INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

University Microfilms International
300 North Zeeb Road
Ann Arbor, Michigan 48106 USA
St. John's Road, Tyler's Green
High Wycombe, Bucks, England HP10 8HR

DECISION RELIABILITY AND CLASSIFICATION VALIDITY FOR DECISION ORIENTED CRITERION-REFERENCED TESTS

by

JANE FAGGEN

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1977

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

5/5/77
date

Chairman of Examining Committee

5/6/77
date

Executive Officer

Dr. Donald A. Rock
Dr. Max Weiner
Supervisory Committee

The City University of New York

Abstract

DECISION RELIABILITY AND CLASSIFICATION VALIDITY
FOR
DECISION ORIENTED CRITERION-REFERENCED TESTS

by

Jane Faggen

Adviser: Professor Alan L. Gross

The purposes of this research are threefold. In the first stage, a statistical model to study the relationships between reliabilities and validities of content-bound criterion-referenced test item pools is developed. Two item parameters are defined: a_i is the probability of a Master answering item i correctly and b_i is the probability of a non-Master answering item i incorrectly. With the use of these item parameters, two useful psychometric properties of the item pools--decision reliability (R_D) and classification validity (V_C)--are developed. The decision-reliability of an item pool is defined as the probability that a randomly selected examinee who is administered two randomly constructed n item tests is classified as either a Master on both tests or as a non-Master on both tests. For a given item pool, there exists a matrix of R_D's, one for each possible pair of test

length and cut score values. The classification validity measures the degree to which Masters score at or above the cut score on a given test and non-Masters score below the cut score. For each R_D matrix, there is a corresponding V_C matrix.

Several interesting theoretical relationships are derived concerning R_D, V_C, and test length, cut score, and the two item parameters. In particular, the theory provides a method for deriving R_D and V_C values as functions of test length, cut score, the item parameters, and the knowledge of the learner state of the examinee--Master or non-Master. A most important relationship between R_D and V_C is derived: R_D is an upper bound to $|V_C|$. Based on this theory, several predictions are made concerning the various relationships between the test length, cut score, R_D, and V_C.

In the second stage, a guide is presented to enable practitioners to generate the decision reliability matrix and the corresponding classification validity matrix associated with a heterogeneous item pool. The use of Bayesian methods is explored to provide estimates of the several parameters of interest.

In the third and final stage, the entire theory and procedures are applied to a mathematics item pool which is currently monitoring student progress in a self-pacing format. The predictions made in stage one are borne out successfully.

Acknowledgements

There are many people who generously contributed their encouragement and support while this research was being carried out. First and foremost, I cannot adequately express my gratitude to my adviser, Dr. Alan L. Gross, who gave his time unreservedly and presented so many intellectual challenges. The innumerable conferences with him provided me with refreshing ideas, inspiration, and continuous guidance.

The two other members of my dissertation committee, Dr. Donald Rock of Educational Testing Service, and Dr. Max Weiner, Director of the Center for Advanced Study in Education (CASE) at the City University of New York (CUNY), offered new insights along the way and assisted me in clarifying my thoughts. Dr. Edward Adams, Research Director, International Business Machines Corporation, read and reacted to an early version of this work. Dr. Carol K. Tittle, Project Director at CASE, and Dr. Richard M. Wolf, Professor of Psychology and Education at Teacher's College, Columbia University, presented constructive comments as readers on my examination committee.

There are many others to whom I am indebted as well. In particular, Dr. Carl E. Helm, formerly Professor of Education at the CUNY Graduate School and currently Professor of Community Medicine and Biomathematics at the College of Medicine and Dentistry of New Jersey-Rutgers Medical School, persuaded me to study educational psychology. Dr. Melvin Novick, Professor of Educational Statistics at the University of Iowa, inspired me to continue a study of Bayesian statistics and psychometrics.

The staff at the Hunter College Mathematics Learning Center and the CUNY Computer Center provided the data for these analyses in usable format and often made unusual arrangements for collecting and assembling the data sets to fit my research needs. Educational Testing Service made time available for my completion of the writing of this dissertation and arranged for the typing and preparation of the manuscript.

Finally, I give my heartfelt thanks to Hattie, John, Margaret, Patricia, Peter and Robert, for without them neither I nor this work could have happened.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Criterion-referenced tests are being applied in ever-widening educational settings to monitor students' progress through a variety of individualized instructional programs. A criterion-referenced measurement is one by which test scores are related to a preselected standard of performance rather than to the distribution of a large normative sample of scores. Instructional management systems require constant decision-making, usually on the basis of test scores, as to which individuals have acquired the minimal standards of proficiency necessary for mastery. To justify the use of these tests in the decision-making process, one must be able to assess the "reliability" and the "validity" of these measures. Unfortunately, the reliability and validity indices, as well as the item analysis methods of classical test theory, are not satisfactory when applied to the criterion-referenced measurement situation, in which test performance is related to the absolute standards of Mastery or non-Mastery rather than to the performance of other students. The assumptions underlying the classical linear test theory model and applicable to norm-referenced tests are not relevant to a measurement scheme in which the proficiency of a specified group of learners is <u>not</u> distributed along a continuum.

The purposes of this research are three-fold. The first

stage is concerned with theory development. Definitions of the statistical parameters decision-reliability (R_D) and classification-validity (V_C) for content-bound criterion-referenced test item pools are introduced, and a relationship between them is derived. An item pool is a large set of items (usually more than 100) matched to behavioral objectives from which individualized tests, containing ten to thirty items, are constructed on a randomized basis. It is from such item pools that mastery tests are generated for Master/non-Mastery decision-making.

The problem of estimating these parameters will be further investigated following a Bayesian approach. In classical statistics, population parameters of distributions are estimated by collecting samples of data, and then calculating sample statistics which are used in estimation procedures. From a classical approach, the collections of sample data are the essential ingredients for parameter estimations. On the other hand, a Bayesian procedure provides a formal quantitative method for incorporating into the analysis, not only the actual data sample itself, but two additional pieces of information: the researcher's prior beliefs about the value of the parameter of interest and the strength of those beliefs. Using Bayesian methods, practitioners first express beliefs about the unknown parameter's value by selecting an appropriate prior distribution function for the parameter.

Researchers "choose" a distribution function with a shape--mean, mode, and variance--which is compatible with their beliefs about the probable values for the parameter and is mathematically tractable. Next, Bayesian analysts express the strength of their beliefs by "equating" this strength to the information available in some hypothetical data sample of a size, say n_H: A high n_H for the hypothetical sample signifies a very strong belief, whereas a small n_H implies the opposite. The Bayesian methods detail exactly how to combine the analysts' prior beliefs concerning the value of the unknown parameter (prior distribution) and the strength of those beliefs (hypothetical data sample size, n_H) with the actual sample statistics (e.g., mean and variance) and the actual sample size, n_A, to produce an updated statement of beliefs. This updated statement of beliefs takes the form of a posterior distribution for the unknown parameter of a given form with a mean, a mode, and a variance, in addition to a posterior n size. When the actual data n_A is large compared to the stipulated n_H of prior beliefs, the Bayesian and the classical results are similar. Provision is also made for non-informative prior beliefs and in this latter case, as well, the classical and the Bayesian estimation procedures yield similar results. In a sense, the classical parameter estimation approach is a special case of the Bayesian methods.

During this first stage, the relationships between R_D and the test length n and the cut-score i are studied. In addition, the effects of n and i on V_C are examined. Several predictions

are made based on these theoretical relationships.

The second stage of the research presents a guide that will enable practitioners—curriculum developers, test constructors, and evaluators--to generate both a decision reliability matrix and a classification validity matrix for a criterion-referenced item pool. A method for estimating an average of the indices of all of the items in the item pool is suggested.

In the third and final stage, the entire theory and procedures are applied to a mathematics item pool which is currently monitoring student progress in a self-pacing format. The predictions made in stage one are borne out successfully. In addition, the resulting analyses indicate to the practitioner the maximum reliabilities and classification validities which can be expected when a particular operating item pool provides the items for criterion-referenced tests of a specified length and cut score.

Chapter 1

Individualized Instructional Programs and Test Use

Relevant educational literature in the area of criterion-referenced testing applied to individualized instructional programs has been concerned with two types of papers: those which describe the individualized programs themselves and those which are addressed to the theoretical and empirical problems of measurement of student performance. A brief review of the rationale and functioning of five instructional models—two widely implemented and three of local design—will be presented in this chapter. A more detailed and extensive treatment of some of the current theoretical formulations for reliability and validity in the criterion-referenced setting will be presented in the next chapter. These surveys will provide a useful perspective from which to view the purpose and the scope of the present research.

Large educational systems such as the Program for Learning in Accordance with Needs (Flanagan, 1967, 1969), and the Individually Prescribed Instruction programs (Glaser, 1968) continue to demonstrate the feasibility of instructional systems that are designed to be adaptive to the unique requirements of individual learners, many of whom have widely varying competencies and needs. Individualization is accomplished in a variety of ways: per-

mitting students some choice in determining the skills they will learn, establishing organizational procedures that permit students to progress at different rates, and developing alternative instructional sequences for teaching skills. The curriculum is specified in terms of behavioral objectives while the goals and instructional resources are individualized for the purpose of maximizing learning. This highly structured instructional framework results in substantial amounts of program time devoted to student testing.

The Program for Learning in Accordance with Needs (PLAN) was developed in the 1960's by the American Institute for Research and the Westinghouse Learning Corporation to provide an innovative approach to correct educational deficiencies revealed by project TALENT (Flanagan, Davis, Dailey, Shaycraft, Orr, Goldberg, & Neyman, Jr., 1964). As of 1974, Westinghouse Learning Corporation is responsible for the monitoring and marketing of Project PLAN materials which are available in the four areas of language arts, mathematics, science, and social studies. Over 151 schools in 27 states enrolling approximately 60,000 students are currently using these materials as well as a medium-sized computer with input/output terminals in each school. The computer installation, operated by Westinghouse Learning Corporation, manages the implementation of PLAN. Each student is assigned a unique program of study consisting of a comprehensive set of educational objectives which are sequenced and modularized. The modules are instructional

packages which consist of about five behavioral objectives and take about two weeks to complete. There are several different instructional approaches for each module, called teacher-learner units (TLU's), which are coded according to: reading level of student, quantity of reading materials, degree to which the student must interact in a social situation, and the variety of learning activities included. At the beginning of the school year on the basis of questionnaire and test information, an individual program of study is prepared for each student. An attempt is made to match the student learning style to an appropriate instructional method in order to maximize student learning. The student's progress is monitored throughout the instructional program at various decision-making nodes by the use of module pretest and posttest scores.

Although PLAN is an imaginative and comprehensive instructional program, including excellent teacher and student materials and a well-developed computer management system, there is still no well-established theory of instruction to guide the construction of meaningful, optimal assignment tables of students to instructional methods. Hambleton (1974), in a review of testing and decision-making procedures for selected individualized instructional programs, has observed that "much remains to be done in the general areas of instructional theory, decision-making procedures, and testing technology before one could fully and successfully implement many of the features of the program (p. 386)."

Individualized Prescribed Instruction (IPI) is a project of

the Learning Research and Development Center (LRDC) at the University of Pittsburgh. For Cooley and Glaser (1969) individualized education represents "the adaptation of instructional procedures to individual requirements (p. 576)." Their conceptualization of an individualized system consists of three major components: (1) educational goals, (2) individual capabilities, and (3) instructional means. The IPI model is more precisely defined by the following sequence of operations (Glaser, 1968):

(1) Specification of the skills to be learned in terms of observable student behavior.

(2) Assessment of an individual's skills upon entry to a course of instruction.

(3) Assignment or election of educational alternatives fitted to the student's entering proficiencies.

(4) Continuous assessment and monitoring of the student's performance and progress.

(5) Completion of available instructional sequences as a function of assessment of student performance and criteria for proficiency.

(6) Collection of data for improving the instructional system.

It is expected that students will work their way through this individualized program in a systematic way.

For successful implementation of this model, a teacher must manage the learning activities of a large number of students with diverse needs and skills. Test data serve as the primary source of information enabling teachers to make differential decisions regarding instruction for individual children. These tests are constructed and used to determine the instructional objectives on which the examinee has met an acceptable performance level. Tests designed for ascertaining whether or not an examinee has met a predetermined behavioral standard or degree of mastery are called criterion-referenced tests. These tests stand in contrast to the more widely known norm-referenced tests, which are designed to produce test scores suitable for ranking examinees on the trait or ability measured by the test.

The teacher uses criterion-referenced test results to place the student within the curriculum sequence at the beginning of the school year as well as to monitor and diagnose the student's continuous learning in the program of study. Placement tests as well as pretests and posttests are available for each unit of instruction. All tests are criterion-referenced in the sense that test performance is compared to a predetermined minimal standard of skill and knowledge competence. All test items are designed to measure specific program objectives. Placement tests provide student gross achievement levels; pretest performance determines the assignment of a particular group of learning tasks within a specified unit; mastery on each objective covered in the unit

posttest is required for student assignment to the next study unit. Hambleton (1974) concludes that "the IPI test model is the most thoroughly developed and comprehensive in scope."

Both PLAN and IPI are large-scale implemented instructional models which incorporate the mastery learning paradigm developed by Bloom (1968) and based in part on Carroll's (1963) "model of learning." According to Carroll's formulation, the learning time for a given student is a complex function of a number of basic aptitudes such as verbal ability, memory ability, and spatial ability. Thus, the essence of individualized instruction is that such a program allows students to learn or reach criterion attainment (mastery) at their own individual rates. The amount of time that students need and, therefore, the degree of mastery that will be attained is affected by the quality of the instruction and the students' ability to understand and profit from the instruction. The amount of time that students actually spend on learning depends on the students' willingness to learn and the time actually allocated for learning.

