(\frac{\partial F_p}{\partial \beta}\right)^2.$$

It may then be said that the estimate $\hat{F}_p$ has an approximate normal distribution with mean F_p and standard deviation (standard error) of $\sigma_\infty(\hat{F}_p) = V_{fp}/\sqrt{m}$. An estimated standard error for $\hat{F}_p$, namely $s_\infty(\hat{F}_p)$, may be obtained by replacing (α,β) by the estimates $(\hat{\alpha},\hat{\beta})$ in the above formula defining $\sigma_\infty(\hat{F}_p)$.

The computations described above also apply to the rate of false negative error. Let F_n and $\hat{F}_n$ be the true and estimated values for this error rate. Then $\sqrt{m}(\hat{F}_n - F_n)$ asymptotically follows a normal distribution with zero mean and with variance

$$V_{fn}^2 = \sigma_{11}\left(\frac{\partial F_n}{\partial \alpha}\right)^2 + 2\sigma_{12} \frac{\partial F_n}{\partial \alpha} \cdot \frac{\partial F_n}{\partial \beta} + \sigma_{22}\left(\frac{\partial F_n}{\partial \beta}\right)^2.$$

In addition, let ρ be the correlation between the estimated false positive and false negative error rates. Then it may be noted that $\rho = \mathrm{cov}(\hat{F}_p, \hat{F}_n)/V_{fp}V_{fn}$ where

$$\mathrm{cov}(\hat{F}_p, \hat{F}_n) = \sigma_{11} \frac{\partial F_p}{\partial \alpha} \frac{\partial F_n}{\partial \alpha} + \sigma_{12}\left(\frac{\partial F_p}{\partial \alpha} \frac{\partial F_n}{\partial \beta} + \frac{\partial F_p}{\partial \beta} \frac{\partial F_n}{\partial \alpha}\right) + \sigma_{22} \frac{\partial F_p}{\partial \beta} \frac{\partial F_n}{\partial \beta}.$$

4. COMPUTATIONS FOR THE PARTIAL DERIVATIVES

The computation of V_{fp}, V_{fn}, and ρ requires the partial derivatives of $Z(\alpha,\beta)$ with respect to α and β. These derivatives, in turn, are based on the partial derivatives of $D(\alpha+x, n+\beta-x; \theta_o)$ and $B(\alpha,\beta)$ with respect to α and β.

4.1. Partial Derivatives of $B(\alpha,\beta)$

With

$$B(\alpha,\beta) = \int_o^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

it may be deduced that

$$\frac{\partial B}{\partial \alpha} = \int_o^1 t^{\alpha-1}(1-t)^{\beta-1} \log t \, dt$$

and that

$$\frac{\partial B}{\partial \beta} = \int_o^1 t^{\alpha-1}(1-t)^{\beta-1} \log (1-t) \, dt.$$

Let Ψ be the Euler psi function as defined and tabled in Abramowitz and Stegun (1968, p. 258, Section 6.3 and Table 6.1). Then according to Gradshteyn and Ryzhik (1965, p. 538, Section 4.253, Formula 1),

$$\frac{\partial B}{\partial \alpha} = B(\alpha,\beta)\big(\Psi(\alpha) - \Psi(\alpha+\beta)\big)$$

and

$$\frac{\partial B}{\partial \beta} = B(\alpha,\beta)\big(\Psi(\beta) - \Psi(\alpha+\beta)\big).$$

Formulae are also available in these texts which are useful in computer programming the psi function. For the present paper, the following steps have been adopted.

1. First the argument of $\Psi(.)$ is reduced to a value in the half closed interval $[1,2)$ by using the formula $\Psi(z+1) = \Psi(z) + 1/z$.

2. If $z = 1$, then $\Psi(1) = -.5772156649$.
3. For $1 < z \leq 1.75$, the following series expansion is used

$$\Psi(1+z) = \Psi(1) + \sum_{n=2}^{\infty} (-1)^n \xi(n) z^{n-1}$$

where $\xi(.)$ is the Riemann zeta function tabulated in Abramowitz and Stegun (1968, p. 811, Table 23.3). If the series is stopped at the term z^{N-1}, the error <u>cannot exceed</u> $\xi(N)z^{N-1} < 1.21z^{N-1}$, $(N \geq 4)$. For this paper, ten significant decimals are adopted for Ψ. The value for N is $-23.21647129/\log z + 1$ which cannot exceed 82.

4. For $1.75 < z < 2$, the four-point Lagrange interpolation is used to compute Ψ on the basis of tabled values of Ψ for $z = 1.745\ (.005)\ 2.010$. Let Ψ_{-1}, Ψ_o, Ψ_1, and Ψ_2 be four consecutive tabled values of Ψ with Ψ_o corresponding to z_o. Then for any p, $0 \leq p \leq 1$,

$$\Psi(z_o + .005p) = \frac{-p(p-1)(p-2)}{6}\Psi_{-1} + \frac{(p^2-1)(p-2)}{2}\Psi_o - \frac{p(p+1)(p-2)}{2}\Psi_1 + \frac{p(p^2-1)}{6}\Psi_2$$

(Abramowitz and Stegun, p. 879, Section 25.2.13). According to these authors (p. 270), this procedure yields ten significant decimals for the psi function.

4.2. Partial Derivatives of $D(\alpha+x, n+\beta-x; \theta_o)$

With

$$D(\alpha+x, n+\beta-x; \theta_o) = \int_o^{\theta_o} t^{\alpha+x-1}(1-t)^{n+\beta-x-1}dt,$$

it may be deduced that

$$\frac{\partial D}{\partial \alpha} = \int_o^{\theta_o} t^{\alpha+x-1}(1-t)^{n+\beta-x-1}\log t\ dt$$

and

$$\frac{\partial D}{\partial \beta} = \int_o^{\theta_o} t^{\alpha+x-1}(1-t)^{n+\beta-x-1}\log (1-t)\ dt.$$

With $x \geq c \geq 1$ and $0 < \theta_o < 1$, the integrating functions for both partial derivatives are <u>continuous</u> with respect to t provided they are taken as zero at $t = 0$. Hence, the process of differentiation under the integral sign is legitimate. Let

$$G(u,v;\theta_o) = \int_o^{\theta_o} t^{u-1}(1-t)^{v-1}\log t\, dt, \quad u > 1,\ v > 0.$$

Then

$$\frac{\partial D}{\partial \alpha} = G(\alpha+x, n+\beta-x; \theta_o)\,.$$

To compute the partial derivative $\partial D/\partial\beta$, let $z = 1 - t$ in the previous integral defining this derivative. It follows that

$$\frac{\partial D}{\partial \beta} = \int_{1-\theta_o}^{1} z^{n+\beta-x-1}(1-z)^{\alpha+x-1}\log z\, dz$$

$$= \int_o^1 z^{n+\beta-x-1}(1-z)^{\alpha+x-1}\log z\, dz - G(n+\beta-x, \alpha+x; 1-\theta_o)\,.$$

From Section 4.1, it may then be deduced that

$$\frac{\partial D}{\partial \beta} = B(n+\beta-x,\alpha+x)\big(\Psi(n+\beta-x) - \Psi(n+\alpha+\beta)\big) - G(n+\beta-x,\alpha+x;1-\theta_o)\,.$$

The computation of $G(u,v;\theta_o)$ is carried out as follows.

1. For $1 < u \le 2$ and $0 < v \le 2$, the 32-point Gaussian-Hermite quadrature is used to integrate the function $t^{u-1}(1-t)^{v-1}$ $\log t$ on the interval $(0,\theta_o)$, then on the two intervals $(0,\theta_o/2)$ and $(\theta_o/2,\theta_o)$. If the <u>relative</u> change between the two resulting G integrals is less than a tolerance error EPS, then the numerical quadrature stops. Otherwise, it will be carried out on the four subintervals $(0,\theta_o/4)$, $(\theta_o/4,\theta_o/2)$, $(\theta_o/2, 3\theta_o/4)$, and $(3\theta_o/4,\theta_o)$ and the resulting integral will be compared with the one obtained via two subintervals. The process continues until the relative change between these integrals is less than EPS. The tolerance error EPS is set at .00005 in this paper.
2. For other values of u and v, the following lemma is used to reduce u and v to two values u' and v' such that $1 < u' \le 2$ and $0 < v' \le 2$.

   <u>Lemma</u>. We have

   $$G(u,v-1;\theta_o) + G(u+1,v;\theta_o) = G(u,v;\theta_o)$$

   and

$$uG(u,v+1;\theta_o) - vG(u+1,v;\theta_o) = H$$

where

$$H = \theta_o^u(1-\theta_o)^v\big((\log\ \theta_o-1)/(u+v)\big) - vD(u,v;\theta_o)/(u+v).$$

Proof. The proof for the first formula is as follows.

$$\begin{aligned}
G(u+1,v;\theta_o) &= \int_o^{\theta_o} t^u(1-t)^{v-1}\log\ t\ dt \\
&= \int_o^{\theta_o}\big(-(1-t)t^{u-1} + t^{u-1}\big)(1-t)^{v-1}\log\ t\ dt \\
&= -\int_o^{\theta_o} t^{u-1}(1-t)^v\log\ t\ dt \\
&\quad + \int_o^{\theta_o} t^{u-1}(1-t)^{v-1}\log\ t\ dt \\
&= -G(u,v+1;\theta_o) + G(u,v;\theta_o)\ .
\end{aligned}$$

As for the second formula, let us integrate in parts the integral

$$G(u,v+1;\theta_o) = \int_o^{\theta_o} t^{u-1}(1-t)^v\log\ t\ dt.$$

Let

$$Y = t^{u-1}(1-t)^v$$

and

$$dZ = \log\ t\ dt.$$

Then

$$dY = \big((u-1)t^{u-2}(1-t)^v dt - vt^{u-1}(1-t)^{v-1}\big)dt$$

and

$$Z = t\ \log\ t - t\ .$$

Hence

$$\begin{aligned}
G(u,v+1;\theta_o) &= YZ\Big|_{t=o}^{\theta_o} - \int_o^{\theta_o} ZdY \\
&= \theta_o^u(1-\theta_o)^v(\log\ \theta_o - 1) \\
&\quad - (u-1)\int_o^{\theta_o} t^{u-1}(1-t)^v\log\ t\ dt \\
&\quad + v\int_o^{\theta_o} t^u(1-t)^{v-1}\log\ t\ dt
\end{aligned}$$

$$+ (u-1)\int_0^{\theta_o} t^{u-1}(1-t)^v dt$$

$$- v\int_0^{\theta_o} t^u(1-t)^{v-1}dt.$$

Algebraic manipulations will yield

$$G(u,v+1;\theta_o) = -(u-1)G(u,v+1;\theta_o) + vG(u+1,v;\theta_o) + H$$

where H is defined in the lemma. The second formula of the lemma is just proved.

The reduction of the range of u and/or v may now be accomplished by using the following recurrence formulae:

$$G(u+1,v;\theta_o) = \big(uG(u,v;\theta_o) - H\big)/(u + v)$$

and

$$G(u,v+1;\theta_o) = \big(vG(u,v;\theta_o) + H\big)/(u + v).$$

4.3. Partial Derivatives of $F_p(\alpha,\beta)$

From the expression

$$F_p = \frac{1}{B(\alpha,\beta)} \sum_{x=c}^{n} \binom{n}{x} D(\alpha+x, n+\beta-x;\theta_o)$$

it follows that

$$\frac{\partial F_p}{\partial\alpha} = \left[\sum_{x=c}^{n} \binom{n}{x} \partial D(\alpha+x,n+\beta-x;\theta_o)/\partial\alpha - F_p \partial B(\alpha,\beta)/\partial\alpha\right]/B(\alpha,\beta)$$

$$= \left[\sum_{x=c}^{n} \binom{n}{x} G(\alpha+x,n+\beta-x;\theta_o)\right]/B(\alpha,\beta) - F_p\big(\Psi(\alpha) - \Psi(\alpha+\beta)\big)$$

and

$$\frac{\partial F_p}{\partial\beta} = \sum_{x=c}^{n} \binom{n}{x}\big(B(\alpha+x,n+\beta-x) - G(n+\beta-x,\alpha+x,\theta_o)\big)/B(\alpha,\beta)$$

$$- F_p\big(\Psi(\beta) - \Psi(\alpha+\beta)\big).$$

The computations may be simplified by noting that

$$D(\alpha+x+1,n+\beta-x-1;\theta_o) = \big(-\theta_o^{\alpha+x}(1-\theta_o)^{n+\beta-x-1}$$

$$+ (\alpha+x)D(\alpha+x,n+\beta-x;\theta_o)\big)/(n+\beta-x-1),$$

and hence

$$G(\alpha+x+1,n+\beta-x-1;\theta_o) = \big(-\theta_o^{\alpha+x}(1-\theta_o)^{n+\beta-x-1}\log\theta_o + D(\alpha+x,n+\beta-x;\theta_o) + (\alpha+x)G(\alpha+x,n+\beta-x;\theta_o)\big)/(n+\beta-x-1)\,.$$

Also,

$$G(n+\beta-x-1,\alpha+x+1;\theta_o) = \big(\theta_o^{n+\beta-x-1}(1-\theta_o)^{\alpha+x}\log\theta_o - D(n+\beta-x-1,\alpha+x+1;\theta_o) + (\alpha+x)G(n+\beta-x,\alpha+x;\theta_o)\big)/(n+\beta-x-1)\,.$$

4.4. Partial Derivatives of $F_n(\alpha,\beta)$

From the expression of F_n in Section 2, namely

$$F_n(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\sum_{y=d}^{n}\binom{n}{y}\int_o^{\xi_o}\xi^{\beta+y-1}(1-\xi)^{n+\alpha-y-1}d\xi,$$

it follows that

$$F_n(\alpha,\beta) = \frac{1}{B(\beta,\alpha)}\sum_{y=d}^{n}\binom{n}{y}D(\beta+y,n+\alpha-y;\xi_o)\,.$$

Hence

$$\frac{\partial F_n}{\partial\beta} = \left[\sum_{y=d}^{n}\binom{n}{y}G(\beta+y,n+\alpha-y;\xi_o)\right]/B(\alpha,\beta) - F_n\big(\Psi(\beta) - \Psi(\alpha+\beta)\big)$$

and

$$\frac{\partial F_n}{\partial\alpha} = \sum_{y=d}^{n}\binom{n}{y}\big(B(\beta+y,n+\alpha-y) - G(n+\alpha-y,\beta+y;\theta_o)\big)/B(\alpha,\beta) - F_n\big(\Psi(\alpha) - \Psi(\alpha+\beta)\big)\,.$$

5. NUMERICAL ILLUSTRATION

Suppose that on a five-item test, the number of students having scores of 0, 1, 2, 3, 4, and 5 are 4, 14, 9, 17, 21, and 26 respectively. Altogether there are $m = 91$ students. It follows that $\bar{x} = 3.264$ and $s = 1.562$. The moment estimates for α and β are $\hat{\alpha} = 1.611$ and $\hat{\beta} = .857$. The estimated covariance matrix of $(\hat{\alpha},\hat{\beta})$ is defined by the elements $\hat{\sigma}_{11} = .18859$, $\hat{\sigma}_{12} = .08318$, and $\hat{\sigma}_{22} = .05035$. Let $\theta_o = .80$ and $c = 4$. The estimated error rates are then $\hat{F}_p = .180$ and $\hat{F}_n = .031$. The values of the partial derivatives evaluated at $(\hat{\alpha}, \hat{\beta})$ are $\partial F_p/\partial\alpha = .02281$, $\partial F_p/\partial\beta = .06926$, $\partial F_n/\partial\alpha = .01229$, and $\partial F_n/\partial\beta = -.01464$. Thus, the estimated standard errors for $\hat{F}_p$ and $\hat{F}_n$ are $s_\infty(\hat{F}_p) = .025$ and $s_\infty(\hat{F}_n) = .003$ respectively. The estimated correlation between $\hat{F}_p$ and $\hat{F}_n$ is $\hat{\rho} = .597$. These data may be of use in estimating other parameters. For example, let γ be the proportion of examinees classified correctly by the test scores. Then an estimate for γ is $\hat{\gamma} = 1 - (\hat{F}_p + \hat{F}_n) = .789$ which is associated with an estimated standard error of $s_\infty(\hat{\gamma}) = \{s_\infty^2(\hat{F}_p) + s_\infty^2(\hat{F}_n) + 2\hat{\rho}s_\infty(\hat{F}_p)s_\infty(\hat{F}_n)\}^{1/2} = \{(.025)^2 + (.003)^2 + 2\times.597\times.025\times.003\}^{1/2} = .061$.

6. TABLES FOR $\hat{F}_p$, V_{fp}, $\hat{F}_n$, V_{fn}, AND $\hat{\rho}$

Tables are presented in Appendix A which facilitate the computations for the false positive and false n -ive error rates, their standard errors of estimate, and their correlation. As indicated previously, this information may serve as the basis for the computation of statistics such as the proportion of correct decisions and its standard error. All computations were carried out via the Amdahl V-6 System with the double precision mode in use whenever feasible.

Input to the tables are (i) number of test items n, (ii) criterion level θ_o, (iii) passing score c, (iv) test mean $\bar{x}$, and (v) the KR21 reliability $\hat{\alpha}_{21}$. It may be noted that if $\tilde{\alpha}$ and $\tilde{\beta}$ are estimates of the parameters α and β other than the moment estimates, then the entries for test mean and KR21 are simply $n\hat{\alpha}/(\hat{\alpha} + \hat{\beta})$ and $n/(n + \hat{\alpha} + \hat{\beta})$ respectively.

For each entry $(n, \theta_o, c, \bar{x}, \hat{\alpha}_{21})$, five values may be read out. They are $\hat{F}_p$, V_{fp}, $\hat{F}_n$, V_{fn}, and $\hat{\rho}$.

The tables are constructed for $n = 5(1)10$ and $\hat{\alpha}_{21} = .10(.10).90$. For each n, the test mean is chosen such that $\bar{x}/n = .10(.10).90$. The criterion level is set at $\theta_o = .60$, .70, and .80, and the passing score is one or two values approximately equal to or larger than $n\theta_o$.

Numerical Example 1

Let $n = 10$, $\theta_o = .6$, and $c = 6$. For $\bar{x} = 5.0$ and $\hat{\alpha}_{21} = .60$, the tables yield the values $\hat{F}_p = .1667$, $V_{fp} = .1858$, $\hat{F}_n = .0504$, $V_{fn} = .0548$, and $\hat{\rho} = .2941$. If the data are obtained from 100 examinees, then the estimated standard errors are $s_\infty(\hat{F}_p) = .1858/10 = .0186$ and $s_\infty(\hat{F}_n) = .0055$. It may be deduced that the proportion of correct decision is .7829 for which the standard error is estimated as .0241.

It may be observed from these tables that the relationship of each of the quantities F_p, V_{fp}, F_n, V_{fn}, and ρ with respect to either $\bar{x}$ or $\hat{\alpha}_{21}$ is rather unpredictable. Hence interpolation for nontabulated entries should be carried out with care since the relationship is obviously not linear. For such a case it is recommended that Lagrange interpolations with three or four points be used whenever possible. Details regarding interpolations of this type may be found in Abramowitz and Stegun (1968, Section 25.2). The four-point Lagrange interpolation has been described in Section 4.1.

Numerical Example 2

Let $n = 10$, $\theta_o = .6$, and $c = 6$, along with $\bar{x} = 4.0$ and $\hat{\alpha}_{21} = .22$. Using th four-point Lagrange interpolation for the false positive error, we have $\Psi_{-1} = .1784$, $\Psi_o = .1883$, $\Psi_1 = .1886$, and $\Psi_2 = .1799$. With $p = (.22-.20)/.1 = .2$, it may be found that the interpolated false positive error is .1891.

7. FINITE-SAMPLE PERFORMANCE OF THE ASYMPTOTIC STANDARD ERRORS

So far only an asymptotic treatment has been presented for the estimates of the false positive and false negative error rates $\hat{F}_p$ and $\hat{F}_n$. An obvious question which needs to be answered is, at what minimum sample size m will the asymptotic standard errors $s_\infty(\hat{F}_p) = V_{fp}/\sqrt{m}$ and $s_\infty(\hat{F}_n) = V_{fn}/\sqrt{m}$ represent adequately the actual standard errors? A theoretical consideration of this issue is rather complex since it involver a joint examination of the spee.

at which $W = \sqrt{m}(\hat{\alpha}-\alpha,\hat{\beta}-\beta)$ converges to its asymptotic bivariate normal distribution and of the adequacy of representing the functions $F_p(\alpha,\beta)$ and $F_n(\alpha,\beta)$ by their Taylor expansions based on the first partial derivatives. Some work regarding the convergence speed of univariate maximum likelihood estimates are summarized in Kendall and Stuart (1967, Vol. 2, p. 46-48). An extension of this work would be needed for any theoretical consideration of the finite-sample behavior of the asymptotic errors.

For this report, simulations employing the IMSL random generator GGUB were used to assess the performance of $s_\infty(\hat{F}_p)$ and $s_\infty(\hat{F}_n)$. An additional issue under study was the degree of bias of $\hat{F}_p$ and $\hat{F}_n$ as estimates of the parameters F_p and F_n. (It may be recalled that both estimates are asymptotically unbiased.)

Five beta-binomial distributions (summarized in Table 1) were used in the simulation study. Four tests consisting of n = 5, 10, 15, and 20 items each were formed by random selection of items from the Comprehensive Tests of Basic Skills, Form S, Level 1, which had been used in a large statewide testing program. The frequency distribution for each of these tests was then altered slightly so that the resulting distribution would conform to almost exactly that of a (marginal) beta-binomial distribution. Relevant information regarding these distributions is listed in Table 1. The other beta-binomial distribution, with α = 8.970 and β = 1.994, is similar to the one used in the Wilcox study (1977).

TABLE 1

Descriptions of the Five Test Data used in the Simulation

Case	Source	n	Mean	SD	α	β	α_{21}	θ_o	c
1	CTBS	5	3.7066	1.5445	1.2515	0.4367	.7476	.5	3
2	CTBS	10	7.4702	2.9435	1.1285	0.3822	.8688	.6	6
3	CTBS	15	8.8630	3.3588	3.3273	2.3039	.7271	.8	12
4	CTBS	20	11.1811	5.1115	1.9115	1.5077	.8540	.6	12
5	Wilcox	10	8.1814	1.6147	8.9703	1.9940	.4770	.8	8

The criterion levels θ_o were chosen to be .5, .6, and .8 and the passing score c is put at $n\theta_o$. The sample size m is set at 25, 50, 100, 200, 400, and 800.

For each situation listed in Table 1, two thousand r lications were used to estimate the means of $\hat{F}_p$ and $\hat{F}_n$, and their finite-sample standard errors of estimate $s_m(\hat{F}_p)$ and $s_m(\hat{F}_n)$. The moment estimates were used when $\hat{\alpha}_{21}$ was positive. For $\hat{\alpha}_{21}$ negative or zero. the procedure used by Wilcox (1977, p. 295) was adopted. In other words, for these situations, the beta-binomial is considered to have degenerated to a binomial distribution (n,λ) where $\lambda = \bar{x}/n$. If $\lambda \geq \theta_o$, only false negative errors may be committed, for which the likelihood is

$$\hat{F}_n = \sum_{x=o}^{c-1} \binom{n}{x} \lambda^x (1-\lambda)^{n-x}.$$

When $\lambda < \theta_o$, only false positive errors may occur with a rate of

$$\hat{F}_p = \sum_{x=c}^{n} \binom{n}{x} \lambda^x (1-\lambda)^{n-x}.$$

The moment estimates receive more attention than the ML estimates in this discussion because (i) they are likely to be used in practical situations, especially where computer facilities are not available, (ii) they are asymptotically equivalent to the maximum likelihood (ML) estimates, and (iii) iteration for ML estimates (which are the best asymptotically normal estimates) is time consuming and may not converge in small samples. (See Zacks (1971, Section 5.2) for additional remarks about ML estimates.) However, simulations for the ML estimates were also conducted to provide comparative data for the ML and moment estimates. (In the rare instances where the ML iteration did not converge, the moment estimates were used.)

Table 2 reports the empirical means of the estimates $\hat{F}_p$ and $\hat{F}_n$. Enclosed within parentheses are the empirical means based on the ML estimates. The data indicate that the means of the moment estimates and the corresponding means of the ML estimates are almost identical when m is at least 50. The degree of bias (as measured by the discrepancy between the empirical means and their

TABLE 2

Empirical Means of the Estimates $\hat{F}_p$ and $\hat{F}_n$
(and of their Maximum Likelihood Counterparts)

Case	Error	Pop. Value	Empirical mean at m = 25	50	100	200	400	800
1	F_p	.040	.037	.038	.039	.040	.040	.040
			(.037)	(.039)	(.039)	(.040)	(.040)	(.040)
	F_n	.061	.059	.060	.060	.060	.060	.061
			(.062)	(.061)	(.061)	(.061)	(.061)	(.061)
2	F_p	.051	.049	.050	.051	.051	.051	.051
			(.050)	(.051)	(.051)	(.051)	(.051)	(.051)
	F_n	.027	.027	.027	.027	.027	.027	.027
			(.028)	(.028)	(.027)	(.027)	(.027)	(.027)
3	F_p	.120	.118	.119	.119	.119	.119	.119
			(.120)	(.119)	(.120)	(.119)	(.120)	(.120)
	F_n	.024	.023	.024	.024	.024	.024	.024
			(.022)	(.023)	(.023)	(.024)	(.024)	(.024)
4	F_p	.078	.078	.078	.078	.078	.078	.078
			(.081)	(.079)	(.079)	(.078)	(.078)	(.078)
	F_n	.041	.041	.041	.041	.041	.041	.041
			(.042)	(.042)	(.042)	(.041)	(.041)	(.041)
5	F_p	.157	.151	.153	.156	.156	.157	.157
			(.149)	(.154)	(.157)	(.156)	(.157)	(.157)
	F_n	.072	.078	.076	.073	.073	.072	.072
			(.080)	(.077)	(.074)	(.073)	(.072)	(.072)

population values) appears noticeable only in some instances when m = 25. In practically all instances, the bias seems negligible.

Table 3 reports the empirical values of $\sqrt{m}\, s_m(\hat{F}_p)$ and $\sqrt{m}\, s_m(\hat{F}_n)$ along with the corresponding values simulated for the ML estimates. The data indicate that for the situations under study, the moment estimates and the ML estimates behave almost identically in terms of sampling variability. The data also show that the asymptotic values $V_{\hat{f}p}$ and V_{fn} tend to underestimate the finite-sample values $\sqrt{m}\, s_m(\hat{F}_p)$ and $\sqrt{m}\, s_m(\hat{F}_n)$. The reader may deduce from the line $\lambda_o = .80$ of Table II of Wilcox (1977) that $\sqrt{m}\, s_m(\hat{F}_p) = .130\times\sqrt{10} = .411$ for m = 10, and $= .072\times\sqrt{30} = .394$ for m = 30. The asymptotic value is .212. Thus the asymptotic standard error tends to be smaller than the actual error. The magnitude of underestimation is substantial when m is small and α_{21} is moderate. (See Case 5

TABLE 3

Empirical Values of $\sqrt{m}\ s_m(\hat{F}_p)$ and $\sqrt{m}\ s_m(\hat{F}_n)$
(and of their Maximum Likelihood Counterparts)

Case	Error	Asymptotic Value	Empirical values at m = 25	50	100	200	400	800
1	F_p	.052	.060	.057	.054	.053	.053	.054
			(.059)	(.056)	(.055)	(.054)	(052)	(.052)
	F_n	.088	.092	.091	.091	.092	.090	.091
			(.092)	(.089)	(.089)	(.088)	(.088)	(.091)
2	F_p	.058	.063	.061	.060	.060	.060	.060
			(.063)	(.059)	(.058)	(.057)	(.058)	(.058)
	F_n	.033	.036	.035	.035	.035	.035	.035
			(.036)	(.035)	(.034)	(.033)	(.034)	(.034)
3	F_p	.102	.117	.109	.105	.103	.101	.103
			(.122)	(.104)	(.106)	(.105)	(.104)	(.102)
	F_n	.040	.039	.041	.039	.041	.040	.040
			(.043)	(.042)	(.041)	(.040)	(.042)	(.041)
4	F_p	.068	.072	.070	.070	.071	.072	.070
			(.076)	(.070)	(.069)	(.069)	(.068)	(.068)
	F_n	.041	.038	.036	.035	.036	.036	.036
			(.039)	(.035)	(.036)	(.036)	(.035)	(.035)
5	F_p	.212	.375	.287	.233	.221	.218	.211
			(.375)	(.264)	(.234)	(.215)	(.205)	(.216)
	F_n	.105	.177	.156	.125	.115	.111	.108
			(.192)	(.168)	(.123)	(.115)	(.111)	(.108)

with m = 25 or 50.) In other situations where α_{21} is reasonably large, the degree of underestimation is not large even with samples of size 25.

8. SUMMARY

This paper describes an asymptotic inferential procedure for the estimates of the false positive and false negative error rates. Formulae and tables are described for the computations of the standard errors. A simulation study indicates that the asymptotic standard errors may be used even with samples of 25 cases as long as the Kuder-Richardson Formula 21 reliability is reasonably large. Otherwise, a large sample would be required.

BIBLIOGRAPHY

Abramowitz, M. & Stegun, I. A. (1968). Handbook of mathematical functions. Washington, D.C.: National Bureau of Standards Applied Mathematics Series 55.

Duncan, G. T. (1974). An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association 69, 50-57.

Gradshteyn, I. S. & Ryzhik, I. M. (1965). Tables of integrals, series, and products. New York: Academic Press.

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics 29, 637-648.

Huynh, H. (1976a). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

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

Huynh, H. (1977a). Two simple classes of mastery scores based on the beta-binomial model. Psychometrika 42, 601-608.

Huynh, H. (1977b). Statistical inference for the kappa and kappa max reliability indices based on the beta-binomial model. Paper presented at the annual meeting of the Psychometric Society. University of North Carolina, June 16-17.

Huynh, H. (1978). Reliability of multiple classification. Psychometrika 43, 317-325.

IBM Application Program, System/360 (1971). Scientific subroutines package (360-CM-03X) Version III, Programmer's manual. White Plains, New York: IBM Corporation Technical Publication Department.

IMSL Library 1 (1977). Houston: International Mathematical and Statistical Libraries.

Keats, J. A. & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika 27, 59-72.

Kendall, M. G. & Stuart, A. (1967). The advanced theory of statistics (Second edition), Volume 2. London: Charles Griffin.

Lord, F. M. (1965). A strong true score theory, with applications. Psychometrika 30, 239-270.

Lord, F. M. (1969). Estimating true-score distribution in psychological testing (an empirical Bayes estimation problem). Psychometrika 34, 259-299.

Pearson, K. (1934). Tables of the incomplete beta function. Cambridge: University Press.

Rao, C. R. (1973). Linear statistical inference and its applications (Second edition). New York: John Wiley.

Wilcox, R. R. (1976). A note on the length and passing score of a mastery test. Journal of Educational Statistics 1, 359-364.

Wilcox, R. R. (1977). Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics 2, 289-307.

Zacks, S. (1971). The theory of statistical inference. New York: John Wiley.

ACKOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not neccessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance and comments of Joseph C. Saunders and Anthony J. Nitko are gratefully acknowledged.

APPENDIX A

Tables of the False Positive Error and Its Standard Error (times $\sqrt{M}$), the False Negative Error and Its Standard Error (times $\sqrt{M}$), and the Correlation Between F_p and F_n
(M = number of subjects, denoted by m in the text)

Input to the tables are (i) number of test items n, (ii) criterion level θ_o, (iii) mastery (passing) score c, (iv) test mean $\bar{x}$, and (v) the KR21 reliability estimate. It may be noted that if $\tilde{\alpha}$ and $\tilde{\beta}$ are estimates of the parameters α and β other than the moment estimates, then the entries for test mean and KR21 are simply $n\tilde{\alpha}/(\tilde{\alpha} + \tilde{\beta})$ and $n/(n + \tilde{\alpha} + \tilde{\beta})$, respectively.

For each entry (n, θ_o, c, $\bar{x}$, $\tilde{\alpha}_{21}$), five values may be read out. They are $\hat{F}_p$, V_{fp}, $\hat{F}_n$, V_{fn}, and $\hat{\rho}$, respectively.

Numerical Example

Let n = 10, θ_o = .60, and c = 6. For $\bar{x}$ = 5.0 and $\hat{\alpha}_{21}$ = .60, the tables yield the values $\hat{F}_p$ = .1667, V_{fp} = .1858, $\hat{F}_n$ = .0504, V_{fn} = .0548, and $\hat{\rho}$ = .2941.

Table of the False Positive Error and its S.E.*SQRT(M), the False Negative Error and its S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 5, Theta Zero: .60, Mastery Score: 3

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.5	.0129	.0184	.0249	.0321	.0384	.0419	.0411	.0347	.0215
	.0881	.1081	.1264	.1321	.1198	.0979	.0801	.0710	.0583
	.0000	.0000	.0000	.0003	.0011	.0026	.0045	.0058	.0050
	.0000	.0000	.0011	.0063	.0145	.0208	.0210	.0157	.0116
	.8848	.9009	.9126	.9025	.8605	.7584	.5915	.5751	.9151
1.0	.0676	.0784	.0890	.0965	.0986	.0943	.0833	.0651	.0386
	.2162	.2289	.2172	.1835	.1506	.1288	.1151	.1012	.0777
	.0000	.0000	.0005	.0022	.0051	.0086	.0115	.0124	.0096
	.0000	.0016	.0125	.0274	.0361	.0357	.0284	.0198	.0164
	.7526	.7898	.7710	.6851	.5356	.3799	.3490	.6038	.9460
1.5	.1740	.1835	.1851	.1776	.1630	.1430	.1182	.0882	.0508
	.3255	.2897	.2405	.2157	.1970	.1755	.1508	.1228	.0885
	.0000	.0008	.0042	.0093	.0146	.0186	.0205	.0193	.0135
	.0009	.0264	.0554	.0627	.0554	.0426	.0303	.0231	.0203
	.6040	.5073	.2413	.0267	-.0338	.0586	.3290	.7330	.9675
2.0	.3221	.3056	.2755	.2421	.2082	.1742	.1392	.1016	.0579
	.3616	.4107	.3941	.3363	.2756	.2218	.1754	.1344	.0934
	.0010	.0088	.0184	.0258	.0302	.0317	.0302	.0254	.0163
	.0607	.1329	.1143	.0828	.0579	.0418	.0328	.0282	.0238
	.0066	-.5539	-.5704	-.4354	-.1844	.1766	.5701	.8560	.9799
2.5	.4346	.3565	.3001	.2549	.2156	.1789	.1427	.1043	.0597
	1.3267	.8222	.5504	.3964	.2980	.2290	.1768	.1341	.0934
	.0235	.0425	.0497	.0511	.0492	.0448	.0385	.0298	.0179
	.4264	.1924	.1057	.0732	.0586	.0494	.0420	.0350	.0268
	-.9434	-.8048	-.4868	-.0663	.3055	.5830	.7820	.9159	.9857
3.0	.2833	.2548	.2299	.2060	.1819	.1565	.1286	.0964	.0564
	.7927	.5035	.3719	.2911	.2341	.1904	.1545	.1227	.0891
	.1160	.0990	.0858	.0744	.0638	.0535	.0429	.0314	.0180
	.3734	.2206	.1541	.1149	.0884	.0689	.0537	.0410	.0288
	-.0492	.1276	.2843	.4323	.5745	.7095	.8317	.9294	.9871
3.5	.0475	.0934	.1148	.1227	.1220	.1145	.1006	.0795	.0483
	.9558	.5065	.3073	.2169	.1699	.1415	.1213	.1036	.0806
	.1451	.1163	.0954	.0792	.0654	.0531	.0414	.0295	.0164
	.4909	.3112	.2044	.1434	.1049	.0786	.0592	.0436	.0292
	-.8742	-.7223	-.4448	-.1015	.2386	.5316	.7585	.9100	.9850
4.0	.0011	.0139	.0327	.0493	.0609	.0664	.0651	.0557	.0358
	.0811	.2505	.2560	.2100	.1603	.1206	.0948	.0813	.0679
	.0670	.0694	.0652	.0582	.0503	.0419	.0330	.0235	.0129
	.1811	.1187	.1065	.0944	.0802	.0664	.0534	.0408	.0273
	.7006	.1539	-.1596	-.1668	-.0239	.2364	.5732	.8535	.9793
4.5	.0000	.0005	.0039	.0105	.0184	.0254	.0293	.0280	.0194
	.0003	.0231	.0749	.1085	.1128	.0967	.0735	.0573	.0492
	.0129	.0180	.0222	.0243	.0240	.0219	.0183	.0135	.0074
	.0880	.0947	.0767	.0568	.0456	.0407	.0368	.0311	.0217
	.9065	.9064	.8432	.6873	.4667	.3415	.4494	.7633	.9691

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 5, Theta Zero: .60, Mastery Score: 4

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.5	.0011	.0023	.0041	.0066	.0093	.0113	.0120	.0106	.0069
	.0143	.0237	.0351	.0425	.0413	.0338	.0257	.0215	.0182
	.0000	.0000	.0001	.0006	.0027	.0068	.0122	.0167	.0157
	.0000	.0001	.0026	.0153	.0373	.0572	.0625	.0498	.0357
	.9487	.9614	.9686	.9641	.9434	.8839	.7387	.6091	.8852
1.0	.0100	.0145	.0197	.0243	.0271	.0275	.0254	.0205	.0124
	.0647	.0802	.0827	.0702	.0541	.0421	.0354	.0312	.0247
	.0000	.0001	.0012	.0054	.0130	.0228	.0321	.0368	.0303
	.0000	.0037	.0297	.0688	.0967	.1028	.0880	.0628	.0497
	.8902	.9246	.9168	.8729	.7727	.6047	.4588	.5653	.9230
1.5	.0383	.0462	.0511	.0520	.0495	.0446	.0376	.0284	.0166
	.1498	.1347	.0972	.0741	.0627	.0550	.0476	.0392	.0287
	.0000	.0019	.0099	.0231	.0378	.0507	.0587	.0581	.0428
	.0019	.0610	.1365	.1655	.1573	.1297	.0961	.0703	.0615
	.8506	.8199	.6593	.4105	.2184	.1710	.3055	.6542	.9534
2.0	.0973	.0972	.0895	.0793	.0685	.0574	.0459	.0335	.0190
	.1945	.1405	.1324	.1140	.0935	.0749	.0588	.0447	.0309
	.0021	.0206	.0451	.0661	.0810	.0888	.0885	.0780	.0525
	.1361	.3228	.3021	.2375	.1771	.1301	.0985	.0823	.0729
	.6031	-.1545	-.3693	-.3240	-.1643	.0976	.4539	.7991	.9728
2.5	.1654	.1324	.1090	.0908	.0755	.0617	.0485	.0350	.0198
	.5614	.3450	.2237	.1556	.1130	.0840	.0629	.0463	.0314
	.0536	.1032	.1269	.1365	.1370	.1301	.1160	.0933	.0583
	1.0248	.5299	.3155	.2175	.1667	.1370	.1174	.1019	.0835
	-.9025	-.7726	-.5290	-.2062	.1389	.4563	.7147	.8937	.9829
3.0	.1220	.1040	.0901	.0781	.0669	.0561	.0450	.0330	.0189
	.4025	.2371	.1652	.1230	.0945	.0737	.0574	.0438	.0305
	.2823	.2565	.2333	.2105	.1870	.1618	.1337	.1007	.0592
	.7452	.4755	.3536	.2791	.2267	.1863	.1532	.1236	.0917
	-.1005	.0816	.2465	.4039	.5552	.6984	.8267	.9282	.9870
3.5	.0216	.0403	.0474	.0487	.0467	.0424	.0361	.0277	.0163
	.4163	.1963	.1115	.0787	.0633	.0535	.0455	.0377	.0280
	.4126	.3398	.2863	.2430	.2050	.1694	.1342	.0970	.0545
	1.2271	.7850	.5350	.3910	.2980	.2319	.1811	.1382	.0956
	-.9293	-.7730	-.4391	-.0234	.3378	.6080	.7997	.9244	.9871
4.0	.0005	.0062	.0140	.0203	.0241	.0253	.0239	.0197	.0122
	.0375	.1082	.1029	.0788	.0570	.0421	.0340	.0296	.0238
	.2655	.2537	.2296	.2015	.1723	.1425	.1119	.0795	.0435
	.3562	.3708	.3618	.3190	.2694	.2222	.1787	.1370	.0919
	.0802	-.4077	-.4606	-.3387	-.0933	.2628	.6363	.8843	.9836
4.5	.0000	.0002	.0017	.0044	.0075	.0099	.0110	.0101	.0067
	.0001	.0104	.0320	.0437	.0427	.0345	.0254	.0204	.0174
	.0902	.0982	.1023	.1004	.0929	.0810	.0656	.0470	.0254
	.2508	.2423	.2056	.1785	.1646	.1526	.1355	.1106	.0751
	.6534	.6331	.4798	.2606	.1258	.1669	.4299	.7997	.9760

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 5, Theta Zero: .70, Mastery Score: 4