Bloom (1968) suggested a method for incorporating the major variables identified by Carroll (1970) into a strategy for mastery learning. He identified some preconditions, indicated the required operating procedures, and suggested possible measurable outcomes. A necessary precondition for such a learning model is a definition of mastery and evidence for determining whether or not students have attained that state. Useful program operating procedures include breaking up a course or subject into smaller units which involve

one or two weeks of learning activity. The ideas of Gagne (1965) and Bloom (1956) provide the basis for forming hierarchies of learning tasks within each unit or module. Brief diagnostic progress tests, based on unit objectives, are administered to students to determine their mastery or non-mastery status and their subsequent learning assignment, either remediation and review or progression to the next program segment. These tests are referenced to an objective and to a standard of proficiency or excellence with respect to that objective and are often called criterion-referenced tests. Achievement standards in the mastery learning milieu emphasize cooperation between students and teachers in learning rather than competition among learners. Affective, as well as cognitive, outcomes are extremely important for any educational strategy. Within the mastery learning scheme particular sets of procedures are selected as those most appropriate for specific students. Students experience the positive reinforcement which comes from frequent indications of continuing self-development. In the long run, it is hoped that these students, who have enjoyed the success associated with such types of early educational environments, will develop an interest in continual learning throughout life.

This mastery learning strategy has also provided the foundation for three smaller scaled individualized instructional programs which are currently operating at two units of the City University of New York. The Multi-Media Tutorial Laboratory at the City College of New York in General Biology 5, an introductory course, was intro-

duced to students in the spring of 1972 as an alternative individualized approach to teaching a laboratory science. In keeping with the multimedia format of the program, the biology laboratory room features TV, cassettes, film loops, and slide projectors, in addition to the traditional pieces of equipment. The laboratory work is divided into 13 units and students have about one and one-half to two weeks to complete each unit. Students receive a step-by-step outline for each unit which indicates a guide to the subject matter and a laboratory performance sequence. Students are scheduled for work in a particular laboratory area, listening to tapes, reading, working on an experiment, performing optional tasks, etc. The Multi-Media Tutorial Laboratory is open five days a week for a minimum of 33 hours and is always staffed by qualified instructors.

Recitations are held by the instructor according to a posted schedule. When students feel they understand the unit objectives, they take a "written" quiz which is administered by computer terminals (Hazeltine 2000) for ease of generating test items from objective based item pools and for ease of scoring and record keeping. Those who score at least 72% go on to the next unit. Those, who do not receive the passing grade, complete further work on the unit and may take a total of three tests on the same unit, if needed.

The other two City University of New York individualized instructional programs are mathematics courses offered at Hunter College. In addition to embodying elements of the mastery learning model into their operating structures, the Hunter College Learning Center

based courses also selected elements from the Keller (1968) personalized system of instruction (PSI). The Keller model, based on an extension of behavior theory principles, requires both specification of terminal behavior (a clear description of what is to be learned) as well as effective management of rewards for desired behavior (study and learning).

Features of any implementation of the Keller plan are administrative strategies and structures designed to apply these theory based principles to the particular educational program of interest. These instructional systems provide an integration of objective specification, instruction procedures, and evaluation techniques. Elements common to all Keller type programs include course content modules, detailed instructional objectives, frequent tests, student proctors, subject-matter mastery, and student determined progress. An excellent critique of personalized instruction is provided by Ryan (1974) and includes a description of the underlying theory, a discussion of several applications and a presentation of a variety of interesting research results relating to academic performance, student attitudes, and student course withdrawal rates.

The adoption of an open enrollment policy by the City University of New York in the fall of 1970 provided the impetus to the Mathematics Department of Hunter College to offer several first year courses, one of which was to be remedial in design. Under open enrollment, any high school graduate in New York City is guaranteed a place in one of CUNY's 18 two or four year colleges. Freshmen

registration in the city system increased from approximately 20,000 students in the fall of 1969 before open enrollment to about 35,000 in the fall of 1970. The fall 1975 registration for freshmen was approximately 40,000. Due to the widely ranging secondary school preparations of the entering students, the various CUNY units adopted strategies for assessing student needs, counseling students into appropriate first year courses, and providing remedial learning experiences, particularly in mathematics and English.

There are two mathematics programs which are currently offered at the Hunter College Mathematics Learning Center and make use of course related item pools. A one-semester remediation course (45.001) is required of all Hunter College entering students who score 30 or below on the computation portion of the California Achievement Test (CAT), Form 5A. Freshmen, who complete this course satisfactorily, earn one college credit; advanced standing students earn zero credit. An introductory one semester mathematics course (45.100) is also available at the Learning Center for all qualified students. The three credits earned in the introductory course can be applied towards the mathematics-science requirements of Hunter College.

In the fall of 1975, approximately 1650 students were utilizing the various components of the Mathematics Learning Center: 116 self-study learning stations, 10 group slide units, slide viewers, audio-cassette players, programmed materials and tutorial assistance. Approximately 200 of these students were registered in a pre-calculus

course (45.120) which does not make use of an item pool for monitoring student progress. The Center is open to all students five and one-half days per week and is continuously staffed by tutors and technicians.

Progress through the remedial course (45.001) is self-paced and students attempt unit pretests and posttests when they feel that they are ready. These unit tests are provided with the student's set of learning materials. A unit usually consists of a daily lesson. Students who pass the posttest proceed to the next unit. Students who do not pass return to specified resources and/or their tutors for help. After completing all of the units which comprise a module or block of units, the student takes a module test. At the present time, there are eight module tests, each available in ten forms. All of the test items are matched to specific behavioral objectives. Items contained in each form were written and selected because they were judged to be parallel in content and difficulty across the ten forms. Two of these tests, Tests 1A and 1B, are computer generated and all test forms were prepared following computer item selection. The passing grades for module tests range from 75% to 84% and each test must be taken as many times as is necessary (i.e., until the student passes the test).

The introductory course (45.100) also provides a self-pacing format for the individual with pretests and posttests. Heavy use is made in this course of slides and tape presentations. The five module tests are each available in 50 parallel forms, all of which

have been computer generated from their respective item pools, following a stratified random sampling plan. The items are matched to a set of specified objectives, and a set of these behavioral objectives is available to each student at the beginning of the course. Currently, students respond to test items on mark-sense cards. Test scores are computed upon test completion and appropriate prescriptive messages are provided for each student. Students who miss items are directed to reinforcing learning materials. Each student is permitted to take each module test twice and receives the higher of the two grades. Every student takes a comprehensive final examination which covers the most important objectives of the course. These designated objectives are also made available to the student prior to the examination. A computer program, operated in conjunction with the mark-sense cards, provides a variety of record keeping functions including item analyses, student scores, objectives missed, etc. The data analyzed in this research come from tests generated from four of the five items pools created for this course.

In summary, all of the programs described--PLAN, IPI, the Multi-Media Tutorial Laboratory, and the Mathematics Learning Laboratory Programs--are individualized in some sense. In addition, all ultimately rely on frequent testing decisions for moving students through the various possible curriculum pathways. All of these tests, whether they are pretest or posttest, multiple choice or short answer, science, language skills, mathematics or social studies, are tightly bound to their respective curricula and are

used to facilitate the efficient monitoring of students through the instructional program. These tests are known as criterion-referenced or domain-referenced tests, and it is to an appraisal of the current state of the art and science of this extremely useful category of tests that we now turn. It may come as a surprise to some to learn that the psychometric information currently available on criterion-referenced testing, used in the critically important decision making phase of individualized instructional programs, is extremely limited.

Chapter 2

A Look at Criterion-Referenced Tests: Reliability and Validity

2.1 Criterion-Referenced Tests and Measurement

Psychometricians and instructional specialists continue to make distinctions between norm-referenced and criterion-referenced measurements. In essence, a norm-referenced score gives information which compares an individual's test performance with that of others. Percentiles and grade equivalents are norm-referenced metrics for comparing an examinee's performance with national or local norms. A principal function of norm-referenced measures is to allow for maximum discrimination among individuals (ranking of examinees) along some continuum. On the other hand, since Glaser (1963) first introduced the subject, there is disagreement in the literature concerning what is meant by both a criterion-referenced test and criterion-referenced measurement. Distinctions between norm-referenced tests and criterion-referenced tests have been presented by Block (1971a), Ebel (1971), Hambleton and Gorth (1971), Hambleton and Novick (1973), Hieronymous (1972), Livingston (1972), and Popham and Husek (1969).

A useful definition applicable to the area of individualized instruction is provided by Glaser and Nitko (1971) who characterize a criterion-referenced test as "a test that is deliberately constructed to yield measurements that are directly interpretable in terms of

specified performance standards [p. 653]." There is no explicit reference to the performance of other persons, although the specification of performance standards is an implicit behavioral reference. The three key elements in the Glaser and Nitko (1971) formulation are a test, a metric, and a criterion or performance standard. Using these concepts as a foundation, criterion-referenced test development procedures are readily formulated. For some particular subject matter area, the curriculum expert unambiguously identifies a well-defined domain of tasks that the learner must perform. The performance standards specified by this group of tasks or behavioral objectives provide the basis for the generation of a domain-referenced item pool: Test items are written and keyed to each objective. Harris (1974a) notes that such an "objectives-based test that is restricted to one or more specific objectives [p. 99]," is a mastery test rather than a general achievement test with respect to its subject matter.

Item construction and selection procedures for mastery tests are dominated by two significant themes consonant with the purposes of this mastery type of criterion-referenced test. The first theme is principally one of content validity. Guion (1977) states that notions concerning content validity, (the source of his discontent!), are neither simple nor trivial. In an effort to clarify some of the recurring ambiguities, he points out that content validity requires a clearly defined, relevant, behavioral content domain which is adequately sampled. The observed sample of behavior (e.g., test score) must be representative of the whole class of behaviors which fall within the content domain boundaries. For mastery test item pools, we shall use the term "domain certification" to encompass the objective and subjec-

tive procedures for assessing the extent to which a match exists between the written items and the instructional program's behavioral objectives. Such a mastery test cannot be content valid, unless all objectives are adequately represented by test items. The second theme is concerned with the item's sensitivity to instruction: a good item will discriminate between those who have benefitted from instruction and those who have not.

Test items may be generated by content specialists or by an item form approach. Hively (1962) introduced the notion of an item form to facilitate the definition of an entire class of test items by the substitution of elements of replacement sets for variable elements in the item forms. Numerical and non-numerical replacement sets (Hively, Patterson, & Page; Osburn, 1968) as well as linguistic transformational rules (Bormuth, 1970) have been employed with item forms. Many item form procedures make use of the computer to generate the test items.

Item selection procedures for norm-referenced tests choose items which maximize differences among individuals. For criterion-referenced (mastery) tests, items which discriminate between those needing instruction and those not needing instruction are selected, with the constraint that there are items on the test which adequately sample each of the stated objectives. Criteria for guiding the development of item sensitivity indices include: (1) the provision of unique information (e.g., the item's ability to discriminate between those who have and have not profited from instruction), (2) ease of computation, and (3) ease of interpretation.

Several criterion-referenced item analysis and selection schemes

are described by Brennan (1974), Brennan and Stolurow (1971), Cox and Vargas (1966), Crehan (1974), Haladyna (1974), Helmstadter (1974), Kosecoff and Klein (1974), Kriewall (1972), Millman (1974), Popham and Husek (1969), Shoemaker (1975), and Woodson (1974). Kriewall (1972) and Popham and Husek (1969) caution against using classical item difficulty and discrimination indices for the identification and elimination of "bad items". Popham and Husek conclude that unless there are at least matched item-objective pairs, any particular mastery test will not yield scores which are readily interpretable in terms of learner proficiency on the particular set of objectives. Millman (1974) points out while empirical item data are relevant for defining and evaluating a domain, they should not be the sole basis for selecting criterion-referenced test items. Shoemaker (1975) emphasizes that "the item universe associated with an instructional program is a critical component of that program and is, indeed, its raison d'être [p. 144]." Woodson (1972) observes that classical theory item analysis methods rely on a sample representative of a population of persons who differ in ability while criterion-referenced test item analysis requires observations at different points on the characteristic (ability) of population of observations. Cox and Vargas (1966) and Brennan and Stolurow (1971) provide useful information for identifying test items which are in need of revision. Kosecoff and Klein (1974) propose two indices which measure the extent to which items "reflect instruction". On the other hand, Haladyna (1974) and Helmstadter (1974) argue that a close correspondence exists between classical estimates of criterion-referenced item discrimination and pre- and post-instruction group difference techniques.

After a suitable collection of such items representative of one or more objectives has been assembled into an item pool, a group of test items is selected by some sampling plan--usually stratified--for a module (several objective) test. The optimal number of objectives to be included on a test of a particular multi-objective module has not yet been satisfactorily resolved (Hambleton & Novick 1973). Next, a measurement scale is selected to yield an absolute--as opposed to relative--interpretation of examinee performance. Specifically, student scores can be interpreted independently of the scores of other examinees.

In individualized instructional programs, test scores often provide the information needed to evaluate the students' mastery or non-mastery of the various modules. Decisions on student placement into the next learning module or into a remedial mode are made contingent upon whether or not the student's score equals or exceeds some predetermined criterion score. The curriculum expert or instructional program manager preselects the appropriate cut score which divides the examinee group into Masters or non-Masters. In practice, prior experience has dictated the choice of cut score, since, as yet, there is no one procedure which is recommended by all practitioners and theorists. The cut score issue has been discussed by Emrick (1971), Kriewall (1969), Millman (1973), and Novick and Lewis (1974).

In summary, a definition of a mastery test has provided the basis for effective item and test construction methods. Furthermore, it was argued that such procedures are of considerable importance in establishing content validity, which is a principal concern for criterion-referenced tests. In the next section, some of the important issues

concerning what may be analagous to construct, concurrent, and predictive validity will be considered.

2.2 Validity: An Overview

A familiar problem in the area of mental measurement concerns the establishment of the validity of an assessment instrument. Validity may be thought of as valuable information which is derived from the test score, the item scores, and/or the relationships between scores on the test of interest and other measurements. Validity coefficients index the goodness of inferences made from test scores. The two basic types of validity inferences are concerned either with internal measures (i.e., the particular trait, characteristic, aptitude, ability, or knowledge the test is purportedly measuring) or with external measures (other related behavior). These validities are not necessarily independent. The "Standards for Educational and Psychological Tests" (1974) provides useful definitions and a discussion of the four interdependent kinds of inferential interpretations--both the internal and the external type--that are usually applied to the use of tests: content validity, construct validity, and the criterion-related validities (predictive and concurrent). The purpose in this section is to present a brief review of contributions to the validation problem with respect to criterion-referenced measurement in the mastery environment.

In Section 2.1, procedures were described for constructing criterion-referenced tests. In particular, the importance of the content validity of the instrument was stressed. Millman (1974) argues

that the validity of a domain-referenced test "can best be assessed by a logical analysis of the domain definition, the item generation scheme and the individual test items [p. 7]." Claims of content validity are based on the degree of matching of the test items (or the collection in the item pool) to the previously adopted behavioral objectives. The items constituting the pool must be congruent to the specified objectives so that a test generated from the item pool, usually following a stratified sampling plan, will be optimum for forming mastery/non-mastery discriminations. Empirical data such as item difficulty and item discrimination indices are more useful for selecting items than for "certifying the domain". If items were eliminated only on a statistical basis, then the item pool might no longer be representative of the domain and test score interpretation would be limited. It is worth noting that a test constructed to make a particular set of mastery/non-mastery decisions will not be useful for a different set of performance criteria. Furthermore, a set of items which maximally discriminates masters from non-masters in a particular group of students may not be a set which yields a comprehensive domain mapping. A good test must sample the whole domain adequately.

Osburn (1968) has observed that what the criterion-referenced

> test is measuring is operationally defined by the universe of content as embodied in the item generating rules. No recourse to response-inferred concepts such as construct validity, predictive validity, underlying factor structure or latent variables is necessary to answer this question [p. 95].

This concept of item-objective matching differs from the usual norm-referenced test construction rationale which postulates a universe of behaviors, only <u>some</u> of which are directly tested by test items.

In a norm-referenced test for which items are collected on the basis of providing a wide dispersion of test scores useful for prediction or selection purposes, the items need not constitute a representative sample of the universe or domain to be tested (Davis & Diamond, 1974). These norm-referenced test items are selected, in part, for their contributions to maximum test score variance and are often items with a difficulty level (p value) of about .5.

Cronbach (1971) has suggested a method for assessing content validity for criterion-referenced tests, which is of limited value since it requires the construction of two tests. Following this approach, two independent test constructors develop a separate domain-referenced test from the same domain specifications. The two tests are given to the same set of examinees and a correlation coefficient between the sets of scores is computed. Other analysis techniques have been proposed by Hempill and Westie (1950), Lu (1971), and Bennan and Light (1973).

Construct validity for mastery tests has been viewed by some researchers as a concern with inferences regarding the proficiency of the learner with respect to the set of matched objectives from which the test items or item pool are constructed (for example, see Brennan, 1974, and Millman, 1974). The construct here is the knowledge which is specified by the stated objectives. Other theorists have proposed models for mastery tests which are based on hypotheses relating the test scores to the learner's true state of knowledge (Harris, 1974; Kriewall, 1972). The applicability of these models is determined in part by a demonstration of this type of construct validity. However a perusal of the various discussions of the validation process indicates

that these analyses are often combinations of content and construct validation and depend upon "the record of how the test was conceived, related to the instructional program, and developed [Harris, 1974a, p. 112]."

Messick (1975) argues against the usefulness of a content validity interpretation in this narrow sense of being restricted to a fixed property of a test rather than to test responses. A view of content validity which incorporates processes underlying test responses as well as the test structures themselves is preferable. Such an approach, he further argues, leads to meaningful interpretations of observed response consistencies. Messick also observes that

> tests per se do not have construct validities --or reliabilities or predictive validities either, for that matter. These are properties of test responses, not of tests, and test responses are a function of the persons making them and of factors in the environmental setting and measurement context [p. 956].

Such a view emphasizes meaningfulness of measures and this meaningfulness is evidenced only in terms of construct validity: valid interpretations of educational instruments must be made only in terms of the attributes and processes underlying task performance and not in terms of test structure variables.

Kriewall (1972) points out that classical latent trait theory does not clearly identify the trait being measured and tests generated following classical approaches do not necessarily have content validity which is obvious for criterion-referenced measurement. Kriewall conceptualizes criterion-referenced performance as a sequence

of Bernouilli trials, each having the same probability of success A_{ak}, the proficiency of the a^{th} student with respect to the k^{th} learning objective, which is defined to be the relative true score of a on all n_k items. The score distribution for a given individual on repeated tests consisting of random samples of size n drawn from learning objective k is given by

$$f(x) = \binom{n}{x} A^x (1 - A)^{n-x}$$

where f(x) is the relative frequency of occurrence of score x for the a^{th} student. Kriewall suggests an estimation of the proportion of items that make up a defined population of tasks that each examinee can answer correctly.

Harris (1974b) introduces the notion of item response homogeneity. He defines a domain such that all items are of equal difficulty for the master student: the probability of passing a randomly chosen item given that another item is passed equals one. Difficulty of an item in this sense is a function of the student's experiences and the student's prior instruction. The concept of master is defined as the proportion (generally not observable) of a population of domain items that an instructed student should be able to answer correctly. For Harris (1974b), since "the ultimate validity question is the question of the extent to which the test sorts students into the correct two categories [p. 109]," data must be accumulated demonstrating a positive relationship of the test decisions to relevent criteria. These criteria include pre-post differences in instructional proficiency (Ozenne, 1971), performance on a transfer task, and degree of subsequent

success.

It will be helpful for the reader to keep in mind the variety of content and construct validity questions which these researchers and theorists are trying to answer without always stating them explicitly: (1) Do the stated objectives exhaustively map the content of interest? (2) Do the items match the stated objectives? (3) Do masters earn perfect or near perfect scores? (4) Do non-masters earn zero or near-zero scores?

Still other practitioners, particularly learning theorists, are concerned with testing hypotheses regarding the hierarchical knowledge structures which are implicit in an ordering of the behavioral objectives. Possible relevant hypotheses include the existence of no structure, a linear hierarchical structure (the knowledge of one element presupposes a prerequisite skill, but the reverse needs not apply), or a complex hierarchy containing an array of either parallel or tree-like branches. In the teaching-learning environment of mastery testing, two types of validation efforts emerge with respect to these structures. The first requires a demonstration of the hierarchical knowledge structure of the domain. What degree of interrelationship, if any, exists between the objectives from a knowledge and from an effective learning point of view? The second investigates the degree of success of a particular instructional sequence which is predicated on an assumed knowledge hierarchy. What is the most effective ordering of presentation of the learning materials to produce maximum learning proficiency? This latter suggested approach is also representative of the third type of validation activity, dealing with the external criterion related validities, because of concern with

success of instruction. However, in this case, the success of a particular sequence is being assessed and this is a construct validation enterprise in the broad sense that has been outlined above.

These content and construct type validities just described are concerned with test score relationships of an internal type: test rationale, item content, domain structure, comprehensiveness of item-objective coverage, and learner proficiency as demonstrated by the mastery test score itself. In contrast, the criterion validities --either concurrent or predictive--provide information about relationships between the particular mastery test of interest and other external measures. These external variables include a variety of achievement level indicators (e.g., pre-post score differences, teacher ratings, final test scores, semester course grades, and general aptitude test scores), measures of retention, measures of transfer, affective measures (e.g., short-term and long-term student interest in the subject matter), and rate of learning the subject matter of the course. For a discussion of validating and predictor variables, as well as validation methods applicable to known or unobserved criterion groups (i.e., with respect to mastery/non-mastery group membership), see Darlington (1970) and Millman (1974).

The validation procedures described above often employ one of three useful types of statistical approaches. Besides the standard correlational methods, analysis of variance procedures (Ozenne, 1971), and decision-theory analysis (Hambleton & Novick, 1973) are useful techniques.

The validity coefficients proposed in this dissertation apply to the item pool itself rather than to any particular criterion-referenced

test. This validity formulation indexes the degree to which learners are correctly categorized (M or NM) on the basis of a given test generated from the item pool. The theory is developed in Chapter 3 and applications of procedures derived from this theory to existing item pools are provided in Chapter 4.

2.3 Reliability: An Overview

Once the criterion-referenced, and in this case, objective-based, test has been developed, it is desirable to associate with the instrument several appropriate indices of reliability which reflect either internal consistency, stability over time, or stability over performance on parallel tests. The reader should keep in mind that while the discussions of reliability which are included in this section are dealing primarily with the reliability of a given criterion-referenced test, the concerns of this dissertation deal with the psychometric properties of objective-based item pools, from which a variety of tests may be drawn. However, the single test reliability investigations provide a variety of interesting solutions to the many existing reliability problems.

Researchers present differing views with respect to the applicability of classical reliability theory to criterion-referenced tests. Haladyna (1974), Klein and Kosecoff (1973), Livingston (1972), and Woodson (1974), maintain that classical theory procedures need not be rejected for analyzing criterion-referenced tests. However, Brennan (1974), Glaser (1963), Hambleton and Novick (1973), Hambleton, Swaminathan, Algina, and Coulson (1975), Millman (1974), and Popham and Husek (1969), argue that the lack of sufficient test score variance

associated with criterion-referenced tests usually renders classical indices inappropriate.

A brief summary of the basic concepts of classical reliability theory will be followed by a look at a group of alternate procedures for criterion-referenced tests. Classical theory (Gulliksen, 1950; Lord & Novick, 1968) reliability as applied to norm-referenced tests is defined either as the squared correlation between true scores and observed scores or the ratio of true score variance to the variance of observed scores. One expects to find differences in true score in any large sample of persons, since the theory assumes that underlying traits have some continuous distribution in the population, that the items in the test represent a sample from this trait or skill continuum, and that there is a monotonic relationship between an individual's possession of the trait and the probability of achieving a given score on the sample of items. Since true scores are unobservable, the theory presents two basic methods for estimating reliability. The first procedure requires two test administrations of the same test or separate administrations of two parallel tests. The second procedure, using only one test administration, is concerned either with an average of item total-test score correlations or with a mean of all possible split-half test score correlations. The underlying basis for all of these estimation procedures is variance. When score variability is maximized, estimates of reliability can be large. Large reliabilities imply that students' rankings on one administration of a test will be replicated on a second test (the same test or a parallel form). Large reliabilities also imply that the test is discriminating well among the examinees, which is an important goal of norm-referenced testing.

On the other hand, for mastery testing purposes, the practitioner is often interested in assigning students to just one of two states (mastery or non-mastery) on the basis of test scores rather than discriminating maximally among the examinees. In this case reliability cannot usefully be based on variance estimates which are necessary small within each learner group (i.e., masters and non-masters). One expects to find mean score differences between pre-instructed and post-instructed groups, but it is not unreasonable for all pre-instructed students to have low scores and all post-instructed students to have perfect or near perfect scores. Since the meeting of a predetermined level of performance is an important instructional goal, which is often met, it is not surprising to find near zero internal reliability estimates (classical) for criterion-referenced tests. Such estimates are based on score variablility which is non-existent or quite small. For such a two classification assignment, high test reliability could be related to the proportion of identical learning group assignments which occurs on the basis of two test scores. Several alternate methods for determining a mastery test's consistency or dependability will be described later in this section.

For mastery tests, the underlying trait can be thought of as knowledge of a specific set of objectives. This "latent" trait is not continuously distributed in the population, but rather may be thought of as two-valued. Furthermore, one is interested in a <u>particular</u> group of students who are about to be instructed. An assumption, for mastery test theory is that students, prior to instruction, have or do not have a minimum acceptable knowledge of the modularized materials. After effective instruction, it is hoped that all will have acquired

this degree of knowledge. The distribution of module knowledge in the relevant student populations is bipolar rather than of the unimodal type, which is often assumed in classical theory. Here, the test items are not thought to be a representative sample from some larger general trait domain but, indeed, are each carefully keyed to specific objectives in the explicitly defined set of learning tasks. The key reliability issues for criterion-referenced tests used in the mastery setting are issues of replication of decision-making and accuracy of decision-making in a two-state mastery/non-mastery system.

Before discussing the various reliability indices which have been introduced in an effort to address the criterion-referenced test model, the issue of errors in testing needs to be mentioned. Errors in testing can conveniently be dimensioned in three ways (Hambleton et al., 1975). First, error definition is related to test use. When tests are used for estimating domain scores or estimating trait scores, it is reasonable to assume that the larger the difference between true score and observed score, the larger the error. For this situation, reliability estimates are based on variances or squares of differences. Thus, mean-squared error functions are most appropriate for quantifying the error. When tests are used primarily for allocation of students to one of two mastery states, errors are made whenever a master is assigned to a non-mastery state or a non-master is assigned to a mastery state. The magnitude of the difference between observed score and mastery score is not important, since the misclassification is an error for both large and small score differences. In this case, a threshold loss function is the more appropriate one to use for "measuring" error. The information provided by such a function reflects the two possible

types of learner classification, correct or incorrect.

A second method of conceptualizing error is related to the way in which the practitioner views probability. From a classical (frequentist) probability point of view, one is concerned with the properties of the observed score distribution. From a subjective or Bayesian stance, one relates the concept of error of measurement to the distribution (posterior) of the true score being estimated. Third, errors of measurement can refer to a group or to an individual and these values are not necessarily the same.

In summary, reliability issues for use of criterion-referenced tests in decision-making are concerned with a fixed, well-defined item pool, a particular group of about-to-be or already-instructed students, replicability of decision-making, correctness of decision making (a validity issue) and the concepts of errors of measurement or errors in decision-making.

2.4 Reliability Indices

The development of suitable reliability indices for criterion-referenced measures has been a frequently addressed statistical issue. Brennan (1974) presents a critical review including comments on index strengths and weaknesses. These indices differ in a variety of substantive ways: (1) relationship to instructional effectiveness; (2) required number of test administrations; (3) reliance on both pre- and posttest scores; (4) interpretations in terms of variance of student scores; (5) usefulness for estimating true scores; and (6) relevance to decision-consistency or decision-accuracy. We shall consider briefly the structures of a variety of these indices,

keeping in mind once again, that the ultimate goal of this proposed research is concerned with the psychometric properties of item pools and not of specific tests. We shall further limit the discussion to indices which are appropriate for criterion-referenced tests used in the mastery setting. For simplicity, the presentation order will be primarily chronological, since the indices differ along many dimensions.