Test Mean	KR21= .100	200	.300	.400	.500	.600	.700	.800	.900
0.5	.0011	.0023	.0041	.0070	.0109	.0149	.0178	.0176	.0125
	.0143	.0237	.0367	.0514	.0610	.0593	.0478	.0365	.0314
	.0000	.0000	.0000	.0001	.0005	.0018	.0043	.0074	.0082
	.0000	.0000	.0001	.0021	.0092	.0209	.0302	.0284	.0187
	.9501	.9509	.9650	.9669	.9606	.9360	.8594	.6854	.8251
1.0	.0100	.0145	.0204	.0275	.0344	.0392	.0400	.0353	.0231
	.0647	.0823	.0999	.1066	.0972	.0783	.0605	.0509	.0435
	.0000	.0000	.0001	.0007	.0028	.0069	.0124	.0170	.0161
	.0000	.0001	.0026	.0136	.0311	.0455	.0479	.0368	.0254
	.8990	.9158	.9275	.9222	.8918	.8141	.6567	.5425	.8727
1.5	.0383	.0474	.0573	.0654	.0697	.0690	.0630	.0509	.0314
	.1507	.1677	.1632	.1363	.1068	.0858	.0737	.0648	.0520
	.0000	.0000	.0009	.0041	.0100	.0175	.0245	.0281	.0233
	.0000	.0027	.0216	.0492	.0676	.0698	.0577	.0394	.0312
	.8997	.8629	.8515	.7877	.6581	.4790	.3581	.5083	.9224
2.0	.0985	.1101	.1165	.1157	.1089	.0975	.0821	.0624	.0368
	.2538	.2268	.1712	.1396	.1233	.1096	.0948	.0782	.0579
	.0000	.0012	.0066	.0157	.0259	.0348	.0401	.0396	.0291
	.0011	.0392	.0909	.1100	.1024	.0815	.0579	.0414	.0375
	.7619	.7176	.5057	.2311	.0637	.0545	.2335	.6426	.9583
2.5	.1991	.1956	.1797	.1595	.1381	.1160	.0930	.0681	.0390
	.2908	.2501	.2470	.2165	.1796	.1450	.1146	.0877	.0611
	.0012	.0129	.0292	.0432	.0529	.0576	.0569	.0496	.0330
	.0786	.2096	.1975	.1512	.1081	.0763	.0571	.0490	.0444
	.4486	-.3180	-.4854	-.4253	-.2470	.0563	.4669	.8238	.9781
3.0	.3000	.2443	.2033	.1708	.1429	.1174	.0926	.0670	.0379
	.9205	.5972	.3986	.2834	.2098	.1537	.1207	.0901	.0616
	.0333	.0666	.0818	.0872	.0864	.0808	.0708	.0560	.0342
	.6796	.3447	.1934	.1280	.0981	.0825	.0722	.0630	.0508
	-.9248	-.6151	-.5697	-.1999	.1991	.5321	.7696	.9175	.9870
3.5	.2029	.1767	.1554	.1362	.1178	.0994	.0800	.0586	.0334
	.6328	.3887	.2799	.2145	.1694	.1355	.1081	.0840	.0587
	.1877	.1653	.1465	.1291	.1122	.0949	.0766	.0563	.0321
	.5358	.3331	.2429	.1884	.1506	.1218	.0982	.0773	.0553
	-.0266	.1658	.3325	.4848	.6255	.7529	.8620	.9441	.9900
4.0	.0263	.0560	.0696	.0739	.0724	.0666	.0570	.0436	.0254
	.6083	.3371	.2006	.1402	.1112	.0948	.0825	.0697	.0517
	.2366	.1933	.1600	.1332	.1101	.0890	.0687	.0482	.0261
	.7130	.4997	.3462	.2526	.1910	.1469	.1129	.0839	.0554
	-.8693	-.7243	-.4250	-.0348	.3381	.6266	.8216	.9368	.9896
4.5	.0002	.0047	.0132	.0213	.0270	.0294	.0283	.0233	.0142
	.0197	.1118	.1367	.1187	.0908	.0678	.0548	.0485	.0369
	.0900	.0931	.0883	.0792	.0681	.0559	.0431	.0298	.0156
	.2374	.1814	.1647	.1539	.1378	.1182	.0967	.0733	.0469
	.6138	.2436	-.0868	-.1407	-.0041	.2937	.6576	.3984	.9863

Table of the False Positive Error and its S.E.*SQRT(M), the False Negative Error and its S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 5, Theta Zero: .80, Mastery Score: 4

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.5	.0011	.0023	.0041	.0071	.0116	.0174	.0234	.0263	.0212
	.0143	.0237	.0368	.0542	.0735	.0857	.0809	.0618	.0500
	.0000	.0000	.0000	.0000	.0000	.0003	.0010	.0023	.0032
	.0000	.0000	.0000	.0001	.0011	.0043	.0094	.0119	.0080
	.9410	.9961	.9681	.9630	.9631	.9536	.9173	.7893	.7400
1.0	.0100	.0145	.0205	.0285	.0384	.0485	.0556	.0549	.0399
	.0647	.0824	.1033	.1248	.1369	.1297	.1052	.0804	.0700
	.0000	.0000	.0000	.0000	.0003	.0012	.0031	.0055	.0064
	.0000	.0000	.0001	.0010	.0049	.0117	.0174	.0166	.0103
	.9394	.9455	.9222	.9271	.9232	.8962	.8137	.6277	.7737
1.5	.0383	.0474	.0586	.0713	.0836	.0920	.0930	.0825	.0552
	.1507	.1709	.1903	.1961	.1790	.1467	.1152	.0973	.0856
	.0000	.0000	.0000	.0003	.0014	.0036	.0068	.0097	.0096
	.0000	.0000	.0009	.0059	.0148	.0227	.0244	.0185	.0120
	.9760	.9437	.8681	.8674	.8353	.7503	.5841	.4519	.8448
2.0	.0986	.1118	.1260	.1380	.1440	.1417	.1296	.1057	.0661
	.2548	.2696	.2635	2278	.1849	.1526	.1333	.1193	.0985
	.0000	.0000	.0003	.0017	.0045	.0084	.0123	.0145	.0123
	.0000	.0007	.0082	.0216	.0318	.0339	.0279	.0181	.0142
	.8124	.7782	.7737	.7083	.5696	.3803	.2494	.4091	.9172
2.5	.2007	.2141	.2216	.2191	.2068	.1861	.1576	.1207	.0717
	.3512	.3291	.2680	.2292	.2098	.1921	.1701	.1432	.1083
	.0000	.0003	.0025	.0067	.0118	.0164	.0194	.0193	.0143
	.0002	.0132	.0386	.0512	.0491	.0387	.0264	.0180	.0173
	.6408	.6136	.4094	.1292	-.0502	-.0693	.1190	.6063	.9625
3.0	.3469	.3412	.3172	.2846	.2485	.2101	.1693	.1243	.0712
	.3826	.3611	.3834	.3555	.3069	.2554	.2068	.1612	.1138
	.0003	.0048	.0124	.0195	.0245	.0270	.0267	.0231	.0152
	.0228	.0906	.0943	.0730	.0505	.0337	.0247	.0222	.0209
	.3417	-.3558	-.5572	-.5239	-.3604	-.0309	.4594	.8470	.9830
3.5	.4841	.4047	.3415	.2893	.2433	.2005	.1583	.1141	.0640
	1.2241	.9085	.6410	.4728	.3607	.2801	.2179	.1654	.1133
	.0134	.0302	.0381	.0407	.0401	.0370	.0319	.0247	.0147
	.3240	.1735	.0911	.0565	.0435	.0382	.0345	.0304	.0241
	-.9401	-.8653	-.5438	-.2240	.2585	.6127	.8245	.9401	.9906
4.0	.2976	.2640	.2352	.2079	.1808	.1528	.1227	.0891	.0498
	.9037	.5804	.4338	.3443	.2811	.2321	.1903	.1501	.1041
	.0942	.0794	.0679	.0581	.0490	.0403	.0315	.0224	.0123
	.3147	.1884	.1336	.1012	.0792	.0628	.0495	.0377	.0255
	.0776	.2704	.4316	.5740	.7002	.8091	.8969	.9593	.9927
4.5	.0182	.0527	.0731	.0821	.0829	.0773	.0663	.0501	.0285
	.6065	.4677	.3049	.2127	.1662	.1437	.1292	.1115	.0814
	.0817	.0687	.0565	.0462	.0373	.0292	.0218	.0146	.0075
	.2148	.1987	.1514	.1152	.0887	.0684	.0517	.0371	.0228
	-.5847	-.6145	-.3945	-.0524	.3344	.6510	.8484	.9502	.9920

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 6, Theta Zero: .60, Mastery Score: 4

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.6	.0026	.0046	.0077	.0120	.0169	.0212	.0233	.0216	.0146
	.0269	.0403	.0567	.0709	.0746	.0661	.0517	.0411	.0345
	.0000	.0000	.0000	.0002	.0009	.0028	.0057	.0085	.0084
	.0000	.0000	.0005	.0046	.0144	.0260	.0317	.0268	.0181
	.9315	.9386	.9551	.9553	.9426	.9032	.8021	.6626	.8704
1.2	.0227	.0299	.0383	.0464	.0522	.0541	.0511	.0424	.0266
	.1117	.1309	.1398	.1281	.1051	.0830	.0681	.0588	.0471
	.0000	.0000	.0003	.0020	.0055	.0107	.0160	.0192	.0164
	.0001	.0007	.0094	.0282	.0455	.0526	.0472	.0338	.0249
	.8744	.8671	.8899	.8615	.7899	.6637	.5266	.5727	.9114
1.8	.0811	.0923	.1004	.1024	.0987	.0900	.0770	.0593	.0355
	.2293	.2208	.1764	.1399	.1185	.1036	.0896	.0743	.0548
	.0000	.0005	.0039	.0105	.0185	.0259	.0309	.0311	.0233
	.0003	.0206	.0608	.0832	.0837	.0707	.0526	.0373	.0308
	.7972	.7603	.6386	.4385	.2669	.2102	.3182	.6369	.9465
2.4	.1903	.1897	.1766	.1585	.1383	.1171	.0947	.0700	.0408
	.2989	.2403	.2323	.2063	.1731	.1408	.1115	.0851	.0590
	.0006	.0091	.0224	.0346	.0434	.0481	.0481	.0424	.0287
	.0469	.1609	.1654	.1331	.0989	.0718	.0535	.0433	.0366
	.4964	-.1332	-.3604	-.3278	-.1706	.0939	.4479	.7877	.9686
3.0	.3098	.2548	.2133	.1800	.1514	.1252	.0997	.0731	.0424
	.8709	.5865	.3954	.2818	.2084	.1570	.1185	.0877	.0598
	.0277	.0579	.0722	.0775	.0771	.0724	.0639	.0510	.0319
	.5905	.3145	.1816	.1226	.0937	.0768	.0645	.0538	.0419
	-.9105	-.7966	-.5491	-.2002	.1628	.4738	.7149	.8855	.9794
3.6	.2149	.1893	.1682	.1489	.1301	.1111	.0909	.0681	.0402
	.6301	.3834	.2799	.2141	.1682	.1335	.1055	.0814	.0575
	.1779	.1559	.1379	.1216	.1059	.0901	.0734	.0549	.0323
	.5184	.3165	.2266	.1725	.1350	.1067	.0840	.0646	.0458
	-.0780	.0991	.2553	.[illegible]024	.5439	.6800	.8068	.9140	.9829
4.2	.0288	.0626	.0792	.0857	.0855	.0804	.0708	.0561	.0345
	.6522	.3694	.2223	.1541	.1189	.0976	.0821	.0684	.0522
	.2329	.1913	.1594	.1337	.1117	.0916	.0722	.0522	.0296
	.6562	.4568	.3109	.2225	.1549	.1246	.0944	.0699	.0471
	-.8719	-.7455	-.4849	-.1435	.2016	.4993	.7324	.8947	.9812
4.8	.0004	.0070	.0191	.0310	.0399	.0447	.0447	.0389	.0255
	.0300	.1434	.1671	.1447	.1128	.0846	.0647	.0534	.0439
	.1091	.1129	.1077	.0978	.0857	.0723	.0578	.0418	.0234
	.2475	.1800	.1596	.1436	.1241	.1040	.0846	.0655	.0445
	.6397	.2285	-.0909	-.1360	-.0263	.2025	.5245	.8251	.9742
5.4	.0000	.0002	.0017	.0054	.0107	.0158	.0193	.0193	.0138
	.0000	.0078	.0361	.0625	.0722	.0660	.0513	.0382	.0317
	.0218	.0287	.0350	.0388	.0395	.0370	.0317	.0239	.0136
	.1194	.1317	.1179	.0930	.0741	.0638	.0573	.0492	.0354
	.8936	.8825	.8443	.7378	.5656	.4188	.4483	.7209	.9608

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 6, Theta Zero: .60, Mastery Score: 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.6	.0002	.0005	.0011	.0022	.0037	.0053	.0063	.0061	.0043
	.0032	.0067	.0124	.0187	.0219	.0204	.0159	.0120	.0100
	.0000	.0000	.0000	.0003	.0018	.0057	.0122	.0193	.0207
	.0000	.0000	.0009	.0087	.0288	.0553	.0728	.0670	.0452
	.9611	.9695	.9800	.9800	.9723	.9475	.8722	.7074	.8264
1.2	.0028	.0047	.0075	.0105	.0131	.0145	.0143	.0123	.0079
	.0242	.0347	.0425	.0415	.0344	.0262	.0202	.0169	.0138
	.0000	.0000	.0006	.0037	.0110	.0222	.0351	.0446	.0408
	.0001	.0012	.0176	.0550	.0943	.1166	.1133	.0865	.0610
	.9198	.9496	.9521	.9356	.8903	.7918	.6305	.5606	.8702
1.8	.0150	.0201	.0245	.0269	.0272	.0256	.0224	.0175	.0106
	.0783	.0824	.0653	.0476	.0367	.0305	.0260	.0218	.0163
	.0000	.0010	.0073	.0204	.0375	.0551	.0690	.0736	.0588
	.0005	.0376	.1168	.1697	.1828	.1662	.1318	.0937	.0744
	.8980	.8985	.8351	.6929	.4995	.3486	.3333	.5516	.9180
2.4	.0490	.0526	.0508	.0465	.0410	.0350	.0284	.0211	.0123
	.1431	.0843	.0701	.0617	.0521	.0425	.0337	.0257	.0178
	.0011	.0168	.0431	.0694	.0909	.1054	.1106	.1026	.0733
	.0834	.3049	.3369	.2926	.2335	.1780	.1325	.1026	.0887
	.7924	.3481	-.0752	-.1727	-.1119	.0531	.3291	.6943	.9529
3.0	.1022	.0831	.0686	.0572	.0476	.0390	.0308	.0224	.0129
	.3036	.2075	.1374	.0957	.0692	.0510	.0377	.0274	.0184
	.0500	.1103	.1442	.1618	.1681	.1648	.1518	.1266	.0826
	1.1078	.6619	.4181	.2925	.2207	.1755	.1453	.1239	.1032
	-.8525	-.7564	-.5606	-.3021	-.0058	.3041	.5992	.8376	.9721
3.6	.0808	.0679	.0583	.0502	.0428	.0358	.0288	.0212	.0123
	.2768	.1584	.1080	.0789	.0596	.0456	.0349	.0251	.0179
	.3413	.3169	.2936	.2697	.2438	.2149	.1812	.1399	.0850
	.8508	.5560	.4181	.3333	.2731	.2265	.1878	.1531	.1158
	-.1869	-.0187	.1411	.3005	.4609	.6199	.7708	.8988	.9803
4.2	.0114	.0237	.0288	.0301	.0292	.0267	.0229	.0177	.0107
	.2502	.1281	.0725	.0500	.0393	.0327	.0273	.0223	.0164
	.5224	.4406	.3770	.3242	.2769	.2318	.1862	.1371	.0791
	1.2797	.8958	.6313	.4711	.3647	.2876	.2271	.1752	.1230
	-.9352	-.8121	-.5234	-.1289	.2399	.5303	.7472	.8983	.9813
4.8	.0001	.0027	.0072	.0113	.0140	.0152	.0148	.0125	.0079
	.0121	.0546	.0600	.0490	.0364	.0267	.0207	.0173	.0140
	.3439	.3315	.3050	.2718	.2356	.1975	.1574	.1137	.0636
	.3972	[illegible]	.4159	.3795	.3288	.2765	.2259	.1758	.1200
	.0258	-.3551	-.4446	-.3588	-.1550	.1634	.5439	.8421	.9765
5.4	.0000	.0001	.0006	.0020	.0038	.0055	.0065	.0063	.0043
	.0000	.0031	.0136	.0225	.0247	.0214	.0161	.0121	.0101
	.1224	.1300	.1347	.1335	.1256	.1117	.0922	.0676	.0374
	.2811	.2805	.2542	.2241	.2041	.1887	.1699	.1419	.0989
	.5333	.5686	.4862	.3261	.1900	.1754	.3584	.7284	.9650

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 6, Theta Zero: .70, Mastery Score: 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.6	.0002	.0005	.0011	.0024	.0044	.0072	.0100	.0112	.0089
	.0032	.0067	.0127	.0219	.0319	.0372	.0339	.0250	.0197
	.0000	.0000	.0000	.0000	.0003	.0014	.0041	.0084	.0110
	.0000	.0000	.0000	.0009	.0058	.0180	.0331	.0382	.0258
	.9640	.9637	.9750	.9794	.9784	.9690	.9360	.8201	.7783
1.2	.0028	.0047	.0077	.0119	.0170	.0217	.0245	.0236	.0167
	.0242	.0351	.0490	.0610	.0639	.0564	.0434	.0328	.0275
	.0000	.0000	.0000	.0004	.0020	.0062	.0130	.0205	.0221
	.0000	.0000	.0010	.0085	.0261	.0472	.0594	.0524	.0335
	.9235	.9308	.9521	.9541	.9432	.9088	.8187	.6522	.8075
1.8	.0150	.0205	.0274	.0345	.0400	.0426	.0414	.0354	.0230
	.0784	.0957	.1056	.0980	.0799	.0615	.0484	.0407	.0335
	.0000	.0000	.0005	.0030	.0089	.0179	.0282	.0358	.0327
	.0000	.0009	.0135	.0419	.0703	.0843	.0790	.0575	.0395
	.9861	.9066	.9111	.8877	.8252	.7049	.5404	.5005	.8682
2.4	.0494	.0592	.0672	.0709	.0702	.0654	.0569	.0447	.0273
	.1684	.1688	.1346	.1016	.0819	.0700	.0604	.0504	.0380
	.0000	.0006	.0051	.0146	.0271	.0399	.0497	.0527	.0418
	.0003	.0250	.0818	.1195	.1265	.1113	.0844	.0576	.0469
	.8542	.8340	.7467	.5657	.3556	.2252	.2511	.5290	.9271
3.0	.1214	.1258	.1202	.1095	.0966	.0825	.0671	.0500	.0293
	.2482	.1658	.1509	.1367	.1169	.0961	.0766	.0587	.0410
	.0096	.0109	.0293	.0478	.0627	.0723	.0750	.0686	.0482
	.0486	.2067	.2330	.1985	.1524	.1110	.0801	.0629	.0562
	.6907	.1661	-.2289	-.2942	-.2082	-.0055	.3281	.7290	.9634
3.6	.2172	.1793	[illegible]	.1256	.1051	.0864	.0685	.0499	.0287
	.5796	.4193	[illegible]	.2035	.1497	.1121	.0841	.0619	.0418
	.0325	.0752	.0986	.1097	[illegible]	.1084	.0979	.0799	.0508
	.7704	.4598	.2755	.1833	[illegible]	.1085	.0921	.0799	.0659
	-.8820	-.8013	-.6085	-.3172	.0363	.3870	.6779	.8788	.9800
4.2	.1559	.1341	.1170	.1020	.0879	.0740	.0596	.0440	.0254
	.4989	.2991	.2112	.1590	.1234	.0970	.0761	.0581	.0402
	.2421	.2179	.1966	.1761	.1554	.1337	.1098	.0824	.0484
	.6425	.4060	.3001	.2357	.1906	.1558	.1270	.1011	.0735
	-.1039	.0848	.2534	.4116	.5615	.7014	.8263	.9259	.9858
4.8	.0154	.0375	.0486	.0528	.0524	.0487	.0422	.0326	.0194
	.4061	.2509	.1511	.1038	.0803	.0669	.0572	.0479	.0355
	.3177	.2661	.2238	.1886	.1576	.1289	.1008	.0719	.0398
	.7817	.6053	.4352	.3239	.2481	.1929	.1496	.1124	.0752
	-.8598	-.7485	-.4864	-.1223	.2505	.5583	.7793	.9182	.9858
5.4	.0001	.0022	.0074	.0133	.0180	.0204	.0203	.0173	.0108
	.0058	.0587	.0867	.0827	.0665	.0499	.0386	.0330	.0266
	.1223	.1263	.1219	.1114	.0975	.0814	.0639	.0448	.0240
	.2759	.2313	.2060	.1923	.1748	.1529	.1276	.0989	.0646
	.5384	.3087	.0114	-.0884	-.0123	.2225	.5740	.8631	.9812

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 6, Theta Zero: .80, Mastery Score: 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.6	.0002	.0005	.0011	.0024	.0047	.0085	.0135	.0179	.0166
	.0032	.0067	.0128	.0227	.0373	.0531	.0600	.0500	.0356
	0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0043
	.0000	.0000	.0000	.0000	.0005	.0031	.0093	.0153	.0120
	.9653	.9659	.9641	.9752	.9774	.9750	.9605	.8979	.7463
1.2	.0028	.0047	.0077	.0123	.0189	.0273	.0356	.0395	.0320
	.0242	.0351	.0499	.0688	.0873	.0953	.0856	.0632	.0493
	.0000	.0000	.0000	.0000	.0002	.0009	.0030	.0065	.0090
	.0000	.0000	.0000	.0004	.0032	.0105	.0198	.0233	.0150
	.9151	.9131	.9443	.9516	.9537	.9444	.907[illegible]	.7858	.7238
1.8	.0150	.0205	.0278	.0373	.0483	.0586	.0647	.0623	.0452
	.0784	.0964	.1173	.1347	.1373	.1212	.0943	.0714	.0608
	.0000	.0000	.0000	.0002	.0010	.0033	.0074	.0122	.0137
	.0000	.0000	.0004	.0037	.0127	.0243	.0315	.0278	.0166
	.9990	.9058	.9093	.9165	.9069	.8692	.7719	.5908	.7653
2.4	.0494	.0597	.0720	.0848	.0950	.0997	.0966	.0832	.0550
	.1686	.1883	.2003	.1884	.1576	.1241	.0992	.0846	.0715
	.0000	.0000	.0002	.0012	.0041	.0089	.0147	.0193	.0180
	.0000	.0002	.0050	.0185	.0340	.0425	.0401	.0282	.0186
	.9990	.8443	.8550	.8341	.7672	.6368	.4591	.4111	.8560
3.0	.1219	.1360	.1482	.1540	.1518	.1418	.1242	.0982	.0605
	.2787	.2800	.2378	.1886	.1572	.1382	.1222	.1042	.0806
	.0000	.0002	.0019	.0063	.0128	.0196	.0251	.0269	.0215
	.0000	.0080	.0346	.0569	.0631	.0557	.0409	.0264	.0223
	.8464	.7519	.6711	.4851	.2617	.1179	.1436	.4732	.9341
3.6	.2466	.2520	.2421	.2227	.1980	.1700	.1390	.1038	.0609
	.3604	.2757	.2645	.2527	.2252	.1912	.1563	.1222	.0867
	.0001	.0041	.0128	.0224	.0304	.0355	.0370	.0336	.0233
	.0129	.0893	.1145	.1005	.0758	.0526	.0363	.0292	.0275
	.5758	.1153	-.2963	-.3877	-.3160	-.1010	.2977	.7586	.9726
4.2	.3979	.3379	.2862	.2427	.2043	.1686	.1336	.0970	.0553
	.8441	.7126	.5174	.3832	.2908	.2237	.1720	.1290	.0878
	.0133	.0356	.0484	.0542	.0552	.0525	.0466	.0372	.0229
	.3740	.2439	.1396	.0866	.0620	.0511	.0454	.0404	.03[illegible]
	-.8941	-.8488	-.6809	-.3649	.0691	.4753	.7553	.9154	.98[illegible]
4.8	.2585	.2273	.2014	.1774	.1539	.1300	.1047	.0765	.0434
	.7921	.4988	.3665	.2862	.2300	.1870	.1513	.1182	.0816
	.1302	.1121	.0975	.0846	.0724	.0603	.0480	.0347	.0194
	.3959	.2419	.1742	.1339	.1060	.0850	.0677	.052[illegible]	.0360
	.0162	.2124	.3780	.5258	.6592	.7770	.8756	.9489	.9904
5.4	.0108	.0381	.0564	.0657	.0678	.0644	.0561	.0431	.0249
	.4112	.3768	.2574	.1799	.1367	.1142	.100[illegible]	.0870	.0641
	.1152	.0997	.0836	.0693	.0566	.0450	.0[illegible]39	.0231	.0121
	.2488	.2497	.1986	.1546	.1209	.0944	.0724	.0527	.0329
	-.4671	-.5964	-.4193	-.1209	.2449	.5805	.8113	.9369	.9896

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 7, Theta Zero: .60, Mastery Score: 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.7	.0005	.0011	.0023	.0044	.0073	.0105	.0130	.0133	.0098
	.0070	.0131	.0224	.0337	.0420	.0423	.0350	.0259	.0209
	.0000	.0000	.0000	.0001	.0007	.0026	.0062	.0107	.0120
	.0000	.0000	.0002	.0028	.0122	.0274	.0400	.0389	.0256
	.9514	.9519	.9716	.9743	.9700	.9528	.9010	.7702	.8306
1.4	.0073	.0109	.0160	.0218	.0271	.0306	.0310	.0273	.0181
	.0494	.0650	.0792	.0816	.0720	.0571	.0439	.0356	.0289
	.0000	.0000	.0002	.0015	.0052	.0115	.0194	.0256	.0239
	.0001	.0003	.0061	.0248	.0487	.0652	.0660	.0508	.0343
	.9023	.9271	.9349	.9255	.8920	.8186	.6906	.6075	.8701
2.1	.0361	.0447	.0528	.0577	.0587	.0559	.0496	.0395	.0245
	.1413	.1501	.1288	.0995	.0781	.0646	.0548	.0456	.0343
	.0000	.0003	.0032	.0103	.0204	.0313	.0401	.0432	.0347
	.0001	.0139	.0573	.0939	.1068	.0990	.0784	.0549	.0417
	.8490	.8595	.8113	.6948	.5328	.3975	.3773	.5727	.9181
2.8	.1077	.1135	.1101	.1015	.0903	.0777	.0637	.0478	.0284
	.2410	.1644	.1400	.1255	.1077	.0889	.0710	.0543	.0376
	.0004	.0083	.0239	.0405	.0542	.0632	.0662	.0610	.0434
	.0315	.1672	.2033	.1806	.1432	.1074	.0768	.0602	.0498
	.7106	.3273	-.0637	-.1646	-.1013	.0748	.3601	.7115	.9530
3.5	.2125	.1766	.1479	.1246	.1046	.0863	.0688	.0507	.0298
	.5349	.3978	.2731	.1945	.1428	.1064	.0792	.0577	.0387
	.0288	.0690	.0914	.1021	.1049	.1014	.0913	.0755	.0488
	.7027	.4349	.2664	.1810	.1351	.1076	.0887	.0735	.0578
	-.8578	-.7749	-.5723	-.2838	.0454	.3639	.6381	.8477	.9712
4.2	.1583	.1371	.1204	.1055	.0915	.0777	.0634	.0476	.0283
	.4857	.2912	.2053	.1540	.1188	.0925	.0717	.0541	.0375
	.2388	.2140	.1927	.1725	.1525	.1316	.1089	.0829	.0500
	.6481	.4054	.2959	.2291	.1819	.1457	.1159	.0899	.0643
	-.1233	.0509	.2068	.3553	.5000	.6412	.7761	.8954	.9777
4.9	.0170	.0411	.0536	.0585	.0589	.0556	.0491	.0391	.0243
	.4301	.2630	.1582	.1081	.0823	.0667	.0553	.0453	.0340
	.3266	.2742	.2315	.1963	.1655	.1369	.1090	.0798	.0462
	.7747	.5890	.4150	.3031	.2279	.1740	.1327	.0988	.0670
	-.8707	-.7688	-.5286	-.1951	.1534	.4581	.7006	.8762	.9763
5.6	.0001	.0035	.0110	.0191	.0257	.0296	.0303	.0268	.0180
	.0107	.0794	.1060	.0974	.0780	.0588	.0441	.0352	.0285
	.1579	.1623	.1563	.1436	.1272	.1085	.0877	.0644	.0368
	.2986	.2396	.2133	.1941	.1703	.1445	.1187	.0928	.0637
	.5542	.2489	-.0501	-.1188	-.0352	.1669	.4729	.7912	.9675
6.3	.0000	.0000	.0007	.0027	.0060	.0097	.0125	.0130	.0097
	.0000	.0025	.0169	.0350	.0449	.0440	.0354	.0257	.0205
	.0330	.0413	.0494	.0551	.0569	.0544	.0477	.0368	.0213
	.1496	.1645	.1570	.1315	.1064	.0900	.0797	.0690	.0509
	.9156	.8510	.8310	.7547	.6201	.4770	.4544	.6780	.9499

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 7, Theta Zero: .60, Mastery Score: 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.7	.0000	.0001	.0003	.0007	.0015	.0024	.0032	.0035	0027
	.0006	.0018	.0041	.0076	.0107	.0116	.0099	.0071	.0056
	.0000	.0000	.0000	.0002	.0012	.0044	.0112	.0204	.0250
	.0000	.0000	.0003	.0046	.0203	.0484	.0759	.0813	.0556
	.9690	.9702	.9855	.9870	.9838	.9717	.9321	.8045	.7812
1.4	.0008	.0015	.0028	.0045	.0062	.0075	.0080	.0073	.0050
	.0083	.0138	.0200	.0226	.0208	.0166	.0124	.0096	.0078
	.0000	.0000	.0003	.0024	.0086	.0200	.0354	.0498	.0504
	.0000	.0004	.0096	.0405	.0839	.1196	.1310	.1097	.0728
	.9389	.9623	.9680	.9614	.9384	.8828	.7624	.6141	.8118
2.1	.0058	.0086	.0116	.0137	.0147	.0146	.0132	.0107	.0067
	.0377	.0456	.0410	.0311	.0231	.0180	.0148	.0123	.0094
	.0000	.0005	.0051	.0168	.0348	.0557	.0751	.0860	.0742
	.0002	.0215	.0923	.1594	.1929	.1921	.1641	.1195	.0864
	.9154	.9327	.9057	.8298	.6931	.5254	.4157	.4997	.8708
2.8	.0242	.0280	.0285	.0269	.0244	.0212	.0175	.0132	.0079
	.0941	.0581	.0407	.0344	.0294	.0244	.0195	.0150	.0104
	.0006	.0129	.0387	.0681	.0950	.1160	.1276	.1243	.0940
	.0477	.2665	.3442	.3275	.2790	.2228	.1684	.1246	.1022
	.8656	.6568	.2559	.0266	-.0129	.0669	.2596	.5915	.9249
3.5	.0619	.0514	.0427	.0357	.0298	.0245	.0194	.0142	.0083
	.1576	.1219	.0832	.0585	.0423	.0310	.0228	.0164	.0109
	.0443	.1110	.1535	.1788	.1916	.1934	.1834	.1581	.1075
	1.1093	.7555	.5047	.3627	.2750	.2157	.1737	.1441	.1204
	-.7663	-.7257	-.5636	-.3498	-.0994	.1816	.4838	.7697	.9574
4.2	.0528	.0439	.0374	.0320	.0272	.0227	.0183	.0135	.0080
	.1872	.1045	.0700	.0504	.0375	.0284	.0214	.0157	.0107
	.3871	.3658	.3442	.3210	.2947	.2640	.2267	.1790	.1121
	.9524	.6240	.4733	.3798	.3131	.2611	.2178	.1790	.1376
	-.2560	-.1053	.0442	.2003	3649	.5363	.7081	.8639	.9715
4.9	.0060	.0138	.0174	.0185	.0181	.0167	.0144	.0113	.0069
	.1472	.0824	.0468	.0317	.0245	.0201	.0166	.0134	.0098
	.6173	.5314	.4610	.4011	.3464	.2933	.2387	.1785	.1055
	1.2565	.9609	.6999	.5330	.4193	.3352	.2679	.2090	.1437
	-.9406	-.8448	-.5976	-.2289	.1428	.4504	.6909	.8686	.9740
5.6	.0000	.0012	.0037	.0062	.0081	.0091	.0091	.0078	.0051
	.0038	.0270	.0343	.0299	.0230	.0169	.0127	.0103	.0082
	.4188	.4055	.3776	.3407	.2988	.2534	.2044	.1498	.0855
	.4201	.4402	.4529	.4259	.3781	.3240	.2688	.2121	.1470
	-.0564	-.3295	-.4362	-.3779	-.2064	.0786	.4538	.7939	.9675
6.3	.0000	.0000	.0002	.0009	.0020	.0030	.0038	.0038	.0028
	.0000	.0009	.0057	.0113	.0139	.0131	.0102	.0074	.0060
	.1567	.1632	.1675	.1665	.1584	.1429	.1199	.0895	.0507
	.3075	.3096	.2930	.2650	.2411	.2222	.2014	.1712	.1221
	.6254	.4923	.4601	.3500	.2305	.1881	.3103	.6553	.9507

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 7, Theta Zero: .70, Mastery Score: 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.7	.0005	.0011	.0023	.0045	.0079	.0128	.0182	.0214	.0179
	.0070	.013[illegible]	.0225	.0363	.0529	.0649	.0638	.0492	.0367
	.0000	.0000	.0000	.0000	.0001	.0004	.0015	.0035	.0049
	.0000	.0000	.0000	.0002	.0016	.0061	.0132	.0169	.0117
	.9531	.9521	.9547	.9709	.9719	.9660	.9421	.8566	.7842
1.4	.0073	.0109	.0161	.0232	.0319	.0406	.0465	.0460	.0339
	.0494	.0652	.0845	.1034	.1120	.1045	.0842	.0631	.0512
	.0000	.0000	.0000	.0001	.0006	.0022	.0052	.0089	.0101
	.0000	.0000	.0002	.0023	.0089	.0187	.0258	.0240	.0150
	.9033	.9317	.9286	.9356	.9304	.9053	.8377	.6988	.7995
2.1	.0361	.0450	.0558	.0672	.0769	.0821	.0806	.0702	.0469
	.1413	.1611	.1755	.1701	.1465	.1172	.0927	.0765	.0624
	.0000	.0000	.0001	.0010	.0033	.0073	.0122	.0160	.0151
	.0000	.0002	.0037	.0149	.0287	.0371	.0362	.0267	.0175
	.8864	.8593	.8699	.8566	.8076	.7115	.5755	.5260	.8576
2.8	.1081	.1212	.1326	.1382	.1367	.1281	.1126	.0896	.0559
	.2608	.2635	.2274	.1830	.1517	.1308	.1129	.0942	.0711
	.0000	.0002	.0017	.0057	.0115	.0177	.0226	.0242	.0194
	.0000	.0071	.0313	.0518	.0580	.0521	.0397	.0267	.0206
	.8348	.7512	.6821	.5277	.3485	.2366	.2662	.5294	.9204
3.5	.2385	.2421	.2316	.2124	.1887	.1623	.1332	.1003	.0600
	.3476	.2746	.2613	.2430	.2122	.1771	.1425	.1097	.0767
	.0002	.0042	.0127	.0218	.0292	.0338	.0351	.0319	.0224
	.0139	.0879	.1097	.0958	.0733	.0527	.0377	.0291	.0247
	.5318	.0814	-.2638	-.3197	-.2249	-.0070	.3378	.7278	.9594
4.2	.3928	.3320	.2816	.2396	.2027	.1685	.1350	.0998	.0587
	.8671	.6859	.4871	.3563	.2675	.2035	.1546	.1147	.0780
	.0146	.0368	.0438	.0540	.0548	.0521	.0464	.0374	.0237
	.3806	.2359	.1368	.0887	.0654	.0529	.0444	.0371	.0289
	-.8964	-.8273	-.6308	-.3093	.0736	.4211	.6884	.8736	.9765
4.9	.2619	.2332	.2089	.1861	.1637	.1405	.1155	.0870	.0517
	.7532	.4726	.3453	.2673	.2124	.1703	.1358	.1055	.0743
	.1327	.1146	.1002	.0875	.0755	.0638	.0516	.0384	.0225
	.4010	.2403	.1697	.1278	.0990	.0776	.0605	.0460	.0320
	-.0721	.1044	.2593	.4047	.5444	.6788	.8042	.9111	.9815
5.6	.0188	.0545	.0767	.0879	.0912	.0881	.0790	.0634	.0392
	.5700	.4290	.2815	.1977	.1498	.1207	.1011	.0849	.0649
	.1451	.1233	.1035	.0868	.0722	.0589	.0460	.0329	.0184
	.3200	.2809	.2053	.1514	.1139	.0868	.0660	.0487	.0322
	-.6873	-.6713	-.4685	-.1826	.1388	.4476	.7052	.8858	.9795
6.3	.0000	.0021	.0092	.0188	.0278	.0341	.0362	.0326	.0217
	.0035	.0652	.1224	.1351	.1206	.0961	.0730	.0587	.0481
	.0330	.0398	.0429	.0423	.0391	.0340	.0276	.0199	.0110
	.1477	.1256	.0911	.0734	.0646	.0575	.0496	.0399	.0270
	.8430	.7845	.5860	.3497	.2307	.2812	.5043	.8036	.9712

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 7, Theta Zero: .70, Mastery Score: 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.7	.0000	.0001	.0003	.0008	.0018	.0034	.0055	.0070	.0062
	.0006	.0018	.0041	.0087	.0155	.0217	.0230	.0180	.0128
	.0000	.0000	.0000	.0000	.0001	.0009	.0035	.0087	.0134
	.0000	.0000	.0000	.0003	.0034	.0140	.0324	.0456	.0340
	.9699	.9696	.9762	.9854	.9861	.9821	.9659	.8988	.7701
1.4	.0008	.0015	.0029	.0051	.0083	.0119	.0148	.0156	.0120
	.0083	.0139	.0223	.0323	.0390	.0386	.0316	.0227	.0178
	.0000	.0000	.0000	.0002	.0014	.0051	.0126	.0227	.0278
	.0000	.0000	.0004	.0049	.0200	.0443	.0657	.0669	.0431
	.9421	.9500	.9646	.9689	.9650	.9479	.8979	.7618	.7592
2.1	.0058	.0087	.0128	.0179	.0227	.0260	.0269	.0243	.0167
	.0377	.0505	.0627	.0654	.0578	.0455	.0342	.0268	.0219
	.0000	.0000	.0002	.0021	.0074	.0170	.0299	.0418	.0421
	.0000	.0003	.0078	.0327	.0663	.0917	.0965	.0768	.0488
	.9339	.9295	.9389	.9306	.8988	.8271	.6921	.5566	.8075
2.8	.0243	.0312	.0381	.0429	.0447	.0434	.0392	.0318	.0201
	.1036	.1146	.1009	.0774	.0590	.0474	.0398	.0331	.0254
	.0000	.0003	.0037	.0126	.0264	.0423	.0568	.0644	.0549
	.0001	.0147	.0674	.1180	.1411	.1363	.1112	.0769	.0560
	.8737	.8864	.8505	.7500	.5853	.4121	.5294	.4676	.8849
3.5	.0726	.0794	.0791	.0743	.0670	.0581	.0480	.0363	.0218
	.1943	.1289	.0991	.0880	.0767	.0642	.0517	.0398	.0278
	.0003	.0086	.0275	.0492	.0689	.0836	.0909	.0870	.0644
	.0278	.1870	.2498	.2350	.1934	.1474	.1066	.0784	.0670
	.7918	.5267	.0912	-.1114	-.1210	-.0046	.2434	.6301	.9430
4.2	.1542	.1296	.1087	.0915	.0767	.0632	.0502	.0368	.0215
	.3521	.2879	.2022	.1451	.1066	.0794	.0591	.0430	.0288
	.0298	.0794	.1105	.1279	.1353	.1341	.1245	.1046	.0690
	.8022	.5554	.3547	.2417	.1765	1372	.1127	.0961	.0802
	-.8086	-.7747	-.6167	-.3795	-.0775	.2582	.5796	.8316	.9709
4.9	.1184	.1008	.0874	.0758	.0651	.0547	.0441	.0327	.0191
	.3886	.2280	.1584	.1176	.0901	.0698	.0540	.0407	.0278
	.2910	.2667	.2443	.2220	.1987	.1733	.1447	.1106	.0666
	.7368	.4712	.3515	.2786	.2272	.1874	.1542	.1240	.0915
	-.1735	.0072	.1745	.3366	.4947	6468	.7874	.9051	.9807
5.6	.0090	.0249	.0338	.0375	.0377	.0354	.0309	.0242	.0146
	.2660	.1839	.1128	.0767	.0582	.0476	.0400	.0332	.0246
	.3962	.3391	.2892	.2463	.2079	.1717	.1358	.0982	.0554
	.3076	.6877	.5119	.3886	.3017	.2371	.1856	.1409	.0954
	-.8497	-.7688	-.5390	-.2004	.1685	.4908	.7351	.8974	.9813
6.3	.0000	.0010	.0042	.0083	.0118	.0141	.0145	.0127	.0081
	.0017	.0302	.0539	.0565	.0479	.0367	.0277	.0227	.0184
	.1566	.1606	.1565	.1451	.1288	.1091	.0868	.0619	.0337
	.3057	.2748	.2456	.2287	.2098	.1861	.1579	.1247	.0831
	.4519	.3216	.0749	-.0445	-.0102	.1730	.4975	.8227	.9749

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 7, Theta Zero: .80, Mastery Score: 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.7	.0000	.0001	.0003	.0008	.0019	.0040	.0077	.0120	.0128
	.0006	.0018	.0041	.0089	.0177	.0306	.0417	.0401	.0269
	.0000	.0000	.0000	.0000	.0000	.0001	.0007	.0025	.0053
	.0000	.0000	.0000	.0000	.0002	.0020	.0082	.0175	.0166
	.9632	.9687	.9772	.9833	.9842	.9841	.9777	.9458	.7947
1.4	.0008	.0015	.0029	.0052	.0092	.0152	.0225	.0281	.0253
	.0083	.0139	.0225	.0354	.0518	.0654	.0667	.0521	.0363
	.0000	.0000	.0000	.0000	.0001	.0007	.0027	.0070	.0114
	.0000	.0000	.0000	.0002	.0019	.0085	.0202	.0289	.0210
	.9365	.9397	.9452	.9643	.9681	.9652	.9474	.8768	.7267
2.1	.0058	.0087	.0130	.0192	.0275	.0368	.0445	.0466	.0365
	.0377	.0507	.0674	.0860	.0981	.0956	.0786	.0565	.0443
	.0000	.0000	.0000	.0001	.0007	.0028	.0074	.0140	.0179
	.0000	.0000	.0001	.0021	.0099	.0236	.0362	.0371	.0225
	.9621	.9990	.9316	.9405	.9393	.9213	.8672	.7222	.7120
2.8	.0243	.0314	.0405	.0512	.0617	.0693	.0713	.0648	.0453
	.1036	.1227	.1410	.1453	.1304	.1047	.0796	.0629	.0527
	.0000	.0000	.0001	.0008	.0035	.0087	.0162	.0234	.0241
	.0000	.0001	.0028	.0144	.0328	.0479	.0512	.0398	.0238
	.9457	.8817	.8952	.8915	.8590	.7816	.6356	.4855	.7868
3.5	.0727	.0849	.0974	.1066	.1100	.1069	.0969	.0792	.0506
	.2073	.2208	.2028	.1633	.1284	.1058	.0907	.0773	.0608
	.0000	.0001	.0014	.0056	.0127	.0216	.0298	.0343	.0294
	.0000	.0044	.0282	.0572	.0727	.0713	.0569	.0372	.0274
	.9990	.8197	.7922	.6907	.5166	.3276	.2359	.4034	.8945
4.2	.1724	.1828	.1820	.1720	.1561	.1364	.1132	.0859	.0515
	.3187	.2430	.1978	.1835	.1665	.1441	.1192	.0938	.0668
	.0001	.0032	.0122	.0239	.0348	.0430	.0470	.0447	.0326
	.0067	.0803	.1258	.1241	.1014	.0741	.0508	.0374	.0341
	.6902	.4634	.0329	-.1953	-.2213	-.0998	.1968	.6601	.9582
4.9	.3215	.2780	.2370	.2016	.1700	.1406	.1118	.0818	.0473
	.5673	.5483	.4127	.3088	.2344	.1794	.1369	.1018	.0688
	.0123	.0391	.0569	.0665	.0700	.0686	.0624	.0511	.0326
	.3947	.3087	.1916	.1222	.0849	.0662	.0568	.0504	.0419
	-.8097	-.8201	-.6866	-.4380	-.0700	.3411	.6763	.8853	.9815
5.6	.2225	.1942	.1711	.1502	.1300	.1098	.0886	.0651	.0373
	.6898	.4269	.3092	.2382	.1890	.1518	.1213	.0939	.0646
	.1668	.1462	.1291	.1134	.0982	.0829	.0668	.0490	.0281
	.4728	.2932	.2138	.1661	.1329	.1076	.0866	.0676	.0472
	-.0411	.1558	.3251	.4782	.6184	.7448	.8536	.9377	.9877
6.3	.0065	.0274	.0433	.0521	.0551	.0532	.0470	.0366	.0215
	.2743	.2991	.2145	.1514	.1132	.0920	.0795	.0684	.0509
	.1508	.1336	.1139	.0956	.0789	.0633	.0483	.0333	.0177
	.2805	.2955	.2444	.1943	.1540	.1217	.0944	.0696	.0441
	-.3589	-.5758	-.4350	-.1720	.1702	.5135	.7725	.9222	.9870