Cox and Graham (1966) suggested a coefficient of reproducibility, which is useful in an instructional setting where objectives are analyzed as being sequential. When the tests of objectives are organized sequentially as well, theoretically, students are capable of answering all items up to a point indicative of their own level of mastery and none beyond. Test items such as these are said to form a Guttman scale. The coefficient measures the degree to which members of the instructional group produce item response patterns of the theoretically predicted form, i.e., all responses are correct up to a particular item and then all responses are incorrect.

Ivens (1970) proposes two indices of agreement useful for indicating stability or equivalence. They are relatively simple to calculate, do not depend on test score variance, but are not helpful in estimating an examinee's true score. Difference scores resulting from two test administrations (test-retest or parallel forms) are computed. The first reliability index is equivalent to the percentage of students with difference scores of a given size or less. The second index is an average of individual item reliabilities. The percentage of examinees whose item scores are the same on a test and a retest represents reliability.

Carver's (1970) method requires either the administration of a

single test form to two groups comparable in terms of instruction, or the administration of two parallel tests to one instructed group. In the former case, the percentage of students meeting the criterion in one group is compared to a similar statistic in the second group. In the latter case, a comparison is made between the percentage of examinees achieving the criterion on both tests.

Ozenne's (1971) sensitivity indices make use of pretest and posttest scores. The first test, suitable for a single instructed group, is a ratio of the variance due to instructional effects divided by the sum of variances due to instructional effects and errors of measurement. The second index, a similar type of treatment variance ratio, is applicable to the case of two treatment groups, of which only one group has been instructed. These indices are primarily measures of instructional effectiveness rather than a type of reliability.

Harris' (1972) index of efficiency can be conceived of as a ratio of true score variance to observed score variance for a single set of mastery test scores. Harris defines the index of efficiency

$$\mu_c^2 = \frac{SS_b}{SS_b + SS_w}$$

where SS_b represents the total score variance between the groups determined by the selected cut score and SS_w represents the within group variance on the k-item test scores. True score for a given subject is defined as the mean observed score of that subject's group (master or non-master). This index measures the extent to which the test sorts students into two categories. Although use of

the index does not require conventional distribution assumptions for item scores or total scores, in some situations the coefficient is a misleading indicator of decision-making accuracy. For example, when all the examinees are masters (non-masters), SS_b may be small implying a small μ_c^2 even though the decision accuracy may be substantial.

Livingston (1972), utilizing classical theory concepts, based a reliability coefficient on squared deviations of scores from the performance standard score rather than the mean of the sample scores. Livingston considers norm-referenced measurements as a special case of criterion-referenced measurement. Specifically, when the cut score or criterion score equals the mean score, the reliability coefficients are the same. Reliability estimates from a single form of the test, correlations between parallel forms, and relationships of reliability to test length are also derived, following the classical theory development. Livingston's coefficient has been criticized (see Harris, 1972) on the grounds that even though the Livingston coefficient is larger than the classical coefficient, this larger coefficient, does not yield a smaller standard error of measurement for a given student. Livingston argues that reliability is a characteristic of a group of scores rather than a single score.

Hambleton and Novick (1973) suggest that the squared-error loss function implicit in a variance ratio reliability estimate is an inappropriate one for the mastery learning situation, where the consistency of only two decisions (M or NM) is required, and test score variances may be small. A loss function may be thought of as a function which weights the errors resulting from inaccurate parameter estimations. The Livingston coefficient is sensitive to score

departures from the criterion score as well as to the differences between group (M, NM) means and the criterion. However, the losses which are of significance are due only to misclassification (based on scores above, equal to, or below the mastery cut score) and not to the magnitude of the test scores themselves. If the number one is assigned to all scores equal to or greater than the cut score, and the number zero is assigned to all observed scores below the cut score, then differences between true learner states and observed learner states can take on only the three numbers -1, 0, and +1. For example, suppose an examinee who is a known master earns a test score which is below the criterion score. The number 0 is assigned to this test score when, in fact, the number 1 is the correct classification number for a master student. The misclassification error equals (0 - 1) or -1. The function which assigns numbers to differences between true and observed learner states is called a threshold loss function. The precise test score value is not of primary interest, but only the learner state with respect to Mastery/non-Mastery that it identifies. Such a threshold loss function seems more appropriate for mastery testing.

The traditional concepts of reliability and validity need to be replaced not only by methods which are sensitive to the misclassifications themselves but also to a quantification of the associated losses. For these reasons, Hambleton and Novick (1973) proposed the adoption of a decision-theoretic approach for developing criterion-referenced measurement indices. Swaminathan, Hambleton, and Algina (1974), following such a decision-theoretic formulation, have defined reliability in terms of the consistency of decisions about mastery

states across repeated test administrations. The coefficient k, introduced by Cohen (1960) and further explored by Huynh (1976) in the context of educational testing, expresses the reliability of a criterion-referenced assessment and measures the agreement over chance alone between the decisions made on two test administrations. Specifically, when p_{ij} represents the proportion of examinees placed in the i^{th} mastery state on the first administration and in j^{th} mastery state on the second administration, we have

$$k = (p_o - p_c) / (1 - p_c),$$

where p_o, the observed proportion of agreements is given by

$$p_o = \sum_{l=1}^{k} p_{ii},$$

p_c, the expected proportion of agreements is given by

$$p_c = \sum_{i=1}^{k} p_{i.} p_{.i}$$

and $p_{i.}$ and $p_{.i}$ represent the proportion of examinees assigned to the mastery state i on the first and second test administrations, respectively. This reliability assessment applies only to a collection of test items which map one specific objective. For the multiobjective criterion-referenced test--as used in practice--there would be multiple reliability coefficients

Huynh (1976) suggested a model for evaluating the kappa reliability index on the basis of a single test administration. Huynh showed

that kappa increases as a function of test length and test score variability. Furthermore, kappa varies with cut score and has lower values at the high and low cut score ranges.

Subkoviak (1976) proposed an alternative, single administration coefficient of agreement which estimates the extent to which students would be assigned to the same mastery state as a result of two test administrations. This model, dealing with a fixed length test, assumes that (1) scores on parallel tests X_1 and X_2 are independently distributed for a fixed person, and (2) the distributions of X_1 and X_2 for a fixed person are identically binomial in form. Implicit in the second assumption is the condition that the probability of a correct response remains constant across items.

Marshall's (1973) index of separation is a measure of the extent to which the test is dependable in its classification of examinees as masters or non-masters. It is based on the assumption that the expected value of a master's score equals the number of items in the test and, for a non-master, the expected value is zero. The index has the additional desirable property of being equal to one when the examinee group consists of all masters, all non-masters, or consists of half masters and half non-masters. For the first two cases above, the classical reliability coefficient is zero, since the test variance is zero.

Marshall and Haertel (1975) propose an index equal to the mean of all possible split-half coefficients of agreement, derived from a single test administration. Coefficient β is the CRT counterpart to Cronbach's coefficient α which is a classical internal consistency index (Cronbach, 1951). Coefficient β is additive in the sense that

it is the mean of component parts made up of each person's score. As test scores depart from the criterion score, the value of β increases. However, β is not dependent on score variance in such a way as to produce a low coefficient when all examinees are masters (non-masters). If total test score variance is high, β is high, but if variance is low, there is no restriction on the range of β. Coefficient β is sensitive to criterion level, total test score, test length, and the mode of the distribution. For a given test and criterion level, β increases as the number of items increases. Coefficient β can be extended to encompass a three classification system as well.

Haladyna (1974) argues that classical theory for estimating reliability through the use of the internal consistency formula may be applied successfully to mastery-based tests. Haladyna further argues that, from a logical point of view, the estimation of true scores is important for both criterion-referenced and norm-referenced tests. In the mastery instructional situation, for examinees whose scores fall close to or at the cut score, two types of errors may occur: (1) true masters may be categorized as non-masters or (2) true non-masters may be assigned to a mastery status. From an empirical point of view, Haladyna presents evidence which supports the practice of applying classical procedures to unrestricted samples (mastery and non-mastery examinees) to establish full-scale reliabilities of criterion-referenced tests. Standard errors of measurement, computed using the reliability coefficient, are useful for the setting of confidence intervals, which in turn permit more accurate decision-making.

Millman (1974) outlines three different data analysis approaches for estimating reliability: (1) the consistency of scores on each of two sets of items drawn from the same domain (randomly parallel tests), (2) the consistency of decisions made from parallel tests by computing the agreement of decisions made on the basis of the test scores, and (3) the consistency of responses to matched items on parallel tests.

Harris' (1975) most recent work conceptualizes a pool of items which has been developed in association with an objective-based instructional program. Student response data to appropriate samples of these items are to be used to study both the instructional program and the test development and interpretation process. Harris outlines three types of studies which should provide insight. First to establish stability, coefficients for items and test scores would be useful. Harris states that "we would also like an estimate of the common difficulty level of the item and we would like to be able to aggregate such estimates to secure a meaningful index to describe the pool of items on the basis of a study of a sample of these items [p. 3]." A second concept is concerned with the equivalence of item difficulties for populations of students with similar prior instructional experience. A third notion deals with a measure of sensitivity to instruction. Instruction can be viewed as experimental manipulation of item difficulty. Harris suggests various sampling procedures, experimental designs, and choices of statistics for use in developing suitable generalizations for a pool of items. This work is currently in progress.

The latter two groups of reliability estimation procedures are not dealing exclusively with a fixed test. Both Millman's (1974)

notion of a randomly parallel test generated from a domain and Harris' (1975) desire to use aggregates of student response scores to tests derived from objective-based item pools lead quite naturally into the concepts of domain or item pool decision reliability, which is the principal concern of this research.

Chapter 3

Theoretical Considerations

3.1 Decision Reliability for Equal Item Parameters

Consider an item pool which contains p items each of which measures a learning objective(s) O. An n-item test is formed by selecting n items, $n \leq p$, randomly from the pool. A student participating in an individualized learning program is presented with such an n-item test. The student's test score equals the number of correct responses c. The student is classified as a Master (M) if the test score is greater than or equal to some predetermined cut score i; otherwise the student is classified as a non-Master (NM). Students who are classified as Masters proceed to the next instructional unit; non-Masters review the failed unit or study parallel materials.

In evaluating the item pool it is important to have an index which assesses the degree to which classification decisions for any given student based on scores on tests constructed from this pool are in agreement. Agreement on two tests implies that identical learner classifications are made for a given subject on the basis of two test scores: the subject is a Master on both tests or the subject is a non-Master. Such an index clearly depends on certain item characteristics, the test length n, the

cut score i, and the characteristics of the learner or the learning group to which the student belongs.

The decision reliability $R_D(n,i,p,N)$ of an item pool is defined as the probability that a randomly selected student who is administered two different randomly constructed n item tests is classified as either a Master on both tests or as a non-Master on both tests. For a given item pool there may exist many R_D's: $R_D(n_f,i,p,N)$ sets for fixed test lengths n_f but varying cut scores, i.e., i_1, i_2, ... , n_f, as well as $R_D(N,i_f,p,N)$ sets for fixed cut scores i_f but varying test lengths, i.e., i_f, i_{f+1}, ... , p. In other words, we can associate an R_D matrix with each item pool. The rows of this matrix represent the possible cut scores and the columns represent possible test lengths. Each cell entry represents R_D for the given pair of cut score and test length values. Theory will be developed concerning the construction of this matrix, its properties, and its usefulness for psychometricians, test constructors, and test users in characterizing item pools.

One can derive an expression for decision reliability in the following manner. Suppose an individual is selected at random from the population N of interest. The learner state, either M or NM, of the examinee is assumed known. First, the examinee is presented with an n-item test T_1 with items U_1, U_2, ..., U_n randomly selected without replacement from the item pool p. Then, the examinee is presented with a second n-item test T_2 with items

V_1, V_2, ..., V_n randomly selected as in T_1. Let X and Y represent the number of items answered correctly by the examinee on T_1 and T_2, respectively. Given the learner state, M or NM, of the examinee, the 2n item scores for all the examinees tested following the above procedure are assumed to be independently and identically distributed variables. Note that the 2n item scores on the U's and the V's are assumed to be independent only within the populations of Masters and non-Masters, but not necessarily across the populations. This independence assumption is similar to the local independence assumption of classical test theory (Lord & Novick, 1968, pp. 360-362). For binary items which are scored either correct or incorrect, the assumption of local independence for Masters can be stated as

$$\text{Prob}(U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n, V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n | M)$$

$$= [\prod_{j=1}^{n} \text{Prob}(U_j = u_j | M)][\prod_{k=1}^{n} \text{Prob}(V_k = v_k | M)] \qquad (3.1.1a)$$

and for non-Masters

$$\text{Prob}(U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n, V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n | NM)$$

$$= [\prod_{j=1}^{n} \text{Prob}(U_j = u_j | NM)][\prod_{k=1}^{n} \text{Prob}(V_k = v_k | NM)] \qquad (3.1.1b)$$

where the u_j's and v_k's are equal to zero for an incorrect response or one for a correct response. Furthermore, for a given cut score i, the independence of the distribution of item scores within the two examinee groups implies the independence of the two total sums X and Y, i.e.,

$$\text{Prob}(X \geq i,\ Y \geq i|M) = \text{Prob}(X \geq i|M)\ \text{Prob}(Y \geq i|M) \qquad (3.1.2a)$$

and

$$\text{Prob}(X < i,\ Y < i|NM) = \text{Prob}(X < i|NM)\ \text{Prob}(Y < i|NM). \qquad (3.1.2b)$$

First, consider the simple case where the probability of a Master answering any item correctly is the same quantity "a" for all items and the probability of a non-Master answering any item incorrectly is the quantity "b." Let x_i equal the response to item i by a randomly selected individual from among the population of Masters. Then the set $\{x_i\}$ are identically distributed if for all a_i, $\text{Prob}(x_i = 1)$ has a constant value a for every item, i.e., $a_i = a$. By similar argument, the set $\{y_i\}$ are identically distributed for non-Masters if for all i, $b_i = b$.[1] For a test of length n, the probabilities of a Master scoring i or greater on tests T_1 and T_2 are given by

$$\text{Prob}(X \geq i|M) = \sum_{j=i}^{n} \binom{n}{j} a^j (1 - a)^{(n-j)} \equiv A \qquad (3.1.3a)$$

---

[1]This assumption of identical response distributions for all items is not necessary for the theoretical development of R_D. The case of unequal item parameters is discussed in Section 3.4.

and

$$\text{Prob}(Y \geq i|M) = \sum_{j=i}^{n} \binom{n}{j} a^j (1 - a)^{(n-j)} \equiv A. \qquad (3.1.3b)$$

For non-Master students, we have

$$\text{Prob}(X < i|NM) = \sum_{j=0}^{i-1} \binom{n}{j} (1 - b)^j b^{(n-j)} \equiv B \qquad (3.1.4a)$$

and

$$\text{Prob}(Y < i|NM) = \sum_{j=0}^{i-1} \binom{n}{j} (1 - b)^j b^{(n-j)} \equiv B. \qquad (3.1.4b)$$

Note that Equations 3.1.3a, 3.1.3b, 3.1.4a, and 3.1.4b result from the fact that X and Y are identically distributed in this simple case where a and b are assumed to be constant for all items. Since X and Y are also independent within a given group, we have for Masters

$$\begin{aligned} \text{Prob}(X \geq i,\ Y \geq i|M) &= \text{Prob}(X \geq i|M)\ \text{Prob}(Y \geq i|M) \\ &= \text{Prob}(X \geq i|M)^2 \\ &= A^2. \end{aligned} \qquad (3.1.5)$$

Similarly, for non-Masters

$$\begin{aligned} \text{Prob}(X < i,\ Y < i|NM) &= \text{Prob}(X < i|NM)^2 \\ &= B^2. \end{aligned} \qquad (3.1.6)$$

And further since

$$\text{Prob}\ (X \geq i|M) + \text{Prob}(X < i|M) = 1 \qquad (3.1.7a)$$

and

$$\text{Prob}(X < i|\overline{M}) + \text{Prob}(X \geq i|\overline{M}) = 1, \qquad (3.1.7b)$$

we have from Equations 3.1.5, 3.1.6, 3.1.7a, and 3.1.7b

$$\text{Prob}(X < i,\ Y < i|M) = \text{Prob}(X < i|M)^2 = (1 - A)^2 \qquad (3.1.8a)$$

and

$$\text{Prob}(X \geq i,\ Y \geq i|\overline{M}) = \text{Prob}\ (X \geq i|\overline{M})^2 = (1 - B)^2. \qquad (3.1.8b)$$

Decision reliability (R_D) for Masters is given by

$$_{M}R_{D}(n,i,p,N) = \text{Prob}(X \geq i,\ Y \geq i|M) + \text{Prob}(X < i,\ Y < i|M) \qquad (3.1.9a)$$

$$= A^2 + (1 - A)^2 \qquad (3.1.9b)$$

$$= 1 - 2A + 2A^2 \qquad (3.1.9c)$$

where A^2 represents the probability of assigning a known Master to the Mastery state on the basis of both tests T_1 and T_2 and $(1 - A)^2$ represents that probability of assigning a known Master to the non-Mastery state. Clearly, the latter situation represents errors in testing but the notion of decision reliability presented in this paper is concerned with <u>consistency</u> of decision-making even though the consistently made decision might be contrary to the true state of Mastery or non-Mastery. The validity of the

decision-making will be dealt with in a later section.

Similarly, decision reliability for non-Masters is given by

$$_{NM}R_D(n_i,i,p,N) = \text{Prob}(X \geq i,\ Y \geq i|NM) + \text{Prob}(X < i,\ Y < i|NM) \tag{3.1.10a}$$

$$= (1 - B)^2 + B^2 \tag{3.1.10b}$$

$$= 1 - 2B + 2B^2 \tag{3.1.10c}$$

where $(1 - B)^2$ represents the probability of assigning a known non-Master to the Mastery state on the basis of both test scores, X and Y, and B^2 represents the probability of assigning a known non-Master to the non-Mastery state. In this case, the former probability, $(1 - B)^2$, represents testing error.

It is worth a brief digression to consider the reliability functions given in Equations 3.1.9c and 3.1.10c. Each function is of the form

$$f(x) = 1 - 2x + 2x^2, \quad 0 \leq x \leq 1. \tag{3.1.11}$$

Differentiating with respect to x gives

$$f'(x) = -2 + 4x \tag{3.1.12}$$

and setting $f'(x) = 0$ and solving for x gives $x = .5$.

Thus, we have a parabolic function which takes a minimum at a value of the argument equal to .5. The implication with respect to the decision-reliabilities is simply that ${}_{M}R_{D}$ and ${}_{NM}R_{D}$ are restricted to a range of 0.5 to 1.00. This is clearly in contrast to the classical notions of reliability which ranges from 0.00 to 1.00. However, in this model the concern is with consistency of two dichotomous decisions, whereas classical reliability is concerned with the consistency of rankings in the sense of measuring the observed score variance. By chance alone, under the hypothesized constraints of the model described above (i.e., with respect to equal $\underline{a}$ values, equal $\underline{b}$ values, and the local or within group item independence) the probability of decision agreement is .5. Therefore, the lower limit of .5, derived above, seems reasonable.

An expression for the unconditional decision-reliability is found by combining ${}_{M}R_{D}$ and ${}_{NM}R_{D}$ and is given by

$$R_D = {}_{M}R_{D}\ \text{Prob}(M) + {}_{NM}R_{D}\ \text{Prob}(NM) \tag{3.1.13}$$

or

$$R_D = {}_{M}R_{D}\ \text{Prob}(M) + {}_{NM}R_{D}\ [1 - \text{Prob}(M)] \tag{3.1.14}$$

since

$$\text{Prob}(M) + \text{Prob}(NM) = 1. \tag{3.1.15}$$

Equations 3.1.13 and 3.1.14 represent the definition of the decision reliability of a criterion referenced item pool for a given length test and a given cut score.

3.2 Classification Validity for Equal Item Parameters

A second useful property of the item pool is a validity index which is a measure of the degree to which Masters score at or above the cut score i on a given test and non-Masters score below the cut score. Let us define item pool classification validity for Masters as

$$ {}_{M}V_{C}^{*} = \text{Prob}(X \geq i|M) \equiv A \qquad (3.2.1) $$

And, similarly, for non-Masters

$$ {}_{NM}V_{C}^{*} = \text{Prob}(X < i|NM) \equiv B. \qquad (3.2.2) $$

The range of values for ${}_{M}V_{C}^{*}$ and ${}_{NM}V_{C}^{*}$ is from 0.00 to +1.00. The unconditional classification validity can be expressed as

$$ V_{C}^{*} = {}_{M}V_{C}^{*}\,\text{Prob}(M) + {}_{NM}V_{C}^{*}\,\text{Prob}(NM) \qquad (3.2.3) $$

where $0 \leq V_{C}^{*} \leq 1$. Classification decision validity V_{C}^{*} is also an index of the degree to which correct learner assignments (M or NM) are made on the basis of a given test administration.

The following considerations provide the motivation for creating a more useful and meaningful validity index which is the result of a

simple linear transformation applied to Equations 3.2.2. and 3.2.3. With respect to Masters, if one flipped a fair coin and made the assignment decision of Master corresponding to a head and non-Master for a tail, one would make the correct assignment decision half of the time, over the long run. The identical argument applies to the non-Master group as well. Thus, a validity index, which reflects the percentage gain over chance (random assignment) due to the item pool testing procedure, is clearly indicated. These transformed validity coefficients are given as

$$ {}_{M}V_{C} = \frac{\text{Prob}(X \geq i|M) - .50}{.50} = \frac{A - .50}{.50} $$

$$ = 2A - 1, \qquad (3.2.4) $$

where $-1 \leq {}_{M}V_{C} \leq +1$ and

$$ {}_{NM}V_{C} = \frac{\text{Prob}(X < i|NM) - .50}{.50} = \frac{B - .50}{.50} $$

$$ = 2B - 1, \qquad (3.2.5) $$

where $-1 \leq {}_{NM}V_{C} \leq +1$.

By combining ${}_{M}V_{C}$ (Equation 3.2.4) and ${}_{NM}V_{C}$ (Equation 3.2.5), we

have an expression for the unconditional validity, which is given by

$$V_C = {}_MV_C \text{ Prob}(M) + {}_{NM}V_C \text{ Prob}(NM) \tag{3.2.6}$$

By use of Equations 3.2.4 and 3.2.5, we have

$$V_C = (2A - 1) \text{ Prob}(M) + (2B - 1) \text{ Prob}(NM) \tag{3.2.7}$$

or

$$V_C = (2A - 1) \text{ Prob}(M) + (2B - 1) [1 - \text{Prob}(M)]. \tag{3.2.8}$$

By substituting ${}_MV_C^*$ for A and ${}_{NM}V_C^*$ for B in Equation 3.2.8, we have

$$V_C = (2{}_MV_C^* - 1) \text{ Prob}(M) + (2{}_{NM}V_C^* - 1) \text{ Prob}(NM)$$

$$= 2V_C^* - 1. \tag{3.2.9}$$

Note that $-1 \leq V_C \leq +1$, while $0 \leq V_C^* \leq +1$. The complete derivation is presented in Appendix 1.

A high negative classification validity is not a good characteristic of an item pool generated test regardless of its absolute value and should not be thought of in the classical sense as high negative correlation. Here a high negative classification validity implies either that the probability for Masters' achieving test scores equal to or greater than the Mastery cut score is less than 0.5 or that the probability for non-Masters' achieving test scores less than the

Mastery cut score is less than 0.5 or a combination of these events. In such cases, the items in the test pool need to be examined and those with small a and b values must be replaced.

3.3 Relationship between R_D and V_C

In classical test theory, the reliability of a test $r_{xx'}$, where x and x' are parallel measurements, is defined as the squared correlation r_{xt}^2 between observed score x and true score t. The reliability or reliability coefficient of a test is also a measure of the ratio of true-score variance to observed score variance, σ_t^2/σ_x^2, and ranges from 0 to 1. On the other hand, the index of reliability refers to the quantity $r_{xt} = \sigma_t/\sigma_x$, the square root of the reliability coefficient. The validity coefficient of a measurement x with respect to a second measurement y is defined as the absolute value of the correlation coefficient $r_{xy} = |\sigma_{xy}/\sigma_x\sigma_y|$. One significant theoretical result of the application of reliability theory is the expression

$$r_{xy} \leq \sqrt{r_{xx'}} = r_{xt} . \qquad (3.3.1)$$

The validity of a test is bounded above by the index of reliability. It is worth noting that the classical reliability coefficient is defined as a squared correlation, whereas the validity coefficient and the index of reliability are simple correlations. Furthermore, although the index of reliability r_{xt} sets an upper bound to the validity of a test, the test's validity with respect to another measure may indeed exceed the coefficient of reliability $r_{xx'}$ of the test!

The arguments presented in Appendix 2 demonstrate there is also a relationship between R_D and V_C: a function of decision reliability sets both an upper and a lower bound to classification validity. The range of V_C values for some R_D is given by

$$-\sqrt{2R_D - 1} \leq V_C \leq \sqrt{2R_D - 1} \,. \qquad (3.3.2)$$

A consideration of some numerical examples will illustrate this relationship. Given R_D, there are exactly four cases which are mutually exhaustive:

Case 1. All students are Masters. Prob(M) = 1.

Case 2. All students are non-Masters. Prob(M) = 0.

Case 3. There are an equal number of Masters and non-Masters. Prob(M) = Prob(NM) = 1/2.

Case 4. There are both Masters and non-Masters and a different number of each.

In Table 1, sets of possible pairs of ${}_{M}R_D$ and ${}_{NM}R_D$ values consistent with given R_D's and the corresponding values for ${}_{max}V_C$ and ${}_{min}V_C$ are presented for Cases 1, 2, and 3. The outside range values for Case 4 situations are also represented by the Case 1 and Case 2 entries. The formulas for calculating the V_C entries in the two columns appear below the table. For Cases 1 and 2, the V_C entries depend only on R_D, whereas for Case 3 the possible R and R pairs consistent with a given R_D produce different values of V_C (i.e., the V_C entries for the given R_D produce different values of V_C (i.e., the V_C entries for the three rows with R_D = .60 are the same for Cases 1 and 2 but differ for Case 3). The single most significant observation to be made from this table is that in all cases $|V_C| \leq \sqrt{2R_D - 1}$.

TABLE 1

V_C Range for Given R_D

R_D	Possible $_MR_D$ and $_{NM}R_D$ Pairs		V_C	
	Largest	Smallest	Cases* 1 and 2	Case 3
.50	.50	.50	0.00	0.00
.51	.52	.50	± 0.14	± 0.10
.55	.60	.50	± 0.32	± 0.23
.60	.60	.60	± 0.45	± 0.45
"	.65	.55	"	± 0.44
"	.70	.50	"	± 0.32
.65	.80	.50	± 0.55	± 0.39
.70	.90	.50	± 0.63	± 0.45
.75	1.00	.50	± 0.71	± 0.50
.80	1.00	.60	± 0.77	± 0.73
.90	1.00	.80	± 0.89	± 0.89
1.00	1.00	1.00	± 1.00	± 1.00

Cases 1 and 2: Prob(M) = 1 and Prob(NM) = 0 or
Prob(M) = 0 and Prob(NM) = 1

$$V_C = \pm \sqrt{2R_D - 1}$$

Case 3: Prob(M) = Prob(NM) = 1/2

$$V_C = \pm \left(\sqrt{2_MR_D - 1} + \sqrt{2_{NM}R_D - 1} \right)$$

*Case 4: Prob(M) ≠ Prob(NM) ≠ 1 or 0
These entries also represent the maximum and minimum values of V_C.