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 8, Theta Zero: .60, Mastery Score: 5

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	.0010	.0021	.0040	.0071	.0115	.0167	.0212	.0226	.0173
	.0125	.0213	.0338	.0496	.0635	.0676	.0591	.0441	.0343
	.0000	.0000	.0000	.0000	.0003	.0011	.0030	.0057	.0068
	.0000	.0000	.0000	.0009	.0048	.0128	.0209	.0220	.0144
	.9414	.9460	.9595	.9664	.9649	.9524	.9132	.8058	.8214
1.6	.0147	.0204	.0278	.0364	.0449	.0509	.0523	.0471	.0322
	.0812	.1000	.1184	.1254	.1157	.0956	.0744	.0591	.0476
	.0000	.0000	.0001	.0006	.0023	.0057	.0102	.0141	.0137
	.0000	.0001	.0020	.0104	.0237	.0346	.0370	.0294	.0190
	.8648	.8960	.9116	.9086	.8835	.8251	.7193	.6294	.8564
2.4	.0679	.0793	.0904	.0977	.0995	.0954	.0855	.0690	.0435
	.2049	.2156	.1957	1595	.1286	.1069	.0905	.0752	.0568
	.0000	.0001	.0013	.0049	.0105	.0169	.0223	.0245	.0200
	.0000	.0048	.0261	.0487	.0592	.0567	.0455	.0316	.0230
	.8056	.8046	.7712	.6731	.5335	.4122	.3890	.5638	.9084
3.2	.1856	.1918	.1861	.1725	.1545	.1337	.1104	.0837	.0506
	.3176	.2462	.2176	.1989	.1736	.1453	.1170	.0898	.0623
	.0001	.0038	.0125	.0223	.0306	.0362	.0381	.0352	.0252
	.0110	.0852	.1150	.1058	.0844	.0630	.0458	.0344	.0275
	.5955	.2680	-.0841	-.1817	-.1183	.0611	.3496	.6996	.9475
4.0	.3421	.2899	.2460	.2094	.1773	.1477	.1187	.0883	.0528
	.7323	.5895	.4192	.3056	.2282	.1723	.1296	.0949	.0638
	.0155	.0404	.0544	.0609	.0623	.0598	.0537	.0439	.0284
	.4134	.2689	.1619	.1079	.0800	.0636	.0521	.0422	.0320
	-.8674	-.7988	-.6005	-.2989	.0480	.3711	.6376	.8402	.9670
4.8	.2400	.2137	.1915	.1709	.1506	.1298	.1075	.0821	.0501
	.6769	.4209	.3051	.2342	.1843	.1460	.1147	.0875	.0612
	.1553	.1352	.1191	.1047	.0912	.0777	.0636	.0481	.0291
	.4582	.2757	.1950	.1466	.1133	.0883	.0682	.0512	.0354
	-.1192	.0505	.2011	.3442	.4842	.6224	.7572	.8811	.9728
5.6	.0197	.0551	.0768	.0879	.0914	.0889	.0807	.0662	.0426
	.5635	.4054	.2613	.1818	.1367	.1084	.0884	.0721	.0549
	.1877	.1593	.1343	.1134	.0952	.0785	.0624	.0458	.0267
	.4042	.3420	.2453	.1784	.1325	.0996	.0747	.0546	.0364
	-.7575	-.7159	-.5177	-.2402	.0705	.3715	.6355	.8427	.9687
6.4	.0001	.0034	.0129	.0250	.0362	.0443	.0476	.0443	.0311
	.0072	.0879	.1417	.1455	.1258	.0995	.0748	.0570	.0455
	.0660	.0740	.0761	.0731	.0668	.0584	.0482	.0361	.0211
	.2016	.1581	.1183	.0997	.0867	.0742	.0616	.0487	.0340
	.7780	.6552	.3804	.1795	.1342	.2211	.4324	.7336	.9545
7.2	.0000	.0000	.0006	.0029	.0073	.0131	.0183	.0207	.0165
	.0000	.0016	.0159	.0411	.0613	.0675	.0593	.0434	.0326
	.0082	.0124	.0175	.0223	.0255	.0263	.0245	.0199	.0121
	.0602	.0773	.0847	.0766	.0618	.0484	.0399	.0343	.0264
	.9041	.9290	.9261	.8964	.8271	.7067	.5894	.6585	.9292

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 8, Theta Zero: .60, Mastery Score: 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	.0001	.0003	.0007	.0016	.0031	.0052	.0072	.0081	.0065
	.0016	.0039	.0081	.0148	.0220	.0255	.0232	.0171	.0128
	.0000	.0000	.0000	.0000	.0005	.0022	.0062	.0122	.0156
	.0000	.0000	.0001	.0016	.0093	.0258	.0447	.0504	.0342
	.9623	.9610	.9789	.9830	.9817	.9732	.9452	.8521	.8060
1.6	.0022	.0039	.0065	.0100	.0139	.0171	.0186	.0174	.0122
	.0196	.0294	.0409	.0480	.0466	.0390	.0296	.0224	.0179
	.0000	.0000	.0001	.0011	.0045	.0114	.0212	.0309	.0317
	.0000	.0001	.0036	.0197	.0468	.0720	.0819	.0691	.0445
	.9195	.9456	.9563	.9539	.9368	.8944	.8024	.8736	.8281
2.4	.0155	.0210	.0271	.0319	.0344	.0343	.0316	.0260	.0167
	.0788	.0923	.0876	.0705	.0539	.0422	.0344	.0283	.0216
	.0000	.0002	.0024	.0093	.0206	.0344	.0473	.0545	.0468
	.0001	.0086	.0488	.0953	.1218	.1235	.1050	.0752	.0528
	.9122	.9060	.8876	.8241	.7099	.5670	.4693	.5460	.8827
3.2	.0590	.0661	.0671	.0638	.0581	.0509	.0423	.0323	.0196
	.1767	.1237	.0906	.0776	.0670	.0561	.0452	.0347	.0241
	.0002	.0069	.0235	.0434	.0617	.0757	.0828	.0797	.0594
	.0194	.1575	.2246	.2191	.1854	.1450	.1075	.0788	.0627
	.8087	.6220	.2582	.0398	.0057	.0988	.3092	.6364	.9326
4.0	.1415	.1195	.1005	.0848	.0712	.0588	.0469	.0347	.0206
	.3120	.2597	.1831	.1313	.0962	.0713	.0527	.0379	.0251
	.0277	.0756	.1060	.1230	.1302	.1292	.1201	.1014	.0677
	.7585	.5392	.3504	.2435	.1808	.1413	.1140	.0933	.0737
	-.7690	-.7415	-.5728	-.3310	-.0420	.2639	.5563	.8026	.9606
4.8	.1141	.0977	.0846	.0736	.0635	.0537	.0437	.0328	.0197
	.3648	.2132	.1474	.1088	.0827	.0635	.0484	.0360	.0245
	.2942	.2688	.2459	.2233	.2000	.1750	.1469	.1136	.0702
	.7609	.4845	.3589	.2815	.2264	.1834	.1474	.1153	.0833
	-.1750	-.0072	.1470	.2972	.4464	.5949	.7402	.8735	.9713
5.6	.0099	.0265	.0356	.0395	.0400	.0379	.0336	.0269	.0169
	.2757	.1832	.1109	.0749	.0564	.0452	.0370	.0299	.0222
	.4197	.3591	.3071	.2629	.2235	.1865	.1499	.1110	.0654
	.8434	.6973	.5082	.3789	.2892	.2233	.1718	.1288	.0879
	-.8705	-.7914	-.5732	-.2519	.0980	.4108	.6645	.8550	.9705
6.4	.0000	.0017	.0062	.0116	.0163	.0193	.0202	.0182	.0125
	.0037	.0427	.0657	.0642	.0530	.0404	.0300	.0232	.0185
	.2114	.2152	.2083	.1931	.1726	.1486	.1213	.0901	.0523
	.3364	.2919	.2638	.2428	.2160	.1854	.1537	.1212	.0841
	.4526	.2347	-.0320	-.1131	-.0492	.1308	.4208	.7532	.9593
7.2	.0000	.0000	.0003	.0014	.0034	.0058	.0079	.0086	.0067
	.0000	.0008	.0077	.0190	.0272	.0286	.0241	.0173	.0133
	.0466	.0559	.0653	.0725	.0755	.0733	.0653	.0513	.0304
	.1782	.1934	.1917	.1637	.1402	.1179	.1030	.0895	.0672
	.9990	.8148	.8081	.7535	.6469	.5167	.4633	.6382	.9361

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 8, Theta Zero: .60, Mastery Score: 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	.0000	.0000	.0001	.0002	.0006	.0011	.0017	.0020	.0016
	.0001	.0004	.0013	.0029	.0050	.0063	.0060	.0044	.0032
	.0000	.0000	.0000	.0001	.0007	.0033	.0098	.0204	.0285
	.0000	.0000	.0001	.0023	.0136	.0397	.0736	.0913	.0669
	.9741	.9732	.9882	.9907	.9894	.9827	.9601	.8730	.7600
1.6	.0002	.0005	.0010	.0019	.0029	.0039	.0044	.0043	.0031
	.0027	.0052	.0089	.0116	.0120	.0103	.0077	.0057	.0045
	.0000	.0000	.0001	.0015	.0065	.0173	.0340	.0526	.0587
	.0000	.0001	.0050	.0283	.0703	.1147	.1407	.1301	.0855
	.9466	.9694	.9766	.9742	.9613	.9280	.8458	.6873	.7596
2.4	.0022	.0036	.0054	.0069	.0079	.0082	.0078	.0065	.0042
	.0172	.0235	.0241	.0197	.0148	.0111	.0087	.0071	.0055
	.0000	.0002	.0034	.0134	.0308	.0536	.0776	.0950	.0882
	.0001	.0118	.0692	.1413	.1909	.2071	.1903	.1454	.0984
	.9254	.9506	.9391	.8964	.8072	.6652	.5156	.4915	.8157
3.2	.0117	.0146	.0157	.0155	.0144	.0127	.0107	.0082	.0050
	.0570	.0400	.0257	.0201	.0169	.0141	.0114	.0088	.0062
	.0003	.0096	.0335	.0641	.0948	.1215	.1397	.1425	.1138
	.0261	.2213	.3321	.3441	.3115	.2609	.2034	.1482	.1142
	.9018	.8029	.5244	.2430	.1140	.1160	.2325	.5049	.8883
4.0	.0369	.0314	.0263	.0221	.0185	.0153	.0121	.0090	.0053
	.0801	.0702	.0497	.0354	.0257	.0189	.0138	.0099	.0065
	.0379	.1074	.1565	.1887	.2081	.2157	.2102	.1867	.1321
	1.0573	.8133	.5720	.4234	.3259	.2557	.2026	.1632	.1350
	-.6144	-.6306	-.5491	-.3689	-.1564	.0904	.3788	.6936	.9379
4.8	.0342	.0281	.0238	.0203	.0172	.0143	.0115	.0086	.0051
	.1249	.0682	.0450	.0320	.0236	.0176	.0132	.0096	.0064
	.4216	.4039	.3850	.3637	.3385	.3077	.2688	.2165	.1396
	1.0284	.6812	.5207	.4203	.3480	.2913	.2438	.2014	.1569
	-.3072	-.1760	-.0400	.1080	.2713	.4503	.6399	.8234	.9606
5.6	.0031	.0080	.0105	.0113	.0112	.0104	.0090	.0071	.0044
	.0851	.0523	.0300	.0200	.0152	.0123	.0101	.0081	.0059
	.6969	.6106	.5362	.4714	.4112	.3519	.2898	.2199	.1327
	1.1342	.9837	.7439	.5779	.4617	.3742	.3023	.2390	.1724
	-.9456	-.8723	-.6626	-.3226	.0468	.3688	.6311	.8352	.9653
6.4	.0000	.0005	.0019	.0034	.0047	.0054	.0055	.0049	.0033
	.0012	.0131	.0193	.0180	.0143	.0107	.0079	.0062	.0049
	.4388	.4743	.4457	.4063	.3600	.3085	.2515	.1868	.1086
	.4306	.4538	.4752	.4590	.4171	.3640	.3066	.2452	.1726
	-.1452	-.3243	-.4345	-.3971	-.2513	.0049	.3680	.7411	.9566
7.2	.0000	.0000	.0001	.0004	.0010	.0017	.0022	.0023	.0018
	.0000	.0002	.0023	.0056	.0077	.0078	.0064	.0045	.0035
	.1923	.1972	.2005	.1992	.1908	.1739	.1479	.1122	.0647
	.3309	.3331	.3233	.2999	.2749	.2531	.2303	.1983	.1444
	.9990	.4131	.4157	.3476	.2511	.1988	.2770	.5860	.9330

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 8, Theta Zero: .70, Mastery Score. 6

Test KR21= Mean	.100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	.0001	0003	.0007	.0016	.0034	.0065	.0106	.0142	.0132
	.0016	.0039	.0082	.0158	.0275	.0399	.0453	.0378	.0260
	.0000	.0000	.0000	.0000	.0000	.0003	.0014	.0040	.0066
	.0001	.0000	.0000	.0001	.0010	.0052	.0141	.0219	.0169
	.9658	.9642	.9619	.9793	.9816	.9795	.9676	.9184	.7981
1.6	.0022	.0039	.0065	.0107	.0166	.0236	.0298	.0322	.0256
	.0196	.0294	.0430	.0596	.0730	.0756	.0651	.0476	.0358
	.0000	.0000	.0000	.0001	.0005	.0021	.0056	.0108	.0138
	.0000	.0000	.0001	.0014	.0075	.0192	.0311	.0333	.0213
	.9205	.9269	.9483	.9561	.9558	.9435	.9053	.7990	.7740
2.4	.0155	.0211	.0285	.0375	.0466	.0534	.0557	.0512	.0361
	.0788	.0968	.1150	.1222	.1127	.0926	.0710	.0548	.0441
	.0000	.0000	.0001	.0007	.0030	.0077	.0143	.0206	.0211
	.0000	.0001	.0024	.0128	.0297	.0442	.0483	.0388	.0239
	.9497	.8958	.9108	.9088	.8840	.8254	.7158	.5980	.8129
3.2	.0591	.0702	.0815	.0898	.0931	.0908	.0825	.0676	.0436
	.1853	.1989	.1837	.1493	.1181	.0964	.0809	.0671	.0512
	.0000	.0001	.0014	.0055	.0124	.0208	.0284	.0324	.0276
	.0000	.0046	.0283	.0562	.0710	.0699	.0569	.0389	.0274
	.9242	.8289	.8037	.7149	.5710	.4221	.3564	.4961	.8876
4.0	.1574	.1664	.1647	.1549	.1403	.1224	.1018	.0777	.0473
	.3006	.2283	.1883	.1714	.1517	.1286	.1044	.0806	.0563
	.0001	.0037	.0132	.0249	.0355	.0433	.0468	.0444	.0325
	.0087	.0874	.1294	.1249	.1021	.0764	.0545	.0400	.0328
	.6774	.4312	.0392	-.1386	-.1322	.0056	.2768	.6571	.9439
4.8	.3037	.2600	.2211	.1881	.1590	.1321	.1059	.0785	.0467
	.5887	.5190	.3782	.2779	.2080	.1571	.1181	.0865	.0581
	.0148	.0431	.0607	.0698	.0728	.0710	.0647	.0534	.0348
	.4356	.3147	.1947	.1283	.0927	.0727	.0599	.0497	.0391
	-.8272	-.7993	-.6354	-.3670	-.0238	.3267	.6243	.8429	.9697
5.6	.2163	.1902	.1688	.1493	.1306	.1116	.0915	.0690	.0413
	.6392	.3932	.2827	.2157	.1690	.1336	.1050	.0802	.0557
	.1765	.1552	.1377	.1217	.1062	.0907	.0743	.0559	.0334
	.4995	.3058	.2197	.1678	.1317	.1043	.0821	.0628	.0440
	-.1046	.0726	.2289	.3758	.5173	.6543	.7840	.8982	.9777
6.4	.0120	.0395	0578	.0675	.0708	.0689	.0621	.0502	.0314
	.4015	.3366	.2249	.1572	.1178	.0937	.0773	.0641	.0486
	.1993	.1725	.1466	.1240	.1040	.0854	.0673	.0486	.0276
	.3767	.3574	.2705	.2031	.1546	.1187	.0907	.0673	.0447
	-.6497	-.6773	-.4938	-.2202	.0976	.4098	.6764	.8702	.9757
7.2	.0000	.0011	.0056	.0126	.0198	.0252	.0276	.0255	.0173
	.0011	.0362	.0815	.0982	.0921	.0752	.0570	.0443	.0359
	.0466	.0548	.0593	.0592	.0554	.0489	.0402	.0296	.0166
	.1774	.1624	.1247	.1003	.0871	.0773	.067[illegible]	.0547	.0377
	.8038	.7707	.6163	.4106	.2788	.2913	.4740	.7722	.9653

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 8, Theta Zero: .70, Mastery Score: 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	.0000	.0000	.0001	.0003	.0007	.0016	.0030	.0044	.0043
	.0001	.0004	.0013	.0033	.0072	.0120	.0149	.0129	.0086
	.0000	.0000	.0000	.0000	.0001	.0006	.0029	.0085	.0154
	.0000	.0000	.0000	.0001	.0019	.0102	.0294	.0500	.0428
	.9769	.9754	.9755	.9887	.9902	.9884	.9796	.9401	.7917
1.6	.0002	.0005	.0010	.0021	.0040	.0064	.0089	.0102	.0085
	.0027	.0052	.0097	.0162	.0226	.0252	.0226	.0164	.0118
	.0000	.0000	.0000	.0001	.0009	.0040	.0115	.0237	.0328
	.0000	.0000	.0001	.0026	.0143	.0388	.0674	.0788	.0540
	.9484	.9507	.9719	.9770	.9761	.9667	.9373	.8411	.7384
2.4	.0022	.0036	.0059	.0091	.0127	.0158	.0174	.0166	.0121
	.0172	.0253	.0350	.0411	.0399	.0332	.0249	.0183	.0145
	.0000	.0000	.0001	.0013	.0058	[illegible]	.0300	.0460	.0508
	.0000	.0001	.0043	.0240	.0585	[illegible]	.1091	.0955	.0592
	.9242	.9432	.9542	.9524	.9350	.8913	.7944	.6358	.7526
3.2	.0118	.0162	.0213	.0256	.0282	.0286	.0268	.0224	.0147
	.0604	.0729	.0712	.0579	.0439	.0337	.0270	.0221	.0171
	.0000	.0002	.0025	.0104	.0243	.0425	.0612	.0741	.0676
	.0000	.0082	.0522	.1092	.1465	.1547	.1358	.0979	.0653
	.8985	.9145	.9010	.8447	.7329	.5729	.4322	.4524	.8341
4.0	.0427	.0494	.0515	.0499	.0461	.0407	.0342	.0263	.0161
	.1417	.1031	.0711	.0588	.0510	.0432	.0352	.0272	.0191
	.0001	.0066	.0246	.0482	.0719	.0916	.1040	.1040	.0809
	.0151	.1596	.2510	.2592	.2278	.1822	.1349	.0958	.0770
	.8432	.7172	.3863	.1022	-.0016	.0350	.2032	.5407	.9161
4.8	.1078	.0926	.0783	.0662	.0556	.0459	.0366	.0270	.0160
	.2095	.1940	.1413	.1026	.0757	.0563	.0417	.0301	.0200
	.0262	.0800	.1179	.1416	.1541	.1569	.1494	.1290	.0881
	.7887	.6274	.4255	.2990	.2197	.1682	.1341	.1116	.0934
	-.6800	-.7355	-.6071	-.4091	-.1523	.1534	.4820	.7768	.9593
5.6	.0892	.0752	.0648	.0560	.0479	.0402	.0325	.0241	.0143
	.2997	.1725	.1182	.0867	.0657	.0504	.0385	.0287	.0194
	.3332	.3101	.2879	.2649	.2401	.2122	.1797	.1398	.0862
	.8203	.5295	.3978	.3173	.2604	.2163	.1793	.1454	.1089
	-.2337	-.0644	.0983	.2615	.4259	.5891	.7453	.8817	.9747
6.4	.0052	.0165	.0233	.0264	.0270	.0256	.0226	.0178	.0109
	.1715	.1332	.0836	.0565	.0423	.0340	.0282	.0232	.0172
	.4701	.4098	.3538	.3043	.2592	.2160	.1725	.1262	.0726
	.8019	.7454	.5746	.4448	.3502	.2782	.2199	.1684	.1155
	-.8396	-.7863	-.5847	-.2709	.0909	.4241	.6891	.8746	.9760
7.2	.0000	.0005	.0023	.0051	.0078	.0096	.0102	.0092	.0061
	.0005	.0154	.0331	.0380	.0341	.0268	.0200	.0158	.0128
	.1923	.1955	.1915	.1793	.1610	.1381	.1112	.0804	.0445
	.3303	.3107	.2820	.2626	.2424	.2173	.1868	.1498	.1017
	.4109	.3041	.1101	-.0118	-.0051	.1379	.4302	.7783	.9674

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 8, Theta Zero: .80, Mastery Score: 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.8	.0000	.0000	.0001	.0003	.0007	.0019	.0043	.0079	.0098
	.0001	.0004	.0013	.0034	.0080	.0168	.0276	.0313	.0212
	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0023	.0061
	.0000	.0000	.0000	.0000	.0001	.0012	.0067	.0184	.0215
	.9684	.9738	.9681	.9825	.9881	.9889	.9858	.9683	.8498
1.6	.0002	.0005	.0010	.0022	.0044	.0083	.0140	.0197	.0199
	.0027	.0052	.0097	.0174	.0293	.0427	.0498	.0427	.0277
	.0000	.0000	.0000	.0000	.0000	.0004	.0023	.0071	.0135
	.0000	.0000	.0000	.0001	.0011	.0064	.0190	.0330	.0277
	.9466	.9597	.9939	.9717	.9761	.9759	.9665	.9249	.7633
2.4	.0022	.0036	.0060	.0098	.0154	.0728	.0303	.0345	.0293
	.0172	.0253	.0369	.0521	.0665	.0720	.0641	.0465	.0331
	.0000	.0000	.0000	.0000	.0004	.0022	.0070	.0151	.0218
	.0000	.0000	.0000	.0011	.0072	.0212	.0385	.0455	.0297
	.9238	.9263	.9446	.9544	.9565	.9476	.9159	.8150	.6960
3.2	.0118	.0162	.0224	.0305	.0396	.0477	.0522	.0501	.0370
	.0604	.0761	.0940	.1060	.1033	.0873	.0662	.0489	.0395
	.0000	.0000	.0000	.0005	.0028	.0081	.0168	.0267	.0302
	.0000	.0000	.0015	.0105	.0295	.0500	.0603	.0518	.0300
	.9990	.9036	.9187	.9219	.9059	.8597	.7559	.5838	.7253
4.0	.0428	.0523	.0631	.0729	.0790	.0800	.0751	.0634	.0420
	.1468	.1647	.1638	.1391	.1089	.0852	.0697	.0584	.0463
	.0000	.0000	.0009	.0046	.0121	.0224	.0335	.0412	.0377
	.0000	.0023	.0216	.0536	.0779	.0843	.0729	.0498	.0328
	.9990	.8593	.8543	.8001	.6840	.5117	.3566	.3889	.8454
4.3	.1190	.1308	.1351	.1314	.1221	.1086	.0916	.0706	.0433
	.2669	.2187	.1620	.1387	.1248	.1093	.0916	.0726	.0520
	.0000	.0024	.0111	.0241	.0377	.0492	.0562	.0559	.0427
	.0033	.0677	.1290	.1420	.1252	.0968	.0678	.0469	.0404
	.7547	.6547	.3362	.0335	-.0906	-.0553	.1479	.5656	.9393
5.6	.2562	.2262	.1945	.1661	.1405	.1166	.0930	.0685	.0401
	.3796	.4150	.3257	.2474	.1885	.1441	.1095	.0808	.0544
	.0109	.0407	.0634	.0772	.0839	.0843	.0787	.0661	.0434
	.3914	.3634	.2430	.1609	.1115	.0837	.0690	.0603	.0510
	-.6566	-.7783	-.6745	-.4724	-.1639	.2228	.5917	.8498	.9752
6.4	.1902	.1648	.1445	.1264	.1093	.0922	.0745	.0550	.0319
	.5973	.3640	.2604	.1983	.1556	.1236	.0978	.0751	.0514
	.2027	.1805	.1613	.1434	.1255	.1071	.0874	.0651	.0380
	.5447	.3416	.2515	.1972	.1592	.1300	.1055	.0831	.0589
	-.0954	.1005	.2723	.4303	.5771	.7119	.8307	.9256	.9847
7.2	.0038	.0196	.0331	.0412	.0445	.0437	.0391	.0309	.0184
	.1806	.2345	.1769	.1266	.0939	.0746	.0632	.0541	.0406
	.1877	.1694	.1464	.1242	.1035	.0838	.0646	.0450	.0242
	.3106	.3355	.2875	.2331	.1872	.1494	.1171	.0873	.0561
	-.2714	-.5545	-.4456	-.2119	.1073	.4508	.7323	.9060	.9840

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 9, Theta Zero: .60, Mastery Score: 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	.0002	.0005	.0013	.0027	.0051	.0086	.0123	.0145	.0121
	.0033	.0070	.0134	.0234	.0350	.0427	.0413	.0314	.0227
	.0000	.0000	.0000	.0000	.0002	.0010	.0033	.0070	.0094
	.0000	.0000	.0000	.0005	.0040	.0130	.0250	.0303	.0208
	.9553	.9576	.9732	.9777	.9782	.9721	.9507	.8781	.8140
1.8	.0050	.0079	.0121	.0178	.0243	.0300	.0332	.0318	.0229
	.0365	.0503	.0666	.0787	.0794	.0692	.0538	.0403	.0316
	.0000	.0000	.0000	.0004	.0022	.0061	.0121	.0184	.0194
	.0000	.0000	.0013	.0089	.0245	.0411	.0491	.0425	.0268
	.9172	.9239	.9409	.9426	.9301	.8959	.8213	.7064	.8260
2.7	.0321	.0405	.0498	.0576	.0620	.0622	.0578	.0482	.0315
	.1291	.1458	.1429	.1205	.0954	.0756	.0615	.0503	.0384
	.0000	.0001	.0011	.0048	.0115	.0202	.0285	.0333	.0289
	.0000	.0032	.0240	.0533	.0726	.0759	.0652	.0465	.0317
	.9990	.8697	.8600	.8060	.7067	.5814	.4928	.5599	.8795
3.6	.1112	.1205	.1217	.1161	.1061	.0933	.0781	.0600	.0369
	.2556	.1980	.1534	.1337	.1169	.0983	.0802	.0618	.0429
	.0001	.0035	.0133	.0259	.0379	.0469	.0515	.0494	.0368
	.0073	.0865	.1371	.1384	.1177	.0913	.0670	.0487	.0376
	.7300	.5630	.2302	.0269	-.0013	.0998	.3191	.6438	.9309
4.5	.2466	.2115	.1798	.1529	.1293	.1076	.0864	.0645	.0388
	.4724	.4206	.3063	.2241	.1667	.1250	.0931	.0673	.0446
	.0162	.0482	.0687	.0797	.0838	.0823	.0757	.0633	.0420
	.4812	.3605	.2305	.1565	.1148	.0896	.0722	.0581	.0442
	-.7792	-.7624	-.5941	-.3363	-.0234	.2951	.5797	.8087	.9589
5.4	.1872	.1640	.1452	.1284	.1122	.0961	.0792	.0603	.0370
	.5497	.3338	.2375	.1794	.1390	.1085	.0839	.0629	.0431
	.2077	.1843	.1648	.1468	.1292	.1113	.0922	.0706	.0435
	.5763	.3552	.2559	.1956	.1534	.1210	.0945	.0715	.0496
	-.1406	.0285	.1791	.3226	.4631	.6023	.7392	.8679	.9680
6.3	.0124	.0385	.0552	.0639	.0668	.0651	.0592	.0486	.0315
	.3879	.3026	.1962	.1352	.1006	.0792	.0640	.0515	.0386
	.2616	.2256	.1922	.1636	.1383	.1148	.0918	.0679	.0402
	.4885	.4445	.3291	.2439	.1835	.1392	.1050	.0771	.0515
	-.7476	-.7308	-.5454	-.2716	.0433	.3497	.6172	.8294	.9641
7.2	.0000	.0018	.0079	.0165	.0248	.0311	.0340	.0321	.0230
	.0027	.0508	.0940	.1024	.0911	.0729	.0546	.0408	.0318
	.0964	.1057	.1086	.1049	.0967	.0852	.0710	.0537	.0319
	.2452	.2078	.1621	.1375	.1201	.1033	.0862	.0685	.0482
	.7162	.6264	.3892	.1985	.1440	.2166	.4119	.7090	.9475
8.1	.0000	.0000	.0003	.0016	.0044	.0085	.0125	.0147	.0121
	.0000	.0005	.0078	.0239	.0396	.0466	.0427	.0316	.0228
	.0125	.0179	.0244	.0309	.0356	.0374	.0355	.0294	.0183
	.0796	.0985	.1099	.1042	.0874	.0696	.0565	.0480	.0374
	-.6261	.9112	.9137	.8924	.8375	.7377	.6223	.6506	.9173

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 9, Theta Zero: .60, Mastery Score: 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	.0000	.0001	.0002	.0005	.0013	.0025	.0039	.0049	.0043
	.0004	.0011	.0028	.0062	.0109	.0146	.0149	.0115	.0080
	.0000	.0000	.0000	.0000	.0003	.0018	.0058	.0130	.0188
	.0000	.0000	.0000	.0009	.0066	.0225	.0460	.0599	.0437
	.9695	.9708	.9840	.9878	.9878	.9830	.9666	.9043	.8003
1.8	.0007	.0014	.0026	.0045	.0070	.0094	.0110	.0110	.0081
	.0072	.0124	.0198	.0265	.0286	.0257	.0200	.0145	.0111
	.0000	.0000	.0001	.0007	.0036	.0105	.0218	.0348	.0391
	.0000	.0000	.0020	.0146	.0418	.0736	.0935	.0868	.0557
	.9353	.9542	.9681	.9687	.9592	.9332	.8704	.7433	.7925
2.7	.0065	.0096	.0136	.0173	.0199	.0208	.0199	.0170	.0113
	.0408	.0527	.0557	.0480	.0373	.0284	.0222	.0178	.0136
	.0000	.0001	.0018	.0079	.0196	.0355	.0523	.0642	.0589
	.0000	.0050	.0388	.0899	.1286	.1421	.1297	.0969	.0642
	.9415	.9315	.9258	.8899	.8146	.6952	.5672	.5502	.8424
3.6	.0315	.0376	.0401	.0396	.0369	.0329	.0278	.0216	.0133
	.1197	.0919	.0627	.0496	.0421	.0354	.0287	.0222	.0155
	.0001	.0055	.0217	.0437	.0661	.0850	.0970	.0973	.0760
	.0112	.1387	.2308	.2460	.2215	.1814	.1378	.0991	.0752
	.8607	.7730	.5173	.2557	.1360	.1514	.2891	.5700	.9067
4.5	.0920	.0793	.0672	.0569	.0478	.0396	.0317	.0235	.0141
	.1768	.1644	.1199	.0870	.0639	.0473	.0348	.0249	.0163
	.0253	.0783	.1158	.1392	.1516	.1542	.1470	.1273	.0878
	.7662	.6201	.4264	.3049	.2282	.1768	.1402	.1127	.0890
	-.6164	-.6950	-.5579	-.3533	-.1026	.1795	.4748	.7513	.9474
5.4	.0807	.0680	.0586	.0507	.0435	.0366	.0298	.0224	.0136
	.2680	.1531	.1040	.0757	.0569	.0432	.0325	.0238	.0160
	.3418	.3175	.2946	.2711	.2459	.2179	.1855	.1458	.0920
	.8577	.5539	.4148	.3287	.2670	.2164	.1772	.1399	.1019
	-.2255	-.0679	.0821	.2324	.3858	.5423	.6993	.8481	.9635
6.3	.0056	.0168	.0234	.0263	.0268	.0255	.0227	.0183	.0116
	.1726	.1254	.0767	.0514	.0383	.0304	.0247	.0197	.0144
	.5077	.4417	.3821	.3301	.2830	.2381	.1929	.1444	.0865
	.8679	.7774	.5858	.4457	.3454	.2699	.2098	.1584	.1089
	-.8715	-.8128	-.6168	-.3108	.0380	.3589	.6249	.8311	.9636
7.2	.0000	.0008	.0034	.0070	.0102	.0125	.0133	.0123	.0085
	.0012	.0225	.0399	.0416	.0355	.0275	.0203	.0153	.0120
	.2676	.2700	.2621	.2445	.2202	.1910	.1573	.1130	.0696
	.3643	.3346	.3085	.2875	.2591	.2250	.1883	.1496	.1049
	.3547	.1980	-.0314	-.1168	-.0674	.0945	.3694	.7119	.9495
8.1	.0000	.0000	.0001	.0007	.0019	.0035	.0050	.0057	.0045
	.0000	.0002	.0034	.0102	.0162	.0183	.0161	.0117	.0086
	.0623	.0722	.0824	.0908	.0949	.0930	.0841	.0671	.0406
	.2049	.2192	.2216	.2027	.1733	.1464	.1268	.1100	.0839
	-.3332	.7761	.7781	.7411	.6564	.5413	.4717	.6029	.9196

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 9, Theta Zero: .70, Mastery Score: 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	.0000	.0001	.0002	.0006	.0014	.0032	[illegible]	.0093	.0097
	.0004	.0011	.0028	.0065	.0136	.0233	.0307	.0287	.0189
	.0000	.0000	.0000	.0000	.0000	.0002	.0013	.0042	.0081
	.0000	.0000	.0000	.0000	.0006	.0041	.0138	.0258	.0228
	.9732	.9716	.9713	.9831	.9868	.9864	.9798	.9506	.8288
1.8	.0007	.0014	.0026	.0043	.0085	.0135	.0189	.0223	.0192
	.0072	.0124	.0206	.0324	.0451	.0522	.0490	.0367	.0256
	.0000	.0000	.0000	.0000	.0003	.0018	.0056	.0122	.0175
	.0000	.0000	.0000	[illegible]	.0058	.0182	.0344	.0421	.0288
	.9402	.9349	.9596	.9673	.9694	.9630	.9401	.8661	.7706
2.7	.0065	.0097	.0143	.0206	.0278	.0343	.0382	.0371	.0275
	.0403	.0545	.0707	.0827	.0828	.0717	.0553	.0406	.0314
	.0000	.0000	.0000	.0005	.0026	.0076	.0156	.0245	.0274
	.0000	.0000	.0014	.0101	.0284	.0484	.0589	.0518	.0314
	.9517	.9203	.9340	.9366	.9238	.8875	.8086	.6779	.7758
3.6	.0315	.0397	.0491	.0574	.0626	.0636	.0599	.0506	.0337
	.1231	.1404	.1400	.1192	.0939	.0735	.0593	.0483	.0370
	.0000	.0000	.0011	.0050	.0125	.0228	.0334	.0404	.0366
	.0000	.0028	.0237	.0563	.0800	.0859	.0749	.0531	.0348
	.8778	.8733	.8673	.8175	.7177	.5786	.4611	.4961	.8504
4.5	.1016	.1120	.1151	.1114	.1030	.0914	.0771	[illegible]	.0370
	.2441	.1946	.1451	.1237	.1090	.0932	.0764	[illegible]	.0414
	.0000	.0030	.0129	.0266	.0404	.0517	.0582	.0573	.0438
	.0051	.0806	.1410	.1496	.1306	.1022	.0741	.0525	.0412
	.7618	.6407	.3272	.0712	-.0114	.0492	.2472	.5920	.9246
5.4	.2299	.2001	[illegible]	.1458	.1233	.1026	.0824	.0613	[illegible]
	.3390	.3022	[illegible]	.2133	[illegible]	.1203	.0899	.0652	.0433
	.0142	.0474	.0707	.0843	.0904	.0903	.0841	.0710	.0475
	.4631	.3868	.2547	.1723	.1239	.0951	.0766	.0630	.0497
	-.7099	-.7606	-.6249	-.3979	-.0959	.2409	.5571	.8079	.9616
6.3	[illegible]	.1531	.1348	.1185	.1031	.0878	.0719	.0543	.0327
	.5347	.3228	.2287	.1723	1334	.1042	.0808	.0609	.0418
	.2199	.1966	.1767	.1579	.1394	.1202	.0995	.0759	.0462
	.5916	.3684	.2682	.2074	.1646	.1317	.1046	.0806	.0569
	-.1419	.0338	.1908	.3400	.4847	.6259	.7615	.8839	.9732
7.2	.0076	.0283	.0431	.0514	.0544	.0533	.0484	.0394	.0249
	.2774	.2600	.1775	.1239	.0920	.0724	.0590	.0483	.0364
	.2569	.2259	.1940	.1655	.1397	.1155	.0916	.0668	.0385
	.4220	.4275	.3345	.2359	.1970	.1524	.1172	.0874	.0585
	-.6166	-.6837	-.518[illegible]	-.2581	.0551	.3704	.6459	.8530	.9713
8.1	.0000	.0005	.0034	.0084	.0139	.0185	.0208	.0197	.0137
	.0003	.0197	.0534	.0703	.0693	.0583	.0443	.0335	.0268
	.0623	.0715	.0772	.0777	.0736	.0658	.0548	.0408	.0232
	2046	.1961	.1588	.1291	.1113	.0985	.0858	.0707	.0496
	.7613	.7473	.626[illegible]	.4490	.3159	.3020	.4491	.7395	.9583

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 9, Theta Zero: .60, Mastery Score: 8

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	.0000	.0000	.0000	.0001	.0002	.0005	.0008	.0011	.0010
	.0000	.0001	.0004	.0011	.0022	.0033	.0035	.0027	.0019
	.0000	.0000	.0000	.0000	.0004	.0024	.0082	.0196	.0310
	.0000	.0000	.0000	.0011	.0087	.0312	.0677	.0968	.0783
	.9778	.9791	.9905	.9929	.9925	.9885	.9746	.9161	.7621
1.8	.0001	.0002	.0004	.0008	.0014	.0020	.0024	.0025	.0019
	.0008	.0019	.0038	.0057	.0066	.0062	.0049	.0034	.0026
	.0000	.0000	.0001	.0009	.0048	.0144	.0314	.0534	.0655
	.0000	.0000	.0026	.0191	.0566	.1050	.1434	.1463	.0991
	.9557	.9724	.9818	.9815	.9738	.9524	.8961	.7567	.7222
2.7	.0008	.0015	.0025	.0035	.0042	.0046	.0045	.0039	.0027
	.0075	.0116	.0134	.0120	.0094	.0070	.0053	.0042	.0032
	.0000	.0001	.0022	.0104	.0265	.0500	.0773	.1008	.1004
	.0000	.0062	.0502	[illegible]	.1808	.2127	.2097	.1697	.1109
	.9340	.9614	.9574	.9317	.8730	.7638	.6111	.5129	.7588
3.6	.0056	.0076	.0086	.0088	.0084	.0076	.0065	.0051	.0032
	.0327	.0263	.0169	.0122	.0099	.0083	.0067	.0052	.0037
	.0001	.0070	.0282	.0586	.0917	.1230	.1474	.1570	.1320
	.0139	.1777	.3083	.3462	.3317	.2909	.2353	.1726	.1253
	.9225	.8743	.6980	.4398	.2524	.1862	.2351	.4402	.8437
4.5	.0217	.0190	.0161	.0136	.0115	.0095	.0076	.0056	.0034
	.0415	.0399	.0294	.0213	.0156	.0115	.0084	.0060	.0039
	.0318	.1014	.1551	.1931	.2185	[illegible]	.2318	.2120	.1556
	.9750	.8416	.6204	.4726	.3708	.2938	.2315	[illegible]	.1476
	-.3596	-.6190	-.5216	-.3695	-.1882	.0258	.2898	[illegible]	.9134
5.4	.0220	.0179	.0151	.0128	.0103	.0090	.0073	.0054	.0032
	.0824	.0441	.0287	.0202	.0148	.0110	.0081	.0058	.0039
	.4467	.4327	.4169	.3982	.3750	[illegible]	.3065	.2517	.1666
	1.0888	.7287	.5614	.4557	.3788	.3179	.2665	.2207	.1737
	-.3446	-.2303	-.1100	.0268	.1839	.3649	.5678	.7776	.9471
6.3	.0016	.0046	.0063	.0069	.0069	.0064	.0056	.0045	.0028
	.0485	.0328	.0191	.0126	.0095	.0076	.0062	.0049	.0035
	.7620	.6781	.6020	.5341	.4702	.4063	.3383	.2602	.1602
	1.0840	.9880	.7677	.6081	.4932	.4051	.3319	.2652	.1939
	-.9503	-.8953	-.7189	-.4094	-.0472	.2859	.5634	.7985	.9550
7.2	.0000	.0002	.0009	.0019	.0027	.0032	.0033	.0030	.0021
	.0004	.0063	.0107	.0107	.0089	.0067	.0049	.0038	.0029
	.5532	.5374	.5086	.4677	.4182	.3617	.2977	.2237	.1321
	.4318	.4563	.4851	.4802	.4463	.3967	.3392	.2750	.1964
	-.2196	-.3339	-.4381	-.4165	-.2916	-.0601	.2874	.6845	.9437
8.1	.0000	.0000	.0000	.0002	.0005	.0009	.0013	.0014	.0011
	.0000	.0001	.0009	.0027	.0042	.0046	.0039	.0028	.0021
	.2288	.2319	.2336	.2315	.2224	.2043	.1758	.1351	.0793
	.3515	.3529	.3472	.3288	.3047	.2812	.2565	.2231	.1654
	.9990	.3632	.3626	.3289	.2564	.2043	.2534	.5227	.9118

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 9, Theta Zero: .70, Mastery Score: 8