3.4 Reliability and Validity for Unequal Item Parameters

Next, consider the cases in which the probability of a Master answering item j correctly is a_j and the probability of a non-Master answering item j incorrectly is b_j, i.e., in general, for items j and k, $a_j \neq a_k$ and $b_j \neq b_k$. The item responses are not identically distributed. In this case there is one entire distribution of a_j's and another distribution of the b_j's. The marginal distributions of the a_j's and the b_j's are assumed to be beta in form, [i.e., $f(a_j)$, $g(b_j)$] with parameters (s,t) and (s',t'), respectively[1].
These situations are clearly more complex than the case considered above in which it was assumed that $a_1 = a_2 = \ldots = a_p = \underline{a}$ and $b_1 = b_2 = \ldots = b_p = \underline{b}$.

Now the probability that a Master will make i or more correct responses to an n-item test is given by

$$\text{Prob}(X \geq i|M) = \int \int \ldots \int \int f(a_1,a_2,\ldots,a_n)\ \text{Prob}(X \geq i|M,a_1,a_2,\ldots,a_n)\ da_1,da_2,\ldots,da_n. \qquad (3.4.1)$$

Since each item is sampled independently, and the a_j values are independent and identically distributed {beta (s,t)}, we may write

$$f(a_1, a_2, \ldots, a_n) = \prod_{j=1}^{n} f(a_j). \qquad (3.4.2)$$

---

[1]The beta distribution is selected since it can take a wide variety of shapes.

Substituting Equation 3.4.2 into Equation 3.4.1 gives

$$\text{Prob}(X \geq i|M) = \int \prod_{j=1}^{n} f(a_j)\ \text{Prob}(X \geq i|M, a_1, a_2, \ldots, a_n)\ da_j. \quad (3.4.3)$$

By similar reasoning, the probability that a non-Master will make less than i correct responses is given by

$$\text{Prob}(X < i|NM) = \int \prod_{j=1}^{n} f(b_j)\ \text{Prob}(X < i|NM, b_1, b_2, \ldots, b_n)\ db_j. \quad (3.4.4)$$

Before we extend the theory, let us consider a particular case for which the number of items n on a test equals five and the cut-score equals four. Equation 3.4.3 may be rewritten as

$$\begin{aligned}\text{Prob}(X \geq 4|M) = \prod_{j=1}^{n} \int f(a_j)\ \{&a_1a_2a_3a_4(1-a_5)+a_1a_2a_3(1-a_4)a_5+\\ &a_1a_2(1-a_3)a_4a_5+a_1(1-a_2)a_3a_4a_5+\\ &(1-a_1)a_2a_3a_4a_5+a_1a_2a_3a_4a_5\}\ da_j. \quad (3.4.5)\end{aligned}$$

Let us set i = 1 and perform the first integration with respect to a_1. We have

$$\text{Prob}(X \geq 4|M) = \prod_{j=2}^{5} \int f(a_j) \{E(a_1)a_2a_3a_4(1-a_5)+E(a_1)a_2a_3(1-a_4)a_5+$$

$$E(a_1)a_2(1-a_3)a_4a_5+E(a_1)(1-a_2)a_3a_4a_5+$$

$$[1-E(a_1)]a_2a_3a_4a_5+E(a_1)a_2a_3a_4a_5\}\, da_j. \qquad (3.4.6)$$

Since a_i ~ beta (s,t)

$$E(a_1) = E(a_2) = \ldots = E(a_p) = \frac{s}{s+t}. \qquad (3.4.7)$$

Thus, by substituting Equation 3.4.7 into Equation 3.4.6, we have

$$\text{Prob}(X \geq 4|M) = \int \prod_{j=2}^{5} f(a_j) \left[\frac{s}{s+t} a_2a_3a_4(1-a_5)+ \frac{s}{s+t} a_2a_3(1-a_4)a_5+\right.$$

$$\frac{s}{s+t}a_2(1-a_3)a_4a_5 + \frac{s}{s+t}(1-a_2)a_3a_4a_5+$$

$$\left.(1-\frac{s}{s+t})a_2a_3a_4a_5+ \frac{s}{s+t} a_2a_3a_4a_5\right] da_j. \qquad (3.4.8)$$

It is easily seen that following successive integrations with respect to a_2, a_3, a_4, and a_5, the a's will be replaced by E(a)'s, which in turn will each be replaced by $\frac{s}{s+t}$. The result of the complete integration gives

$$\text{Prob}(X \geq 4|M) = \left(\frac{s}{s+t}\right)^4 \left(1-\frac{s}{s+t}\right)+\left(\frac{s}{s+t}\right)^4\left(1-\frac{s}{s+t}\right)+\left(\frac{s}{s+t}\right)^4\left(1-\frac{s}{s+t}\right)$$

$$+\left(\frac{s}{s+t}\right)^4\left(1-\frac{s}{s+t}\right)+\left(\frac{s}{s+t}\right)^4\left(1-\frac{s}{s+t}\right)+\left(\frac{s}{s+t}\right)^5$$

$$= 5\left(\frac{s}{s+t}\right)^4\left(1-\frac{s}{s+t}\right)+\left(\frac{s}{s+t}\right)^5. \qquad (3.4.9)$$

This result can be generalized to an n-item test with a cut score i such that

$$\text{Prob}(X \geq i|M) = \binom{n}{i}\left(\frac{s}{s+t}\right)^i\left(1-\frac{s}{s+t}\right)^{(n-i)} + \binom{n}{i+1}\left(\frac{s}{s+t}\right)^{(i+1)}\left(1-\frac{s}{s+t}\right)^{[n-(i+1)]}$$

$$+\ldots+n\left(\frac{s}{s+t}\right)^{(n-1)}\left(1-\frac{s}{s+t}\right)+\left(\frac{s}{s+t}\right)^n,$$

or we may write

$$A \equiv \text{Prob}(X \geq i|M) = \sum_{j=i}^{n} \binom{n}{j}\left(\frac{s}{s+t}\right)^j\left(1-\frac{s}{s+t}\right)^{n-j}. \qquad (3.4.10)$$

Similarly, it can be shown that

$$B \equiv \text{Prob}(X < i|NM) = \sum_{j=0}^{i-1} \binom{n}{j}\left(\frac{s'}{s'+t'}\right)^{n-j}\left(1-\frac{s'}{s'+t'}\right)^j. \qquad (3.4.11)$$

It is interesting to note that the results expressed in Equations 3.6.10 and 3.6.11 would also be obtained if all of the a_i (b_i) item parameters in the item pool were equal to $\frac{s}{s+t}$ ($\frac{s'}{s'+t'}$), instead of being different values from the distribution. In other words, Prob($X \geq i|M$) is given by a binomial distribution with parameter $\frac{s}{s+t}$. It also follows that Prob($X < i|NM$) is given by a cumulative binomial distribution with parameter $\frac{s'}{s'+t'}$. Finally, we see that expressions for R_D and V_C are of exactly the same form for heterogeneous item pools (i.e., $\frac{s}{s+t}$ is substituted for $\underline{a}$ and $\frac{s'}{s'+t'}$ is substituted for $\underline{b}$). Note that the number of items in the item pool does not enter into these equations. The only requirement is that the number of items in the item pool must be large compared to the length of the test.

3.5 A Bayesian Approach to the Estimation of the Unknown Parameter R_D

In a Bayesian analysis one can express prior beliefs concerning an unknown parameter before one collects data. Prior beliefs about the decision reliability R_D for a given item pool, test length, cut score, and learner group can be represented by a type of beta distribution. The density of the general beta distribution with parameters p and q is given by

$$f(x) = \frac{x^{(p-1)}\ (1 - x)^{(q-1)}}{B(p,q)} \quad \text{for } 0 \leq x \leq 1 \text{ and } p,\ q > 0, \quad (3.5.1)$$

where

$$B(p,q) = \frac{(p - 1)!\ (q - 1)!}{(p + q - 1)!} \qquad (3.5.2)$$

When p and q are not integers the factorials (p-q)! and (q-1)! are defined by the relations

$$(p - 1)! = \Gamma(p)$$

and

$$(q - 1)! = \Gamma(q),$$

where $\Gamma(p)$ and $\Gamma(q)$ are gamma functions.

Thus,

$$B(p,q) = \frac{\Gamma(p)\ \Gamma(q)}{\Gamma(p + q)} . \qquad (3.5.3)$$

However, the beta distribution described in Equation 3.5.1 is not appropriate for expressing prior beliefs about R_D, since the range of possible R_D values has already been shown to be $.5 \leq R_D \leq 1$, whereas for the beta distribution the range of the argument is $0 \leq x \leq 1$.

To achieve the required truncated range for the argument, the expression for f(x) given in Equation 3.5.1. must be rescaled in order to maintain a probability distribution, for which the area under the density function equals one.

By definition, the distribution function of a continuous random variable x is defined to be

$$F(x) = \text{Prob}(X \leq x) = \int_{-\infty}^{x} f(t)\ dt. \qquad (3.5.4)$$

From Equations 3.5.4 and 3.5.1, we have

$$F(.50) = \int_{0}^{.50} f(t)\ dt. \qquad (3.5.5)$$

Let us define W'(x) such that

$$W'(x) = \int_{x}^{1} f(t)\ dt = 1 - F(x) \qquad (3.5.6)$$

We may now express a prior probability distribution for R_D in the form of the following truncated beta distribution. By combining the results of Equations 3.5.1 and 3.5.6, we have

$$p'(R_D) = \frac{f(R_D)}{W'(R_D)}, \quad .5 \leq R_D \leq 1,$$

$$= 0\ , \text{ elsewhere,} \qquad (3.5.7)$$

and where

$$f(R_D) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} R_D^{(a-1)} (1 - R_D)^{(b-1)},$$

and $\underline{a}$ and $\underline{b}$ are the parameters of the beta distribution, and the rescaling factor

$$W'(.50) = \int_{.50}^{1} f(R_D)\, dR_D = 1 - F_{a,b}(.50).$$

The expectation or expected value of a continuous random variable X is defined by

$$E(X) = \int x\, f(x)\, dx. \qquad (3.5.8)$$

Applying Equation 3.5.8 to R_D and using Equation 3.5.7, we have the expectation of R_D prior to data collection

$$E'(R_D) = \frac{1}{W'(.50)} \int_{.50}^{1} R_D\, f(R_D)\, dR_D$$

$$= \frac{\int_{.50}^{1} R_D \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} R_D^{(a-1)} (1-R_D)^{(b-1)}\, dR_D}{\int_{.50}^{1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} R_D^{(a-1)} (1-R_D)^{(b-1)}\, dR_D}$$

$$= \frac{\int_{.50}^{1} \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} R_D^{\,a} (1-R_D)^{(b-1)}\, dR_D}{\int_{.50}^{1} \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} R_D^{(a-1)} (1-R_D)^{(b-1)} dR_D}$$

$$= \frac{\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \int_{.50}^{1} R_D^{\,a} (1 - R_D)^{(b-1)}\, dR_D}{\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \int_{.50}^{1} R_D^{(a-1)} (1-R_D)^{(b-1)}\, dR_D}. \qquad (3.5.9)$$

Since the coefficients in the numerator and the denominator cancel, we have

$$E'(R_D) = \frac{\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} [1 - F_{(a+1,b)}(.50)]}{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} [1 - F_{(a,b)}(.50)]}, \qquad (3.5.10)$$

where

$$F_{(a+1,b)}(.50) = \int_0^{.50} \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} R_D^{\,a} (1 - R_D)^{(b-1)} dR_D .$$

Since $\Gamma(a) = (a - 1)!$ and $\Gamma(a + 1) = a!$, then

$$\Gamma(a + 1) = a\Gamma(a). \tag{3.5.11}$$

Applying Equation 3.5.11 to 3.5.10, we have

$$E'(R_D) = \frac{\frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} [1 - F_{(a+1,b)}(.50)]}{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} [1 - F_{(a,b)}(.50)]}$$

$$= \frac{a}{a + b} \frac{[1 - F_{(a+1,b)}(.50)]}{[1 - F_{(a,b)}(.50)]} . \tag{3.5.12}$$

In the numerical analysis which is presented in Chapter 4 a non-informative prior distribution for R_D is assumed in the range $.50 \leq R_D \leq 1.00$. The resulting probability distribution takes the form of Equation 3.5.7 where $a = b = 1$.

In the next stage of a Bayesian analysis, data are collected and incorporated into the expression for a prior distribution on the

parameter of interest to produce a posterior distribution. For example, in this case the parameter of interest is R_D, which, in a given sample of size n subjects, is estimated by the number of agreements in decision-making x on the basis of two test scores divided by n, i.e., the proportion of agreements in the given sample. When the prior distribution is a member of the beta family, and the probability that x agreements occur for n subjects is a binomial distribution with parameter R_D, the posterior distribution is

$$p''(R_D \mid n,x) = \frac{\frac{(a+b+n)}{(a+x)\quad(b+n-x)} R_D^{(a+x-1)} (1-R_D)^{(n+b-x-1)}}{W''(.50)} \qquad (3.5.13)$$

where

$$W''(.50) = 1 - F_{(a+x,n+b-x)}(.50).$$

The expectation or expected value for the posterior distribution of R_D is calculated by applying the methods of Equations 3.5.8 - 3.5.12 and

$$E''(R_D) = \frac{\frac{a+b}{(a+b+n)} \left[1 - F_{(a+x+1,n+b-x)}(.50)\right]}{W''(.50)}. \qquad (3.5.14)$$

A numerical example illustrating these methods will be presented in Chapter 4.

3.6 Predictability and Robustness

Scientific models can be evaluated on both theoretical and practical bases. A model is theoretically sound if all of the deductions which are derived follow logically from the basic assumptions made prior to the theory development in combination with any accepted axioms and proven theorems of logic and mathematics. The assumptions made in developing the decision reliability and classification validity indices are the following:

(1) Test items are randomly selected from the item pool without replacement.

(2) The item scores for all subjects are independently and identically distributed variables within the populations of Masters and non-Masters, but not necessarily across populations.

(3a) The item parameters for items j and k are equal (i.e., $a_j = a_k$ and $b_j = b_k$) for all items in the item pool (see Section 3.2).

or

(3b) The item parameters for items j and k are unequal (i.e., $a_j \neq a_k$ and $b_j \neq b_k$) and the marginal distribution of the a_j's and b_k's are of the beta form with parameters (s,t) and (s',t'), respectively.