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	.0000	.0000	.0000	.0001	.0003	.0007	.0016	.0027	.0030
	.0000	.0001	.0004	.0012	.0032	.0064	.0093	.0091	.0059
	.0000	.0000	.0000	.0000	.0000	.0004	.0023	.0080	.0169
	.0000	.0000	.0000	.0000	.0010	.0071	.0253	.0517	.0515
	.9808	.9797	.9799	.9903	.9925	.9919	.9866	.9622	.8268
1.8	.0001	.0002	.0004	.0009	.0019	.0035	.0053	.0066	.0060
	.0008	.0019	.0040	.0078	.0126	.0159	.0157	.0119	.0080
	.0000	.0000	.0000	.0000	.0006	.0030	.0101	.0236	.0369
	.0000	.0000	.0000	.0014	.0098	.0323	.0654	.0875	.0656
	.9594	.9537	.9763	.9819	.9825	.9770	.9585	.8921	.7429
2.7	.0008	.0015	.0027	.0046	.0071	.0095	.0112	.0113	.0086
	.0075	.0122	.0188	.0247	.0265	.0236	.0182	.0129	.0098
	.0000	.0000	.0001	.0009	.0044	.0133	.0289	.0484	.0586
	.0000	.0000	.0022	.0168	.0493	.0890	.1166	.1123	.0709
	.9336	.9482	.9637	.9650	.9550	.9269	.8589	.7126	.7127
3.6	.0056	.0083	.0117	.0152	.0176	.0187	.0182	.0157	.0107
	.0339	.0442	.0478	.0419	.0326	.0245	.0189	.0150	.0116
	.0000	.0001	.0017	.0083	.0217	.0412	.0634	.0817	.0797
	.0000	.0044	.0388	.0964	.1448	.1663	.1568	.1192	.0753
	.8897	.9319	.9289	.8961	.8218	.6934	.5360	.4696	.7797
4.5	.0249	.0303	.0331	.0333	.0315	.0284	.0242	.0189	.0118
	.0982	.0798	.0537	.0410	.0345	.0293	.0241	.0187	.0131
	.0001	.0049	.0213	.0457	.0722	.0965	.1143	.1191	.0971
	.0079	.1307	.2409	.2719	.2543	.2135	.1633	.1148	.0864
	.8741	.8147	.5970	.3131	.1364	.0995	.1951	.4688	.8826
5.4	.0745	.0655	.0560	.0476	.0401	.0332	.0266	.0197	.0118
	.1261	.1286	.0977	.0722	.0536	.0399	.0295	.0212	.0139
	.0224	.0781	.1215	.1510	.1690	.1762	.1719	.1523	.1074
	.7446	.6761	.4848	.3518	.2624	.2004	.1566	.1267	.1054
	-.4627	-.6818	-.5847	-.4181	-.1988	.0728	.3914	.7163	.9449
6.3	.0667	.0558	.0478	.0411	.0351	.0295	.0238	.0177	.0106
	.2292	.1297	.0877	.0637	.0478	.0363	.0275	.0203	.0136
	.3639	.3478	.3266	.3040	.2786	.2492	.2139	.1690	.1066
	.8937	.5817	.4396	.3522	.2903	.2423	.2021	.1652	.1252
	-.2844	-.1283	.0270	.1881	.3561	.5289	.7000	.8556	.9675
.2	.0050	.0109	.0161	.0186	.0192	.0184	.0164	.0131	.0081
	.1090	.0954	.0615	.0415	.0307	.0243	.0199	.0162	.0120
	.5379	.4764	.4159	.3611	.3100	.2605	.2100	.1553	.0907
	.7749	.7808	.6231	.4919	.3927	.3154	.2517	.1946	.1350
	-.8299	-.7017	-.6252	-.3353	.0172	.3581	.6418	.8497	.9699
8.1	.0000	.0002	.0013	.0031	.0051	.0065	.0072	.0066	.0045
	.0001	.0077	.0200	.0253	.0240	.0194	.0145	.0111	.0089
	.2288	.2308	.2264	.2136	.1936	.1677	.1365	.0999	.0561
	.3513	.3396	.3141	.2935	.2724	.2463	.2140	.1740	.1201
	.3444	.2692	.1241	.0098	-.0005	1116	.5720	.7312	.9584

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items: 9, Theta Zero: .80, Mastery Score: 8

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
0.9	.0000	.0000	.0000	.0001	.0003	.0009	.0024	.0052	.0074
	.0000	.0001	.0004	.0012	.0035	.0089	.0176	.0236	.0172
	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0021	.0067
	0000	.0000	.0000	.0000	.0000	.0007	.0052	.0181	.0262
	.9768	.9747	.9917	.9739	.9900	.9916	.9902	.9798	.8945
1.8	.0001	.0002	.0004	.0009	.0021	.0045	.0087	.0138	.0155
	.0003	.0019	.0040	.0083	.0159	.0268	.0359	.0345	.0219
	.0000	.0000	.0000	.0000	.0000	.0003	.0018	.0068	.0153
	.0000	.0000	.0000	.0000	.0006	.0046	.0170	.0354	.0348
	.9479	.9300	.9246	.9769	.9811	.9821	.9768	.9510	.8105
2.7	.0008	.0015	.0027	.0049	.0086	.0141	.0205	.0254	.0233
	.0075	.0122	.0195	.0304	.0433	.0523	.0510	.0386	.0254
	.0000	.0000	.0000	.0000	.0003	.0017	.0063	.0156	.0254
	.0000	.0000	.0000	.0006	.050	.0181	.0387	.0525	.0378
	.9456	9403	.9534	.9632	.9668	.9625	.9428	.8741	.7105
3.6	.0056	.0083	.0123	.0180	.0252	.0326	.0379	.0385	.0300
	.0339	.0455	.0603	.0741	.0786	.0710	.0553	.0393	.0299
	.0000	.0000	.0000	.0003	[illegible]	.0072	.0167	.0292	.0361
	.0000	.0000	.0007	.0073	.0252	.0494	.0668	.0636	.0375
	.8903	.9517	.9338	.9401	.9325	.9038	.8321	.6765	.6831
4.5	.0249	.0319	.0405	.0493	.0562	.0594	.0579	.0505	.0345
	.1001	.1179	.1263	.1147	.0922	.0707	.0552	.0448	.0355
	.0000	.0000	.0006	.0037	.0110	.0224	.0359	.0471	.0461
	.0000	.0011	.0158	.0478	.0790	.0941	.0880	.0637	.0387
	.9990	.8353	.8900	.8609	.7859	.6505	.4790	.4122	.7911
5.4	.0813	.0926	.0992	.0996	.0948	.0860	.0737	.0577	.0361
	.2148	.1910	.1396	.1101	.0956	.0837	.0708	.0565	.0406
	.0000	.0017	.0098	.0234	.0392	.0539	.0644	.0667	.0532
	.0015	.0546	.1258	.1538	.1459	.1192	.0864	.0581	.0466
	.8030	.7579	.5528	.2613	[illegible]	.0176	.1370	.4859	.9153
6.3	.2020	.1823	.1584	.1360	[illegible]	.0961	.0770	.0570	.0338
	.2639	.3102	.2547	.1970	.1513	.1158	.0877	.0645	.0431
	.0093	.0410	.0680	.0861	.0963	.0992	.0948	.0816	.0552
	.3709	.4059	.2912	.2005	.1405	.1034	.0821	.0702	.0598
	-.4041	-.7205	-.6496	-.4835	-.2238	.1260	.5064	.8092	.9677
7.2	.1616	.1391	.1214	.1059	.0914	.0771	.0623	.0462	.0270
	.5146	.3091	.2187	.1649	.1282	.1009	.0791	.0602	.0411
	.2371	.2139	.1934	.1737	.1536	.1324	.1092	.0824	.0490
	.6120	.3872	.2871	.2267	.1844	.1516	.1241	.0987	.0708
	-.1457	.0468	.2196	.3818	.5350	.6780	.8068	.9126	.9813
8.1	.0023	.0140	.0252	.0324	.0358	.0356	.0323	.0259	.0156
	.1176	.1821	.1447	.1053	.0779	.0608	.0506	.0429	.0324
	.2252	.2062	.1804	.1546	.1299	.1060	.0824	.0580	.0316
	.3384	.3698	.3270	.2703	.2198	.1771	.1400	.1055	.0686
	-.2091	-.5335	-.4530	-.2443	.0534	.3925	.6916	.8885	.9806

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items:10, Theta Zero: .60, Mastery Score: 6

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	.0004	.0010	.0021	.0041	.0076	.0126	.0183	.0221	.0191
	.0058	.0111	.0197	.0327	.0488	.0613	.0621	.0490	.0345
	.0000	.0000	.0000	.0000	.0001	.0005	.0017	.0038	.0055
	.0001	.0000	.0000	.0002	.0016	.0062	.0133	.0175	.0124
	.9494	.9449	.9521	.9713	.9733	.9693	.9520	.8922	.8148
2.0	.0095	.0139	.0199	.0279	.0371	.0456	.0508	.0494	.0364
	.0585	.0755	.0952	.1117	.1154	.1039	.0830	.0621	.0478
	.0000	.0000	.0000	.0002	.0010	.0031	.0066	.0105	.0115
	.0001	.0000	.0004	.0038	.0121	.0223	.0283	.0254	.0159
	.8872	.8992	.9217	.9273	.9192	.8910	.8276	.7231	.8151
3.0	.0566	.0674	.0794	.0898	.0960	.0964	.0901	.0758	.0502
	.1325	.1994	.1987	.1742	.1423	.1149	.0939	.0766	.0582
	.0000	.0000	.0005	.0024	.0061	.0113	.0164	.0196	.0173
	.0000	.0011	.0111	.0282	.0412	.0448	.0392	.0281	.0185
	.9026	.8234	.8222	.7769	.6888	.5768	.4961	.5558	.8684
4.0	.1784	.1883	.1886	.1799	.1650	.1457	.1225	.0948	.0590
	.3181	.2673	.2195	.1957	.1736	.1484	.1214	.0939	.0652
	.0000	.0017	.0072	.0148	.0222	.0279	.0308	.0297	.0222
	.0026	.0447	.0791	.0833	.0717	.0557	.0406	.0292	.0220
	.6275	.4821	.1798	-.0055	-.0269	.0796	.3046	.6317	.9244
5.0	.3643	.3170	.2724	.2337	.1990	.1667	.1349	.1014	.0619
	.6066	.5753	.4330	.3235	.2446	.1858	.139[illegible]	.1018	.0676
	.0090	.0290	.0[illegible]22	.0491	.0515	.0504	.046	.0383	.0254
	.2868	.2284	.1450	.0968	.0704	.0548	.0441	.0350	.0259
	-.7948	-.7867	-.6245	-.3606	-.0333	.2941	.5772	.8017	.9543
6.0	.2603	.2335	.2106	.1891	.1677	.1455	.1215	.0939	.0586
	.7163	.4484	.3267	.2518	.1986	.1575	.1234	.0935	.0647
	.1381	.1195	.1049	.0921	.0800	.0682	.0560	.0427	.0263
	.4124	.2454	.1720	.1283	.0983	.0759	.0580	.0429	.0290
	-.1455	.0183	.1650	.3055	4440	.5824	.7206	.8538	.9624
7.0	.0134	.0475	.0725	.0874	.0943	.0944	.0881	.0743	.0496
	.4589	.4162	.2889	.2053	.1535	.1196	.0953	.0760	.0572
	.1523	.1338	.1146	.0976	.0824	.0684	.0547	.0406	.0242
	.2589	.2580	.1969	.1468	.1101	.0829	.0619	.0450	.0297
	-.5209	-.6531	-.5067	-.2767	-.0016	.2859	.5594	.7947	.9550
8.0	.0000	.0016	.0086	.0197	.0318	.0421	.0484	.0477	.0357
	.0015	.0508	.1124	.1361	.1303	.1099	.0846	.0619	.0468
	.0407	.0490	.0541	.0550	.0526	.0476	.0406	.0313	.0190
	.1539	.1401	.1046	.0810	.0673	.0570	.0477	.0382	.0272
	.8419	.8060	.6662	.4768	.3450	.3232	.4265	.6678	.9315
9.0	.0000	.0000	.0002	.0015	.0049	.0103	.0166	.0210	.0186
	.0000	.0003	.0066	.0252	.0480	.0629	.0630	.0491	.0337
	.0031	.0055	.0088	.0128	.0165	.0188	.0190	.0166	.0108
	.0290	.0424	.0550	.0583	.0521	.0418	.0323	.0261	.0206
	.9346	.9485	.9530	.9439	.9153	.8532	.7461	.6845	.8948

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items:10, Theta Zero: .60, Mastery Score: 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	.0000	.0001	.0004	.0010	.0023	.0044	.0071	.0092	.0084
	.0008	.0021	.0050	.0104	.0183	.0257	.0278	.0224	.0152
	.0000	.0000	.0000	.0000	.0001	.0009	.0033	.0079	.0120
	.0001	.0000	.0000	.0003	.0030	.0120	.0273	.0380	.0283
	.9655	.9620	.9717	.9840	.9852	.9820	.9693	.9214	.8212
2.0	.0016	.0030	.0052	.0086	.0130	.0175	.0208	.0211	.0161
	.0150	.0234	.0349	.0464	.0517	.0483	.0388	.0281	.0211
	.0000	.0000	.0000	.0003	.0019	.0060	.0133	.0220	.0253
	.0000	.0000	.0007	.0070	.0233	.0446	.0595	.0565	.0358
	.9220	.9339	.9574	.9607	.9543	.9332	.8821	.7755	.8032
3.0	.0146	.0200	.0267	.0332	.0380	.0400	.0386	.0333	.0224
	.0748	.0913	.0975	.0874	.0703	.0545	.0426	.0339	.0259
	.0000	.0000	.0008	.0044	.0118	.0223	.0337	.0417	.0385
	.0000	.0020	.0203	.0536	.0817	.0930	.0854	.0635	.0411
	.8875	.9062	.9065	.8764	.8103	.7061	.5935	.5782	.8481
4.0	.0645	.0737	.0778	.0766	.0718	.0643	.0546	.0426	.0266
	.1908	.1572	.1138	.0921	.0790	.0669	.0547	.0423	.0294
	.0000	.0030	.0132	.0281	.0435	.0564	.0644	.0642	.0498
	.0045	.0811	.1502	.1658	.1500	.1216	.0911	.0650	.0483
	.8046	.7277	.4878	.2445	.1351	.1608	.3112	.5943	.9107
5.0	.1727	.1507	.1288	.1097	.0928	.0772	.0621	.0465	.0282
	.2947	.2892	.2170	.1602	.1193	.0892	.0661	.0474	.0310
	.0159	.0535	.0804	.0967	.1044	.1051	.0988	.0844	.0576
	.5174	.4430	.3002	.2093	.1539	.1186	.0942	.0751	.0572
	-.6312	-.7145	-.5757	-.3547	-.0772	.2253	.5188	.7723	.9489
6.0	.1427	.1232	.1080	.0947	.0823	.0701	.0576	.0438	.0270
	.4359	.2588	.1809	.1347	.1030	.0794	.0606	.0447	.0301
	.2583	.2333	.2115	.1907	.1697	.1478	.1238	.0960	.0601
	.6842	.4297	.3144	.2437	.1935	.1544	.1219	.0930	.0649
	-.1720	-.0063	.1437	.2883	.4313	.5747	.7160	.8520	.9622
7.0	.0077	.0264	.0390	.0457	.0480	.0469	.0428	.0352	.0230
	.2598	.2209	.1448	.0991	.0731	.0572	.0459	.0365	.0270
	.3392	.2966	.2553	.2190	.1862	.1555	.1253	.0934	.0560
	.5525	.5370	.4103	.3099	.2363	.1810	.1376	.1015	.0680
	-.7416	-.7466	-.5752	-.3083	.0083	.3203	.5937	.8133	.9586
8.0	.0000	.0009	.0048	.0106	.0167	.0215	.0240	.0230	.0167
	.0010	.0286	.0609	.0706	.0648	.0526	.0395	.0290	.0221
	.1318	.1417	.1449	.1406	.1303	.1156	.0970	.0741	.0445
	.2822	.2531	.2059	.1768	.1553	.1344	.1128	.0901	.0637
	.6712	.5844	.3805	.2030	.1442	.2058	.3879	.6812	.9389
9.0	.0000	.0000	.0001	.0008	.0026	.0054	.0084	.0102	.0088
	.0000	.0002	.0037	.0136	.0250	.0315	.0302	.0228	.0159
	.0181	.0246	.0324	.0404	.0467	.0496	.0478	.0403	.0256
	.1001	.1198	.1341	.1316	.1144	.0928	.0750	.0628	.0493
	.8788	.8915	.8981	.8838	.8396	.7555	.6462	.6439	.9031

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items:10, Theta Zero: .60, Mastery Score: 8

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	.0000	.0000	.0001	.0002	.0005	.0012	.0021	.0029	.0028
	.0001	.0003	.0009	.0025	.0052	.0081	.0092	.0076	.0050
	.0000	.0000	.0000	.0000	.0002	.0013	.0052	.0132	.0215
	.0001	.0000	.0000	.0004	.0044	.0186	.0445	.0667	.0536
	.9751	.9725	.9831	.9907	.9912	.9884	.9780	.9360	.8101
2.0	.0002	.0005	.0010	.0020	.0035	.0051	.0065	.0068	.0054
	.0025	.0050	.0092	.0140	.0169	.0165	.0134	.0096	.0070
	.0000	.0000	.0000	.0005	.0028	.0093	.0214	.0373	.0458
	.0000	.0000	.0011	.0103	.0354	.0710	.1005	.1026	.0677
	.9459	.9559	.9754	.9773	.9718	.9549	.9113	.8027	.7683
3.0	.0027	.0043	.0067	.0093	.0114	.0125	.0124	.0110	.0075
	.0200	.0285	.0335	.0313	.0253	.0192	.0145	.0113	.0086
	.0000	.0000	.0012	.0065	.0179	.0351	.0552	.0719	.0704
	.0000	.0028	.0295	.0805	.1284	.1541	.1508	.1188	.0760
	.9131	.9470	.9473	.9261	.8758	.7839	.6553	.5751	.8008
4.0	.0165	.0210	.0226	.0242	.0232	.0211	.0181	.0143	.0090
	.0762	.0651	.0443	.0328	.0268	.0224	.0183	.0142	.0099
	.0001	.0042	.0193	.0423	.0677	.0911	.1082	.1130	.0923
	.0062	.1165	.2252	.2612	.2497	.2142	.1679	.1206	.0871
	.8923	.8513	.6873	.4490	.2748	.2220	.2925	.5167	.8756
5.0	.0587	.0518	.0443	.0377	[illegible]	.0264	.0212	.0158	.0096
	.1014	.1018	.0771	.0568	.0420	.0312	.0229	.0163	.0106
	.0223	.0779	.1212	.1507	.1685	.1756	.1713	.1522	.1083
	.7386	.6762	.4903	.3614	.2745	.2129	.1669	.1318	.1032
	-.3679	-.6326	-.5306	-.3579	-.1414	.1122	.3981	.6951	.9310
6.0	.0562	.0469	.0401	.0345	.0295	.0248	.0201	.0151	.0092
	.1933	.1081	.0724	.0521	.0387	.0291	.0217	.0157	.0104
	.3810	.3591	.3373	.3141	.2884	.2587	.2231	.1780	.1148
	.9402	.6141	.4641	.3707	.3034	.2501	.2047	.1629	.1197
	-.2712	-.1257	.0171	.1652	.3209	.4846	.6537	.8194	.9543
7.0	.0032	.0105	.0151	.0173	.0178	.0170	.0152	.0123	.0079
	.1060	.0844	.0525	.0350	.0258	.0203	.0163	.0129	.0094
	.5877	.5191	.4538	.3955	.3417	.2897	.2367	.1790	.1088
	.8574	.8293	.6457	.5014	.3945	.3123	.2454	.1869	.1295
	-.8737	-.8331	-.6587	-.3704	-.0252	.3032	.5820	.8048	.9555
8.0	.0000	.0004	.0019	.0041	.0063	.0080	.0087	.0081	.0058
	.0004	.0116	.0238	.0265	.0235	[illegible]	.0136	.0101	.0077
	.3251	.3254	.3161	.2964	.2686	.2346	.1946	[illegible]	.0880
	.3852	.3673	.3461	.3267	.2984	.2621	.2214	.1774	.1255
	.3049	.1471	-.0439	-.1278	-.0891	.0580	.3191	.6683	.9380
9.0	.0000	.0000	.0000	.0003	.0010	.0020	.0031	.0037	.0031
	.0000	.0001	.0015	.0053	.0095	.0115	.0107	.0078	.0055
	.0799	.0903	.1008	.1099	.1148	.1132	.1034	.0837	.0515
	.2296	.2423	.2471	.2328	.2044	.1744	.1504	.1301	.1004
	.9990	.7549	.7425	.7206	.6541	.5540	.4777	.5722	.9003

Table of the False Positive Error and its S.E.*SQRT(M), the False Negative Error and its S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items:10, Theta Zero: .70, Mastery Score: 7

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	.0000	.0001	.0004	.0010	.0024	.0051	.0098	.0153	.0167
	.0008	.0021	.0050	.0106	.0208	.0357	.0490	.0485	.0322
	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0041
	.0001	.0000	.0000	.0000	.0002	.0015	.0060	.0125	.0118
	.9670	.9644	.9626	.9757	.9827	.9832	.9783	.9546	.8468
2.0	.0016	.0030	.0052	.0088	.0146	.0226	.0315	.0377	.0334
	.0150	.0234	.0355	.0522	.0710	.0837	.0815	.0631	.0430
	.0000	.0000	.0000	.0000	.0001	.0007	.0025	.0059	.0089
	.0000	.0000	.0000	.0002	.0022	.0079	.0164	.0214	.0150
	.9283	.9238	.9350	.9558	.9599	.9562	.9375	.8760	.7800
3.0	.0146	.0201	.0274	.0370	.0480	.0584	.0650	.0637	.0482
	.0748	.0926	.1131	.1299	.1323	.1180	.0936	.0692	.0526
	.0000	.0000	.0000	.0002	.0011	.0034	.0075	.0123	.0141
	.0000	.0000	.000[illegible]	.0039	.0127	.0235	.0300	.0270	.0162
	.8998	.9142	.9072	.9146	.9062	.8753	.8069	.6927	.7731
4.0	.0645	.0760	.0890	.100[illegible]	.1083	.1096	.1033	.0879	.0593
	.1933	.2107	.2116	.1870	.1527	.1226	.0999	.0814	.0521
	.0000	.0000	.0004	.0022	.0060	.0113	.0170	.0209	.0191
	.0000	.0008	.0098	.0266	.0403	.0447	.0394	.0279	.0178
	.9990	.8172	.8194	.7758	.6844	.5597	.4586	.5010	.8452
5.0	.1848	.1962	.1987	.1915	.1772	.1576	.1334	.1038	.0651
	.3252	.2791	.2249	.1990	.1784	.1545	.1278	.0995	.0695
	.0000	.0013	.0062	.0135	.0210	.0272	.0306	.0301	.0229
	.0016	.0369	.0724	.0799	.0702	.0546	.0392	.0276	.0211
	.6504	5346	.2405	.0139	-.0486	.0293	.2451	.5959	.9214
6.0	.3742	.3298	.2850	.2450	.2089	.1750	.1415	.1062	.0646
	.5491	.5717	.4442	.3370	.2569	.1962	.1483	.1085	.0723
	.0070	.0256	.0389	.0463	.0493	.0487	.0449	.0375	.0249
	.2463	.2186	.1419	.0937	.0666	.0512	.0414	.0335	.0254
	-.7472	-.7908	-.6558	-.4146	-.0868	.2649	.5734	.8088	.9584
7.0	.2651	.2370	.2132	.1908	.1687	.1459	.1212	.0930	.0572
	.7450	.4670	.3405	.2627	.2076	.1651	.1301	.0994	.0690
	.1318	.1139	.0999	.0875	.0759	.0645	.0527	.0398	.0242
	.3923	.2338	.1043	.1230	.0947	.0735	.0566	.0423	.0288
	-.1218	.0463	.1952	.3365	.4748	.6115	.7458	.8714	.9685
8.0	.0084	.0370	.0607	.0758	.0834	.0844	.0789	.0661	.0432
	.3397	.3802	.2818	.2052	.1543	.1201	.0961	.0778	.0594
	.1262	.1142	.0990	.0848	.0716	.0592	.0471	.0344	.0200
	.2037	.2106	.1709	.1319	.1013	.0778	.0592	.0438	.0291
	-.1846	-.5443	-.4412	-.2348	.0258	.3098	.5864	.8189	.9634
9.0	.0000	.0005	.0037	.0105	.0190	.0270	.0322	.0321	.0235
	.0001	.0191	.0648	.0972	.1053	.0950	.0752	.0557	.0432
	.0181	.0243	.0297	.0326	.0328	.0307	.0265	.0204	.0119
	.1000	.1076	.0904	.0[illegible]3	.0547	.0459	.0395	.0331	.0239
	.3965	.8897	.8391	.7291	.5833	.4757	.4962	.7028	.9447

Table of the False Positive Error and its S.E.*SQRT(M), the False Negative Error and its S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items:10, Theta Zero: .70, Mastery Score: 8

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	.0000	.0000	.0001	.0002	.0006	.0016	.0035	.0060	.0071
	.0001	.0003	.0009	.0026	.0064	.0131	.0201	.0213	.0141
	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0042	.0095
	.0001	.0000	.0000	.0000	.0003	.0030	.0127	.0284	.0292
	.9764	.9745	.9726	.9858	.9900	.9903	.9864	.9681	.8633
2.0	.0002	.0005	.0010	.0021	.0043	.0077	.0119	.0153	.0143
	.0025	.0050	.0094	.0169	.0267	.0348	.0359	.0283	.0186
	.0000	.0000	.0000	.0000	.0002	.0014	.0053	0131	.0210
	.0000	.0000	.0000	.0005	.0043	.0162	.0357	.0497	.0372
	.9508	.9472	.9634	.9749	.9774	.9740	.9594	.9084	.7845
3.0	.0027	.0043	.0071	.0111	.0164	.0218	.0259	.0266	.0208
	.0200	.0292	.0414	.0534	.0584	.0540	.0429	.0307	.0226
	.0000	.0000	.0000	.0004	.0021	.0071	.0161	.0277	.0336
	.0000	.0000	.0008	.0076	.0256	.0498	.0674	.0648	.0401
	.9189	.9273	.9484	.9532	.9465	.9231	.8671	.7490	.7511
4.0	.0165	.0221	.0291	.0362	.0417	.0442	.0431	.0376	.0259
	.0775	.0940	.1014	.0919	.0741	.0571	.0444	.0351	.0269
	.0000	.0000	.0008	.0043	.0121	.0238	.0372	.0477	.0458
	.0000	.0016	.0188	.0532	.0847	.0990	.0925	.0687	.0428
	.9990	.9016	.9039	.8755	.8085	.6956	.5623	.5194	.8114
5.0	.0643	.0741	.0793	.0791	.0749	.0677	.0579	.0455	.0287
	.1880	.1615	.1166	.0924	.0791	.0677	.0559	.0436	.0305
	.0000	.0024	.0120	.0272	.0438	.0586	.0687	.0702	.0559
	.0028	.0704	.1449	.1685	.1567	.1284	.0955	.0666	.0497
	.8144	.7578	.5438	.2802	.1264	.1144	.2423	.5369	.9016
6.0	.1711	.1517	.1307	.1118	.0948	.0790	.0635	.0475	.0287
	.2571	.2754	.2147	.1615	.1215	.0915	.0681	.0491	.0323
	.0131	.0497	.0784	.0968	.1066	.1089	.1037	.0895	.0615
	.4665	.4476	.3129	.2183	.1580	.1297	.0946	.0766	.0604
	-.5172	-.7090	-.6032	-.4101	-.1463	.1672	.4897	.7690	.9519
7.0	.1415	.1213	.1065	.0931	.0807	.0685	.0560	.0423	.0257
	.4414	.2618	.1830	.1353	.1045	.0808	.0620	.0462	.0313
	.2613	.2370	.2155	.1947	.1736	.1512	.1265	.0977	.0605
	.6762	.4268	.3143	.2454	.1967	.1589	.1272	.0988	.0702
	-.1817	-.0094	.1475	.2987	.4474	.5940	.7366	.8681	[illegible]
8.0	.0047	.0201	.0319	.0387	.0415	.0409	.0374	.0306	.0196
	.1883	.1980	.1387	.0969	.0715	.0557	.0449	.0364	.0273
	.3159	.2815	.2442	.2099	.1783	.1483	.1184	.0870	.0508
	.4560	.4882	.3945	.3074	.2396	.1869	.1446	.1085	.0731
	-.5891	-.6904	-.5431	-.2958	.0117	.3295	.6139	.8345	.9664
9.0	.0000	.0003	.0020	.0055	.0097	.0134	.0156	.0151	.0107
	.0001	.0105	.0344	.0496	.0515	.0446	.0343	.0254	.0199
	.0799	.0897	.0964	.0976	.0932	.0841	.0708	.0533	.0308
	.2294	.2261	.1918	.1586	.1365	.1204	.1052	.0873	.0621
	.8150	.7162	.6228	.4706	.3425	.3113	.4234	.7064	.9501

Table of the False Positive Error and its S.E.*SQRT(M), the False Negative Error and its S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items:10, Theta Zero: .80, Mastery Score: 8

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	.0000	.0000	.0001	.0002	.0006	.0017	.0044	.0097	.0146
	.0001	.0003	.0009	.0026	.0067	.0158	.0314	.0446	.0343
	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0007	.0026
	.0000	.0000	.0000	.0000	.0000	.0002	.0017	.0068	.0107
	.9718	.9662	.9634	.9594	.9874	.9885	.9880	.9796	.9112
2.0	.0002	.0005	.0010	.0022	.0045	.0090	.0167	.0265	.0307
	.0025	.0050	.0094	.0173	.0304	.0492	.0666	.0670	.0436
	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0026	.0062
	.0000	.0000	.0000	.0000	.0001	.0014	.0061	.0140	.0144
	.9464	.9656	.9364	.9674	.9730	.9756	.9724	.9512	.8351
3.0	.0027	.0043	.0071	.0114	.0183	.0282	.0402	.0499	.0467
	.0200	.0292	.0420	.0597	.0808	.0970	.0971	.0763	.0498
	.0000	.0000	.0000	.0000	.0001	.0006	.0023	.0062	.0104
	.0000	.0000	.0000	.0001	.0015	.0064	.0151	.0215	.0156
	.9372	.9732	.9259	.9470	.9525	.9511	.9351	.8773	.7369
4.0	.0165	.0221	.0297	.0400	.0526	.0656	.0754	.0767	.0605
	.0775	.0951	.1160	.1361	.1442	.1333	.1072	.0777	.0581
	.0000	.0000	.0000	.0001	.0007	.0027	.0066	.0119	.0150
	.0000	.0000	.0002	.0023	.0092	.0196	.0277	.0266	.0155
	.9990	.9119	.9026	.9115	.9081	.8833	.8197	.6877	.7032
5.0	.0643	.0760	.0897	.1037	.1146	.1195	.1160	.1013	.0701
	.1896	.2038	.2185	.2029	.1695	.1348	.1077	.0877	.0689
	.0000	.0000	.0002	.0013	.0043	.0092	.0150	.0198	.0192
	.0000	.0003	.0052	.0184	.0327	.0400	.0374	.0267	.0161
	.9990	.8216	.8334	.8086	7367	.6120	.4664	.4330	.8041
6.0	.1810	.1949	.2025	.2005	.1998	.1719	.1476	.1162	.0733
	.3257	.2999	.2390	.2017	.1808	.1607	.1370	.1098	.0787
	.0000	.0006	.0039	.0098	.0169	.0233	.0277	.0283	.0222
	.0003	.0201	.0531	.0676	.0643	.0516	.0366	.0247	.0194
	.7019	.6364	.4258	.1582	-.0001	-.0070	.1535	.5261	.9197
7.0	.3829	.3476	3048	.2640	.2258	.1893	.1528	.1141	.0685
	.4292	.5335	.4517	.3573	.2792	.2170	.1664	.1234	.0831
	.0038	.0185	.0313	.0394	.0434	.0438	.0410	.0346	.0230
	.1622	.1912	.1341	.0888	.0608	.0451	.0365	.0308	.0247
	-.5402	-.7680	-.6867	-.4964	-.1898	.1979	.5635	.8251	.9667
8.0	.2741	.2440	.2185	.1947	.1712	.1470	.1209	.0913	.0546
	.7931	.4995	.3659	.2839	.2259	.1814	.1448	.1123	.0783
	.1203	.1035	.0903	.0787	.0678	.0572	.0463	.0344	.0203
	.3614	.2161	.1526	.1150	.0893	.0701	.0548	.0416	.0285
	-.0758	.1014	.2561	.4003	.5382	.6705	.7946	.9026	.9779
9.0	.0025	.0196	.0389	.0533	.0616	.0640	.0603	.0500	.0314
	.1443	.2841	.2520	.1965	.1499	.1162	.0938	.0785	.0608
	.0785	.0767	.0692	.0603	.0513	.0422	.0330	.0235	.0129
	.1734	.1360	.1222	.1031	.0845	.0683	.0540	.0408	.0267
	.4953	-.1424	-.2467	-.1395	.0673	.3393	.6279	.8561	.9745

Table of the False Positive Error and its
S.E.*SQRT(M), the False Negative Error and its
S.E.*SQRT(M), and the Correlation between FP and FN
Number of Items:10, Theta Zero: .80, Mastery Score: 9

Test Mean	KR21= .100	.200	.300	.400	.500	.600	.700	.800	.900
1.0	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0034	.0056
	.0000	.0000	.0001	.0004	.0015	.0046	.0109	.0174	.0141
	-.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0018	.0070
	.0000	.0000	.0000	.0000	.0000	.0004	.0039	.0171	.0304
	.9784	.9741	.9635	.9704	.9920	.9934	.9928	.9863	.9262
2.0	.0000	.0000	.0001	.0004	.0010	.0024	.0053	.0095	.0120
	.0003	.0007	.0016	.0038	.0085	.0164	.0251	.0272	.0178
	.0000	.0000	.0000	.0000	.0000	.0002	.0014	.0064	.0167
	.0000	.0000	.0000	.0000	.0003	.0031	.0145	.0362	.0418
	.9530	.9751	.9496	.9800	.9844	.9860	.9830	.9661	.8541
3.0	.0003	.0006	.0012	.0024	.0047	.0086	.0138	.0186	.0184
	.0032	.0057	.0100	.0172	.0274	.0368	.0395	.0319	.0199
	.0000	.0000	.0000	.0000	.0002	.0013	.0055	.0155	.0285
	.0000	.0000	.0000	.0003	.0033	.0149	.0371	.0577	.0464
	.9370	.9402	.9594	.9692	.9735	.9717	.9589	.9112	.7424
4.0	.0027	.0042	.0067	.0105	.0159	.0221	.0274	.0294	.0242
	.0184	.0264	.0375	.0501	.0580	.0562	.0459	.0322	.0230
	.0000	.0000	.0000	.0002	.0016	.0062	.0160	.0308	.0416
	.0000	.0000	.0004	.0049	.0207	.0468	.0708	.0743	.0459
	.9990	.9990	.9442	.9519	.9490	.9304	.8801	.7520	.6636
5.0	.0143	.0192	.0257	.0331	.0397	.0438	.0444	.0400	.0283
	.0663	0819	.0939	.0914	.0769	.0593	.0448	.0350	.0274
	.0000	.0000	.0004	.0029	.0097	.0217	.0373	.0522	.0543
	.0000	.0006	.0111	.0411	.0769	.1004	.1013	.0780	.0455
	.9990	.9023	.9126	.8974	.8482	.7474	.5874	.4579	.7377
6.0	.0551	.0649	.0723	.0750	.0732	.0677	.0591	.0470	.0299
	.1677	.1606	.1212	.0910	.0751	.0648	.0550	.0442	.0319
	.0000	.0012	.0083	.0221	.0396	.0572	.0713	.0768	.0640
	.0007	.0425	.1180	.1599	.1626	.1402	.1056	.0707	.0527
	.8402	.8179	.6903	.4538	.2225	.1089	.1527	.4256	.8863
7.0	1579	.1458	.1281	.1108	.0945	.0789	.0635	.0472	.0283
	.2010	.2300	.1977	.1561	.1210	.0930	.0704	.0516	.0344
	.0078	.0401	.0708	.0931	.1071	.1130	.1104	.0972	.0676
	.3395	.4360	.3341	.2393	.1709	.1249	.0963	.0802	.0683
	-.0746	-.6418	-.6135	-.4787	-.2593	.0509	.4250	.7641	.9588
8.0	.1367	.1169	.1016	.0884	.0761	.0642	.0519	.0386	.0227
	.4412	.2615	.1832	.1370	.1056	.0824	.0641	.0484	.0329
	.2692	.2459	.2247	.2036	.1817	.1582	.1318	.1007	.0609
	.6748	.4299	.3206	.2545	.2082	.1723	.1421	.1140	.0828
	-.1919	-.0050	.1674	.3327	.4918	.6430	.7817	.8987	.9776
9.0	.0013	.0100	.0192	.0255	.0286	.0290	.0266	.0215	.0132
	.0758	.1401	.1174	0872	.0645	.0497	.0406	.0341	.0259
	.2629	.2435	.2153	.1861	.1576	.1296	.1014	.0721	.0397
	.3631	.3936	.3626	.3052	.2512	.2042	.1628	.1237	.0814
	-.1701	-.5140	-.4586	-.2716	.0065	.3383	.6505	.8697	.9768

APPENDIX B

SUBROUTINE ERRFPN

This subroutine computes the false positive error estimate and its standard error, the false negative error estimate and its standard error, and the correlation between the two estimates. The beta-binomial distribution is used as the vehicle for computations.

Disclaimer: The computer program hereafter listed has been written with care and tested extensively under a variety of conditions using tests with 60 or fewer items. The author, however, makes no warranty as to its accuracy and functioning, nor shall the fact of its distribution imply such warranty.

INFERENCE FOR ERROR RATES

```
      SUBROUTINE ERRFPN(N,A,B,M,TT,IM,FP,SEFP,FN,SEFN,RHO)
C.......................................................................
C
C     THIS SUBROUTINE COMPUTES THE FALSE POSITIVE ERROR ESTIMATE AND ITS
C     STANDARD ERROR, THE FALSE NEGATIVE ERROR ESTIMATE AND ITS STANDARD
C     ERROR, AND THE CORRELATION BETWEEN THE TWO ESTIMATES.  THE BETA-
C     BINOMIAL DISTRIBUTION IS USED AS THE VEHICLE FOR COMPUTATIONS.
C
C     INPUT DATA ARE:
C
C     N....NUMBER OF ITEMS
C     A....ALPHA OF THE BETA DISTRIBUTION
C     B....BETA OF THE BETA DISTRIBUTION
C     M....NUMBER OF EXAMINEES
C     TT...THETA ZERO, THE CRITERION LEVEL SET IN THE TRUE SCORE
C     IM...TEST CUTOFF SCORE (MASTERY SCORE)
C
C     A, B, AND TT ARE IN THE DOUBLE PRECISION FORMAT.
C
C     OUTPUT DATA ARE:
C
C     FP....FALSE POSITIVE ERROR ESTIMATE
C     SEFP..STANDARD ERROR OF FP
C     FN....FALSE NEGATIVE ERROR ESTIMATE
C     SEFN..STANDARD ERROR OF FN
C     RHO...CORRELATION BETWEEN FP AND FN
C
C     ALL OUTPUT DATA ARE IN THE DOUBLE PRECISION FORMAT.
C
C     THE SUBROUTINE IS SET UP FOR TESTS WITH UP TO 60 ITEMS.
C     FOR LONGER TESTS, SIMPLY CHANGE THE DIMENSIONS OF DF(.), DA(.),
C     AND DB(.) TO DF(N+1), DA(N+1), AND DB(N+1).
C
C     EXTERNAL SUBROUTINES REQUIRED: DQG32 OF SSP
C                                    MDBETA OF IMSL
C
C.......................................................................
      DOUBLE PRECISION A,B,TZ,BETA,GFCT,DFCT,U,V,DX,ONE,F,PSI,GA,GB,Y1,
     *Y2,Y3,Y4,VMONE,Z1,Z2,BB,DF(61),DA(61),DB(61),FP,SEFP,Z3,FN,SEFN,
     * H1,H2,H3,E(2),S(2),TT,P1,BA,PA,B1,W1,W2,RHO
      EXTERNAL BETA,BI,GFCT,DFCT,PSI
C
      ONE=1.D0
      Y1=BETA(A,B)
      Y2=PSI(A+B)
      Y3=PSI(A)-Y2
      Y4=PSI(B)-Y2
      P1=PSI(DFLOAT(N)+A+B)
      CALL NEHY2(N,A,B,DF)
      CALL        VARAB(N,A,B,H1,H2,H3,M,DF,DA,DB)
C
C     SET UP FOR FALSE POSITIVE ERRORS
      TZ=TT
      IC=IM
      U=A+DFLOAT(IC)
      V=B+DFLOAT(N-IC)
      W1=0.
      W2=0.
C
      DO 40 L=1,2
C
      F=ONE-TZ
      DX=DFCT(U,V,TZ)
      GA=GFCT(U,V,TZ)
      GB=GFCT(V,U,F)
C
```

```
      BB=BI(N,IC)
      E(L)=DX*BB
      DFPA=GA*BB
C
      BA=BETA(U,V)
      PA=PSI(V)
      DFPB=(BA*(PA-P1)-GB)*BB
C
C
      IF(IC.EQ.N) GO TO 30
C
   10 IZ=N-IC
      DO 15 I=1,IZ
      IX=IC+I
      VMONE=V-ONE
      Z1=-(TZ**U)*F**VMONE
      Z2=Z1*DLOG(TZ)
      Z3=(F**VMONE)*(TZ**U)*DLOG(F)
C
      CA=(Z2+DX+U*GA)/VMONE
C
C
      DX=(Z1+U*DX)/VMONE
C
      BB=BB*(N-IX+1)/IX
C
      V=V-ONE
      BA=BA*U/V
C
      GB=(Z3-(BA-DX)+U*GB)/VMONE
C
      U=U+ONE
      PA=PA-ONE/V
C
C
      E(L)=E(L)+BB*DX
      DFPA=DFPA+BB*GA
      DFPB=DFPB+BB*(BA*(PA-P1)-GB)
   15 CONTINUE
   30 IF(L.EQ.1) GOTO 35
C
C       INTERCHANGE DFPA AND DFPB FOR FALSE NEGATIVE ERROR
C
      F=DFPA
      DFPA=DFPB
      DFPB=F
C
   35 E(L)=E(L)/Y1
      DFPA=DFPA/Y1-E(L)*Y3
      DFPB=DFPB/Y1-E(L)*Y4
      W1=W1+DFPA
      W2=W2+DFPB
C
C
      S(L)=(H1*DFPA**2+H2*DFPB**2+2*H3*DFPA*DFPB)**.5D0
C
C       SET UP FOR FALSE NEGATIVE ERRORS
      TZ=ONE-TT
      IC=N-IM+1
      U=B+DFLOAT(IC)
      V=A+DFLOAT(N-IC)
C
   40 CONTINUE
C
      FP=E(1)
      FN=E(2)
```

```
      SEFP=S(1)
      SEFN=S(2)
      RHO =(H1*W1**2+H2*W2**2+2.*H3*W1*W2-S(1)**2-S(2)**2)/(S(1)*S(2)*2)
C
C
      RETURN
      END
      DOUBLE PRECISION FUNCTION BI(N,M)
      BI=1
      IF(M*(N-M).EQ.0) GOTO 20
      MM=MIN(N,N-M)
      DO 15 J=1,MM
   15 BI=BI*(N-J+1)/J
   20 RETURN
      END
      SUBROUTINE NEHY2(N,A,B,F)
      DOUBLE PRECISION A,B,F(1),Z1,Z2
      Z1=DFLOAT(N)+B
      Z2=Z1+A
      K=0
      F(1)=1.D0
      DO 5 I=1,N
5     F(1)=F(1)*(Z1-DFLOAT(I))/(Z2-DFLOAT(I))
10    KP1=K+1
      KP2=K+2
      F(KP2)=F(KP1)*DFLOAT(N-K)*(A+DFLOAT(K))/
     *                         (DFLOAT(KP1)*(Z1-DFLOAT(KP1)))
      K=K+1
      IF(K-N) 10,15,15
15    RETURN
      END
      SUBROUTINE VARAB(N,A,B,VA,VB,VAB,M,F,DA,DB)
      DIMENSION F(1),DA(1),DB(1)
      DOUBLE PRECISION A,B,DA,DB,F,B11,B12,B22,D,VA,VB,VAB
      CALL      DERLAB(N,A,B,DA,DB)
      B11=0.D0
      B12=0.D0
      B22=0.D0
      NP1=N+1
      DO 15 I=1,NP1
      B11=B11+DA(I)*DA(I)*F(I)
      B12=B12+DA(I)*DB(I)*F(I)
15    B22=B22+DB(I)*DB(I)*F(I)
      B11=B11*M
      B12=B12*M
      B22=B22*M
      D=B11*B22-B12*B12
      VA=B22/D
      VB=B11/D
      VAB=-B12/D
      RETURN
      END
      SUBROUTINE DERLAB(N,A,B,DA,DB)
      DIMENSION DA(1),DB(1)
      DOUBLE PRECISION A,B,DA,DB,Z1,Z2
      DOUBLE PRECISION ONE
      ONE=1.D0
      DA(1)=0.D0
      DB(1)=0.D0
      Z1=DFLOAT(N)+B
      Z2=Z1+A
      NP1=N+1
C
      DO 5 I=1,N
      DA(1)=DA(1)-ONE/(Z2-DFLOAT(I))
5     DB(1)=DB(1)+ONE/(Z1-DFLOAT(I))
```

```
      DB(1)=DB(1)+DA(1)
C
      DO 10 I=1,N
      IP1=I+1
      IX=I-1
      DA(IP1)=DA(I)+ONE/(A+DFLOAT(IX))
10    DB(IP1)=DB(I)-ONE/(Z1-DFLOAT(I))
      RETURN
      END
      DOUBLE PRECISION FUNCTION PSI(X)
      DOUBLE PRECISION X,A,P,ZETA(99),Y(54),PSI1,PM1,PP1,PM2,P2M1
C
      ZETA(2) =1.64493406684822643647D0
      ZETA(3) =1.20205690315959428540D0
      ZETA(4) =1.08232323371113819152D0
      ZETA(5) =1.03692775514336992633D0
      ZETA(6) =1.01734306198444913971D0
      ZETA(7) =1.00834927738192282684D0
      ZETA(8) =1.00407735619794433938D0
      ZETA(9) =1.00200839282608221442D0
      ZETA(10)=1.00099457512781808534D0
      ZETA(11)=1.00049418860411946456D0
      ZETA(12)=1.00024608655330804830D0
      ZETA(13)=1.00012271334757848915D0
      ZETA(14)=1.00006124813505870483D0
      ZETA(15)=1.00003058823630702049D0
      ZETA(16)=1.00001528225940865187D0
      ZETA(17)=1.00000763719763789976D0
      ZETA(18)=1.00000381729326499984D0
      ZETA(19)=1.00000190821271655394D0
      ZETA(20)=1.00000095396203387280D0
      ZETA(21)=1.00000047693298678781D0
      ZETA(22)=1.00000023845050272773D0
      ZETA(23)=1.00000011921992596531D0
      ZETA(24)=1.00000005960818905126D0
      ZETA(25)=1.00000002980350351465D0
      ZETA(26)=1.00000001490155482837D0
      ZETA(27)=1.00000000745071178984D0
      ZETA(28)=1.00000000372533402479D0
      ZETA(29)=1.00000000186265972351D0
      ZETA(30)=1.00000000093132743242D0
      ZETA(31)=1.00000000046566290650D0
      ZETA(32)=1.00000000023283118337D0
      ZETA(33)=1.00000000011641550173D0
      ZETA(34)=1.00000000005820772088D0
      ZETA(35)=1.00000000002910385044D0
      ZETA(36)=1.00000000001455192189D0
      ZETA(37)=1.00000000000727595984D0
      ZETA(38)=1.00000000000363797955D0
      ZETA(39)=1.00000000000181898965D0
      ZETA(40)=1.00000000000090949478D0
      ZETA(41)=1.00000000000045474738D0
      ZETA(42)=1.00000000000022737368D0
C
      Y(1) =.2436449038D0
      Y(2) =.2474724535D0
      Y(3) =.2512859559D0
      Y(4) =.2550855103D0
      Y(5) =.2588712154D0
      Y(6) =.2626431686D0
      Y(7) =.2664014664D0
      Y(8) =.2701462043D0
      Y(9) =.2738774769D0
      Y(10)=.2775953776D0
      Y(11)=.2812999992D0
      Y(12)=.2849914333D0
```

```
      Y(13)=.2880697707D0
      Y(14)=.2923351012D0
      Y(15)=.2959875138D0
      Y(16)=.2996270966D0
      Y(17)=.3032539307D0
      Y(18)=.3068681205D0
      Y(19)=.3104697335D0
      Y(20)=.3140588602D0
      Y(21)=.3176355846D0
      Y(22)=.3211999895D0
      Y(23)=.3247521572D0
      Y(24)=.3282921691D0
      Y(25)=.3318201056D0
      Y(26)=.3353360467D0
      Y(27)=.3388400713D0
      Y(28)=.3423322577D0
      Y(29)=.3458126835D0
      Y(30)=.3492814255D0
      Y(31)=.3527385596D0
      Y(32)=.3561841612D0
      Y(33)=.3596183049D0
      Y(34)=.3630410646D0
      Y(35)=.3664525136D0
      Y(36)=.3698527244D0
      Y(37)=.3732417688D0
      Y(38)=.3766197179D0
      Y(39)=.3799866424D0
      Y(40)=.3833426119D0
      Y(41)=.3866876959D0
      Y(42)=.3900219627D0
      Y(43)=.3933454805D0
      Y(44)=.3966583163D0
      Y(45)=.3999605371D0
      Y(46)=.4032522088D0
      Y(47)=.4065333970D0
      Y(48)=.4098041664D0
      Y(49)=.4130645816D0
      Y(50)=.4163147060D0
      Y(51)=.4195546030D0
      Y(52)=.4227843351D0
      Y(53)=.4260039643D0
      Y(54)=.4292135520D0
C
      A=X
      IF(X.LT.1.D0) A=A+1.D0
      PSI1=-.5772156649D0
C
      IF(A.GT.1.D0)GO TO 5
C
      PSI=PSI1
      RETURN
C
    5 PSI=0.D0
C
      IF(A.LT.2.D0)GO TO 20
C
   10 A=A-1.D0
      PSI=PSI+1.D0/A
      IF(A.LT.2.D0)GO TO 20
      GO TO 10
C
   20 IF(A.GT.1.75D0)GO TO 35
      IF(A.GT.1.D0) GOTO 21
      PSI=PSI+PSI1
      RETURN
C
```

```
   21 A=A-1.D0
      L=-23.21647129D0/DLOG(A)+1
      IF(L.LT.2)L=2
      M=MIN0(L,42)
C
      DO 25 N=2,M
   25 PSI=PSI+(-1)**N*ZETA(N)*A**(N-1)
      PSI=PSI+PSI1
      IF(M.EQ.L) GOTO 40
C
      M1=M+1
      DO 30 N=M1,L
      ZETA(N)=(ZETA(N-1)+1.D0)*.5D0
   30 PSI=PSI+(-1)**N*ZETA(N)*A**(N-1)
      GOTO 40
C
   35 P=(A-1.745D0)*200.D0
      IZ=DINT(P+1.D-10)
      IF(IZ.LT.1) IZ=1
C
      P=P-DFLOAT(IZ)
      IZ=IZ+1
C
      IF(P.NE.0.D0) GOTO 37
C
      PSI=Y(IZ)
      GOTO 40
C
   37 PM1=P-1.D0
      PP1=P+1.D0
      PM2=P-2.D0
      P2M1=PM1*PP1
      PSI=-P*PM1*PM2/6.D0*Y(IZ-1)+P2M1*PM2/2.D0*Y(IZ)-
     &P*PP1*PM2/2.D0*Y(IZ+1)+P*P2M1/6.D0*Y(IZ+2)+PSI
C
   40 IF(X.LT.1.0) PSI=PSI-1.D0/X
      RETURN
      END
      DOUBLE PRECISION FUNCTION GFCT(U,V,TZ)
      EXTERNAL FCT,DFCT
      DOUBLE PRECISION U,V,TZ,VP,UP,DFCT,ONE,H,XL,XU,FCT,Y,Y1,YHOLD,EPS
      DOUBLE PRECISION DX,TWO
      COMMON UP,VP
      TWO=2.D0
C
C
      IER=0
      XL=0.D0
      XU=TZ
      ONE=1.D0
      EPS=.00005
      KL=15
      IU=U-TWO
      IF(U.LE.TWO) IU=0
      UP=U-DFLOAT(IU)
      IV=V-TWO
      IF(V.LE.TWO) IV=0
      VP=V-DFLOAT(IV)
C
      DX=DFCT(UP,VP,TZ)
C
      IF(U.LT.ONE) UP=UP+ONE
C
      CALL DQG32(XL,XL,FCT,YHOLD)
C
      DO 6 J=2,KL
```

```
      Y=0.D0
      ML=2**J
      H= TZ/DFLOAT(ML)
C
      DO 5 I=1,ML
      XL=DFLOAT(I-1)*H
      XU=XL+H
      CALL DQG32(XL,XU,FCT,Y1)
    5 Y=Y+Y1
      IF(DABS((Y-YHOLD)/YHOLD).LE.EPS) GOTO 7
    6 YHOLD=Y
C
      IER=1
C
    7 GFCT=Y
C
      IF(IER.NE.0)WRITE(6,100)U,V,TZ,ML,EPS
  100 FORMAT(' ERROR IN GFCT AT U,V,THETA ZERO = ',3F10.5/
     *' AFTER',I9,' PARTITIONS, A TOLERANCE ERROR OF',F9.6,' CANNOT BE R
     *EACHED' /' COMPUTATIONS CONTINUED')
C
      IF(U.GE.ONE) GOTO 9
      UP=UP-ONE
      YHOLD=TZ**UP*(ONE-TZ)**VP
      H=YHOLD*(DLOG(TZ)-ONE/(UP+VP))-DX*VP/(UP+VP)
      GFCT=(UP+VP)*GFCT/UP+H/UP
C
    9 IF(IU.EQ.0) GO TO 20
C
      DO 10 I=1,IU
      YHOLD=TZ**UP*(ONE-TZ)**VP
      H=YHOLD*(DLOG(TZ)-ONE/(UP+VP))-DX*VP/(UP+VP)
      GFCT=(UP*GFCT-H)/(UP+VP)
      DX=(-YHOLD+UP*DX)/(UP+VP)
   10 UP=UP+ONE
C
   20 IF(IV.EQ.0) RETURN
C
      DO 30 I=1,IV
      YHOLD=TZ**U*(ONE-TZ)**VP
      H=YHOLD*(DLOG(TZ)-ONE/(U+VP))-DX*VP/(U+VP)
      GFCT=(GFCT*VP+H)/(U+VP)
      DX=(YHOLD+VP*DX)/(U+VP)
   30 VP=VP+ONE
C
      RETURN
      END
      DOUBLE PRECISION FUNCTION DFCT(A,B,TZ)
      EXTERNAL BETA
      DOUBLE PRECISION A,B,TZ,BETA
C
      AA=A
      BB=B
      TZZ=TZ
      CALL MDBETA(TZZ,AA,BB,P,IER)
C
      IF(IER.NE.0) WRITE(6,100)A,B,TZ,IER
  100 FORMAT('0',' ERROR IN BDTR, A B TZ IER ARE ',3F20.10,I5)
      DFCT=DBLE(P)*BETA(A,B)
      RETURN
      END
      DOUBLE PRECISION FUNCTION BETA(X,Y)
      DOUBLE PRECISION A,B,CON,X,Y,F
      F=5.D0
      A=X
      B=Y
```

```
      CON=1.D0
      IF(A.LE.F) GOTO 2
    1 A=A-1.D0
      CON=CON*A/(A+B)
      IF(A.LE.F) GOTO 2
      GOTO 1
    2 IF(B.LE.F) GOTO 4
    3 B=B-1.D0
      CON=CON*B/(A+B)
      IF(B.LE.F) GOTO 4
      GOTO 3
    4 BETA=DGAMMA(A)*DGAMMA(B)/DGAMMA(A+B)*CON
      RETURN
      END
      DOUBLE PRECISION FUNCTION FCT(T)
      COMMON U,V
      DOUBLE PRECISION T,U,V
      FCT=0.D0
      IF(T.EQ.0.D0) RETURN
      IF(T.EQ.1.D0) RETURN
C
      FCT=T**(U-1.D0)*(1.D0-T)**(V-1.D0)*DLOG(T)
      RETURN
      END
```

13

RELATIONSHIP BETWEEN DECISION ACCURACY AND DECISION CONSISTENCY IN MASTERY TESTING

Huynh Huynh
Joseph C. Saunders

University of South Carolina

ABSTRACT

In mastery testing, decision accuracy refers to the proportion of examinees who are classified correctly, in one of several achievement categories, by test data. Decision consistency expresses the extent to which decisions agree across two test administrations. Based on twelve cases involving a wide range of α_{21} reliabilities, it was found that decision accuracy and decision consistency were almost perfectly related.

1. INTRODUCTION

In classical measurement theory and practice, the reliability of a set of measurements (often, albeit unfortunately, referred to as the reliability of a test) is typically defined as the ratio of true-score variance to observed-score variance. The assumptions of classical test theory imply reliability can also be viewed as the correlation between two sets of parallel measurements

This paper has been distributed separately as RM 80-8, August, 1980.

(Lord & Novick, 1968). Capitalizing upon this property, several writers (Carver, 1970; Hambleton & Novick, 1973; Huynh, 1976c; Subkoviak, 1976) have proposed that reliability (of decisions) in mastery testing be considered from the standpoint of decision consistency (i.e., consistency of individual decisions across two test administrations). It has also been argued (Huynh, 1976b, [for the case of Q=1]; Livingston & Wingersky, 1979; van der Linden & Mellenbergh, 1978; Subkoviak & Wilcox, 1978; Wilcox, 1977) that the quality of the decision-making process would be more appropriately assessed via the agreement between decisions based on test data and those based on true scores, had these been known. Such agreement, in its simplest form, may be expressed as the proportion of examinees who are correctly classified by the test scores. This quantity will be referred to as decision accuracy in subsequent sections of this paper. In a slightly different form, it has been called a validity coefficient by Berk (1976). Decision accuracy, in this context, presumes that false positive and false negative errors are weighted equally. When the weights (losses or utilities) are not equal, then coefficients based on decision theory, such as ε (Huynh, 1976b), δ (van der Linden & Mellenbergh, 1978), or γ (Wilcox, 1978) may be more appropriate. However, decision consistency regards both types of inconsistent decision as being of equal severity. Thus, only the case involving equal (and constant) losses will be considered in this paper, so that comparisons might be anchored in the same framework.

The purpose of this paper is to study the relationship between decision consistency and decision accuracy for a variety of situations involving mastery tests. For reason of computational simplicity, the study is restricted to test score distributions which follow a beta-binomial form.

2. COMPUTATIONAL PROCEDURES

Let x and θ denote the observed and true score for a subject, and let c and θ_o denote the corresponding passing scores for mastery classification. In addition, let y be the observed score for the same subject on a second (parallel) test administration. The raw index of decision consistency is defined as p_{xy} = $Pr(x<c,y<c) + Pr(x \geq c,y \geq c)$, and an index of decision accuracy may be taken as $p_{x\theta} = Pr(x<c,\theta<\theta_o) + Pr(x \geq c,\theta \geq \theta_o)$. (Other indices similar to Cohen's kappa may also be used; however, since the marginal probabilities of the mastery and nonmastery categories as defined by the test scores x and y, and by the true score θ are identical or almost identical, any relationship between the p indices would hold for the kappa indices.)

When the test data can be described via a beta-binomial model, both indices p_{xy} and $p_{x\theta}$ may be computed via formulae, tables, and computer programs reported in Huynh (1979a, 1979b, 1980b, 1980c). Additionally, in the context of decision-making, it seems logical to select a (test) passing score c which reflects the true cutoff score θ_o and the two (equal and constant) losses under consideration. When the beta-binomial model holds, the value c may be obtained via the incomplete beta functions (Huynh, 1976a). Let n be the number of items, and α and β be the two parameters of the beta distribution. Then the Bayesian passing score is the smallest integer c at which the incomplete beta function $I(\alpha+c,n+\beta-c;\theta_o)$ is less than or equal to .5. In most instances involving minimax decisions (Huynh, 1980b), the value of c is very close to $n\theta_o$; this simple expression will be used throughout this paper.

3. DATA BASE

Two sets of test data were used in this study, one fictitious and the other derived from responses to the Science Research Associates Mastery Tests (SRA, 1974, 1975). The fictitious data set consists of eight beta-binomial distributions, each of which was selected to yield a testing situation in which the α_{21} reliability was low or moderate. Table 1 contains descriptions of these cases.

TABLE 1

A Comparison of Decision Accuracy and
Decision Consistency based on
Moderately Reliable Beta-Binomial Test Scores

Case	Shape	n	μ	σ	α_{21}	θ_o	c	$p_{x\theta}$	p_{xy}
1	Unimodal	5	3.125	1.301	.385	.5	3	.768	.687
2	Symmetric	5	2.500	1.279	.294	.5	3	.693	.605
3	Unimodal	10	8.000	1.706	.500	.7	7	.845	.799
4	J-Shaped	10	9.000	1.500	.667	.7	7	.941	.921
5	Unimodal	20	12.000	3.024	.500	.7	14	.773	.678
6	Unimodal	20	16.000	2.646	.571	.7	14	.868	.821
7	Unimodal	30	16.000	3.801	.500	.8	24	.979	.964
8	J-Shaped	30	29.250	1.319	.600	.8	24	.993	.990

Table 2 describes the second data set which consists of four SRA-compiled tests. The SRA data were obtained from the South Carolina State Department of Education. The data, consisting of the item responses of approximately 3000 sixth grade students for the SRA Mathematics (form X) and SOBAR Reading (form L) tests, were collected in a field testing conducted in the spring of 1978. Artificial subtests of 10, 20, 30, and 40 items were created from the SRA data by random selection of items from sets of homogeneous objectives.

TABLE 2

Description of the SRA Mastery Tests Data

Case	Subject Area	Number of Items	Mean	S.D.	α_{21}
9	Reading	10	7.016	2.391	.704
10	Reading	20	12.268	4.787	.835
11	Math	30	15.666	5.901	.812
12	Math	40	19.552	7.439	.840

4. RESULTS AND DISCUSSION

The data regarding decision accuracy and decision consistency are reported in the right side on Table 1 for the fictitious data set and in Table 3 for the SRA-compiled tests. In all situations under consideration, p_{xy} is smaller than $p_{x\theta}$; the ratio of p_{xy} to $p_{x\theta}$ averages about .96. However, the correlation between the two indices is .993, which represents an almost perfect linear relationship. For the 12 cases under study, decision accuracy relates to decision consistency via the empirical formula $p_{x\theta} = .25 + .75p_{xy}$.

TABLE 3

A Comparison of Decision Accuracy and Decision Consistency Based on Real Data

Case	True Cutoff θ_o	Test Cutoff c	Decision Accuracy	Decision Consistency
9	.50	5	.894	.858
	.70	7	.828	.780
10	.50	10	.892	.852
	.70	14	.870	.826
11	.50	15	.863	.812
	.70	21	.893	.853
12	.50	20	.872	.823
	.70	28	.922	.892

This study indicates that there is little difference between the indices of decision accuracy and decision consistency in terms of ranking the quality of different test-based decision-making processes. Decision accuracy can be predicted with very little error from decision consistency. The relationship between the two indices thus parallels that of the two approaches to classical reliability discussed in the introduction to this paper.

The basic result of this study casts doubt on the conjecture by Mellenbergh and van der Linden (1979, p. 263) that "the consistency of decisions is not related in the same way to the association between decisions and true states as consistency of measurements as related to the reliability coefficient." The very basic assumption which underlies our conclusion is that the test passing score must reflect in some way the true cutoff score and the various losses which are incorporated in the decision-making process. If this assumption is tenable, any comparison between decision accuracy and decision consistency would have no useful meaning if the test passing score and the true cutoff score were selected independently of each other. The counterexample presented by Mellenbergh and van der Linden (1979, p. 263) seems to reflect this type of selection. In addition, the above conjecture appears to be contradicted by the theoretical results reported by Huynh (1976c, 1978a), namely the fact that under fairly general assumptions, the raw agreement index and the kappa index for decision consistency are increasing functions of the classical reliability. Thus, both these indices of decision consistency across two test administrations reflect the nature of the relationship between true scores and observed scores.

It should be pointed out that the indices of decision accuracy and of decision consistency are defined for a set of test scores collected from the administration of a test to a group of examinees. Both indices thus represent internal characteristics of the data. As may be recalled, the decision accuracy index considered in this paper presumes that losses associated with incorrect decisions are equal (and constant); it should be replaced by appropriate efficiency indices when losses do not have this simple form. In this case, the Huynh efficiency indices (Huynh, 1975, 1976b, 1980a), the δ index proposed by van der Linden and Mellenbergh (1978), or the Wilcox γ index (1978) might be used. Because losses are often defined as a function of the true ability (which is typically estimated from test data), all these indices actually represent the internal characteristics of the data; they do not appear to be reflective of any other trait which might relate to the test itself. Decision accuracy and other similar efficiency indices seem to act as counterparts of reliability in classical test theory.

Finally, it may be noted that in many practical situations, losses are very hard to assess, and loss-based coefficients may not be useful. For example, procedures for setting passing scores are often based on an examination of the test items or on a consideration of the objectives underlying the test. For situations in which these procedures are appropriate, only the test passing score is available for the evaluation of the internal characteristics of the test data; hence decision consistency may very well be the only characteristic of the data which could feasibly be used to assess reliability. The argument seems convincing that decisions based on test data would not be acceptable if they could not be replicated to a satisfactory degree by use of the data collected from another test administration. The practical implications of this study seemly contradict the assertion by Mellenbergh and van der Linden that "decision consistency and reliability are not equivalent concepts" (1979, p. 270). Based on the results of this study, it appears that decision consistency acts very much like a counterpart of classical test reliability.

BIBLIOGRAPHY

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education 45, 4-9.

Carver, R. P. (1970). Special problems in measuring changes in psychometric devices. In Evaluative research: Strategies and methods. Pittsburgh: American Institute for Research.

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

Huynh, H. (1975). Statistical problems in binary classification. Paper presented at the annual meeting of the American Statistical Association, Atlanta.

Huynh, H. (1976a). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

Huynh, H. (1976b). On mastery scores and efficiency of criterion-referenced tests when losses are partially known. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Huynh, H. (1976c). Reliability of multiple classifications. Paper presented at Psychometric Society Meeting, Chapel Hill.

Huynh, H. (1978). Reliability of multiple classifications. Psychometrika 43, 317-325.

Huynh, H. (1979a). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics 4, 231-246.

Huynh, H. (1979b). Computation and inference for two reliability indices in mastery testing based on the beta-binomial model. Research Memorandum 78-2, Publication Series in Mastery Testing. University of South Carolina College of Education.

Huynh, H. (1980a). Assessing efficiency of decisions in mastery testing. Research Memorandum 80-5, Publication Series in Mastery Testing. University of South Carolina College of Education.

Huynh, H. (1980b). A nonrandomized minimax solution for passing scores in the binomial error model. Psychometrika 45, 167-182.

Huynh, H. (1980c). Statistical inference for false positive and false negative error rates in mastery testing. Psychometrika 45, 107-120.

Livingston, S. A. & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. Journal of Educational Measurement 16, 247-260.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company.

Mellenbergh, G. J. & van der Linden, W. J. (1979). The internal and external optimality of decision based on tests. Applied Psychological Measurement 3, 257-273.

SRA (1974). Mastery mathematics, level F, form X. Chicago: Science Research Associates.

SRA (1975). Mastery SOBAR reading, level F, form L. Chicago: Science Research Associates.

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

Subkoviak, M. J. & Wilcox, R. R. (1978). Estimating the probability of correct classification in mastery testing. Paper presented at the annual meeting of the American Educational Research Association, Toronto.

van der Linden, W. J. & Mellenbergh, G. J. (1978). Coefficients for tests from a decision theoretic point of view. Applied Psychological Measurement 2, 119-134.

Wilcox, R. R. (1977). Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics 2, 289-307.

Wilcox, R. R. (1978). A note on decision theoretic coefficients for tests. Applied Psychological Measurement 2, 609-613.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no endorsement should be inferred.

PART FIVE

Efficiency of Decisions

14

A NOTE ON DECISION-THEORETIC COEFFICIENTS FOR TESTS

Huynh Huynh

University of South Carolina

ABSTRACT

A modification is suggested for the decision-theoretic coefficient δ proposed by van der Linden and Mellenbergh. Under reasonable assumptions, the modified index varies from 0 to 1 inclusive. It is argued that in many practical applications of mastery testing, coefficients such as δ are not readily available, and consistency of decisions may serve as evidence of the quality of the decision-making process.

1. INTRODUCTION

Coefficients for tests (or strictly speaking, for a set of measurements) derived from decision theory have been formulated in a variety of ways (Huynh, 1975, 1976; van der Linden & Mellenbergh, 1978). These coefficients are based on the reduction in the proportion of expected loss (or Bayes risk) which would result from using test scores in the decision-making process. The efficiency coefficient proposed by Huynh is defined as $\varepsilon = (R^* - R_o)/R^*$ where R_o is the expected opportunity loss associated

---

This paper has been distributed separately as RM 80-4, July, 1980.

with the best use of test scores. The denominator R* is the minimum of a similar loss which would be incurred if decisions were based on information having no relationship to the true ability of the individual subject. (It may be noted that the opportunity losses associated with perfect information, i.e., when decisions are always correct, are zero.) Using the notion of monotone decisions along with the assumption of monotone likelihood ratio for the test score density, Huynh was able to prove that the efficiency index ε ranges between 0 and 1 inclusive. The lowest value 0 occurs when test information is unrelated to the ability of the subject, and the upper bound 1 is reached when test scores reveal faithfully the ability of the subject.

The decision-theoretic coefficient proposed by van der Linden and Mellenbergh (1978) is defined as $\delta = (R_n - R_B)/(R_n - R_c)$, where R_B represents the expected loss associated with the use of test scores. R_c and R_n, on the other hand, are the expected losses for situations in which the test contains complete and no information about the true scores, respectively. These losses are not necessarily opportunity losses. As defined, the coefficient δ is 0 when test scores are unrelated to true ability, and reaches the value 1 when test scores contain complete information about true ability. However, as noted by van der Linden and Mellenbergh (1978), the coefficient δ may not always lie within the interval defined by 0 and 1. To overcome this deficiency, Wilcox (1978) proposed that R_n and R_c be replaced with the upper and lower bounds of the expected loss R_B. His index γ, then, will range between 0 and 1. However, it is not known if these bounds have direct interpretations in terms of the degree of relationship between test score and true ability.

The purpose of this note is to modify the index δ slightly, and to describe the situations in which the resulting index falls between 0 and 1. The assumptions are presented only for the case of binary (mastery versus nonmastery) classification; however, they may be generalized in a fairly simple manner to situations involving more than two classification categories.

2. GENERAL CONSIDERATIONS

Consider a population of subjects for whom the true ability θ is distributed according to the density $p(\theta)$ with Ω as range. If there is only one subject in the population, then $p(\theta)$ represents the prior density in the context of Bayesian statistics. Let x represent the observed test score and $f(x|\theta)$ be its conditional density with the real line as the range. Let a_1 be the action of denying mastery status (the nonmastery category) and a_2 be the action of granting mastery (the mastery category). Following the notation used in Ferguson (1967, chapter 6), let $L(\theta,a_1)$ and $L(\theta,a_2)$ be the losses associated with the actions a_1 and a_2. In most formulations of mastery testing, it is usually assumed that there exists a true cutoff ability θ_o such that action a_1 is better than action a_2 when $\theta < \theta_o$ and the reverse is true when $\theta \geq \theta_o$. To be consistent with these assumptions, the losses would have to satisfy the following inequalities: $L(\theta,a_1) \leq L(\theta,a_2)$ for $\theta < \theta_o$ and $L(\theta,a_1) \geq L(\theta,a_2)$ for $\theta \geq \theta_o$. Under these conditions, the binary decision problem involving the actions a_1 and a_2 is said to be <u>monotone</u>.

In practical situations, however, mastery/nonmastery decisions are usually based on observed test data. In general, it seems reasonable that mastery should be granted if the test score x is high, and nonmastery should be presumed if the test score is low. In order that this type of classification be optimum in most decision-theoretic contexts, it is traditionally assumed that the conditional density $f(x|\theta)$ has <u>monotone likelihood ratio</u>. This condition is fulfilled for test models involving the exponential, Poisson, normal, negative binomial, gamma, and beta distributions, and in general, distributions belonging to the one-parameter exponential family (Ferguson, 1967, p. 208-209). In addition, the assumption of monotone likelihood ratio for $f(x|\theta)$ implies (Lehmann, 1966; Dykstra, Hewett, & Thompson, 1973, p. 679, definition) that x is positive likelihood ratio dependent upon θ. This result, in turn, implies that x and θ are stochastically increasing in sequence (Dykstra et al., Theorem 2); that is, the conditional distribution of x, $F(x|\theta)$ is nonincreasing in θ. Thus, when the monotone likelihood ratio assumption is fulfilled, the probability that a subject achieves a test score of x or lower is greater for subjects with lower ability.

When $f(x|\theta)$ has monotone likelihood ratio, it is best to declare mastery if the test score x is at least c, and declare nonmastery if the test score x is smaller than c. The expected loss (or Bayes risk) associated with the cutoff test score c is

$$R = \int_\Omega \int_{-\infty}^{c} L_1(\theta,a_1)f(x|\theta)p(\theta)dxd\theta$$

$$+ \int_\Omega \int_{c}^{\infty} L_2(\theta,a_2)f(x|\theta)p(\theta)dxd\theta,$$

or

$$R = \int_\Omega L_1(\theta,a_1)\ Pr(x<c|\theta)p(\theta)d\theta$$

$$+ \int_\Omega L_2(\theta,a_2)\ Pr(x\geq c|\theta)p(\theta)d\theta. \tag{1}$$

Consider now the first extreme case where x carries no information about θ, i.e., when x and θ are independent. For this situation, the two probabilities $Pr(x<c|\theta)$ and $Pr(x\geq c|\theta)$ are free of θ, and the expected loss may be written as

$$R_n = [\int_\Omega L_1(\theta,a_1)d\theta]Pr(x<c)$$

$$+ [\int_\Omega L_2(\theta,a_2)d\theta]Pr(x\geq c). \tag{2}$$

The relationship between R and R_n may be stated as follows.

Theorem 1. Let $L_1(\theta,a_1)$ be nondecreasing in θ and $L_2(\theta,a_2)$ be nonincreasing in θ. In addition, let $f(x|\theta)$ have monotone likelihood ratio. Then $R \leq R_n$.

Proof. Equation (1) may be written as

$$-R = E_\theta[-L_1(\theta,a_1)\ Pr(x<c|\theta)]$$

$$+ E_\theta[L_2(\theta,a_2)\ \{-Pr(x\geq c|\theta\}].$$

All the functions $-L_1(\theta,a_1)$, $Pr(x<c|\theta)$, $L_2(\theta,a_2)$, and $-Pr(x\geq c|\theta)$ are nonincreasing in θ, hence (Dykstra et al., 1973, p. 678)

$$-R \geq -[E_\theta L_1(\theta,a_2)]\, E_\theta Pr(x<c|\theta)$$
$$-[E_\theta L_2(\theta,a_2)]\, E_\theta Pr(x\geq c|\theta),$$

or

$$-R \geq -R_n. \quad \text{Q.E.D.}$$

The assumptions regarding the variations of $L_1(\theta,a_1)$ and $L_2(\theta,a_2)$ with respect to a_1 and a_2 seem intuitively justified. The denial of mastery status probably should cause less harm to a subject with lower ability than to someone with higher ability. Granting mastery status, on the other hand, should entail lesser consequences for a high ability subject than to someone with lower ability.

Consider now the second extreme case where the test score x reveals fully the ability θ of the subject. It appears reasonable to impose a strictly increasing function relating x to θ. Let θ_c be the image of the test cutoff score c on the true ability scale θ. Then it may be deduced that $P(x<c|\theta) = 1$ when $\theta < \theta_c$ and 0 when $\theta \geq \theta_c$. On the other hand, $P(x\geq c|\theta) = 0$ when $\theta < \theta_c$ and 1 otherwise. Thus, under the assumption of complete information, the expected loss as expressed in (1) will be equal to

$$\int_{-\infty}^{\theta_c} L_1(\theta,a_1)p(\theta)d\theta + \int_{\theta_c}^{+\infty} L_2(\theta,a_2)p(\theta)d\theta.$$

Under the monotone-decision conditions imposed previously on the loss functions, it may be shown that this loss is minimized when $\theta_c = \theta_o$. Hence the minimum complete-information expected loss may be taken as

$$R_c = \int_{-\infty}^{\theta_o} L_1(\theta,a_1)p(\theta)d\theta + \int_{\theta_o}^{\infty} L_2(\theta,a_2)p(\theta)d\theta. \quad (3)$$

Theorem 2. Under the monotone-decision assumptions, the expected loss R, computed at any test cutoff score, and the minimum complete-information expected loss, R_c, satisfy the inequality $R_c \leq R$.

Proof. Consider the expected loss R of (1) which can be written as

$$R = \int_{-\infty}^{\theta_o} L_1(\theta,a_1)\ \Pr(\theta<c|\theta)p(\theta)d\theta + \int_{-\infty}^{\theta_o} L_2(\theta,a_2)\ \Pr(\theta\geq c|\theta)p(\theta)d\theta$$

$$+ \int_{\theta_o}^{\infty} L_1(\theta,a_1)\ \Pr(\theta<c|\theta)p(\theta)d\theta + \int_{\theta_o}^{\infty} L_2(\theta,a_2)\ \Pr(\theta\geq c|\theta)p(\theta)d\theta.$$

When $\theta < \theta_o$, $L_1(\theta,a_1) \leq L_2(\theta,a_2)$ and when $\theta \geq \theta_o$, $L_2(\theta,a_1) \leq L_1(\theta,a_2)$. By noting that $\Pr(x<c|\theta) + \Pr(x\geq c|\theta) = 1$, it may then be verified that $R \geq R_c$. Q.E.D.

The following corollary is immediate.

Corollary. Let the loss $L_1(a_1,\theta)$ be nondecreasing in θ, the loss $L_2(a_2,\theta)$ be nonincreasing in θ, and let the graphs of these functions cross at a given point within the positive-probability range of θ. In addition, let $f(x|\theta)$ have monotone likelihood ratio. Then the index $\delta = (R - R_c)/(R_n - R_c)$ in which R_c is the minimum complete information expected loss will be between 0 and 1 inclusive.

3. RATIONALE FOR THE USE OF MINIMUM EXPECTED LOSS

The use of the minimum expected loss for the case of a strictly increasing relationship between x and θ guards against the seeming contradiction in which the use of perfectly reliable test data would cause more harm than the use of less-than-perfectly reliable test data.

The bounds R_n and R_c for the expected loss R have fairly straight-forward psychometric interpretations. The lower limit R_n would occur if nonmastery and mastery status were randomly assigned to examinees regardless of the test scores, keeping the proportion of nonmasters equal to that of examinees having test scores smaller than c, and the proportion of masters equal to that of examinees having a test score of c or greater. The upper limit R_c corresponds to the best use of completely reliable test data.

It may be noted that both bounds (R_n and R_c) are easy to compute, given the quantities $p(\theta)$, $f(x|\theta)$, $L_1(\theta,a_1)$ and $L_2(\theta,a_2)$. Thus, the index δ as defined in this note may be estimated in a fairly straight-forward manner for most situations involving the use of test data to make decisions. This represents an advantage over the Wilcox γ (Wilcox, 1978, p. 610) which seems to involve rather complex calculations.

4. SOME ADDITIONAL REMARKS

As additional remarks regarding the index δ proposed by van der Linden and Mellenbergh (1978), some departures appear apparent between its formulation and the various illustrations. The authors argued that their index δ seemed more realistic than the coefficient ε defined in Huynh (1976) because δ was defined on any chosen cutoff score while the ε index relied on the optimum cutoff score. But, in both illustrations based on squared-error and linear losses, the optimum cutoff score was used in order to reach the conclusion that the δ index was equal to the classical reliability index. In addition, δ was presented as a coefficient that represented the optimality of decisions (p. 133). Thus the use of a less-than-optimal cutoff score in the formulation of δ seemed to contradict the very characteristic which δ was thought to embrace.

Finally, the use of any decision-theoretic coefficient for tests presumes the availability of the losses (or utilities) associated with the various actions. In a number of practical situations, however, decisions regarding cutoff scores are not based on losses because they are not readily quantified or because the decision-maker is not willing to use them. In many instances, for example, cutoff scores are derived from an examination of item content or a consideration of the educational objectives. For these cases, the decision-theoretic coefficients as described in this paper are not available and the consistency of various decisions across two test administrations may serve as evidence of the quality of the decision-making process. It may be argued that decisions regarding success or failure for each subject may not be acceptable if they cannot be replicated to a reasonable extent on a second test administration. It is cautioned, of course, that test-retest consistency for decisions does not necessarily imply that the corresponding decisions are reflective of the purposes that the decision-maker has in mind. This line of reasoning is reminiscent of the well-accepted fact that in measurement, reliability is a necessary but not a sufficient condition for validity.

BIBLIOGRAPHY

Dykstra, R. L., Hewett, J. E., & Thompson, W. A., Jr. (1973). Events which are almost independent. Annals of Statistics 1, 674-681.

Ferguson, T. S. (1967). Mathematical statistics: A decision-theoretic approach. New York: Academic Press.

Huynh, H. (1975). Statistical problems in binary classification. Paper presented at the annual meeting of the American Statistical Association, Atlanta.

Huynh, H. (1976). On mastery scores and efficiency of criterion-referenced tests when losses are partially known. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lehmann, E. L. (1966). Some concepts of dependence. Annals of Mathematical Statistics 37, 1137-1153.

van der Linden, W. J. & Mellenbergh, G. J. (1978). Coefficients for tests from a decision theoretic point of view. Applied Psychological Measurement 2, 119-134.

Wilcox, R. R. (1978). A note on decision theoretic coefficients for tests. Applied Psychological Measurement 2, 609-613.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance of Joseph C. Saunders is gratefully acknowledged.

15

ASSESSING EFFICIENCY OF DECISIONS IN MASTERY TESTING

Huynh Huynh

University of South Carolina

ABSTRACT

Two indices are proposed for assessing the efficiency of decisions in mastery testing. The indices are generalizations of the raw agreement index and the kappa index. Both express the reduction in the proportion of average loss (or the gain in utility) resulting from the use of test scores to make decisions. Empirical data are presented which show little discrepancy between estimates based on the beta-binomial and compound binomial models for one index.

1. INTRODUCTION

A primary purpose of mastery testing is to classify examinees in several achievement or ability categories. Typically, there are two such categories, which are often referred to as mastery (ready, competent, or instructed) and nonmastery (nonready, incompetent, or uninstructed) groups. Ideally, these categories are defined on the basis of the true ability (θ) of the subjects; however, in reality,

---

This paper has been distributed separately as RM 80-5, July, 1980.

observed test scores are used to make mastery/nonmastery decisions. Since observed test data are often fallible, decisions based thereupon are less than completely accurate or efficient.

In the simplest formulation of mastery testing (Hambleton & Novick, 1973; Huynh, 1976a), the categories of true mastery and true nonmastery are defined respectively by the conditions $\theta \geq \theta_o$ and $\theta < \theta_o$, θ_o being a constant referred to as a criterion level by Hambleton and Novick and a true mastery score by Huynh. A test is given, and the observed test score x is obtained for each individual examinee. A suitable test passing (cutoff, mastery) score c will be chosen, and the examinee will be granted or denied mastery status if the observed test score x is such that $x \geq c$ or $x < c$. The two combinations ($\theta < \theta_o$; $x < c$) and ($\theta \geq \theta_o$; $x \geq c$) represent correct decisions; they entail no (opportunity) losses in the decision process. The other two possible combinations correspond to a false positive error ($\theta < \theta_o$; $x \geq c$) and a false negative error ($\theta \geq \theta_o$; $x < c$). Some form of loss function, such as constant, linear, or squared error loss, is typically assigned to each of these errors in most decision-theoretic formulations of mastery testing (Hambleton & Novick, 1973; Huynh, 1976a, 1980b; van der Linden & Mellenbergh, 1977).

Given various parameters defining the decision situation (such as θ_o; the number of test items; the losses incurred by misclassification; and, when available, prior information regarding the individual examinee or the group of examinees), a test passing score may be determined by minimizing either the average loss (Bayesian or empirical Bayes passing score) or the maximum loss (minimax passing score). For example, where classification errors are weighted equally (e.g., when the false positive loss and the false negative loss are identical), an optimum passing score may be determined by minimizing the sum of the probabilities of making such errors. Details regarding the determination of passing scores may be found in Huynh (1976a, 1980b).

Once a passing score has been set for a test, an obvious question concerns the extent to which the test itself contributes to the quality of the decision-making process. The question may be answered in a variety of ways. For example, if the test scores are used to identify students who need instructional remediation, then the detection of poor achievers (nonmasters) is important, and therefore a substantial false positive error rate may not be acceptable. In this context, a mastery test may be considered as effective or efficient if it yields a small false positive error rate. In most situations, however, some combination of false positive error, false negative error, and their corresponding losses would be desirable in assessing the efficiency of using test scores to make decisions regarding individual examinees.