The theoretical assessment of the model is made without evaluating the reasonableness of these assumptions. A pure theorist merely asks: "Can the claimed deductions be derived logically from these assumptions?". If the answer is negative, then new theory must be deduced. If the answer is in the affirmative, then the model is

theoretically sound. However, such theoretical soundness does not guarantee usefulness in an applied setting.

Now, let us consider the item pool response model from a practical point of view, specifically in terms of its applicability to item pools developed for an introductory mathematics course. This one semester course, Mathematics 45.100, offered at the Mathematics Learning Center of Hunter College, is divided into five modules. Each module, in turn, is subdivided into ten or more objectives. Associated with each module is an item pool from which 50 different parallel tests are generated by computer following a stratified random sampling plan. The number of strata in a given pool is equal to the number of objectives in that pool. The number of items per stratum (objective) selected for a given test is consonant with the relative importance of that objective as determined by the curriculum expert.

In testing the practical usefulness of the model, one can reasonably ask two questions:

(1) Do the assumptions 1, 2, and 3a (or 1, 2, and 3b) adequately and correctly describe the real world?

(2) For the cases in which the assumptions do not fit behavior in the real world, are predictions based on the logical deductions nevertheless useful for the practitioner?

With respect to the first question, in the Hunter program, the items for a given test are not randomly selected but are selected following a stratified random sampling plan.[1] Also, only a fixed number

---

[1] Each item pool contains from four to sixteen items matched to a given objective and the number of objectives per pool varies from 12 to 15. The number of items per objective that appears on any test is one or two.

(i.e., 50 in course 45.100) of tests is prepared in this way and one of these tests is then randomly assigned to each student. The independence of item scores within populations is doubtful. Item content may be hierarchically related implying that the conditional probability of answering one item correctly given that a second was answered correctly may be different from the simple probability of answering the first correctly. The mathematical counterpart of this statement for Masters is:

$$\text{Prob(Item 1 is correct} \mid \text{Item 2 is correct)} \neq \text{Prob(Item 1 is correct)}.$$

Also, information contained in one item might contribute to a correct response to a second item. Finally, with respect to the item parameter distributions, it is clear that the $\underline{a}$ and $\underline{b}$ values are not equal for all items. Further, beta distributions for the a's and b's are not exact representations.

Let us turn to the second question. How "good" is the model despite violations of the three assumptions? One model confirmation strategy requires the theorist to derive predictions based on the model. Next, these predictions are tested out on data collected by the practitioner (or experimentalist) in an applied setting. The first set of predictions of interest in these studies concern the relationships between the item pool decision reliabilities R_D, test length n, and cut score i. The second prediction describes the relationship between the decision reliabilities R_D and the corresponding classification validities V_C.

Prediction 1 R_D as a function of n, i, and a.

Test length n and cut score i do not have linear relationships with R_D. The following discussion will support these conclusions by exploring first the algebraic relationships between ${}_MR_D$ and A, and then the algebraic relationships between A and a, i, and n, in turn. At the conclusion of these explorations, specific predictions for the R_D, i, a, and n relationships will be made.

From Figure 1, a graph of ${}_MR_D$ as a function of A where

$$ {}_MR_D = 1 - 2A + 2A^2 \qquad (3.1.9c) $$

and

$$ A = \text{Prob}(X \geq i \mid M) = \sum_{j=1}^{n} \binom{n}{j} a^j (1-a)^{(n-j)}, \qquad (3.1.3a) $$

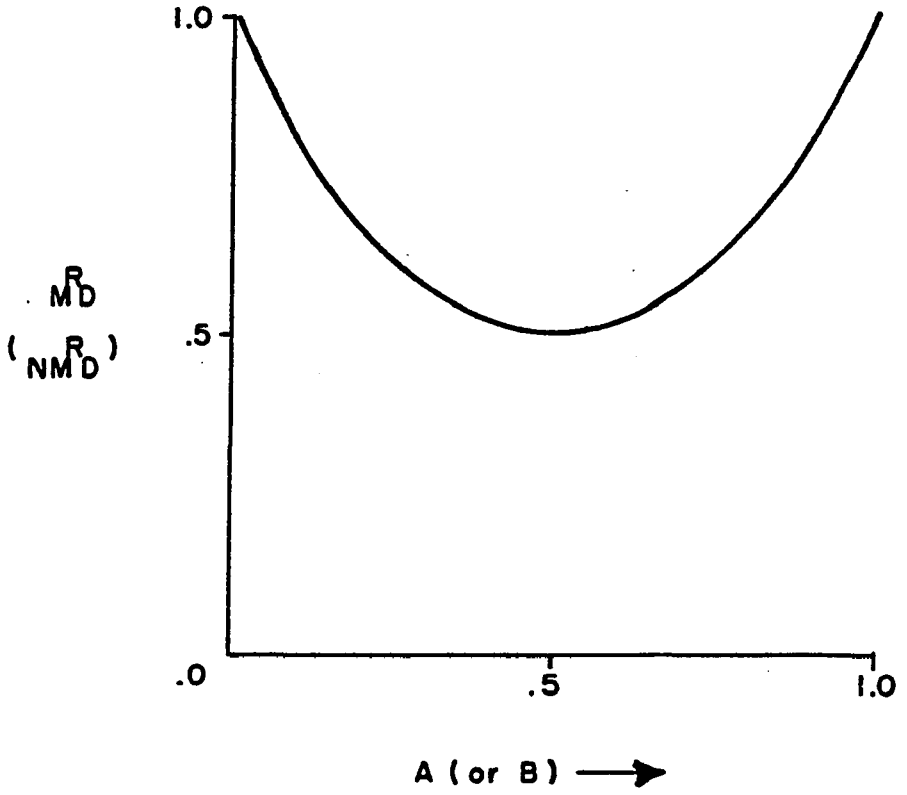
we see the ${}_MR_D$ approaches maximum values as A approaches one or zero and achieves a minimum value for A = 1/2. The relationship is a parabolic one. Similar arguments apply to ${}_{NM}R_D$ and B. Additionally, we recall that

$$ R_D = {}_MR_D \, \text{Prob}(M) + {}_{NM}R_D \, \text{Prob}(NM), \qquad (3.1.3) $$

and, further, that each term to the right of the '=' sign is positive or zero. Thus, it is extremely unlikely that the two terms on the right-hand side of Equation 3.1.3 would combine to form an expression which implies a linear relationship between R_D and A.

FIGURE I

M_D^R as a Function of A



$$M_D^R = 1 - 2A + 2A^2$$

$$A = \text{Prob}(X \geq i \mid M) = \sum_{j=i}^{n} \binom{n}{j} a^j (1-a)^{n-j}$$

$$B = \text{Prob}(X < i \mid NM) = \sum_{j=0}^{i-1} \binom{n}{j} (1-b)^j b^{n-j}$$

The relationships between n, a, and i with A are also not linear and can be studied with the aid of the curves in Figure 2 which depict the cumulative probability functional values (A in Equation 3.1.3a) for two 10-item tests (a = .7 and .8), and for five 20-item tests (a = .5, .6, .7, .8 and .9). The values from which these curves were plotted are tabled in several mathematics handbooks (e.g., Beyer, 1966). From these curves we can investigate three relationships of interest: (1) A and n, (2) A and a, and (3) A and i. Since A is a function of three parameters, n, i, and a, one method for exploring the relationship between any two requires that the remaining two parameters be kept fixed. For example, given a = .7 (Figure 2), every A value for a given cut score i is greater for the (n = 20, a = .7) curve than for the (n = 10, a = .7) curve. The (n = 20, a = .7) curve is "above" the (n = 10, a = .7) curve. Similarly, the (n = 20, a = .8) curve is "above" the (n = 10, a = .8) curve. Furthermore, the distances between these pairs of curves differ from cut score (i value) to cut score. Thus, we can conclude that the relationship between A and n is not linear and for a given a and i, A increases as n increases. Similar comparisons between the (n = 10, a = .7) and (n = 10, a = .8) curves and between the (n = 20, a = .6) and (n = 20, a = .9) curves indicate that for a given a and i, A is larger for larger a, and the relationship is also not linear. Finally, since not one graph in Figure 2 is a straight line, the relationship between A and i for a given n and a is not linear and A decreases as i increases.

These theoretical relationships can also be explored by a direct computational approach (Tables 2 and 3). The entries in Table 2 represent values of A (or B), the cumulative probability function for

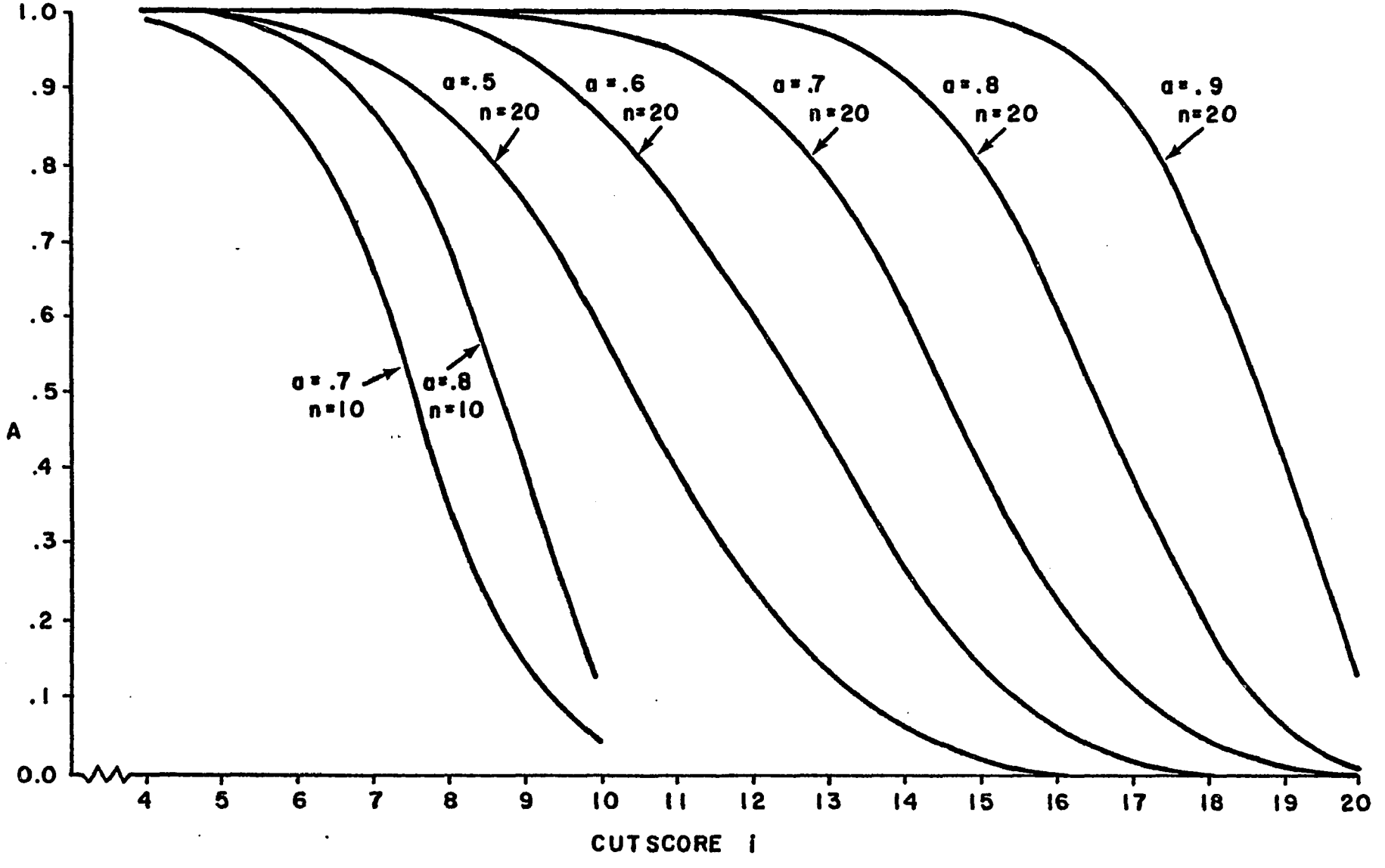


FIGURE 2-CUMULATIVE PROBABILITY FUNCTIONS

TABLE 2

Cumulative Probability Values*

$$A = \text{Prob}\ (X \geq i) = \sum_{j=1}^{n} \binom{n}{j} a^j\ (1-a)^{n-j}$$

i \ a	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
1	.401	.651	.803	.893	.944	.972	.987	.994	.997	.999
2	.086	.264	.456	.624	.756	.851	.914	.954	.977	.989
3	.012	.070	.180	.322	.474	.617	.738	.833	.900	.945
4	.001	.013	.050	.121	.224	.350	.486	.618	.734	.828
5	.000	.002	.010	.033	.078	.150	.249	.367	.496	.623
6	.000	.000	.001	.006	.020	.047	.095	.166	.262	.377
7	.000	.000	.000	.001	.004	.011	.026	.055	.102	.172
8	.000	.000	.000	.000	.000	.002	.005	.012	.027	.055
9	.000	.000	.000	.000	.000	.000	.001	.002	.005	.011
10	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001

i \ a	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
1	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	.989	.995	.998	.999	1.000	1.000	1.000	1.000	1.000	1.000
3	.945	.973	.988	.995	.998	1.000	1.000	1.000	1.000	1.000
4	.828	.898	.945	.974	.989	.996	.999	1.000	1.000	1.000
5	.623	.738	.834	.905	.953	.980	.994	.999	1.000	1.000
6	.377	.504	.633	.751	.850	.922	.967	.990	.998	1.000
7	.172	.266	.382	.514	.650	.776	.879	.950	.987	.999
8	.055	.100	.167	.262	.383	.526	.678	.820	.930	.988
9	.011	.023	.046	.086	.149	.244	.376	.544	.736	.914
10	.001	.003	.006	.013	.028	.056	.107	.197	.349	.599

*n = 10

TABLE 3

Theoretical Values for ${}_{M}R_{D}$ (${}_{NM}R_{D}$)

$${}_{M}R_{D} = A^2 + (1-A)^2 = 1-2A + 2A^2$$

or

$${}_{NM}R_{D} = B^2 + (1-B)^2 = 1-2B + 2B^2$$

i \ a	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
1	.519*	.546	.684	.808	.894	.945	.973	.988	.995	.998
2	.843	.611	.504	.531	.631	.746	.843	.912	.955	.979
3	.977	.869	.705	.563	.501	.527	.614	.721	.821	.897
4	.998	.975	.905	.787	.652	.545	.500	.528	.609	.715
5	1.000	.997	.980	.937	.856	.745	.627	.535	.500	.530
6	1.000	1.000	.997	.987	.961	.910	.828	.723	.614	.530
7	1.000	1.000	1.000	.998	.993	.979	.949	.896	.817	.715
8	1.000	1.000	1.000	1.000	.999	.997	.990	.976	.947	.897
9	1.000	1.000	1.000	1.000	1.000	1.000	.999	.997	.991	.979
10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.998

i \ a	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
1	.998	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	.979	.991	.997	.999	1.000	1.000	1.000	1.000	1.000	1.000
3	.897	.947	.976	.990	.997	.999	1.000	1.000	1.000	1.000
4	.715	.817	.896	.949	.979	.993	.998	1.000	1.000	1.000
5	.530	.614	.723	.828	.910	.961	.987	.997	1.000	1.000
6	.530	.500	.535	.626	.745	.856	.937	.980	.997	1.000
7	.715	.609	.528	.500	.545	.652	.787	.905	.975	.998
8	.897	.821	.721	.614	.527	.501	.563	.705	.869	.977
9	.979	.955	.912	.843	.746	.631	.531	.504	.611	.843
10	.998	.995	.988	.973	.945	.894	.808	.684	.546	.519

*minimum R_D values for each cut score are underlined

10-item tests[1] for pairs of i (cut score) and a (probability of answering an item correctly) values. Looking along any row (i.e., i = 7), one observes that A increases (from .000 to .999) as a increases (from .05 to .95). A column pattern (e.g., a = .75) shows A decreasing (from 1.000 to .056) as cut score i increases.