2. REVIEW OF LITERATURE

The consideration of decision efficiency was introduced by Huynh (1975, 1976c) for the case involving constant losses. Let R_o be the expected loss associated with the best use of test data and R^*_{min} be the smallest expected loss encountered in the case of no relationship between true ability and test score. Huynh's efficiency coefficient, defined as $\varepsilon = 1-R_o/R^*_{min}$, was interpreted as the proportion of reduction in random loss which would result from the best use of test data in the decision-making process. Under fairly general conditions regarding the nature of test data, Huynh proved that ε was included between 0 and 1. The lower bound occurs when there is no relationship between test score and true ability; the upper bound is reached when there is a perfect increasing relationship between these two variables.

The concept of decision efficiency was later extended under a slightly different form by van der Linden and Mellenbergh (1978) and Mellenbergh and van der Linden (1979). These writers proposed the use of the coefficient $\delta = (R_n-R_B)/(R_n-R_c)$, which may be written equivalently as $\delta = 1 - (R_B-R_c)/(R_n-R_c)$, a form similar to Huynh's original ε. In these formulae, R_B represents the expected loss associated with any predetermined test passing score; R_c and R_n are the expected losses encountered in situations in which the test scores contain complete and no information about the true score, respectively. As shown by van der Linden and Mellenburgh, there is a direct relationship between δ and the classical reliability index when δ is computed for linear losses at the optimum test passing score. In addition, the two special values $\delta = 0$ and $\delta = 1$ have the same meaning as ε. However, van der Linden and Mellenbergh correctly stated that their proposed δ may not always be included between 0 and 1, as would be typically desirable in the formulation of indices to be used in educational and psychological measurement. Huynh (1980c) proposed a revised δ in which R_c represented the expected loss associated with the best use of completely infallible data and proved that $0 \leq \delta \leq 1$ under fairly general conditions. Wilcox (1978) had also advanced a modification of δ; his index γ ranged between 0 and 1. However, these boundary values of γ did not appear to bear direct interpretations in terms of the relationship between test scores and true ability.

Livingston and Wingersky (1979) proposed the assessment of the quality of pass/fail decisions (mastery testing) on the basis of the probabilities of making correct and incorrect decisions and on the basis of an efficiency index involving these probabilities and the corresponding utilities. The issue of errors in decisions has been considered at length in the literature (Hambleton & Novick, 1973; Huynh, 1976a; Wilcox, 1977). In addition, the Livingston-Wingersky index varies from -1 to +1, a range which often complicates the interpretation of the index. Estimates for the various quantities considered by these authors are based on the compound binomial model, which typically requires the responses of at least 1000 examinees. The requirement seems quite stringent in many cases involving field testing or the use of mastery tests. (Actually, as can be seen later in this paper, the Livingston-Wingersky index relates directly to the raw efficiency index ε_2; there is little difference between estimates of ε_2 based on the compound binomial and beta-binomial models.)

The purpose of this paper is to provide a general formulation of decision efficiency in mastery testing, to provide illustrations based on the beta-binomial model, to describe ways to estimate the proposed efficiency indices, and to report data comparing estimates based on the compound binomial and beta-binomial models.

Figure I provides the motivation for the general formulation of decision efficiency as presented in the subsequent section. Let us consider the simplest case in which the losses encountered by both the false positive and false negative errors are constant and equal (and are set at Q). With the cell probabilities p_{ij} as previously defined, the expected loss (Bayes risk) in using test scores to make decisions is equal to

$$R = Q(p_{01} + p_{10}) \, . \tag{1}$$

Let us presume now that there is no relationship between ability and test score x, hence mastery/nonmastery decisions are based on a random process independent of the examinee's ability. For this situation, the loss is expected to be

$$R_e = Q(p_{.1}p_{0.} + p_{.0}p_{1.}) \, . \tag{2}$$

This quantity will be referred to as random-decision risk. In addition, over all possible values for θ_o and c, the worst decision would occur when a true master is always denied mastery status and a true nonmaster is always granted mastery status. For these extreme situations, the risk stands at the maximum $R_m = Q$. Under fairly general conditions (see Section 3), it may be verified that $R \leq R_e$.

From the three expected losses R, R_e, and R_m, two efficiency indices may be formulated. First, $R_e - R$ represents the amount of reduction in the random-decision risk which could be achieved by using test data. Hence, an index of decision efficiency may be defined via the ratio

$$\varepsilon_1 = (R_e - R)/R_e \tag{3}$$

which is the extent to which the reliance on test scores will reduce the expected loss which would be encountered if no test data (or completely fallible data) were used in the decision situation defined by θ_o and c. From Equations (1) and (2), it may be deduced that

$$\varepsilon_1 = (P-P_c)/(1-P_c)$$

where $P = p_{00} + p_{11}$ and $P_c = p_{0.}p_{.0} + p_{1.}p_{.1}$. This index, ε_1, is actually the kappa index proposed by Cohen (1960) and studied

Figure I provides the motivation for the general formulation of decision efficiency as presented in the subsequent section. Let us consider the simplest case in which the losses encountered by both the false positive and false negative errors are constant and equal (and are set at Q). With the cell probabilities p_{ij} as previously defined, the expected loss (Bayes risk) in using test scores to make decisions is equal to

$$R = Q(p_{01} + p_{10}) . \tag{1}$$

Let us presume now that there is no relationship between ability and test score x, hence mastery/nonmastery decisions are based on a random process independent of the examinee's ability. For this situation, the loss is expected to be

$$R_e = Q(p_{.1}p_{0.} + p_{.0}p_{1.}) . \tag{2}$$

This quantity will be referred to as random-decision risk. In addition, over all possible values for θ_o and c, the worst decision would occur when a true master is always denied mastery status and a true nonmaster is always granted mastery status. For these extreme situations, the risk stands at the maximum $R_m = Q$. Under fairly general conditions (see Section 3), it may be verified that $R \leq R_e$.

From the three expected losses R, R_e, and R_m, two efficiency indices may be formulated. First, $R_e - R$ represents the amount of reduction in the random-decision risk which could be achieved by using test data. Hence, an index of decision efficiency may be defined via the ratio

$$\varepsilon_1 = (R_e - R)/R_e \tag{3}$$

which is the extent to which the reliance on test scores will reduce the expected loss which would be encountered if no test data (or completely fallible data) were used in the decision situation defined by θ_o and c. From Equations (1) and (2), it may be deduced that

$$\varepsilon_1 = (P-P_c)/(1-P_c)$$

where $P = p_{00} + p_{11}$ and $P_c = p_{0.}p_{.0} + p_{1.}p_{.1}$. This index, ε_1, is actually the kappa index proposed by Cohen (1960) and studied extensively in the context of mastery testing by Swaminathan, Hambleton, and Algina (1975) and Huynh (1976b, 1978, 1979a).

A second efficiency index may also be formulated, using R and R_m. It is

$$\varepsilon_2 = (R_m - R)/R_m . \qquad (4)$$

This index represents the extent to which the use of test scores will reduce the maximum risk which is common to all situations. From Equation (1), it may be verified that

$$\varepsilon_2 = p_{00} + p_{11} = P .$$

Thus ε_2 is simply the combined probability of making a correct decision. In the context of reliability of mastery tests, ε_2 (or P) is often referred to as the raw agreement index (Subkoviak, 1976; Huynh, 1979a).

With the rationale for ε_1 and ε_2 as stated, a general formulation of decision efficiency will now be presented.

4. A GENERAL FORMULATION OF DECISION EFFICIENCY

Let θ be the true ability of a given examinee and Ω be its range. For the binomial error model (Lord & Novick, 1968, ch. 23), θ may be taken as the proportion of items in a large item pool that the examinee is expected to answer correctly, and the range Ω is the interval $[0,1]$. Let x be the test score observed for the examinee, and let x be distributed according to the conditional density $f(x|\theta)$. In addition, let $p(\theta)$ be the density of θ.

A referral task (Huynh, 1976a) is assumed to exist and is used as an external criterion for the determination of a passing score. The task is defined operationally via a nondecreasing function $s(\theta)$ which describes the probability that an examinee with true ability θ will succeed in completing the task. As noted in the author's previous writing (Huynh, 1976a, 1980b), the referral task may be real or hypothetical. For example, in individualized instructional programs where a student proceeds from one content unit to the next (presumably more complex) unit, each succeeding unit may serve as a referral task for the previous unit. In other situations, where no hierarchy can be logically or empirically assumed to hold, a consensus on what constitutes an acceptable level of performance may be translated into a hypothetical referral task. To be specific, let us suppose that there exists a constant θ_o such that mastery is equivalent to the condition $\theta \geq \theta_o$ and nonmastery is described by the inequality $\theta < \theta_o$. The corresponding referral task is operationally defined by the nonincreasing function $s(\theta) = 0$ for $\theta < \theta_o$ and $s(\theta) = 1$ for $\theta \geq \theta_o$.

On the basis of the observed test score x and by relying on a decision rule c, the examinee will be classified in the mastery status (action a_1) or in the nonmastery status (action a_2). Let $C_f(\theta)$ be the opportunity loss incurred in granting mastery status to an examinee who will eventually fail to perform the referral task (a false positive error). Likewise, let $C_s(\theta)$ be the loss associated with the denial of mastery to someone who will succeed in completing the task (a false negative error). In most practical situations, action a_1 is taken when $x \geq c$, and action a_2 is taken where $x < c$. Here, the constant c is referred to as a test passing (cutoff, mastery) score.

Within the decision framework as stated, the expected loss (Bayes risk) associated with the passing score c is given as

$$R = \int_\Omega C_s(\theta)s(\theta)Pr(x<c|\theta)p(\theta)d\theta + \int_\Omega C_f(\theta)\{1-s(\theta)\}Pr(x\geq c|\theta)p(\theta)d\theta. \quad (5)$$

When the test score x is discrete, the integration sign in each of the two terms on the right side of (5) is to be replaced by the summation (Σ) sign. For the special 0-1 form for $s(\theta)$ as defined previously, the Bayes risk is given as

$$R = \int_{\theta_o}^{\infty} C_s(\theta)Pr(x<c|\theta)p(\theta)d\theta + \int_{-\infty}^{\theta_o} C_f(\theta)Pr(x\geq c|\theta)p(\theta)d\theta . \quad (6)$$

In both Equations (5) and (6), the two separate terms on the right define the individual Bayes risk for the false negative error and the false positive error.

Consider now the situation where test data do not reflect the ability of the examinees and therefore are useless in the decision-making process. For such a case, there would be no relationship between ability θ and test score x; in other words, θ and x would be independent of each other. The expected loss may now be written as

$$R_e = \left(\int_\Omega C_s(\theta)s(\theta)p(\theta)d\theta\right)\Pr(x<c) + \left(\int_\Omega C_f(\theta)\{1-s(\theta)\}p(\theta)d\theta\right)\Pr(x\geq c)\ , \tag{7}$$

and, for the special 0-1 case for $s(\theta)$, as

$$R_e = \left(\int_{\theta_o}^{\infty} C_s(\theta)p(\theta)d\theta\right)\Pr(x<c) + \left(\int_{\theta_o}^{\infty} C_f(\theta)p(\theta)d\theta\right)\Pr(x\geq c)\ . \tag{8}$$

Let $p = \Pr(x\geq c)$ so that $1-p = \Pr(x<c)$. Then for the situation in which no relationship exists between x and θ, the decision process is carried out by randomly assigning individuals to mastery and nonmastery categories according to the proportions p and 1-p, respectively. As in the previous section, the Bayes risk R_e will be referred to as the random-decision risk, or simply, random risk.

It may be verified from Equation (5) that the Bayes risk R cannot exceed the quantity

$$R_m = \int_\Omega C_s(\theta)s(\theta)p(\theta)d\theta + \int_\Omega C_f(\theta)\left(1-s(\theta)\right)p(\theta)d\theta\ . \tag{9}$$

This risk is encountered when mastery/nonmastery decisions based on test data are always incorrect, that is, a true master is always denied mastery status and a true nonmaster is always granted mastery status.

With the three risks R, R_e, and R_m as defined, the two decision efficiency indices ε_1 and ε_2 may now be written as

$$\varepsilon_1 = 1-R/R_e \tag{10}$$

and

$$\varepsilon_2 = 1-R/R_m\ . \tag{11}$$

Since ε_1 is a generalization of the corrected-for-chance kappa index, it seems appropriate to refer to it as the corrected-for-chance efficiency index. Likewise, with ε_2 as a general case of the raw agreement index, it may be referred to as the raw efficiency index.

Just as in the case of kappa and P, there are fundamental differences between ε_1 and ε_2. The ε_2 index is formulated on the basis of the baseline risk R_m which expresses the worst possible risk which could occur in the decision-making process. This risk is incurred when decisions regarding mastery/nonmastery are always incorrect. Thus ε_2 equals 1 when decisions are always correct and reaches the minimum 0 where decisions are always incorrect.

On the other hand, ε_1 assumes the random risk R_e to be the baseline risk and expresses the extent to which the use of test scores will reduce this random risk. As is the case of kappa, ε_1 reveals the magnitude by which the test scores will improve the effectiveness of the decision-making process beyond the level which could be expected from random classification. (The random assignment of examinees to the mastery and nonmastery categories, however, keeps intact the proportions of masters and of nonmasters as defined by the observed test score frequencies.) Thus ε_1 attains the maximum value of 1 when decisions are always correct. It will be equal to zero when the decision-making process is carried out by random classification (i.e., when test scores have no relationship with the ability of the examinees).

It should be clear from the above elaboration that decision efficiency depends not only on the characteristics of the test (as reflected in the dependency between x and θ), but also on the particular circumstances under which the test scores are used to make decisions regarding the individual examinees. Such circumstances are reflected in the referral success function $s(\theta)$, the two loss functions $C_s(\theta)$ and $C_f(\theta)$, and the prior or group ability density $p(\theta)$.

To complete this section, it may be noted that under all circumstances $0 \leq \varepsilon_2 \leq 1$ and $\varepsilon_1 \leq \varepsilon_2$. In addition, since the referral success function $s(\theta)$ enters in the definition of R and R_e, but not in that of R_m, it is expected that $s(\theta)$ will have more influence on ε_1 than on ε_2. Thus, in the simplest formulation of mastery testing which involves the true mastery score θ_o, this score θ_o will probably have more bearing on ε_1 than on ε_2.

5. CONDITIONS UNDER WHICH ε_1 IS POSITIVE

In the most general situation, ε_1 may be negative. This section will describe the conditions under which this index is positive.

From the definition of losses presented at the beginning of Section 3, it seems reasonable to assume that both $s(\theta)$ and $C_f(\theta)$

are nondecreasing and that $C_s(\theta)$ is nonincreasing. In fact, if the referral task is chosen appropriately, then examinees of higher ability should be more likely to succeed in performing the task than those of low ability. In addition, the denial of mastery status should cause less harm for subjects with low ability than for those with high ability. Likewise, granting mastery status to a low ability examinee would cause more harm than granting mastery to a high ability examinee. Thus, it seems sensible to assume that $C_s(\theta)s(\theta)$ is nondecreasing with respect to θ and that $C_f(\theta)\big(1-s(\theta)\big)$ is nonincreasing with respect to θ.

Now let us focus on the relationship between ability θ and test score x. If the test is reasonably well constructed, then the probability $\Pr(x<c|\theta)$ is nonincreasing in its argument θ. In other words, examinees with low ability are <u>more</u> likely to get low test scores than those with high ability. This assumption is tenable if the density $f(x|\theta)$ belongs to the monotone likelihood ratio (Esary, Proschan, & Walkup, 1967; Dykstra, Hewett, & Thompson, 1973). It follows from Theorem 1 of Dykstra <u>et al</u>. that

$$\int_\Omega C_s(\theta)s(\theta)\Pr(x<c|\theta)p(\theta)d\theta$$
$$< \Big(\int_\Omega C_s(\theta)s(\theta)p(\theta)d\theta\Big)\Big(\int_\Omega \Pr(x<c|\theta)p(\theta)d\theta\Big)\ . \tag{12}$$

The last integral is simply the unconditional probability $\Pr(x<c)$. By using the same theorem, it may be verified that

$$\int_\Omega C_f(\theta)\big(1-s(\theta)\big)\Pr(x\geq c|\theta)p(\theta)d\theta < \{\int_\Omega C_f(\theta)\big(1-s(\theta)\big)p(\theta)d\theta\}\Pr(x\geq c)\ . \tag{13}$$

It follows that, at each test passing score c, $R \leq R_c$, and hence $0 \leq \varepsilon_1 \leq 1$.

6. AN ILLUSTRATION BASED ON THE BETA-BINOMIAL MODEL WITH CONSTANT LOSSES AND 0-1 REFERRAL SUCCESS

Consider now the simple case in which the test score x obtained from the administration of an n-item test to a subject with ability θ is distributed according to the binomial density

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x = 0,1,\ldots,n\ . \tag{14}$$

In addition, let it be assumed that the subject comes from a population of examinees for whom the ability θ is distributed according to the beta density

$$p(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1 . \tag{15}$$

Then the unconditional distribution of the test score x is defined by the negative hypergeometric density

$$f(x) = \frac{\binom{n}{x} B(\alpha+x, n-x+\beta)}{B(\alpha,\beta)} . \tag{16}$$

Let θ_o be the minimum passing level in the ability continuum, and let $C_f(\theta) = 1$ and $C_s(\theta) = Q$. In other words, Q is the ratio of the constant loss due to a false negative error to the one produced by a false positive error. The two Bayes risks R and R_e may now be computed via the following formulae:

$$R = Pr(\theta<\theta_o, x\geq c) + Q\ Pr(\theta\geq\theta_o, x\leq c-1) \tag{17}$$

and

$$R_e = Pr(\theta<\theta_o)Pr(x\geq c) + Q\ Pr(\theta\geq\theta_o)Pr(x\leq c-1) . \tag{18}$$

The two probabilities listed in (17) may be obtained from tables of the incomplete beta function (Pearson, 1934), by use of the formulae presented in Huynh (1976a, p. 71), or from tables and a computer program documented in Huynh (1979b, 1980a). The two probabilities in Equation (18), on the other hand, may be secured by applications of the inductive formulae reported in Huynh (1976b). It may also be noted that $R_m = Pr(\theta<\theta_o) + Q\ Pr(\theta\geq\theta_o)$.

Numerical Example 1

Consider the situation in which a 10-item test is administered to a group of examinees and the resulting test scores have a mean of $\mu = 7.00$ and a KR21 index of $\alpha_{21} = .40$. From the formulae in Huynh (1976a), it may be deduced that the parameters defining the beta true ability are $\alpha = (-1 + 1/\alpha_{21})\mu = 10.5$ and $\beta = -\alpha + n/\alpha_{21} - n = 4.5$. Let $\theta_o = .60$, $c = 8$, and $Q = .50$. Then, by using the tables reported in Huynh (1979b), the rates of false positive error and of false negative error may be found to be

$$Pr(\theta<\theta_o, x\geq c) = .0173$$

and

$$Pr(\theta\geq\theta_o, x<c) = .3955 .$$

Hence the Bayes risk in using the test scores to make decisions is

$R = .0173 + .50 \times .3955 = .2151$. On the other hand, $Pr(x<c) = .5713$ and $Pr(\theta<\theta_o) = .1931$, and hence $R_e = .1931 \times .4287 + .50 \times .8069 \times .5713 = .3133$. In addition, $R_m = .1931 + .50 \times .8069 = .5966$. The decision efficiency indices are $\varepsilon_1 = 1-.1931/.3133 = .384$ and $\varepsilon_2 = 1-.2151/.5966 = .639$

7. DECISION EFFICIENCY FOR THE BETA-BINOMIAL MODEL WITH POWER LOSSES AND 0-1 REFERRAL SUCCESS

Consider now the beta-binomial model along with the special 0-1 referral success and the losses defined by

$$C_f(\theta) = (\theta_o-\theta)^{p_1} \text{ for } \theta < \theta_o \qquad (19)$$
$$= 0 \qquad \text{for } \theta \geq \theta_o$$

and

$$C_s(\theta) = Q(\theta-\theta_o)^{p_2} \text{ for } \theta \geq \theta_o \qquad (20)$$
$$= 0 \qquad \text{for } \theta < \theta_o.$$

Then, apart from the denominator $B(\alpha,\beta)$, the Bayes risk at the test passing score c is given as

$$R = Q \int_{\theta_o}^{1} (\theta-\theta_o)^{p_1}\theta^{\alpha-1}(1-\theta)^{\beta-1} \sum_{x=0}^{c-1} \binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta \qquad (21)$$
$$+ \int_0^{\theta_o} (\theta_o-\theta)^{p_2}\theta^{\alpha-1}(1-\theta)^{\beta-1} \sum_{x=c}^{n} \binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta .$$

Similarly, apart from the denominator $B(\alpha,\beta)$, the random-decision Bayes risk is given as

$$R_e = Q\left(\int_{\theta_o}^{1} (\theta-\theta_o)^{p_1}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta\right)\left(\sum_{x=0}^{c-1} f(x)\right) \qquad (22)$$
$$+ \left(\int_0^{\theta_o} (\theta_o-\theta)^{p_2}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta\right)\left(\sum_{x=c}^{n} f(x)\right),$$

and the maximum risk as

$$R_m = Q \int_{\theta_o}^{1} (\theta-\theta_o)^{p_1}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$$
$$+ \int_0^{\theta_o} (\theta_o-\theta)^{p_2}\theta^{\alpha}(1-\theta)^{\beta-1}d\theta . \qquad (23)$$

When p_1 (or p_2) is an integer such as in the case of linear or quadratic losses, the integrals in (21), (22), and (23) which involve p_1 (or p_2) may be computed via the incomplete beta function (Pearson, 1934) and the recurrence formula described as follows. Let

$$D(u,v;\theta_o) = \int_0^{\theta_o} r^{u-1}(1-t)^{v-1}dt \qquad (24)$$
$$= B(u,v)I(u,v;\theta_o) .$$

Then

$$D(u+1,v-1;\theta_o) = \left(-\theta_o^u(1-\theta_o)^{v-1} + uD(u,v;\theta_o)\right)/(v-1) . \qquad (25)$$

The computations for R, R_e, and R_m are simplified considerably when losses are of the linear form. The Bayes risk R of Equation (21) may now be written as

$$R = \frac{Q}{B(\alpha,\beta)}\int_{\theta_o}^{1}\left(\theta^{\alpha+1-1}(1-\theta)^{\beta-1}-\theta_o\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)\sum_{x=0}^{c-1}\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta$$

$$+ \frac{1}{B(\alpha,\beta)}\int_0^{\theta_o}\left(\theta_o\theta^{\alpha-1}(1-\theta)^{\beta-1}-\theta^{\alpha+1-1}(1-\theta)^{\beta-1}\right)\sum_{x=c}^{n}\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta.$$

Let $F_n(n,\alpha,\beta,\theta_o,c)$ and $F_p(n,\alpha,\beta,\theta_o,c)$ denote the false negative and false positive error rates associated with the beta true ability distribution with parameters α and β. By noting that

$$B(\alpha+1,\beta) = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} = \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha B(\alpha,\beta)}{\alpha+\beta},$$

it may be verified that the Bayes risk R is given as

$$R = Q\left(\frac{\alpha}{\alpha+\beta} F_n(n,\alpha+1,\beta,\theta_o,c) - \theta_o F_n(n,\alpha,\beta,\theta_o,c)\right) \qquad (26)$$
$$+ \theta_o F_p(n,\alpha,\beta,\theta_o,c) = \frac{\alpha}{\alpha+\beta} F_p(n,\alpha+1,\beta,\theta_o,c) .$$

Formulae, tables, and a computer program are available (Huynh, 1979a, 1980a) for the computation of the false positive and false negative error rates.

As for R_e and R_m, they may be expressed via the incomplete beta function as follows:

$$R_e = Q\left\{\frac{\alpha}{\alpha+\beta}\left(1-I(\alpha+1,\beta;\theta_o)\right) - \theta_o\left(1-I(\alpha,\beta;\theta_o)\right)\right\} \qquad (27)$$
$$\cdot \left(\sum_{x=0}^{c-1} f(x)\right) + \left(\theta_o I(\alpha,\beta;\theta_o) - \frac{\alpha}{\alpha+\beta} I(\alpha+1,\beta;\theta_o)\right) \cdot \left(\sum_{x=c}^{n} f(x)\right),$$

and

$$R = Q\left\{\frac{\alpha}{\alpha+\beta}[1-I(\alpha+1,\beta;\theta_o)] - \theta_o[1-I(\alpha,\beta;\theta_o)]\right\} \tag{28}$$

$$+ \theta_o I(\alpha,\beta;\theta_o) - \frac{\alpha}{\alpha+\beta} I(\alpha+1,\beta;\theta_o) .$$

Numerical Example 2

For the basic data described in the first numerical example, the use of linear losses ($p_1 = p_2 = 1$) will result in the Bayes risks $R = .02165$, $R_e = .03865$, and $R_m = .07118$. Hence the values of the efficiency indices are $\varepsilon_1 = 1-.02165/.03865 = .440$ and $\varepsilon_2 = 1-.02165/.07118 = .696$.

8. RELATIONSHIP BETWEEN ε_2 AND THE LIVINGSTON-WINGERSKY EFFICIENCY INDEX

Recently, Livingston and Wingersky (1979) proposed an index of efficiency for situations in which the consequences of granting or denying mastery status are expressed in terms of utility. For the simplest case involving linear and opposite utility, the utility of granting mastery status is $\theta-\theta_o$ and the utility of denying mastery status is $\theta_o-\theta$. Here θ is the true ability of the examinee, and θ_o is a given constant. As before, let x be the observed test score and c be the test passing score. The efficiency index proposed by Livingston and Wingersky (1979) is the ratio

$$e = \frac{\Sigma(\theta-\theta_o)\text{sign}(x-c)}{\Sigma|\theta-\theta_o|} \tag{29}$$

where the summation sign (Σ) is extended over all examinees. This index reaches the maximum value of 1 when decisions based on test data are always correct and the minimum value of -1 when these decisions are always incorrect.

We will show that a linear relationship exists between the Livingston-Wingersky efficiency index e and the raw efficiency index ε_2 computed from the corresponding (opportunity) loss functions. These loss functions are expressed as

$$C_f(\theta) = 2(\theta_o-\theta) \quad \text{for} \quad \theta < \theta_o$$

$$= 0 \qquad\qquad \text{for} \quad \theta \geq \theta_o ,$$

and

$$C_s(\theta) = 2(\theta-\theta_o) \quad \text{for} \quad \theta \geq \theta_o$$
$$= 0 \qquad\qquad \text{for} \quad \theta < \theta_o \, .$$

Then the raw efficiency index ε_2 is given as

$$\varepsilon_2 = \frac{\sum\limits_{\theta \geq \theta_o} \sum\limits_{x \geq c} (\theta-\theta_o) + \sum\limits_{\theta < \theta_o} \sum\limits_{x < c} (\theta_o-\theta)}{\Sigma|\theta-\theta_o|} \, . \tag{30}$$

With the losses as defined, it will now be shown that $e = 2\varepsilon_2-1$. In fact, apart from the denominator $\Sigma|\theta-\theta_o|$, the quantity $2\varepsilon_2-1$ is equal to

$$2 \sum_{\theta \geq \theta_o} \sum_{x \geq c} (\theta-\theta_o) + 2 \sum_{\theta < \theta_o} \sum_{x < c} (\theta_o-\theta) -$$
$$\left(\sum_{\theta \geq \theta_o} \sum_{x \geq c} (\theta-\theta_o) + \sum_{\theta \geq \theta_o} \sum_{x < c} (\theta-\theta_o) + \sum_{\theta < \theta_o} \sum_{x \leq c} (\theta_o-\theta) + \sum_{\theta < \theta_o} \sum_{x \geq c} (\theta_o-\theta) \right)$$
$$= \sum_{x \geq c} \left(\sum_{\theta \geq \theta_o} (\theta-\theta_o) + \sum_{\theta < \theta_o} (\theta-\theta_o) \right) - \sum_{x < c} \left(\sum_{\theta \geq \theta_o} (\theta-\theta_o) + \sum_{x \geq \theta_o} (\theta-\theta_o) \right)$$
$$= \Sigma(\theta-\theta_o)\text{sign}(x-c) \, .$$

This quantity defines the numerator of the Livingston-Wingersky efficiency index. Thus the relationship $e = 2\varepsilon_2-1$ holds for linear and opposite utilities. For other opposite utilities which define the Livingston-Wingersky general index of efficiency, and with the corresponding (opportunity) loss functions, it may also be verified that the same relationship will hold.

As a passing remark to end this section, it may be noted that Livingston and Wingersky (1979, p. 258) appear to imply that "if examinees' chances of passing the test were completely unrelated to their true scores, the efficiency index would have an expected value of zero." This assertion regarding e apparently is not complete, as may be seen from the following argument. If there is complete independence between true ability θ and observed score x, then it may be verified that at each given pair (θ_o, c), the numerator of e in (26) is given as

$$\Sigma(\theta-\theta_o)\text{sign}(x-c) = \left(\Sigma(\theta-\theta_o)\right)\Pr(x \geq c) - \left(\Sigma(\theta-\theta_o)\right)\Pr(x < c) \, .$$

Hence, when $\Sigma(\theta-\theta_o) \neq 0$, e is 0 if and only if the test passing score c is set up such that half of the subjects will pass and the other half will fail. (This observation also holds for situations in which the action of granting mastery and the action of denying mastery have opposite utilities other than opposite linear ones.)

9. ESTIMATION PROCEDURES BASED ON THE BETA-BINOMIAL AND COMPOUND BINOMIAL ERROR MODELS

The estimation of the decision efficiency indices ε_1 and ε_2 may be carried out on the basis of the observed test data if reasonable assumptions can be made regarding the functional forms of the conditional probability $Pr(x<c|\theta)$ and of the density $p(\theta)$ of the true ability.

When the beta-binomial error model (Lord & Novick, 1968, ch. 23) is appropriate, the estimation of decision efficiency under constant or power losses may be carried out via the formulae described in Sections 6 and 7. In using these formulae, the parameters α and β of the beta distribution are to be replaced by their corresponding estimates based on sample data. A commonly used set of estimates is the moment estimates which are obtained as follows. Let $\bar{x}$ and s be the mean and standard deviation of the test scores, and let the KR21 reliability be defined as

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left[1 - \frac{\bar{x}(n-\bar{x})}{ns^2}\right]. \tag{31}$$

Then the moment estimates of α and β are given as

$$\hat{\alpha} = (-1 + 1/\hat{\alpha}_{21})\bar{x} \tag{32}$$

and

$$\hat{\beta} = -\hat{\alpha} + n/\hat{\alpha}_{21} - n . \tag{33}$$

While the beta-binomial model has been found to fit several test score distributions reasonably well (Keats & Lord, 1962; Duncan, 1974), and to provide useful results in mastery testing (Huynh, 1976a, 1976b, 1977, 1979, 1980a), the compound binomial error model (Lord, 1965, 1969) has been advocated as a more realistic model for the description of actual test data. Livingston and Wingersky (1979) used the latter model to obtain estimates for the false positive and false negative error rates, estimates for decision accuracy (proportion of examinees who are correctly classified), and estimates of the decision efficiency index e under linear and opposite utilities. A basic feature of the estimation process is the use of Lord's Method 20 (Lord, 1969) as implemented by Wingersky, Lees, Lennon, and Lord (1969). Its use is recommended for data from at least 1000 examinees.

In small-scale testing programs such as those associated with field testing for mastery tests or those conducted at the school-district level, the requirement of 1000 examinees cannot be easily fulfilled. In addition, the data presented in Wilcox (1977) seem to indicate that as far as error rates (and therefore efficiency under constant losses) are concerned, the use of the more complex compound binomial model instead of the simple beta-binomial model does not improve substantially the accuracy of the estimates.

This section will compare estimates of ε_2 based on the beta-binomial model with those computed from the compound binomial model as implemented by Livingston and Wingersky (1979). (These authors proposed the use of the index e which is $2\varepsilon_2-1$.) For the case of constant and equal losses, the estimate for ε_2 is simply the sum of the two probabilities of making a correct decision. Hence, in using the output described by Livingston and Wingersky, the compound binomial estimate for ε_2 may be obtained by summing the probabilities which appear in the two cells "Should Pass/Will Pass" and "Should Fail/Will Fail." For the first output reported in Figure 1 of the Livingston-Wingersky paper, this estimate is 55.9% + 24.3% = 80.2% or .802. The output also reports the compound binomial estimate for the efficiency index e under linear and opposite utilities. The (raw) efficiency index ε_2, in turn, may be deduced from e via the formula $\varepsilon_2 = (1+e)/2$. For the output just referenced, the value of e is 0.81, hence the estimate for ε_2 is $(1+0.81)/2 = .905$.

The compound binomial estimates for efficiency index ε_2 under constant and linear losses with $Q = 1$ (or under constant and linear, but opposite utilities) were derived from the computer programs provided by Livingston and Wingersky. The corresponding estimates based on the beta-binomial model were obtained via the computer program listed in Appendix A. The comparison of the two sets of estimates was made using the basic test data summarized in Table 1. These data were extracted from the Comprehensive Tests of Basic

Skills data file collected in the 1978 South Carolina statewide testing program. In this table, s_d^2 represents the variance of the item difficulty (defined as the proportion of examinees who correctly answered the item).

TABLE 1

Description of Test Data Used to Compare the Beta-Binomial and Compound Binomial Estimates of ε_2

Case	n	Mean	S.D.	$s^2_{diff}(\times 10^4)$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}_{21}$
A	10	7.2315	2.6888	64.87	1.7693	0.6774	.8034
B	15	8.6247	3.1932	301.61	3.9433	2.9148	.6862
C	20	16.1621	3.8987	97.93	3.1278	0.7427	.8379
D	30	18.0707	6.3192	202.90	3.2300	2.1323	.8484
E	40	23.5658	8.3406	281.87	3.1258	2.1799	.8829
F	50	30.4848	10.7558	205.92	2.8152	1.8022	.9155

Table 2 reports the estimates of ε_2 for a variety of combinations of θ_o and c. The data reveal only negligible discrepancies between the beta-binomial estimates and those based on the compound binomial model. Since the beta-binomial estimates only require estimation of the two parameters of the beta distribution, they may be safely obtained from the responses of a small or moderate sample of examinees. For a sample of this type, estimation via the compound binomial model may not be appropriate.

TABLE 2

Estimates of ε_2 Based on the Beta-Binomial (BB) and Compound Binomial (CB) Models

Case	θ_o	c	Opposite & Constant Utility		Opposite & Linear Utility	
			BB	CB	BB	CB
A	.70	7	.874	.893	.948	.950
B	.70	10	.792	.798	.898	.905
C	.70	14	.912	.923	.972	.975
D	.80	24	.901	.906	.977	.980
E	.80	32	.920	.917	.985	.985
F	.80	40	.925	.934	.987	.990

10. COMPUTER PROGRAM

A FORTRAN IV program which provides an analysis of decision efficiency for the case of constant and linear losses is listed in Appendix A. For each problem, the input data are to be "keypunched" on three cards detailed as follows.

First Card

This card contains the title of the problem, keypunched between columns 1 and 80.

Second Card

This card provides data on number of items (n), the alpha (α) and beta (β) parameters of the true ability distribution, the true mastery score (θ_o), the test passing score (c), and the loss ratio (Q). These must be keypunched according to the format (I5, 2F10.5, F5.3, I5, F5.2).

For example, the efficiency analysis described in numerical examples 1 and 2 may be performed via the computer program using the following two input cards.

```
                         1    1    2    2    3    3    4
Column:        1...5... 0....5....0....5....0....5....0
First card:    AN EXAMPLE OF DECISION EFFICIENCY ANALYSIS
Second card:      10  10.5       4.5       .60    8  .50
```

Table 3 lists the output for this problem.

Several problems may be performed in one run by stacking the input cards together.

11. SUMMARY

This paper describes two indices which pertain to the efficiency of decisions in mastery testing. The indices are generalizations of the raw agreement index and the kappa index. Both express the reduction in proportion of losses (or the gain in proportion of utility) resulting from the use of test scores to make decisions. Empirical data reveal only negligible discrepancies between the beta-binomial and compound binomial estimates for these indices.

TABLE 3

An Output of the Computer Program

```
ANALYSIS OF DECISION EFFICIENCY BASED ON THE
BETA-BINOMIAL MODEL. THE TITLE OF THIS PROBLEM IS:
AN EXAMPLE OF DECISION EFFICIENCY ANALYSIS
INPUT DATA ARE:

    NUMBER OF ITEMS .....          10
    ALPHA ...............    10.50000
    BETA ................     4.50000
    THETA ZERO ..........     0.60000
    TEST PASSING SCORE ..           8
    LOSS RATIO Q ........     0.50000

FOUR-CELL TABLE WITH PROBABILITIES

    SHOULD FAIL AND WILL FAIL .....      0.1758
    SHOULD PASS AND WILL PASS .....      0.4113
    SHOULD FAIL BUT WILL PASS
    (A FALSE POSITIVE ERROR) ......      0.0173
    SHOULD PASS BUT WILL FAIL
    (A FALSE NEGATIVE ERROR) ......      0.3955

FOR LINEAR LOSSES, THE OUTPUT ARE:

    RISK FOR USING TEST SCORES ..   0.02165
    RANDOM-DECISION RISK ........   0.03865
    MAXIMUM RISK ................   0.07118

DECISION-EFFICIENCY INDICES:

    CORRECTED-FOR-CHANCE INDEX ... E1 =  0.440
    NO CORRECTION FOR CHANCE
    (RAW) INDEX .................. E2 =  0.696

** NORMAL END OF PROGRAM **
   PROGRAM WRITTEN BY
   HUYNH HUYNH
   COLLEGE OF EDUCATION
   UNIVERSITY OF SOUTH CAROLINA
   COLUMBIA, SOUTH CAROLINA 29208
   MAY 1980
```

BIBLIOGRAPHY

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

Duncan, G. T. (1974). An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association 69, 50-57.

Dykstra, R. L., Hewett, J. E. & Thompson, W. A., Jr. (1973). Events which are almost independent. Annals of Statistics 1, 674-681.

Esary, J. D. & Proschan, F. (1970). A reliability bound for systems of maintained, interdependent components. Journal of the American Statistical Association 65, 329-338.

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

Huynh, H. (1975). Statistical problems in binary classification. Paper presented at the annual meeting of the American Statistical Association, Atlanta.

Huynh, H. (1976a). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

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

Huynh, H. (1976c). On mastery scores and efficiency of criterion-referenced tests when losses are partially known. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Huynh, H. (1978). Reliability of multiple classifications. Psychometrika 43, 317-325.

Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics 4, 231-246.

Huynh, H. (1980a). Statistical inference for false positive and false negative error rates in mastery testing. Psychometrika 45, 107-120.

Huynh, H. (1980b). A nonrandomized minimax solution for passing scores in the binomial error model. Psychometrika 45, 167-182.

Huynh, H. (1980c). A note on decision-theoretic coefficients for tests. Research Memorandum 80-5, Publication Series in Mastery Testing. University of South Carolina College of Education.

Keats, J. A. & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika 27, 59-72.

Livingston, S. A. & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. Journal of Educational Measurement 16, 247-260.

Lord, F. M. (1965). A strong true score theory, with applications. Psychometrika 30, 239-270.

Lord, F. M. (1969). Estimating true-score distribution in psychological testing (an empirical Bayes estimation problem). Psychometrika 34, 259-299.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Co.

Mellenbergh, G. J. & van der Linden, W. J. (1979). The internal and external optimality of decisions based on tests. Applied Psychological Measurement 3, 257-273.

Pearson, K. (1934). Tables of the incomplete beta function. Cambridge: University Press.

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

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

van der Linden, W. J. & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. Applied Psychological Measurement 1, 593-599.

van der Linden, W. J. & Mellenbergh, G. J. (1978). Coefficients for tests from a decision theoretic point of view. Applied Psychological Measurement 2, 119-134.

Wilcox, R. R. (1977). Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics 2, 289-307.

Wilcox, R. R. (1978). A note on decision theoretic coefficients for tests. Applied Psychological Measurement 2, 609-61 .

Wingersky, M. S., Lees, D. M., Lennon, V. & Lord, F. M. (1969). A computer program for estimating true-score distributions and graduating observed-score distributions. Research Memorandum 69-4. Princeton, New Jersey: Educational Testing Service.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance of Joseph C. Saunders is gratefully acknowledged.

APPENDIX A

A Computer Program for the Analysis of the Efficiency of Decisions in Mastery Testing Based on the Beta-Binomial Model

Disclaimer: The computer program hereafter listed has been written with care and tested extensively under a variety of conditions using tests with 50 or fewer items. The author, however, makes no warranty as to its accuracy and functioning, nor shall the fact of its distribution imply such warranty.

```
C     A COMPUTER PROGRAM FOR THE COMPUTATION OF DECISION-EFFFICIENCY
C     WITH CONSTANT OR LINEAR LOSSES AND WITH BETA-BINOMIAL TEST DATA.
C     CONSTANT LOSSES INCLUDE CONSTANT UTILITIES, AND LINEAR LOSSES
C     INCLUDE LINEAR AND OPPOSITE UTILITIES.
C
C     INPUT DATA ARE:
C
C     FIRST CARD: TITLE OF THE PROBLEM (ENTER ANYTHING YOU WANT)
C
C     SECOND CARD: ENTER THE FOLLOWING INFORMATION, USING THE FORMAT
C     (I5,2F10.5,F5.2,I5,F5.2)
C        N .... NUMBER OF TEST ITEMS
C        A .... ALPHA PARAMETER OF THE BETA DISTRIBURION
C        B .... BETA DISTRIBUTION OF THE BETA DISTRIBUTION
C        TT ... THETA ZERO (MINIMUM TRUE SCORE FOR PASSING)
C        IM ... TEST PASSING SCORE
C        Q .... LOSS RATIO
C
C     SEVERAL PROBLEMS MAY BE RUN CONSECUTIVELY BY STACKING THE INPUT
C     CARDS TOGETHER.
C
C     SUBROUTINE REQUIRED: THE BDTR OF THE SCIENTIFIC SUBROUTINE
C     PACKAGE.
C
      DOUBLE PRECISION A,B,TT,FP,FN,FP1,FN1,SUM
      DIMENSION W(20)
    1 READ(5,100,END=99) W
  100 FORMAT(20A4)
      WRITE(6,200) W
  200 FORMAT('1','ANALYSIS OF DECISION EFFICIENCY BASED ON THE'/
     *T2,'BETA-BINOMIAL MODEL. THE TITLE OF THIS PROBLEM IS:'/T2,20A4)
      READ(5,110) N,A,B,TT,IM,Q
  110 FORMAT(I5,2F10.5,F5.2,I5,F5.2)
C
      WRITE(6,230) N,A,B,TT,IM,Q
  230 FORMAT(T2,'INPUT DATA ARE:'//
     *        T6,'NUMBER OF ITEMS .....',I10/
     *        T6,'ALPHA................',F10.5/
     *        T6,'BETA ................',F10.5/
     *        T6,'THETA ZERO ..........',F10.5/
     *        T6,'TEST PASSING SCORE...',I10/
     *        T6,'LOSS RATIO Q ........',F10.5//)
      CALL ERRFPN(N,A,B,TT,IM,FP,FN)
      CALL ERRFPN(N,A+1.D0,B,TT,IM,FP1,FN1)
      CALL MDBETA(TT,A,B,P1,IER)
      CALL MDBETA(TT,A+1.D0,B,P2,IER)
      ZZ=A/(A+B)
      R=Q*(ZZ*FN1-TT*FN)+TT*FP-ZZ*FP1
      AA=Q*(ZZ*(1.-P2)-TT*(1.-P1))
      BB=TT*P1-ZZ*P2
      RM=AA+BB
      CALL NEHY3(N,A,B,IM,SUM)
      RE=AA*SUM+BB*(1.-SUM)
      E1=1.-R/RE
      E2=1.-R/RM
      P1=SUM-FN
      P2=1.-SUM - FP
      WRITE(6,236) P1,P2,FP,FN
  236 FORMAT(T2,'FOUR-CELL TABLE WITH PROBABILITIES'//
     *        T6,'SHOULD FAIL AND WILL FAIL .....',F10.4/
     *        T6,'SHOULD PASS AND WILL PASS .....',F10.4/
     *        T6,'SHOULD FAIL BUT WILL PASS '/
     *        T6,'(A FALSE POSITIVE ERROR).......',F10.4/
     *        T6,'SHOULD PASS BUT WILL FAIL'/
     *        T6,'(A FALSE NEGATIVE ERROR) ......',F10.4//
     *        T2,'FOR LINEAR LOSSES, THE OUTPUT ARE:'//)
```

```
      WRITE(6,240) R,RE,RM,E1,E2
  240 FORMAT(T6,'RISK FOR USING TEST SCORES...',F10.5/
     *       T6,'RANDOM-DECISION RISK ........',F10.5/
     *       T6,'MAXIMUM RISK ................',F10.5//
     *       T2,'DECISION-EFFICIENCY INDICES:'//
     *       T6,'CORRECTED-FOR-CHANCE INDEX ... E1 = ',F6.3/
     *       T6,'NO CORRECTION FOR CHANCE'/
     *       T6,'(RAW) INDEX .................. E2 = ',F6.3)
      GOTO 1
   99 WRITE(6,150)
  150 FORMAT(T2,'** NORMAL END OF PROGRAM **'/
     *       T2,'    PROGRAM WRITTEN BY'/
     *       T2,'    HUYNH HUYNH'/
     *       T2,'    COLLEGE OF EDUCATION'/
     *       T2,'    UNIVERSITY OF SOUTH CAROLINA'/
     *       T2,'    COLUMBIA, SOUTH CAROLINA 29208'/
     *       T2,'    MAY 1980')
      STOP
      END
      SUBROUTINE ERRFPN(N,A,B,TT,IM,FP,FN)
      DOUBLE PRECISION A,B,TZ,BETA,DFCT,U,V,DX,ONE,Y1,
     *VMONE,BB,DF(61),FP,FN,
     *E(2),TT,P1,BA,BI
      EXTERNAL BETA,BI,DFCT
C
      ONE=1.D0
      Y1=BETA(A,B)
C     SET UP FOR FALSE POSTITIVE ERRORS
      TZ=TT
      IC=IM
      U=A+DFLOAT(IC)
      V=B+DFLOAT(N-IC)
      DO 40 L=1,2
C
      F=ONE-TZ
      DX=DFCT(U,V,TZ)
      BB=BI(N,IC)
      E(L)=DX*BB
C
      BA=BETA(U,V)
C
      IF(IC.EQ.N) GO TO 30
C
   10 IZ=N-IC
      DO 15 I=1,IZ
      IX=IC+I
      VMONE=V-ONE
      Z1=-(TZ**U)*F**VMONE
C
      DX=(Z1+U*DX)/VMONE
C
      BB=BB*(N-IX+1)/IX
C
      V=V-ONE
      BA=BA*U/V
C
      U=U+ONE
C
      E(L)=E(L)+BB*DX
   15 CONTINUE
   30 IF(L.EQ.1) GOTO 35
C
C       INTERCHANGE DFPA AND DFPB FOR FALSE NEGATIVE ERROR
C
   35 E(L)=E(L)/Y1
C       SET UP FOR FALSE NEGATIVE ERRORS
```

```
      TZ=ONE-TT
      IC=N-IM+1
      U=B+DFLOAT(IC)
      V=A+DFLOAT(N-IC)
C
   40 CONTINUE
C
      FP=E(1)
      FN=E(2)
C
      RETURN
      END
C
      DOUBLE PRECISION FUNCTION BI(N,M)
      BI=1
      IF(M*(N-M).EQ.0) GOTO 20
      MM=N
      IF(N.GT.(N-M)) MM=N-M
      DO 15 J=1,MM
   15 BI=BI*(N-J+1)/J
   20 RETURN
      END
C
      SUBROUTINE NEHY3(N,A,B,IM,SUM)
      DOUBLE PRECISION A,B,F,Z1,Z2,SUM
      Z1=DFLOAT(N)+B
      Z2=Z1+A
      K=0
      F=1.D0
      DO 5 I=1,N
    5 F=F*(Z1-DFLOAT(I))/(Z2-DFLOAT(I))
      SUM=F
   10 KP1=K+1
      IF(KP1.GE.IM) RETURN
      F=F*DFLOAT(N-K)*(A+DFLOAT(K))/
     *                          (DFLOAT(KP1)*(Z1-DFLOAT(KP1)))
      SUM=SUM+F
      K=K+1
      GOTO 10
      END
C
      DOUBLE PRECISION FUNCTION DFCT(A,B,TZ)
      EXTERNAL BETA
      DOUBLE PRECISION A,B,TZ,BETA
C
      CALL MDBETA(TZ,A,B,P,IER)
C
      IF(IER.NE.0) WRITE(6,100)A,B,TZ,IER
  100 FORMAT('0',' ERROR IN BDTR, A B TZ IER ARE ',3F20.10,I5)
      DFCT=DBLE(P)*BETA(A,B)
      RETURN
      END
      DOUBLE PRECISION FUNCTION BETA(X,Y)
      DOUBLE PRECISION A,B,CON,X,Y,F
      F=5.D0
      A=X
      B=Y
      CON=1.D0
      IF(A.LE.F) GOTO 2
    1 A=A-1.D0
      CON=CON*A/(A+B)
      IF(A.LE.F) GOTO 2
      GOTO 1
    2 IF(B.LE.F) GOTO 4
    3 B=B-1.D0
      CON=CON*B/(A+B)
```

```
      IF(B.LE.F) GOTO 4
      GOTO 3
    4 BETA=DGAMMA(A)*DGAMMA(B)/DGAMMA(A+B)*CON
      RETURN
      END
C
      SUBROUTINE MDBETA(X,A,B,P,IER)
      DOUBLE PRECISION A,B,X,BETA
      EXTERNAL BETA
      IF(A.GT..5 .AND. B.GT..5) GOTO 10
      IF(A.GT..5 .AND. B.LT..5) GOTO 20
      IF(A.LT..5 .AND. B.GT..5) GOTO 30
C             OTHERWISE BOTH A AND B ARE SMALLER THAN .5
      AA=A+1.
      BB=B+1.
      XX=X
      CALL BDTR(XX,AA,BB,P,D,IER)
      P=X**A*(1.D0-X)**B/(A*BETA(A,B))+X**B*(1.D0-X)**(A+1.D0)/
     *                          (B*BETA(A+1.D0,B)) + P
      RETURN
   10 AA=A
      BB=B
      XX=X
      CALL BDTR(XX,AA,BB,P,D,IER)
      RETURN
   20 AA=A
      BB=B+1.
      XX=X
      CALL BDTR(XX,AA,BB,P,D,IER)
      P=X**B*(1.D0-X)**A/(B*BETA(A,B))+ P
      RETURN
   30 AA=A+1.
      BB=B
      XX=X
      CALL BDTR(XX,AA,BB,P,D,IER)
      P=X**A*(1.D0-X)**B/(A*BETA(A,B)) + P
      RETURN
      END
```

PART SIX

TEST SENSITIVITY

16

ASSESSING TEST SENSITIVITY IN MASTERY TESTING

Huynh Huynh

University of South Carolina

A preliminary version of this paper was presented as part of the symposium "Approaches to test design for the assessment of the effectiveness of educational programs" sponsored by the American Educational Research Association at its annual meeting in Boston, April 7-11, 1980.

ABSTRACT

This paper addresses the concept of test sensitivity within the context of mastery testing. It is argued that correlation-based indices may not be appropriate for the assessment of test sensitivity. Global assessment of test sensitivity may be carried out via indices such as p-max or δ-max. Local measures of sensitivity may be described via a two-parameter logistic model. Procedures are described to check the tenability of test sensitivity on the basis of observed test data.

1. INTRODUCTION

Educational tests which are used for student or program evaluation are often described using terms such as "criterion-referenced," "domain-referenced," or "mastery" tests (Harris, Alkin, and Popham, 1974; Berk, 1980). It is important to note, however, that these different labels often refer to different aspects of the same process; depending on the context, all might be used to describe the same test. For example, test items can be deliberately constructed (or selected from an item bank) to reflect specific educational

This paper has been distributed separately as RM 80-7, August, 1980.

objectives; the resulting test scores are referenced to these objectives for interpretation and may then be used to assess the competency or mastery status of the individual student with respect to each of the objectives. For reasons of specificity, the term mastery testing will be used in this paper. By mastery testing, it is meant that, at the end of the testing process test scores are used to make decisions regarding the individual student. In most testing for instructional purposes (such as formative testing or basic skills assessment programs) and for certification (in the professions or in minimum competency testing programs), there are two decision categories based on test scores, namely mastery and nonmastery. Students with high test scores are granted mastery status (in the domain of performances or educational objectives underlying the test) and perhaps are permitted to move to a more advanced or complex instructional unit. Other students with low scores will be placed in the nonmastery category and will perhaps be provided with the opportunity of remedial instruction.

In the light of the above discussion, it appears clear that a mastery test is most useful if it can differentiate students who have mastered the educational objectives from those who have not. The extent to which the test fulfills this specific requirement will be referred to as instructional sensitivity (Harris, 1977; Haladyna and Goid, 1980). Of course, the concept of test sensitivity cannot be detached from the unique purposes and/or circumstances for which the test scores are to be used.

Another situation in which the concept of test sensitivity is called upon involves the use of test scores for admission or placement purposes. Here, decisions are made on whether or not the test scores show sufficient evidence that the student or applicant has the prerequisite skills or knowledge for a successful performance in the training or instructional program. For example, admission to a statistics course may require a minimal level of performance in arithmetic; hence arithmetic test scores may be used as a criterion for admission to such a course. In this case, test sensitivity may be framed within the context of predictive validity; a test may be said to be sensitive to the content of a course to the extent that test scores can separate those who, given effective instruction, will succeed in the course from the others who will not.

The purpose of this paper is to address the concept of test sensitivity within the context of mastery testing (Huynh, 1976), and to propose new ways to assess the degree to which a test is sensitive to the particular purpose for which it is intended.

2. POSSIBLE MISUSE OF CORRELATION TO ASSESS TEST SENSITIVITY

A variety of designs has been proposed to assess test sensitivity. Most involve the use of two contrasting groups of test scores. For example, a pretest-posttest design may be in order if there are reasons to assume that instruction is effective. In other words, a mastery test is given prior to instruction and again at the completion of instruction. The mastery test is sensitive to the instructional objectives to the extent that the distribution of pretest scores and that of posttest scores can be separated from each other. Another contrasting groups design involves the use of an uninstructed group and an instructed group. This design is appropriate for a test to be used to admit students to a course; in this case the instructed group would consist of students who have successfully completed the course and the uninstructed group would be formed of students who have failed the course.

How should test sensitivity be assessed on the basis of the separation between the test score distributions of the two contrasting groups? Is the point biserial correlation an appropriate choice for test sensitivity? (The reader may note that this correlation may be obtained by assigning the dummy code $X = 0$ to the lower score group and $X = 1$ to the higher score group and then by computing the Pearson correlation between X and the test scores.) Correlation, typically, is influenced by the variability in the test scores, yet test score variation usually does not play a major role in mastery testing (Millman and Popham, 1974). To substantiate this point, let a mastery test be such that all pretest scores are below the score of 20 and all posttest scores are above this score 20. Then, for classification purposes, a passing score of 20 would be selected. It should take no imagination to see that the test is completely sensitive (i.e., completely separates the pretest score distribution from the posttest score distribution). Yet, the point biserial correlation between the dummy code X and the test scores will change according to the means and standard deviations of the pretest and posttest scores. Following are two examples based on contrasting groups of ten subjects each.

Pretest		Posttest		Point
Mean	S.D.	Mean	S.D.	Biserial
14.10	2.21	23.00	2.68	.88
10.40	5.52	31.00	12.05	.74

3. A SIMPLE ALTERNATIVE TO POINT BISERIAL CORRELATION

The above numerical illustration clearly indicates that the use of point biserial correlation (or of similar indices) may not be appropriate if the distribution of the pretest scores or that of the posttest scores shows a large degree of variability. Unfortunately, it is a common experience that the pretest scores tend to show substantial variation. This is probably true for the case involving an uninstructed group, as well. (This occurs mainly because of random guessing and differences in input student characteristics.)

Thus, alternatives to point biserial correlation may be needed to assess test sensitivity in the use of test scores to make educational decisions. There are a variety of ways to approach the issue. For example, something like the maximum raw agreement index (p-max) may be appropriate. This index is very simple to conceptualize and to compute. At each possible cutoff score, compute the raw agreement index p between the grouping categories (pretest versus posttest, uninstructed versus instructed) and the decisions based on the test data (nonmastery versus mastery). Then search for the maximum of these raw agreement indices. This maximum p value corresponds to the situation in which the test scores are put to the best use. For both data sets in the previous illustration, the maximum of p (or p-max) is exactly 1.

FIGURE I

Configuration of Decisions Based on Contrasting-Group Data

Contrasting groups \ Test data	Nonmastery	Mastery	Marginal sum
Posttest (instructed) i=1	n_{10}	n_{11}	n_1
Pretest (uninstructed) i=0	n_{00}	n_{01}	n_0
	(j=0)	(j=1)	

↑ cutoff score

Figure I depicts the configuration of decisions based on contrasting-group data. Let the index i take the value 0 when the individual test score belongs to the pretest (or uninstructed) group, and the value 1 when the test score belongs to the posttest (or instructed) group. On the other hand, let the index j be 0 when the test score is smaller than the cutoff score c (nonmastery status), and 1 when the test score is at least c (mastery status). The number of test scores in the combined contrasting groups in each (i,j)-cell will be denoted as n_{ij}. In addition, let $n_0 = n_{00} + n_{01}$ be the number of pretest (uninstructed) scores and $n_1 = n_{10} + n_{11}$ be the number of posttest (instructed) scores. For the pretest-posttest design with no dropouts (experimental mortality), $n_0 = n_1$. For the most general situation, particularly when the instructed-uninstructed design is contemplated, n_0 and n_1 are not typically equal.

With the notation as defined, the p index at each cutoff score is given as

$$p = \frac{1}{2}\left(\frac{n_{11}}{n_1} + \frac{n_{00}}{n_0}\right), \tag{1}$$

and the p-max index is simply the maximum of p when the cutoff score varies in its range of possible scores.

Numerical Illustration 1

Let $n_{00} = 5$, $n_{01} = 10$, $n_{10} = 15$, and $n_{11} = 20$. Then $n_0 = 15$ and $n_1 = 35$. Hence $p = .452$.

Numerical Illustration 2

Table 1 reports the frequency distributions of the pretest and the posttest scores of 50 students on a four-item test. The p indices are listed as follows.

Cutoff score	1	2	3	4
p-index	.67	.76	.77	.64

From this list, it may be deduced that p-max is .77.

TABLE 1

Frequency Distributions of Pretest and Posttest Data for Fifty Students

Test score	Pretest frequency	Posttest frequency
0	20	3
1	10	1
2	8	7
3	7	20
4	5	19

The p-max index does not take directly into account changes within individual students from pretesting to posttesting. Other indices may be more appropriate, particularly for the pretest-posttest design. Harris (1977), for example, argues that in studies of item sensitivity, an appropriate index would involve the difference between the proportion of students who have learned the item and the proportion of those who have forgotten it. The first proportion is the probability of responding correctly on the posttest, given that the student responded incorrectly on the pretest. The second proportion represents the probability of responding incorrectly following instruction, given that the response prior to instruction was correct. This index was referred to as the Index of Departure from Symmetry (δ). To use this index for the assessment of test sensitivity, a cutoff score c may be selected, and students are then classified into the two categories of mastery and nonmastery. A δ index may then be computed, considering nonmastery as an incorrect response and mastery as a correct one. Then, the maximum of δ, δ-max, may be determined by locating the maximum of δ when the cutoff score c varies within its range of possible values. For both sets of data considered in Section 2, the δ-max indices are exactly 1.

Figure II depicts the configuration of decisions based on pretest and posttest data. With c as a cutoff score, each student is classified twice, once based on pretest data and again based on posttest data. Let $i = 0$ (for nonmastery) and 1 (for mastery) be the decision based on pretest data, and $j = 0$ or 1 for the decision based on posttest data. In addition, let n_{ij} be the number of students in each (i,j)-cell, $n_0 = n_{00} + n_{01}$ be the number of students who fail the pretest, and $n_1 = n_{10} + n_{11}$ be the number of students who pass the pretest. Then the index δ is defined a

$$\delta = \frac{n_{01}}{n_0} - \frac{n_{10}}{n_1}. \tag{2}$$

As previously stated, δ-max is the maximum value that δ can take within the range of possible cutoff scores.

FIGURE II

Configuration of Decisions Based on Pretest-Posttest Data

Pretest \ Posttest	Nonmastery	cutoff score ↓	Mastery	Marginal sum
Mastery	n_{10}		n_{11}	n_1
cutoff score →				
Nonmastery	n_{00}		n_{01}	n_0

Numerical Illustration 3

Table 2 reports the bivariate pretest-posttest frequency distribution of 50 students on a four-item test. At the cutoff score 3, the cell and pretest marginal frequencies are given as $n_{00} = 8$, $n_{01} = 30$, $n_{10} = 3$, and $n_{11} = 9$; $n_0 = 38$ and $n_1 = 12$. Hence the δ index is $\delta = .539$. At all possible cutoff scores, the δ indices are listed as follows.

Cutoff score	1	2	3	4
δ-index	.833	.867	.539	-.400

From the list it may be deduced that δ-max is .867.

TABLE 2

Bivariate Frequency of Pretest-Posttest Data

		Posttest score					
		0	1	2	3	4	
	4	0	0	3	1	1	5
Pretest	3	0	0	0	2	5	7
score	2	0	0	0	4	4	8
	1	2	0	1	3	4	10
	0	1	1	3	10	5	20
		3	1	7	20	19	50

4. AN OVERALL APPROACH TO TEST SENSITIVITY

It may now be pointed out that point biserial correlation, r-max, δ-max, and other similar indices provide only a global (overall) measure of test sensitivity. They do not provide an assessment of the extent to which the test is sensitive at a particular ability or test score level or in a given range of ability. For example, it is well known that one test may provide a smaller error of measurement than another; however, its relative efficiency with respect to the other test varies as a function of examinee ability level (Lord, 1974). The same situation may appear in test sensitivity. It is conceivable that a test is able to separate two contrasting groups more effectively at one level of ability than at another.

Consider now the case of instructional sensitivity. If the test items faithfully reflect the objectives underlying the instructional unit, then a posttest score (or the score of a student who has completed the unit) is more likely to be high than a pretest score (or the score of a noninstructed student). Let the qualifier "success" be applied to any posttest score and "failure" to any pretest score. The following definitions apply to test sensitivity.

Definition 1

Let $s(\theta)$ be the probability of success at the ability (or test score) level θ. A test is said to be sensitive to the instructional unit (or to the task for which the test is used as a predictor) in a range of ability if $s(\theta)$ is nondecreasing (but not a constant uniformly) within this range.

The function $s(\theta)$ may take any shape, as long as it is nondecreasing. As defined, $s(\theta)$ is reminiscent of the concept of item characteristic curve (Lord & Novick, 1968) and of the notion of referral success (Huynh, 1976). The second notion is more relevant to the psychometric foundation of mastery testing.

Now, at the ability level θ, a test is more sensitive if the probability $s(\theta)$ changes sharply at this point. The following definition applies to the case where $s(\theta)$ has a derivative.

Definition 2

Let $\xi(\theta)$ denote the derivative of $s(\theta)$ with respect to θ. This derivative is said to be the test sensitivity at the ability level θ.

It follows from the second definition that test sensitivity is a non-negative function since $s(\theta)$ is nondecreasing. It may be noted that $\xi(\theta)$ acts like the density of a cumulative distribution function; hence estimation procedures associated with density functions (Wegman, 1974) would be applicable to $\xi(\theta)$.

5. TEST SENSITIVITY AND ITEM INFORMATION

Within the context of mastery testing (Huynh, 1976), a two-parameter logistic form has been proposed for $s(\theta)$, namely

$$s(\theta) = \frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}}, \tag{3}$$

where $\alpha > 0$ and β are suitably chosen constants. The test sensitivity function is now given as

$$\xi(\theta) = s'(\theta) = \frac{\alpha e^{\alpha(\theta-\beta)}}{\left[1+e^{\alpha(\theta-\beta)}\right]^2} = \alpha s(\theta)\left(1-s(\theta)\right) . \tag{4}$$

Let $P(\theta) = s(\theta)$ and $Q(\theta) = 1-s(\theta)$. Then it may be verified that

$$\xi(\theta) = \frac{\left(P'(\theta)\right)^2}{P(\theta)Q(\theta)} . \tag{5}$$

The quantity on the right of this expression represents the information provided by a test item for which the item characteristic curve is $P(\theta) = s(\theta)$ (Birnbaum, 1968, p. 454).

6. STATISTICAL INFERENCE REGARDING TEST SENSITIVITY AS A MONOTONE REGRESSION PROBLEM

Consider now a range of ability (or test score) in which a test is suspected to be sensitive to a given instructional unit or to a task which it is intended to predict. An inferential procedure will now be presented for checking the hypothesis that $s(\theta)$ is nondecreasing.

Let the mentioned range of ability be partitioned into k mutually exclusive and exhaustive sets, namely $A_1, A_2, \ldots, A_k$ in such a way that the number of test scores in each of the k categories in the combined pretest-posttest or instructed-noninstructed sample are as nearly equal as possible. Let $n_1, n_2, \ldots, n_k$ be the number of test scores which fall into each of the A sets, and let $\hat{s}_i$ be the corresponding proportion of students belonging to the success category.

Under the assumption that $s(\theta)$ is nondecreasing, the sample proportions $\hat{s}_i$ must be adjusted if necessary to reflect this preimposed trend. This may be done via the Pool-Adjacent-Violator algorithm described in Barlow, Bartholomew, Bremner, and Brunk (1972). In essence, whenever two consecutive sample values $\hat{s}_i$ and $\hat{s}_{i+1}$ are in the unexpected direction (decreasing), they are taken as the weighted average $(n_i\hat{s}_i + n_{i+1}\hat{s}_{i+1})/(n_i + n_{i+1})$. This common value will then be compared with $\hat{s}_{i+1}$. If these two quantities are not in the expected direction, then the $\hat{s}_i$, $\hat{s}_{i+1}$, and $\hat{s}_{i+2}$ values will be taken as equal, and equal to their weighted average.

Once the set of monotone-adjusted $\hat{s}_i^*$ has been obtained, the standard chi square test for association in a 2 × k contingency table may be applied. The null hypothesis (independence) corresponds to the case where $s(\theta)$ is a constant for all the A cells; the alternative (dependence) expresses the nondecreasing nature of $s(\theta)$ with respect to θ. The use of the standard chi square test in this case was suggested by Bartholomew (1959) and Shorack (1967) for the case where the n_i are equal. Presumably the test should hold approximately when the n_i are nearly equal.

Numerical Illustration 4

Table 3 presents detailed computations for the chi square test based on the data of Table 1. In this table, the A categories are taken as the test score levels of 0, 1, 2, 3, and 4. As explained previously, at each score, n_i denotes the total number of cases, $\hat{s}_i$ the unadjusted proportion of success, and $\hat{s}_i^*$ the monotone-adjusted proportion of success. Thus, at the same test score, the monotone-adjusted number of cases is $n_i\hat{s}_i^*$ for success and $n_i(1-\hat{s}_i^*)$ for failure. The corresponding expected frequencies are n_ip and $n_i(1-p)$ where p is the proportion of success in the combined sample of test scores. (In the case of pretest-posttest, $p = \frac{1}{2}$.) The value of χ^2 is now

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i\hat{s}_i^* - n_ip)^2}{n_ip} + \sum_{i=1}^{k} \frac{\left(n_i(1-\hat{s}_i^*) - n_i(1-p)\right)^2}{n_i(1-p)}$$

$$= \sum_{i=1}^{k} \frac{n_i(\hat{s}_i^*-p)^2}{p} + \sum_{i=1}^{k} \frac{n_i(\hat{s}_i^*-p)^2}{1-p}$$

$$= \frac{2}{p(1-p)} \sum_{i=1}^{k} n_i(\hat{s}_i^*-p)^2 . \qquad (6)$$

With the data of Table 1, the n_i are equal to 23, 11, 15, 27, and 24 at the test scores of 0, 1, 2, 3, and 4. The adjusted frequencies of success are 2.71, 1.29, 7.00, 20.00, and 19.00. In addition, p = .5. Hence $\chi^2 = 17.18$. With a standard chi-square distribution of k-1 = 5 degrees of freedom, the upper tail probability at this observed χ^2 value is smaller than .01. Hence the hypothesis of test sensitivity is supported by the test data.

TABLE 3

An Example of the Adjusted Chi Square Test for Test Sensitivity

Ability/ Test score (θ_i)	n_i	s_i (×100)	s_i^* (×100)	Cell frequency: Adjusted observed†	Cell frequency: Expected	Chi-square contribution
0	23	13.04	11.76	2.71	11.50	6.72
1	11	9.09	11.76	1.29	5.50	3.22
2	15	46.67	46.67	7.00	7.50	0.03
3	27	74.07	74.07	20.00	13.40	3.13
4	24	79.17	79.17	19.00	12.00	4.08
Total	100			100	100	$\chi^2 = 17.18$††

† computed as $n_i s_i^*$

†† df = 4; $p < .01$

7. ESTIMATING TEST SENSITIVITY VIA THE TWO-PARAMETER LOGISTIC MODEL

Let it be assumed now that the function $s(\theta)$ can be satisfactorily represented by the two-parameter logistic curve

$$s(\theta) = \frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}},$$

and hence the test sensitivity curve will take the form $\xi(\theta) = \alpha s(\theta)\left(1-s(\theta)\right)$.

There are at least two ways to estimate the two parameters α and β, namely the minimum logit square method and the maximum likelihood (ML) procedure. The first procedure is less elegant than the second one; however, the computations are much less demanding.

To apply the minimum logit square technique, let p_i be the natural logarithm of the ratio $\hat{s}_i/(1-\hat{s}_i)$. (Preferably, the log of the ratio $\hat{s}_i^*/(1-\hat{s}_i^*)$ should be used.) Let θ_i denote the typical ability of the test score category A_i. Then, at each i

$$p_i = \alpha(\theta_i-\beta),$$

hence α and β may be estimated via standard linear regression technique. They are given as

$$\alpha = \frac{N\Sigma\theta_i p_i - (\Sigma\theta_i)(\Sigma p_i)}{N\Sigma\theta_i^2 - (\Sigma\theta_i)^2}, \tag{7}$$

and

$$\beta = \frac{\alpha\Sigma_{\ \ i} - \Sigma p_i}{N\alpha}. \tag{8}$$

In these formulae, N is the number of cases in the combined sample. Strictly speaking, the procedure does not work if $\hat{s}_i = 0$ or 1 for some score category, since p_i would then be equal to $-\infty$ or $+\infty$. To proceed with the estimation, however, it has been recommended (Berkson, 1953) that $\hat{s}_i$ be set to a small constant when it is exactly zero, and a number near 1 when it is actually one.

A more direct procedure to estimate α and β would be an application of the ML principle. To do this, let θ_i denote a test score in the combined sample and u_i be 1 for the success category and 0 for the failure category. Then, assuming local independence for the success/failure classification, the likelihood function for the combined sample may be written as

$$L = \prod_{i=1}^{N} \left(s(\theta_i)\right)^{u_i}\left(1-s(\theta_i)\right)^{1-u_i}$$

$$= \prod_{i=1}^{N} \frac{e^{\alpha(\theta_i-\beta)u_i}}{1+e^{\alpha(\theta_i-\beta)}}.$$

Hence the log of the likelihood function will take the form

$$\log L = \sum_{i=1}^{N} \alpha(\theta_i-\beta)u_i - \sum_{i=1}^{N} \log\left(1+e^{\alpha(\theta_i-\beta)}\right). \tag{9}$$

The partial derivatives of log L with respect to α and β are given as

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{N} (\theta_i-\beta)u_i - \sum_{i=1}^{N} \frac{(\theta_i-\beta)e^{\alpha(\theta_i-\beta)}}{1+e^{\alpha(\theta_i-\beta)}} \tag{10}$$

and

$$\frac{\partial \log L}{\partial \beta} = -\sum_{i=1}^{N} \alpha u_i + \sum_{i=1}^{N} \frac{\alpha e^{\alpha(\theta_i-\beta)}}{1+e^{\alpha(\theta_i-\beta)}}. \tag{11}$$

By setting these two partial derivatives to zero, the values for α and β may be found. The process will lead to the following simpler equations:

$$G(\alpha,\beta) = \sum_{i=1}^{N} \frac{e^{\alpha(\theta_i-\beta)}}{1+e^{\alpha(\theta_i-\beta)}} - \sum_{i=1}^{N} u_i = 0 , \tag{12}$$

and

$$F(\alpha,\beta) = \sum_{i=1}^{N} \frac{\theta_i e^{\alpha(\theta_i-\beta)}}{1+e^{\alpha(\theta_i-\beta)}} - \sum_{i=1}^{N} \theta_i u_i = 0 . \tag{13}$$

Equations (12) and (13) may be solved via iteration procedures such as the Newton-Raphson process. The process requires the following partial derivatives:

$$G'_\alpha = \sum_{i=1}^{N} (\theta_i-\beta)s(\theta_i)\left(1-s(\theta_i)\right) , \tag{14}$$

$$G'_\beta = -\alpha \sum_{i=1}^{N} s(\theta_i)\left(1-s(\theta_i)\right) , \tag{15}$$

$$F'_\alpha = \sum_{i=1}^{N} \theta_i(\theta_i-\beta)s(\theta_i)\left(1-s(\theta_i)\right) , \tag{16}$$

and

$$F'_\beta = -\alpha \sum_{i=1}^{N} \theta_i s(\theta_i)\left(1-s(\theta_i)\right) . \tag{17}$$

Let α_o and β_o be two starting values for α and β. Then the Newton-Raphson iterated values α_1 and β_1 satisfy the linear equations

$$\begin{cases}(\alpha_1-\alpha_0)G'_\alpha(\alpha_o,\beta_o) + (\beta_1-\beta_0)G'_\beta(\alpha_o,\beta_o) = -G(\alpha_o,\beta_o)\\ (\alpha_1-\alpha_0)F'_\alpha(\alpha_o,\beta_o) + (\beta_1-\beta_0)F'_\beta(\alpha_o,\beta_o) = -F(\alpha_o,\beta_o) .\end{cases} \tag{18}$$

Hence α_1 and α_2 are given as

$$\alpha_1 = \alpha_o - \left(G(\alpha_o,\beta_o)F'_\beta(\alpha_o,\beta_o) - F(\alpha_o,\beta_o)G'_\beta(\alpha_o,\beta_o)\right)/\Delta$$

and

$$\beta_1 = \beta_o + \left(G(\alpha_o,\beta_o)F'_\alpha(\alpha_o,\beta_o) - F(\alpha_o,\beta_o)G'_\alpha(\alpha_o,\beta_o)\right)/\Delta$$

where $\Delta = G'_\alpha(\alpha_o,\beta_o)F'_\beta(\alpha_o,\beta_o) - F'_\alpha(\alpha_o,\beta_o)G'_\beta(\alpha_o,\beta_o)$.

The iteration process continues until convergence is assured to a satisfactory degree.

Numerical Illustration 5

For the data of Table 1, the logit square procedure based on the unadjusted proportions $\hat{s}_i$ yields the estimates $\hat{\alpha}$ = .982 and $\hat{\beta}$ = 2.397. The maximum likelihood procedure results in the estimates $\tilde{\alpha}$ = .947 and $\tilde{\beta}$ = 2.244.

Within the logistic model the traditional asymptotic likelihood ratio test may be used to check the hypothesis of test sensitivity. The log likelihood associated with ML estimation for α and β is equal to $\log L(\tilde{\alpha},\tilde{\beta})$, where log L is given in Equation (9). When the test shows no sensitivity, then the probability $s(\theta_i)$ is uniformly equal to $p_o = n_0/(n_0+n_1)$. (This probability is equal to ½ for the pretest-posttest design.) The corresponding log likelihood is given as $\log L_o = n_0 \log p_o + n_1 \log (1-p_o)$. The asymptotic likelihood ratio test is carried out via the quantity $\chi^2 = \log L(\tilde{\alpha},\tilde{\beta}) - \log L_o$ which is distributed approximately as a chi square distribution with one degree of freedom. With the data referred to in Numerical Illustration 5, for example, it was found that $\log L(\tilde{\alpha},\tilde{\beta}) = -51.718$, and $\log L_o = -69.315$. Hence $\chi^2 = 17.597$, which corresponds to an upper tail probability of less than .01. Thus, the data show strong evidence of test sensitivity.

Appendix A provides a listing of a computer program for the computations described in this section.

8. SUMMARY

This paper has discussed test sensitivity in mastery testing. Arguments have been presented to show that correlation-based indices may not be appropriate for assessing the sensitivity of mastery tests. Instead, indices such as p-max or δ-max are advocated for the global assessment of test sensitivity, while local measures of sensitivity may be obtained using a two-parameter logistic model. Finally, procedures are described to test the tenability of the hypothesis of test sensitivity on the basis of observed test data.

BIBLIOGRAPHY

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. (1972). Statistical inference under order restrictions. New York: John Wiley & Sons.

Bartholomew, D. J. (1959). A test for homogeneity of ordered alternatives. Biometrika 46, 36-48.

Berk, R. A. (Ed.) (1980). Criterion-referenced measurement: The state of the art. Baltimore, Maryland: Johns Hopkins Press.

Berkson, J. (1953). A statistical precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. Journal of the American Statistical Association 48, 565-599.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Co.

Haladyna, T. & Goid, G. (1980). The role of instructional sensitivity in the empirical review of criterion-referenced test items. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Harris, C. W., Alkin, M. C. & Popham, W. J. (Eds.) (1974). Problems in criterion-referenced measurement. Los Angeles: Center for the Study of Evaluation, University of California at Los Angeles.

Harris, C. W., Pearlman, A. P. & Wilcox, R. R. (1977). Achievement test items: Methods of study. Los Angeles: Center for the Study of Evaluation, University of California at Los Angeles.

Huynh, H. (1976). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

Lord, F. M. (1974). Quick estimates of the relative efficiency of two tests as a function of ability level. Journal of Educational Measurement 11, 247-254.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Co.

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

Shorack, G. R. (1967). Testing against ordered alternatives in model I analysis of variance: Normal theory and nonparametric. Annals of Mathematical Statistics 38, 1740-1753.

Wegman, E. J. (1972). Nonparametric probability density estimation: I. A summary of available methods. Technometrics 14, 533-546.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, principal investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance of Joseph C. Saunders is gratefully acknowledged.

APPENDIX A

Listing of a Computer Program for Assessing Test Sensitivity via the Two-Parameter Logistic Model

Disclaimer: The computer program hereafter listed has been written with care and tested extensively under a variety of conditions using tests with 60 or fewer items. The author, however, makes no warranty as to its accuracy and functioning, nor shall the fact of its distribution imply such warranty.

```
C     THIS PROGRAM COMPUTES THE MAXIMUM LIKELIHOOD ESTIMATES OF THE
C     ALPHA AND BETA PARAMETERS WHICH FORM THE BASIS FOR ASSESSING
C     TEST SENSITIVITY.
C     INPUT DATA ARE LISTED AS FOLLOWS.
C       FIRST CARD:  TITLE CARD (ENTER ANYTHING YOU WANT.)
C       SECOND CARD: NUMBER (M) OF TEST SCORE/ABILITY LEVELS (I5)
C       THIRD CARD:  FORMAT CARD FOR EACH OF THE M FOLLOWING CARDS
C       M CARDS:     EACH CONTAINS THE TEST SCORE LEVEL, THE FREQUENCY
C                    OF THE PRETEST/UNINSTRUCTED GROUP, AND THE
C                    FREQUENCY OF THE POSTTEST/INSTRUCTED GROUP. EACH
C                    CARD IS TO BE KEYPUNCHED ACCORDING TO THE FORMAT
C                    ENTERED VIA THE THIRD CARD.
C     SEVERAL PROBLEMS MAY BE PERFORMED IN ONE RUN BY STACKING THE
C     INPUT CARDS TOGETHER.
C     THIS PROGRAM IS WRITTEN FOR TESTS WITH UP TO 61 LEVELS OF TEST
C     SCORE OR ABILITY. FOR LONGER TESTS, REDIMENSION T AND N TO BE
C     T(M) AND N(M), M BEING THE NUMBER OF LEVELS.
C
      DIMENSION T(61),N(61),FCT(20)
      DOUBLE PRECISION A,B,EA,EB,EPS,DELTA
      EPS=.00001
      NTOT=0
      SU=0.
      STU=0.
      ST=0.
      ST2=0.
      SR=0.
      STR=0.
    5 READ(5,95,END=99) FCT
   95 FORMAT(20A4)
      WRITE(6,195) FCT
  195 FORMAT(T2,'ANALYSIS OF TEST SENSITIVITY VIA THE LOGISTIC MODEL'/
     &       T2,'**************************************************'/
     &       T2,'TITLE OF THIS PROBLEM IS:'/T2,20A4)
      READ(5,96) M
   96 FORMAT(I5)
      WRITE(6,196) M
  196 FORMAT(T2,'NUMBER OF TEST SCORE/ABILITY LEVELS:',I5)
      READ(5,97) FCT
   97 FORMAT(20A4)
      WRITE(6,197) FCT
  197 FORMAT(T2,'INPUT FORMAT FOR FREQUENCY DATA:'/T2,20A4)
      WRITE(6,198)
  198 FORMAT(T2,'FREQUENCY DISTRIBUTION'/
     &       T2,'  SCORE   PRETEST/UNINSTRUCTED    POSTTEST/INSTRUCTED'
     &      /T2,'  LEVEL          GROUP                      GROUP'/
     &       T2,'  *****   ****************        ***************  ')
      DO 20 K=1,M
      READ(5,FCT) T(K),NLOWER,NUPPER
      WRITE(6,200) T(K),NLOWER,NUPPER
  200 FORMAT(T2,F8.2,T21,I3,T44,I3)
      N(K)=NLOWER+NUPPER
      NTOT=NTOT+N(K)
      R=FLOAT(NUPPER)/FLOAT(N(K))
      SU=SU+NUPPER
      STU=STU+T(K)*NUPPER
      R=AMAX1(.01,R)
      R=AMIN1(.99,R)
      R=ALOG(R/(1.-R))
      ST=ST+T(K)
      ST2=ST2+T(K)**2
      SR=SR+R
   20 STR=STR+T(K)*R
      A=(M*STR-ST*SR)/(M*ST2-ST*ST)
      B=(A*ST-SR)/(M*A)
      WRITE(6,215) A,B
  215 FORMAT(T2,'STARTING VALUES BASED ON MINIMUM LOGIT'/
     &       T17,'ALPHA = ',F10.5/T17,'BETA =  ',F10.5)
   30 CALL NEWTON(M,N,T,SU,STU,A,B,EA,EB)
      DELTA=DMAX1(EA,EB)
      IF(DABS(DELTA).LT.EPS) GOTO 40
      A=A+EA
      B=B+EB
```

```
      GOTO 30
   40 WRITE(6,220) A,B
  220 FORMAT(T2,'FINAL RESULTS: ALPHA = ',F10.5/
     *       T2,'               BETA  = ',F10.5//)
      H1=A*(STU-B*SU)
      P=SU/NTOT
      DO 50 I=1,K
   50 H1=H1-N(I)*DLOG(1.+DEXP(A*(T(I)-B)))
      H0=SU*ALOG(P)+(NTOT-SU)*ALOG(1.-P)
      CHISQ=H1-H0
      WRITE(6,221) H1,H0,CHISQ
  221 FORMAT(T2,'LOG OF THE LIKELIHOOD FUNCTION'/
     &       T2,'  WITH TEST SENSITIVITY: ',F10.5/
     &       T2,'  NO TEST SENSITIVITY..: ',F10.5/
     &       T2,'CHI-SQUARE STATISTIC ....: ',F10.5/
     &       T2,'WITH ONE DEGREE OF FREEDOM.')
      GOTO 5
   99 WRITE(6,225)
  225 FORMAT(T2,'   **NORMAL END OF JOB**'/
     &       T2,'   PROGRAM WRITTEN BY HUYNH HUYNH'/
     &       T2,'   COLLEGE OF EDUCATION'/
     &       T2,'   UNIVERSITY OF SOUTH CAROLINA'/
     &       T2,'   COLUMBIA, SOUTH CAROLINA 29208'/
     &       T2,'   JULY 1980')
      STOP
      END
C
      SUBROUTINE NEWTON(K,N,T,SU,STU,A,B,EA,EB)
      DIMENSION N(1),T(1)
      DOUBLE PRECISION S,G,F,GA,GB,FA,FB,D,E,P,A,B,EA,EB
      G=-SU
      F=-STU
      FA=0.D0
      FB=0.D0
      GA=0.D0
      GB=0.D0
C
      DO 10 I=1,K
      E=DEXP(A*(T(I)-B))
      P=E/(E+1.D0)
      S=P*(1.D0-P)
      G=G+P*N(I)
      F=F+N(I)*T(I)*P
      GA=GA+N(I)*(T(I)-B)*S
      GB=GB-A*S*N(I)
      FA=FA+N(I)*T(I)*(T(I)-B)*S
   10 FB=FB-A*T(I)*S*N(I)
      D=GA*FB-FA*GB
      EA=-(G*FB-F*GB)/D
      EB=(G*FA-F*GA)/D
      RETURN
      END
```

PART SEVEN

TEST DESIGN

17

SELECTING ITEMS AND SETTING PASSING SCORES FOR MASTERY TESTS BASED ON THE TWO-PARAMETER LOGISTIC MODEL

Huynh Huynh

University of South Carolina

Presented at the Informal Meeting on Model-Based Psychological Measurement sponsored by the Office of Naval Research, Iowa City, Iowa, August 17-22, 1980.

ABSTRACT

Three issues in mastery testing are considered, using a minimax decision framework, based on the two-parameter logistic model. The issues are: (1) setting passing scores, (2) assessing decision efficiency, and (3) selecting items to maximize decision efficiency. The losses or disutilities under consideration have a constant or normal ogive form. It is found that, in the context of minimax decisions, the item selection procedure based on maximum information may not provide the best decision efficiency.

1. INTRODUCTION

A primary purpose of mastery testing is to classify each examinee in one of several achievement (or ability) categories. Typically there are two such categories, commonly labeled mastery and nonmastery. Let θ be the ability or trait being measured. On the θ scale, the status of mastery is defined by the condition $\theta \geq \theta_o$, and that of nonmastery by $\theta < \theta_o$, where θ_o is a prespecified constant often referred to as a true mastery score. (As can be seen

This paper has been distributed separately as RM 80-6, August, 1980.

later, the postulated existence of θ_o is justified when the losses or utilities associated with the decision problem fulfill fairly reasonable assumptions.) In most practical situations, however, θ is not known, and mastery/nonmastery decisions are usually based on the responses of the examinee to a relevant set of items. Three issues thus emerge, which deal with (1) scoring item responses, (2) setting a test passing score, and (3) selecting test items which serve best (in some sense) the process of classification (mastery testing).

Within the context of Bayesian decision theory as applied to the case of constant losses, and considering tolerable limits on the probabilities of making false positive (α) and false negative (β) errors, Birnbaum (1968) and Lord (1980) have given considerable attention to the three issues mentioned above. The treatment developed by Birnbaum does not seem to lead to an easy generalization to situations involving other than constant losses, and the discussion by Lord, at times, moves from Bayesian decision theory to confidence interval estimation without a strong link of continuity.

The purpose of this paper is to provide a consideration of the aforementioned issues in mastery testing, using a minimax decision framework. Consideration is restricted to a two-parameter logistic model in which a sufficient statistic exists for the estimation of ability. A minimax treatment of mastery testing which involves the simple binomial error model may be found in Huynh (1980), and in Wilcox (1976) in another form.

2. SUFFICIENCY, MONOTONE LIKELIHOOD RATIO, AND MONOTONE DECISION PROBLEMS

Consider a test consisting of n items (indexed by $i = 1,2,\ldots,n$) for which the item response u_i of an examinee with ability θ follows a two-parameter logistic model with item difficulty b_i and item discrimination a_i. It is well known that the composite test score $x = \sum_{i=1}^{n} a_i u_i$ is a sufficient statistic for estimating θ, and that the conditional density $f(x|\theta)$ has the monotone likelihood ratio property (Birnbaum, 1968, sec. 19.4). Sufficiency implies

(Ferguson, 1967, p. 120, Theorem 1) that any decision problem focusing on θ may be simply based on the test score x since the set of decision rules based on x forms an essentially complete class. In other words, for any decision rule based on the vector of responses $(u_1, u_2, \ldots, u_n)$, there is always a decision rule based on x which performs at least as well as the given rule in terms of risk (or expected loss).

Consider now the action (a_1) of granting mastery status and the action (a_2) of denying mastery status to an examinee with ability θ. Let $L_1(\theta)$ and $L_2(\theta)$ be the losses (disutilities) associated with the two actions a_1 and a_2. In practical situations, it seems reasonable to assume that $L_1(\theta)$ is nonincreasing in θ and $L_2(\theta)$ is nondecreasing in θ. In other words, granting mastery status should cause less harm to an examinee with high ability than to someone with low ability. The reverse should hold for the act of denying mastery status. When the graphs of $L_1(\theta)$ and $L_2(\theta)$ do not cross, either action a_1 or action a_2 is uniformly better than the other at all ability levels θ; hence the choice for the best course of action would be either a_1 or a_2 regardless of the observed test score x. This "degenerate" case does not represent a typical use of test data; hence it seems reasonable to assume that the graphs of $L_1(\theta)$ and $L_2(\theta)$ cross at least at one point. Due to the nondecreasing nature of the difference $L_2(\theta) - L_1(\theta)$, crossing can occur only once. Hence, there exists one ability level θ_o such that $L_1(\theta) \geq L_2(\theta)$ for $\theta < \theta_o$ and $L_1(\theta) \leq L_2(\theta)$ for $\theta > \theta_o$. Under these conditions, the decision problem is said to be monotone (Ferguson, 1967, chap. 6). It may then be noted that, in terms of loss, action a_1 is best when $\theta > \theta_o$, and action a_2 is best when $\theta < \theta_o$.

Within the monotone decision problem as stated and with the monotone likelihood ratio property for the density $f(x|\theta)$, it is well known (Ferguson, 1967, p. 286; Zacks, 1971, ch. 9) that the search for an optimum decision rule may be restricted to the (essentially complete) class of decision rules defined by $a_1 = \{x;\ x \geq c\}$ and $a_2 = \{x;\ x < c\}$, where c is a suitable test passing score. At each potential passing score c, the expected loss is

$$R(c;\theta) = L_1(\theta)P(x \geq c|\theta) + L_2(\theta)P(x<c|\theta) . \quad (1)$$

A minimax passing score c_o is the score which minimizes the maximum of $R(c;\theta)$ with respect to θ. (For the sake of simplicity, it is assumed that the search for maximum and minimum can be accomplished.)

Consider now the maximum $G(\theta)$ of the two losses $L_1(\theta)$ and $L_2(\theta)$. It is given as $G(\theta) = L_1(\theta)$ for $\theta < \theta_o$, and $G(\theta) = L_2(\theta)$ for $\theta \geq \theta_o$. The expected loss $R(c;\theta)$ may now be written as

$$R(c;\theta) = G(\theta) + \big(L_2(\theta) - L_1(\theta)\big)P(x<c|\theta)$$

for $\theta < \theta_o$, and as

$$R(c;\theta) = G(\theta) + \big(L_1(\theta) - L_2(\theta)\big)P(x \geq c|\theta)$$

for $\theta \geq \theta_o$. The quantity $C_f(\theta) = L_2(\theta) - L_1(\theta)$, $\theta < \theta_o$, represents the opportunity loss due to a false negative error, and the quantity $C_s(\theta) = L_2(\theta) - L_1(\theta)$, $\theta < \theta_o$, denotes the opportunity loss due to a false positive error. Opportunity losses are zero when correct decisions, namely the two combinations $(\theta<\theta_o, x<c)$ and $(\theta \geq \theta_o, x \geq c)$, are made. Thus, as indicated in this special case, solutions for a monotone decision problem may be found by looking at the original losses, or at the corresponding opportunity losses. Additional examples of this duality may be found in elementary textbooks such as Schlaifer (1969).

Due to the duality as presented, both losses and opportunity losses will be considered in the remaining part of this paper. Thus, for opportunity losses $C_f(\theta)$ will be taken as zero when $\theta \geq \theta_o$, and $C_s(\theta)$ as zero when $\theta < \theta_o$. In all other cases, both $C_f(\theta)$ and $C_s(\theta)$ are nonnegative, with $C_f(\theta)$ being nonincreasing and $C_s(\theta)$ nondecreasing in θ.

3. MINIMAX PASSING SCORE AND DECISION EFFICIENCY

The risk $R(c;\theta)$ may now be written as follows:

$$R(c;\theta) = \begin{cases} C_f(\theta)P(x \geq c|\theta) & \text{for } \theta < \theta_o \\ C_s(\theta)P(x<c|\theta) & \text{for } \theta \geq \theta_o . \end{cases} \quad (2)$$

Now let

$$L_1(c) = \sup_{\theta<\theta_o} C_f(\theta)P(x \geq c|\theta) \quad (3)$$

and

$$L_2(c) = \sup_{\theta \geq \theta_o} C_s(\theta)P(x<c|\theta) . \tag{4}$$

Then the maximum (or supremum) of $R(c;\theta)$ over θ is

$$M(c) = \max\{L_1(c),L_2(c)\} .$$

The optimum (minimax) passing score is the test score c_o at which $M(c)$ is minimized. The minimum (or infinimum) value of $M(c)$, henceforth denoted as R_o, is traditionally referred to as the <u>minimax value</u> of the decision problem (Ferguson, 1967, p. 33).

Consider now the extreme case where the score x does not reveal the true ability θ, e.g., when x and θ are stochastically independent. Let

$$C_f^* = \sup_{\theta<\theta_o} C_f(\theta)$$

and

$$C_s^* = \sup_{\theta \geq \theta_o} C_s(\theta) .$$

In the case where both C_f^* and C_s^* are finite, the minimax passing score c^* satisfies the equation

$$C_f^* P(x \geq c^*) = C_s^* P(x<c^*) .$$

In other words, when there is no relationship between x and θ, it is best to randomly assign mastery with a probability of $C_s^*/(C_s^* + C_f^*)$ and nonmastery with a probability of $C_f^*/(C_s^* + C_f^*)$. The minimax value of the decision situation is then

$$R^* = C_f^*C_s^*/(C_f^* + C_s^*) . \tag{5}$$

It may be recalled that opportunity losses are zero when the decisions are correct. Hence, when the test score x reveals fully the nature of the ability θ, the minimax value is zero. This observation along with the nature of R_o and R^* suggests the use of the quantity $\eta = (R^*-R_o)/R^*$ as an index to measure the efficiency of using test scores in making mastery/nonmastery decisions. This efficiency index measures the extent to which the best use of test data will reduce the amount of risk which would be expected had the test data not been used at all. It is a function of the opportunity losses $C_f(\theta)$ and $C_s(\theta)$, and of the item parameters a_i and b_i.

As defined, the efficiency index η is computable only when both C_f^* and C_s^* are finite. This means that the opportunity losses $C_s(\theta)$ and $C_f(\theta)$ are <u>not</u> allowed to drift out of bounds when θ goes to infinity. Hence, efficiency is not defined for linear or quadratic losses if these are expressed as a direct function of θ. However, as Novick and Lindley (1978) point out, it seems sensible to demand that losses or utilities be bounded, at least in the context of educational and psychological testing. This assumption will be made throughout the remaining part of this paper.

With the efficiency index now defined, the design of a mastery test may be accomplished by deciding on the number of test items, n, and selecting the test items such that the resulting efficiency index would be equal or nearly equal to a specified level.

It seems intuitively true that as the number of test items increases, the efficiency index will increase. However, when the situation permits, a short test is preferable to a lengthy one. Hence, a balance seems appropriate between efficiency and test length. As a passing remark, one may express the latter trait as a function of n, say $\ell(n)$, and then search for an n value at which the product of $\ell(n)$ with the efficiency index $\eta(n)$ is maximized.

4. DESIGNING A MASTERY TEST FOR THE CASE OF CONSTANT LOSSES

For technical reasons which should be apparent from the work of Birnbaum (1968, ch. 19), the case of constant losses in minimax decision problems may be represented by the following functions:

$$C_f(\theta) = \begin{cases} 1 \text{ if } \theta \leq \theta_o \\ 0 \text{ if } \theta > \theta_o , \end{cases} \tag{6}$$

and

$$C_s(\theta) = \begin{cases} Q \text{ if } \theta_o + \varepsilon \leq \theta \\ 0 \text{ if } \theta < \theta_o + \varepsilon , \end{cases} \tag{7}$$

where Q is a constant. The region $\{\theta;\theta_o<\theta<\theta_o+\varepsilon\}$ is an indifference zone. For any examinee whose true ability falls within this range, it does not matter whether action a_1 or a_2 is taken. The constant Q is the ratio of the false negative error to false positive error. (It may also be said simply that the false negative error and the false positive error are weighted according to the ratio Q ÷ 1.)

The risk $R(c;\theta)$ of Equation (2) may now be expressed as follows:

$$R(c;\theta) = \begin{cases} P(x\geq c|\theta) & \text{for } \theta\leq\theta_o \\ QP(x<c|\theta) & \text{for } \theta\geq\theta_o+\varepsilon . \end{cases} \tag{8}$$

As elaborated in Section 2, the conditional density $f(x|\theta)$ belongs to the monotone likelihood ratio family. It follows from Dykstra, Hewett, and Thompson (1973) that x and θ are stochastically increasing in sequence; hence the maximum value of $P(x<c|\theta)$ occurs at $\theta = \theta_o+\varepsilon$ and that of $P(x\geq c|\theta)$ occurs at $\theta = \theta_o$. Thus the expressions $L_1(c)$, $L_2(c)$, and $M(c)$ of Equations (3), (4), and (5) become

$$L_1(c) = P(x\geq c|\theta = \theta_o), \tag{9}$$

$$L_2(c) = QP(x<c|\theta = \theta_o+\varepsilon) , \tag{10}$$

and

$$M(c) = \max\{L_1(c),L_2(c)\} .$$

It may be noted that, as a function of c, $L_1(c)$ is nonincreasing and varies from 1 to 0. As for $L_2(c)$, it is nondecreasing and varies from 0 to Q. If the test score x can be assumed to be continuous, then the minimum of M(c) will occur at c_o where $L_1(c_o) = L_2(c_o)$.

Consider now the special case where $\varepsilon = 0$. Then the minimax passing score c_o satisfies the equation

$$P(x\geq c_o|\theta = \theta_o) = QP(x<c_o|\theta = \theta_o) ,$$

or

$$P(x<c_o|\theta = \theta_o) = 1/(Q+1) .$$

The minimax value of the decision problem is $R_o = Q/(Q+1)$ regardless of the nature of the items which form the test. In addition, the minimax value encountered when the test data are not used is

$R^* = Q/(Q+1)$; thus the decision efficiency index η is zero. (This conclusion is consistent with the observation by Wilcox (1977) that when $\eta = 0$, the process of randomly assigning an examinee to mastery and nonmastery status, each with a probability of .5, would encounter no more maximum error than any attempt to use test data.) Thus, when there is no indifference zone separating masters and nonmasters on the ability scale, there is no way to design a test which will add any efficiency to the minimax decision-making process. For this reason, the constant ε shall be assumed to be strictly positive in the remaining part of this section.

As may be seen from Equations (9) and (10), $L_1(c)$ decreases from 1 to 0 and $L_2(c)$ increases from 0 to Q when the passing score c spans the range of possible values. If the test score can be assumed to be continuous, then the minimax passing score c_o is the one at which $L_1(c) = L_2(c)$. Otherwise, c_o is one (or both) of the two scores which lie nearest to the location at which the graphs of $L_1(c)$ and $L_2(c)$ meet. As before, the minimax passing score is the test score at which M(c) is the smallest.

5. APPROXIMATE SOLUTION FOR MINIMAX PASSING SCORE FOR CONSTANT LOSSES

Let the test now consist of n items. Each item is associated with a characteristic function defined by the probability that the item response u_i is correct, namely

$$p_i(\theta) = \frac{e^{a_i(\theta-b_i)}}{1 + e^{a_i(\theta-b_i)}}. \tag{11}$$

Let the (composite) test score be $x = \sum_{i=1}^{n} a_i u_i$. The mean and the variance of the test score x are given respectively as

$$\mu(\theta) = \sum_{i=1}^{n} a_i p_i(\theta) \tag{12}$$

and

$$\sigma^2(\theta) = \sum_{i=1}^{n} a_i p_i(\theta) q_i(\theta), \tag{13}$$

where $q_i(\theta) = 1 - p_i(\theta)$.

When there are a sufficient number of items forming the test, the conditional distribution of x, given θ, may be approximated by the normal distribution with mean $\mu(\theta)$ and standard deviation $\sigma(\theta)$.

The minimax passing score c_o now satisfies the equation

$$P(x \geq c_o | \theta = \theta_o) = QP(x < c_o | \theta = \theta_o + \varepsilon) . \tag{14}$$

Let $\Phi(.)$ denote the cumulative distribution function of a unit normal variable (with zero mean and unit variance). Then c_o is the solution of the equation

$$1 - \Phi\left(\frac{c_o - \mu(\theta_o)}{\sigma(\theta_o)}\right) = Q\Phi\left(\frac{c_o - \mu(\theta_o + \varepsilon)}{\sigma(\theta_o + \varepsilon)}\right) . \tag{15}$$

This equation may be solved numerically via the Newton-Raphson iteration process. To do this, let the function H be defined as

$$H(c) = \Phi\left(\frac{c - \mu(\theta_o)}{\sigma(\theta_o)}\right) + Q\Phi\left(\frac{c - \mu(\theta_o + \varepsilon)}{\sigma(\theta_o + \varepsilon)}\right) - 1 . \tag{16}$$

The derivative of H with respect to c is given as

$$H'(c) = \frac{1}{\sigma(\theta_o)}\phi\left(\frac{c - \mu(\theta_o)}{\sigma(\theta_o)}\right) + \frac{Q}{\sigma(\theta_o + \varepsilon)}\phi\left(\frac{c - \mu(\theta_o + \varepsilon)}{\sigma(\theta_o + \varepsilon)}\right) \tag{17}$$

where $\phi(.)$ is the density of the unit normal variable. In other words,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} . \tag{18}$$

To proceed with the Newton-Raphson process, a starting value c_1 for the passing score must be found. This may be taken as the average of the two c values at which

$$\Phi\left(\frac{c - \mu(\theta_o)}{\sigma(\theta_o)}\right) = \frac{1}{1+Q} \tag{19}$$

and

$$\Phi\left(\frac{c - \mu(\theta_o + \varepsilon)}{\sigma(\theta_o + \varepsilon)}\right) = \frac{1}{1+Q} . \tag{20}$$

Once c_1 has been computed, the updated c_2 value is given as

$$c_2 = c_1 - H(c_1)/H'(c_1) .$$

Using c_2 as a starting value, the updated c_3 value may be found. The process will end when the change in the c value is sufficiently small.

Numerical Illustration

Let a test consist of ten items with parameters listed as follows:

Item	1	2	3	4	5	6	7	8	9	10
a_i	3.0	1.0	1.0	0.6	0.6	0.3	0.3	0.2	0.2	0.1
b_i	-2.0	-2.0	-1.5	-1.0	0.3	0.6	0.8	2.0	3.0	5.0

In addition, let $\theta_o = 1.2$, $\varepsilon = 1.0$, and $Q = 2$. Then $\mu(\theta_o) = 6.2875$, $\sigma(\theta_o) = .7795$, $\mu(\theta_o+\varepsilon) = 6.5424$, and $\sigma(\theta_o+\varepsilon) = .6943$. The unit normal z score at which $\phi(z) = 1/(1+Q) = 1/3$ is $z = .432$, hence the starting value for the Newton-Raphson process is $c_1 = 6.7333$. The first updated value is $c_2 = 6.1280$. If a tolerance error of .00001 is acceptable, then the iteration process ends at the solution $c_o = 6.1487$. At this minimax passing score, the minimax value of the decision problem is $R_o = M(c_o) = P(x \geq c_o | \theta_o) = .5707$. With $R^* = Q/(1+Q) = 2/3$, the efficiency index η is $1 - R_o/R^* = .1440$.

6. AN ITEM SELECTION PROCEDURE FOR CONSTANT LOSSES

Consider now the task of selecting n items for a test from a pool consisting of N items. (Conceptually, N may be infinite.) Which items should be selected? Lord (1980) proposes that items should be selected in such a way that the item responses would show the highest degree of information at θ_o (for the case where $\varepsilon = 0$). While it appears clear that there is a direct relationship between test information and the reduction of decision errors, it seems desirable to base the selection of test items on the efficiency index η, which is derived from (minimax) decision theory in a more direct way than is test information.

Since the efficiency index is $\eta = 1-R_o/R^*$ and since R^* is constant, the highest efficiency would occur when the minimax value R_o is at its minimum. When the test score can be assumed to be continuous, R_o is either $P(x \geq c_o | \theta=\theta_o)$ or $QP(x<c_o | \theta=\theta_o+\varepsilon)$. Thus, the selection of the items must be such that these two quantities are simultaneously as small as possible.

Except for the case of equal item difficulties and equal item discriminations, the probabilities which define the minimax value R_o involve the item parameters in a rather complex manner. Hence, the optimum selection of items would require the complete enumeration of all the $\binom{N}{n}$ possible item combinations. The number of combinations may be very large; thus, for large-item pools, optimality in selection of items does not appear to justify the computing costs at the present time.

An approximate solution for item selection may be obtained by noting that, at each passing score c, $P(x \geq c_o | \theta=\theta_o)$ is an increasing function of each individual probability $p_i(\theta_o)$, and that $QP(x<c_o | \theta=\theta_o+\varepsilon)$ is an increasing function of each individual component $Qq_i(\theta_o+\varepsilon)=Q(1-p_i(\theta_o+\varepsilon))$. Hence, at each c, the maximum value M(c) would be small if $p_i(\theta_o)$ and $Q(1-p_i(\theta_o+\varepsilon))$ are simultaneously small. (This cannot be true if $\varepsilon=0$.) Hence, the selection of items may be accomplished as follows. (i) For each item i, compute the maximum δ_i of $p_i(\theta_o)$ and $Q(1-p_i(\theta_o+\varepsilon))$. (ii) Select the n items for which the δ_i values are the smallest.

Numerical Illustration

With the item parameters documented in the numerical illustration found in Section 5, the δ_i values are given as follows:

Item	1	2	3	4	5	6	7	8	9	10
δ_i	1.00	.96	.94	.79	.63	.76	.79	.98	1.08	1.14

Thus, if five items are to be selected for the decision situation under consideration, they would be the ones indexed by the numbers 3, 4, 5, 6, and 7. The efficiency index computed from the normal approximation is η = .1411. It may be interesting to note that the selection procedure based on maximum information (at $\theta_o+\frac{\varepsilon}{2}$) would result in the items with numbers 4, 5, 6, 7, and 8. The efficiency index for this selection is .1163. To gain some insight in the selection procedure based on δ, a random selection of items was conducted and resulted in the items 1, 3, 4, 8, and 10. The corresponding efficiency index was found to be .1086.

The numerical illustration seems to indicate that the procedure based on maximum item information may not be the best way to select test items in the context of minimax decision theory. In addition, though this procedure and the one based on minimum δ value appear to select a fair number of common items, the δ procedure seems to be more consistent with the minimax decision approach to mastery testing.

7. A COMPUTER PROGRAM FOR THE CASE OF CONSTANT LOSSES

Appendix A provides the listing of a FORTRAN computer program which is written for the analysis of decisions based on the minimax principle. Input data to the program are (i) a title card; (ii) a card providing the data for n, θ_o, $\theta_o + \varepsilon$, and Q, (iii) an input format card for reading each pair (a_i, b_i); and (iv) n cards of item parameters. For example, the input data for the numerical example of Section 5 is listed in Table 1. Table 2 lists the output of the program.

TABLE 1

An Example of Input Data

```
AN EXAMPLE OF MINIMAX DECISION ANALYSIS
        10    1.20000    2.20000    2.00000    .43200
(2F10.5)
   3.0      -2.0
   1.0      -2.0
   1.0      -1.5
   0.6      -1.0
   0.6       0.3
   0.3       0.6
   0.3       0.8
   0.2       2.0
   0.2       3.0
   0.1       5.0
```

8. AN APPROXIMATE SOLUTION FOR MINIMAX PASSING SCORES UNDER NORMAL LOSSES

Novick and Lindley (1978) indicated that in most practical applications, a more realistic form of utility (and consequently, of the loss function) would be the normal ogive family. Let $\psi(x) = e^x/(1+e^x)$ be the logistic function. Then (Haley, 1952, p. 7) $\psi(1.7z)$ and the unit normal distribution $\Phi(z)$ differ by less than .01 uniformly in z. For this reason, and for the computational

TABLE 2

An Example of Output from the Computer Program

```
MINIMAX DECISION ANALYSIS FOR THE TWO-PARAMETER
LOGISTIC MODEL. TITLE OF THIS PROBLEM IS:
AN EXAMPLE OF MINIMAX DECISION ANALYSIS
NUMBER OF ITEMS ..............  10

INDIFFERENCE ZONE ON THE ABILITY THETA SCALE
     LOWER LIMIT (THETA-ZERO).   1.20000
     UPPER LIMIT (THETA-ZERO
               PLUS EPSILON).   2.20000

LOSS RATIO Q .................   2.00000
TOLERANCE ERROR ....... .....   0.00001

ITEM PARAMETERS
ITEM ID     DISCR.      DIFF.

     1       3.000      -2.000
     2       1.000      -2.000
     3       1.000      -1.500
     4       0.600      -1.000
     5       0.600       0.300
     6       0.300       0.600
     7       0.300       0.800
     8       0.200       2.000
     9       0.200       3.000
    10       0.100       5.000

NORMAL APPROXIMATION FOR TEST SCORES
AT LIMITS OF INDIFFERENCE ZONE

LOWER LIMIT : MEAN .....     6.288
              S.D. .....     0.694

UPPER LIMIT : MEAN .....     6.542
              S.D. .....     0.694

MINIMAX VALUES
   WITH USE OF TEST SCORES .....   0.57067
   WITH NO USE OF TEST SCORES ..   0.66667

FINAL RESULTS

FINAL MINIMAX PASSING SCORE   6.14872
DECISION EFFICIENCY .......   0.14400
```

simplicity associated with the logistic function, the two functions $\Phi(z)$ and $\psi(1.7z)$ will be used interchangeably in this section.

The normal (or logistic) form for the two loss functions (disutilities) $L_1(\theta)$ for action a_1 and $L_2(\theta)$ for action a_2 may be written as

$$L_1(\theta) = 1/\left(1+e^{\alpha_1(\theta-\beta_1)}\right) \tag{21}$$

and

$$L_2(\theta) = Qe^{\alpha_2(\theta-\beta_2)}/\left(1+e^{\alpha_2(\theta-\beta_2)}\right). \tag{22}$$

In these expressions, α_1 and α_2 are positive constants. Constant losses correspond to the degenerate case in which $\beta_1 = \beta_2$ and $\alpha_1 = \alpha_2 = \infty$.

Now let θ_o be the solution of $L_1(\theta_o) = L_2(\theta_o)$. This quantity may be obtained via a typical Newton-Raphson iteration process. Given θ_o, the opportunity losses are given as follows:

$$C_s(\theta) = \begin{cases} L_2(\theta)-L_1(\theta) & \text{for } \theta \geq \theta_o \\ 0 & \text{for } \theta < \theta_o , \end{cases} \tag{23}$$

and

$$C_f(\theta) = \begin{cases} 0 & \text{for } \theta \geq \theta_o \\ L_1(\theta)-L_2(\theta) & \text{for } \theta < \theta_o . \end{cases} \tag{24}$$

At each potential passing score c, the risk $R(c;\theta)$ of Equation (2) is equal to

$$R(c;\theta) = \begin{cases} \left(L_1(\theta)-L_2(\theta)\right)P(x \geq c|\theta) & \text{for } \theta < \theta_o \\ \left(L_2(\theta)-L_1(\theta)\right)P(x < c|\theta) & \text{for } \theta \geq \theta_o . \end{cases} \tag{25}$$

Consider first the situation where $\theta < \theta_o$. At $\theta = \theta_o$, $\left(L_1(\theta)-L_2(\theta)\right)P(x \geq c|\theta)$ is zero. As θ approaches $-\infty$, this (positive) quantity moves to 0. Hence there exists a value θ_1 at which this function reaches a maximum. Let $L_1(c)$ be this maximum. Likewise, let $L_2(c)$ be the maximum of $\left(L_2(\theta)-L_1(\theta)\right)P(x<c|\theta)$ when $\theta \geq \theta_o$. Then $M(c) = \max\{L_1(c), L_2(c)\}$, and the minimax passing score is the test score c_o at which $M(c)$ is the smallest.

Given c, both $L_1(c)$ and $L_2(c)$, and hence $M(c)$, may be obtained via numerical procedures such as the Newton-Raphson iteration process. The process is rather involved; however, it can be simplified by replacing the two probabilities $P(x \geq c|\theta)$ and $P(x<c|\theta)$ by two appropriate logistic functions. Let $\mu(\theta)$ and $\sigma(\theta)$ be the mean and

standard deviation described in Section 5. Then, approximately,

$$P(x<c|\theta) = e^y/(1+e^y)$$

and

$$P(x\geq c|\theta) = 1/(1+e^y)$$

where $y = 1.7\{c-\mu(\theta)\}/\sigma(\theta)$. By using these logistic expressions, the two derivatives with respect to θ which form the basis for the Newton-Raphson process will involve only rational forms of the exponential functions, and thus can be obtained without undue difficulty.

The location of the test score c_o at which the maximum risk $M(c)$ is minimized is somewhat tedious, since the algebraic form of $M(c)$ as a function of c is not known explicitly. Hence numerical procedures such as the Newton-Raphson iteration may not be applicable. It may be noted, however, that the test score x varies from 0 to the maximum of $x_m = \sum_{i=1}^{n} a_i$ via only a finite number of points. (When all item discriminations are equal, x can take only n+1 points; these may be taken conveniently as 0,1,2,...,n.) The location of the minimax passing score c_o may now be accomplished by computing the value of $M(c)$ at several equally spaced points in the interval $(0,x_m)$, and then by selecting the point at which $M(c)$ is the smallest. A refinement of this approach may be carried out by plotting $M(c)$ against c, and then by drawing a smooth curve through the points $\{c,M(c)\}$. The place at which the smooth curve is peaked may then be taken as the minimax passing score.

9. ITEM SELECTION UNDER NORMAL LOSSES

The item selection process described in Section 6 for the case of constant losses may be generalized to normal losses as follows:

1. For each item, compute the maximum risk defined as

$$\delta_i = \max_{\theta} \{L_1(\theta)p_i(\theta) + L_2(\theta)\{1-p_i(\theta)\}\} \tag{26}$$

where

$$p_i(\theta) = \exp\{a_i(\theta-b_i)\}/\{1+\exp\{a_i(\theta-b_i)\}\}.$$

2. Then select the n items which show the highest δ values.

10. SUMMARY

This paper provides a minimax decision framework in which three issues in mastery testing based on the two-parameter logistic model are approached. The issues deal with setting passing scores, assessing decision efficiency, and selecting it s to maximize decision efficiency. The losses or disutilities under consideration have constant or normal ogive form. It is found that, within the context of minimax decisions, the item selection procedure based on maximum information may not provide the best decision efficiency.

BIBLIOGRAPHY

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Co.

Dykstra, R. L., Hewett, J. E., & Thompson, W. A , Jr. (1973). Events which are almost independent. Annals of Statistics 1, 674-681.

Ferguson, T. S. (1967). Mathematical statistics: A decision-theoretic approach. New York: Academic Press.

Haley, D. C. (1952). Estimation of the dosage mortality relationship when the dose is subject to error. Technical Report No. 15. Stanford University Applied Mathematics and Statistics Laboratory.

Huynh, H. (1980). A nonrandomized minimax solution for passing scores in the binomial error model. Psychometrika 45, 167-182.

Lord, F. M. (1980). Practical applications of item response theory to practical testing problems. Hillsdale, New Jersey: Erlbaum (in press).

Novick, M. R. & Lindley, D. V. (1978). The use of more realistic utility functions in educational applications. Journal of Educational Measurement 15, 181-191.

Schlaifer, R. (1969). Analysis of decisions under uncertainty. New York: McGraw-Hill.

Wilcox, R. R. (1976). A note on the length and passing score of a mastery test. Journal of Educational Statistics 1, 359-364.

Zacks, S. (1971). The theory of statistical inference. New York: Wiley.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance of Joseph C. Saunders is gratefully acknowledged.

APPENDIX A

A Computer Program for Minimax Decision Analysis for the Two-Parameter Logistic Model under Constant Losses

Disclaimer: This program has been written with care and tested under a variety of conditions. The author, however, makes no warranty as to its accuracy and functioning, nor shall the fact of its distribution imply such warranty.

```
C*****************************************************************
C      A FORTRAN PROGRAM FOR THE COMPUTATION OF MINIMAX PASSING SCORE
C      AND DECISION EFFICIENCY FOR THE TWO-PARAMETER LOGISTIC MODEL
C      WITH CONSTANT LOSSES WHICH ARE EQUAL TO ZERO OVER A SELECTED
C      INDIFFERENCE ZONE.  THE NORMAL APPROXIMATION IS USED TO DESCRIBE
C      THE CONDITIONAL DISTRIBUTION OF THE TEST SCORE AT EACH ABILITY
C      LEVEL, HENCE THE PROGRAM IS APPROPRIATE WHEN THE NUMBER OF TEST
C      ITEMS IS SUFFFICIENTLY LARGE.
C
C      INPUT DATA CARDS ARE:
C          FIRST CARD:  TITLE OF THE PROBLEM. ENTER ANYTHING YOU WANT.
C          SECOND CARD: ENTER THE FOLLOWING DATA, USING THE FORMAT
C              (I10,3F10.5)
C                N ... NUMBER OF ITEMS
C                T1... LOWER LIMIT OF THE INDIFFERENCE ZONE
C                T2 .. UPPER LIMIT OF THE INDIFFERENCE ZONE
C                Q ... LOSS RATIO
C          THIRD CARD:  INPUT FORMAT FOR THE READING OF EACH PAIR O
C                ITEM PARAMETERS. AN EXAMPLE IS (2F10.5).
C          FOLLOWING IN THE INPUT DECK ARE N CARDS, EACH CARD
C          CONTAINING THE DISCRIMINATION AND DIFFICULTY OF ONE
C          ITEM, KEYPUNCHED IN THAT ORDER.
C
C     THE PROGRAM IS SET UP FOR TESTS WITH UP TO 200 ITEMS. IF THERE
C     ARE MORE THAN 200 ITEMS, SIMPLY CHANGE THE DIMENSIONS OF A AND B
C     IN THE FOLLOWING DIMENSION STATEMENT TO A(N) AND B(N).
C*****************************************************************
      DIMENSION A(200),B(200),FCT(20)
    5 READ(5,95,END=99) (A(I),I=1,20)
   95 FORMAT(20A4)
      WRITE(6,195) (A(I),I=1,20)
  195 FORMAT('1','MINIMAX DECISION ANALYSIS FOR THE TWO-PARAMETER'/
     *        T2,'LOGISTIC MODEL. TITLE OF THIS PROBLEM IS:'/T2,20A4)
      READ(5,100) N,T1,T2,Q
  100 FORMAT(I10,3F10.5)
      TOL=.00001
      READ(5,95) FCT
      WRITE(6,200) N,T1,T2,Q,TOL
  200 FORMAT(T2,'NUMBER OF ITEMS .............',I4//
     *       T2,'INDIFFERENCE ZONE ON THE ABILITY THETA SCALE'/
     *       T2,'     LOWER LIMIT (THETA-ZERO).',F10.5/
     *       T2,'     UPPER LIMIT (THETA-ZERO  '/
     *       T2,'                 PLUS EPSILON).',F10.5//
     *       T2,'LOSS RATIO Q .................',F10.5/
     *       T2,'TOLERANCE ERROR ..............',F10.5//
     *       T2,'ITEM PARAMETERS'/
     *       T2,'ITEM ID     DISCR.       DIFF.'/)
      DO 10 I=1,N
      READ(5,FCT) A(I),B(I)
      P1=EXP(A(I)*(T1-B(I)))
      P1=P1/(1.+P1)
      P2=EXP(A(I)*(T2-B(I)))
      P2=Q*(1.-P2/(1.+P2))
      D=P1
      IF(P1.LT P2) D=P2
      FOR=EXP(A(I)*((T1+T2)/2-B(I)))
      FOR=A(I)*FOR/((1+FOR)**2)
   10 WRITE(6,220) I,A(I),B(I)
  220 FORMAT(T4,I4,F12.3,F12.3)
      CALL SCORE(N,A,B,T1,T2,TOL,Q,CZERO,ETA)
      WRITE(6,230) CZERO,ETA
  230 FORMAT(//T2,'FINAL RESULTS'//
     *       T2,'FINAL MINIMAX PASSING SCORE',F10.5/
     *       T2,'DECISION EFFICIENCY ......',F10.5//)
      GOTO 5
   99 WRITE(6,245)
  245 FORMAT(T2,'** NORMAL END OF JOB **'/
     *       T2,'    PROGRAM WRITTEN BY'/
     *       T2,'    HUYNH HUYNH'/
     *       T2,'    COLLEGE OF EDUCATION'/
     *       T2,'    UNIVERSITY OF SOUTH CAROLINA'/
     *       T2,'    COLUMBIA, SOUTH CAROLINA 29208'/
     *       T2,'    JULY 1980')
```

```
      STOP
      END
C
      SUBROUTINE SCORE(N,A,B,T1,T2,TOL,Q,CZERO,ETA)
      DIMENSION A(1),B(1)
      AA=1./6.28318**.5
      P=1./(Q+1.)
      CALL NORMAL(P,CZERO)
      XM1=0.
      XM2=0.
      SD1=0.
      SD2=0.
      DO 10 I=1,N
      P1=EXP(A(I)*(T1-B(I)))
      P1=P1/(1.+P1)
      P2=EXP(A(I)*(T2-B(I)))
      P2=P2/(1.+P2)
      XM1=XM1+A(I)*P1
      XM2=XM2+A(I)*P2
      SD1=SD1+A(I)*P1*(1.-P1)
   10 SD2=SD2+A(I)*P2*(1.-P2)
      SD1=SD1**.5
      SD2=SD2**.5
      WRITE(6,200) XM1,SD2,XM2,SD2
  200 FORMAT(/T2,'NORMAL APPROXIMATION FOR TEST SCORES'/
     *       T2,'AT LIMITS OF INDIFFERENCE ZONE'//
     *       T2,'LOWER LIMIT : MEAN .....',F10.3/
     *       T2,'                S.D. .....',F10.3//
     *       T2 'UPPER LIMIT : MEAN .....',F10.3/
     *       T2,'                S.D. .....',F10.3/)
C
      CZERO=(XM1+XM2+(SD1+SD2)*CZERO)/2.
C
C     WRITE(6,205) CZERO
C 205 FORMAT(T2,'STARTING CZERO',F10.5)
   20 Z1=(CZERO-XM1)/SD1
      Z2=(CZERO-XM2)/SD2
      H=.5*ERFC(-.7071068*Z1)+Q*.5*ERFC(-.7071068*Z2)-1.
      HP=AA*(1./SD1 *EXP(-Z1**2/2)+Q/SD2 *EXP(-Z2**2/2))
      D=H/HP
      IF(ABS(D).LT.TOL) GOTO 30
      CZERO=CZERO-D
C     WRITE(6,210) CZERO
C 210 FORMAT(T2,'UPDATED CZERO ',F10.5)
      GOTO 20
   30 RZERO=Q*.5*ERFC(-.7071068*Z2)
      RSTAR=Q/(Q+1.)
      WRITE(6,220) RZERO,RSTAR
  220 FORMAT(T2,'MINIMAX VALUES'/
     *       T2,'  WITH USE OF TEST SCORES .....',F10.5/
     *       T2,'  WITH NO USE OF TEST SCORES ..',F10.5)
C
      ETA=1.-RZERO/RSTAR
      RETURN
      END
      SUBROUTINE NORMAL(P,X)
      D=P
      IF(D-.5) 9,9,8
    8 D=1.-D
    9 T2=ALOG(1./(D*D))
      T=SQRT(T2)
      X=T-(2.515517+0.802835*T+0.010328*T2)/(1.0+1.432788*T+0.189269*T2
     &     +0.001308*T*T2)
      IF(P-0.5) 10,10,11
   10 X=-X
   11 RETURN
      END
```

A View on the Future of Mastery Testing

A VIEW ON THE FUTURE OF MASTERY TESTING

Anthony J. Nitko

University of Pittsburgh

These remarks were made as part of the symposium "First year of the Mastery Testing Project. Technical advances, applications, and conjectures" at the annual meeting of the American Educational Research Association, Boston, April 7-11, 1980.

As is pointed out in the Overview, the Mastery Testing Project has made important strides in solving several psychometric problems associated with setting cutting scores on tests for the purpose of making mastery decisions. It has been encouraging that the research has taken as its central concern making effective and consistant decisions. This focus has contributed to the reformulation of testing issues in the decision context--away from the traditional view of the measurement of individual differences and toward a view of classification decisions within the context of instruction.

A second encouraging aspect which contributes to a future view of mastery testing is the project's use of the binomial error model and the beta-binomial distribution. In the past, most testers have applied decision theoretic statistical methods to a normal distribution model, assuming that both measurement error and ability are distributed normally. The Mastery Testing Project has broken with this tradition. In a formal and rigorous way, the project has shown that other assumptions about the mathematical form of human behavior can be plausible. Thus, solutions to testing and classification problems can be modeled on distributions other than the normal distribution. Eventually, this work will help to dispel the enchantment of test users with the nineteenth century view that human abilities are "naturally" normally distributed. Unleashed from the constraints of a Gaussian view, new vistas of human accomplishments are possible in the future.

The strong true score model adopted by the Mastery Testing Project has helped to advance a broader view of what it means to have a "reliable" test. This means that in the future test developers will be more concerned with the consistency of decisions made using test scores than they have in the past. Further, wider use of the raw agreement and kappa indices are to be expected. In addition, since these indices have a broader application than in mastery testing alone, and since their statistical form has been rigorously traced by the studies of the Mastery Testing Project, there should be a spillover of the technical knowledge gained in this project to other areas.

The Mastery Testing Project has focused on only one view of what it means to be a master. The findings of the studies reported here will give tremendous creditability to this one view of mastery because they have put it on a technically rigorous psychometric foundation. In this view of mastery, a "master" is one who can perform correctly more of essentially the same kind of task. What is to be learned is conceived of essentially as a large domain of test items. The test administrator selects a random (or representatively random) sample of items from this domain and administers them to the examinee. This tester's interest is in estimating either the number or percentage of the tasks in the domain to which the examinee can respond correctly.

This is a useful model for a number of learning objectives, especially at an elementary, minimal competence level. But the model tends to equate mastery with information store and to limit this store to verbal information. This view is appropriate, for example, when estimating the proportion of simple addition facts known, or number of three digit, two addend arithmetic problems that can be solved.

In the future, one can speculate that such a view will not be applicable to other important learning problems. Cognitive psychologists, for example, have studied the differences between "expert" and "novice" performers of complex, problem solving tasks. They find that experts differ from novices on qualitative attributes, not just on the amount of information stored. For example, on inductive reasoning tasks, Pelligreno and Glaser (1979) found that competent performers have (a) better management of memory, (b) better knowledge of the constraints in a given problem solving situation, and (c) better representation of the structure or organization of the knowledge base that is relevant to the problem at hand.

Teaching and learning directed toward this latter, more cognitive view of what it means to have competence or mastery, is quite different than the "domain of tasks" view currently adopted by most educationists. In the future, we can expect that the cognitive view will offer insights into how to diagnose learning problems and design teaching qualitative aspects of competence, not just its quantitative aspects.

But these newer cognitive views of mastery are not yet ready to be applied. A great deal of research remains to be done before the state of knowledge is at a level where application to test development is possible. Thus, the lag between these psychological views and development of psychometric theory is to be expected and we cannot fault the Mastery Testing Project for not attending to these issues. It is the nature of the beast, that psychometric theorists have to wait until psychological problems are better formulated before attempting to apply quantitative methods to their solutions. Perhaps at the end of the fourth year of the Mastery Testing Project, it can be reported that Huynh and his colleagues have applied their tremendous talents to the measurement of a new kind of mastery or expertise.

BIBLIOGRAPHY

Pelligrino, J. W. & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. Intelligence 3, 187-214.