However, the parameter A is not of primary interest, but rather $_{M}R_{D}$ and $_{NM}R_{D}$ which have been defined previously:

$$_{M}R_{D} = 1 - 2A + 2A^{2} \qquad (3.1.9c)$$

$$_{NM}R_{D} = 1 - 2B + 2B^{2} \qquad (3.1.10c)$$

Table 3 presents values for $_{M}R_{D}$ ($_{NM}R_{D}$) for tests of length 10. Each entry represents $_{M}R_{D}$ ($_{NM}R_{D}$) for a given cut score i and a given probability of a correct (incorrect) response a (b). (The patterns observed in Table 3 may also be derived by combining the information presented in Table 2 and Figure 2.) These R_{D} patterns, which result directly from the theory itself, suggest the following specific predictions:

1.1 $_{M}R_{D}$ as a function of i

1.1.1 For a = .05, $_{M}R_{D}$ is an increasing function of cut score i. For example, the $_{M}R_{D}$ entries in the column headed a = .05 increase from .52 for a cut score of 1 to

---

[1] Ten-item tests have been selected, since the empirical studies which are presented in Chapter 4 are based on characteristics of responses to tests of such a length.

1.00 for a cut score of 10.

1.1.2 For $\underline{a}$ = .95, $_{M}R_{D}$ is a <u>decreasing function</u> of cut score i. For example, under the a = .95 column, the $_{M}R_{D}$ entries decrease from 1.00 to .52.

1.1.3 For $\underline{a}$ ranging from .15 to .85, the $_{M}R_{D}$ values suggest a <u>U-shaped relationship</u> with cut score i. For example, for a = .75, $_{M}R_{D}$ first decreases from 1.00 at cut score i = 1 to .50 for cut score i = 8 and then increases again to .89 for cut score 1 = 10.

1.2 <u>$_{M}R_{D}$ as a function of $\underline{a}$</u>

1.2.1 For i = 1, $_{M}R_{D}$ is an <u>increasing function</u> of $\underline{a}$.

1.2.2 For i = 10, $_{M}R_{D}$ is a <u>decreasing function</u> of $\underline{a}$.

1.2.3 For i ranging from 2 to 9 the <u>relationship</u> between $_{M}R_{D}$ and $\underline{a}$ is <u>U-shaped</u>. For example, for i = 7, $_{M}R_{D}$ decreases from 1.00 for a = .05 to .50 for a = .65 and then increases to .998 for a = .95.

The discussions above of Figures 2 and 3 lead to the following prediction:

1.3 <u>$_{M}R_{D}$ as a function of n</u>

Since A increases as n increases (Figure 3), and $_{M}R_{D}$ has a parabolic (U-shaped) relationship with A, then $_{M}R_{D}$ has a <u>U-shaped relationship</u> with n.

At this point, the reader should be somewhat comforted to learn that these seemingly complex dependencies will be unravelled, perhaps more

understandably, in the discussion of the empirical results which will be described in Chapter 4.

Prediction 2 Relationship between R_D and V_C

The theoretical development in Section 3.3 produced the important and interesting result that $|V_C| \leq R_D$, specifically, $-\sqrt{2R_D - 1} \leq V_C \leq +\sqrt{2R_D + 1}$. Thus, the model predicts that in an applied setting R_D sets an upper bound to V_C.

If Predictions 1 and 2 are not confirmed by particular data analyses, then the model has no utility for the practitioner in the setting in which it is applied. For such item pool testing situations, one could conclude that the violation of the model's assumptions would be of too large a magnitude: the model would not be a good approximation.

On the other hand, if these relationships are supported by the data analyses, then some evidence has been accumulated in support of the model's utility in an applied setting. The confirmation of a model's utility ultimately requires many such successes! Let us now turn to a careful examination of an interesting data set and test the model.

Chapter 4

Application of the Theory to Existing Item Pools

This chapter provides the practitioner with a detailed description for generating both the decision reliability matrix (R_D-post) and a validity index empirically as well as means for assessing the robustness of the theoretical model using these empirical data. A step by step procedure is presented for calculating reliabilities from test item data and validities from test scores. These procedures are consonent with the theoretical development which was presented in Chapter 2. Various logical deductions have been made from the theoretical relationships concerning decision reliability, test length, cut score, and classification validity (Section 3.6). A comparison of the empirical data with these theoretical predictions will lead to informed judgments concerning the usefulness, effectiveness, and efficiency of the model in characterizing test item pools from which individualized tests are randomly generated.

4.1 Procedure for Generating the Decision Reliability Matrix

Let us consider one procedure, based on the theoretical model which has been presented in Chapter 3, for producing the decision reliability estimation matrix (R_D-post) of the module item pool. There are three different R_D matrices which are needed in this procedure: R_D-pre, R_D-data, and R_D-post. The construction of each will be

described in detail. To illustrate the process, the methods described will be applied to items matched to Module 5 objectives. The number of items and the number of objectives in each module pool, as well as the number of test items per objective are presented in Table 4. The data for Module 5 test scores consist of an (N x n) array I and an (N x 1) vector S, where N represents the number of students and n represents the number of test items (see Figure 3). Each entry in matrix I (I_{kj}) is the item score (1 for a correct response and 0 for an incorrect response) for the k^{th} subject on the j^{th} item. The (N x 1) vector S contains the test scores (i.e., sum of the item scores) for for each subject.

Let us suppose that 300 students (N = 300) have each taken a different 20 (n = 20) item test T_k (k = 1 to 300) generated randomly from the Module 5 pool. One set of estimates (a vector of length 10) of decision reliabilities (R_D-post entries under the column which corresponds to a length of 10) for tests of length 10 (n/2 = 10), with mastery cut scores varying from 1 to 10 (i = 1 to i = n/2), can be made in the following way:

(1) The items from a given 20 item test T_k (n = 20) for a given subject N_k are randomly assigned to one of two subtests, T_{k1} and T_{k2}, so that each subtest contains 10 (n/2 = 10) of the original 20 items of test T_k.

(2) The total subtest scores, S_{k1} and S_{k2}, on the subtest T_{k1} and T_{k2}, respectively, are computed ($0 \leq S_{k\ell} \leq 10$, ℓ = 1 or 2).

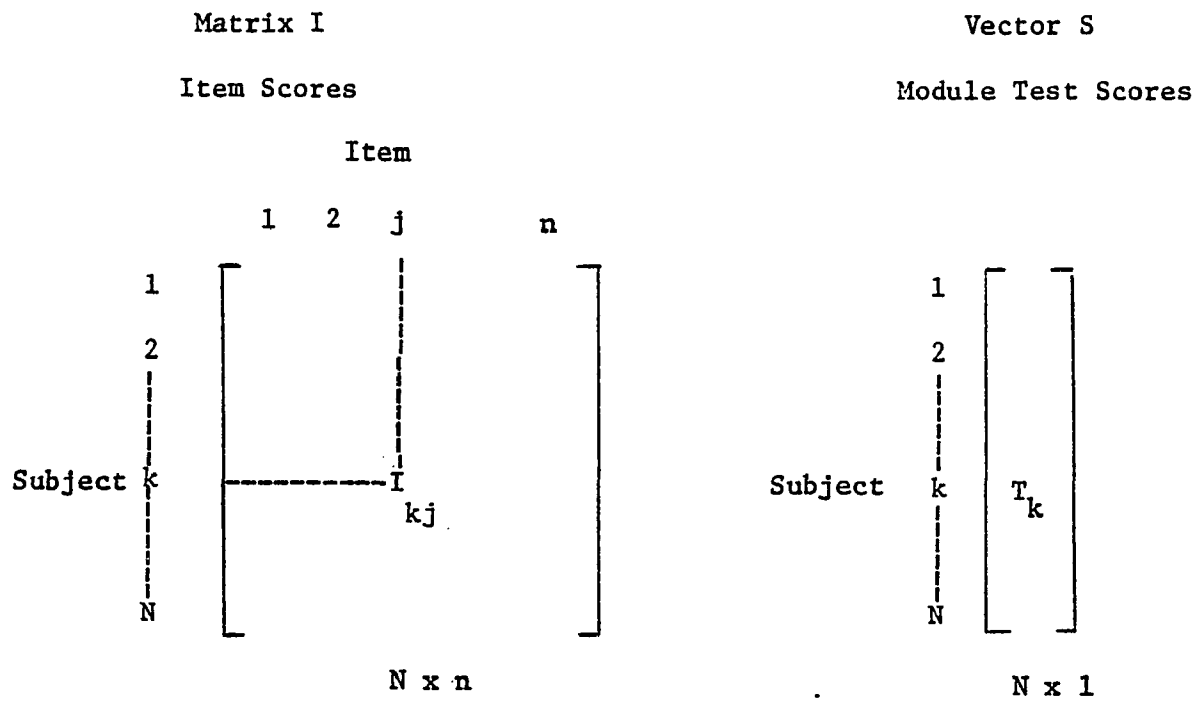
(3) For a given cut score i (i = 1, 2,..., 10), subject N_k is designated a Master if $S_{k\ell} \geq i$ and a non-Master if $S_{k\ell} < i$.

TABLE 4

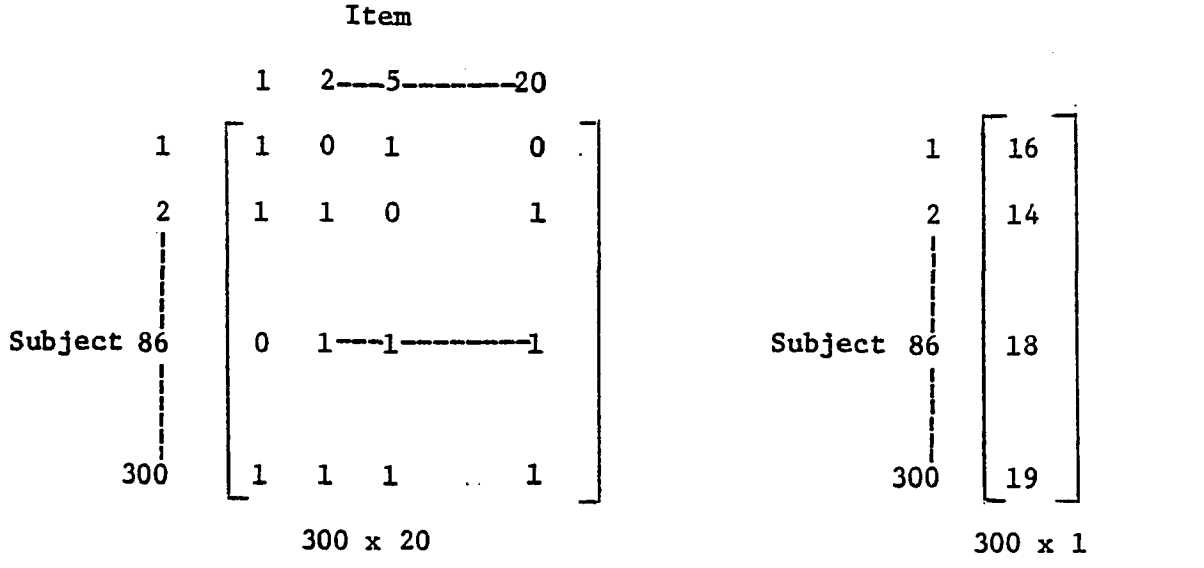
Item Content of Module Item Pools

Objective Number	Number of Items									
	Module 1		Module 2		Module 3		Module 4		Module 5	
	Pool	Test	Pool	Test	Pool	Test	Pool	Test	Pool	Test
1	14	2	7	2	10	2	10	1	15	1
2	12	2	8	2	11	2	10	1	10	1
3	8	1	10	1	9	1	10	1	7	1
4	9	2	6	2	6	1	9	2	10	1
5	18	2	7	1	7	2	10	1	10	2
6	8	1	8	2	5	1	10	2	10	2
7	8	2	10	2	15	2	12	2	10	2
8	4	1	9	1	8	1	10	1	10	2
9	8	2	6	2	10	1	20	2	5	1
10	14	2	8	1	10	2	20	2	8	1
11	10	2	10	2	9	1	10	1	10	2
12	8	1	6	1	10	1	10	2	16	2
13	-	-	6	1	7	1	6	2	-	-
14	-	-	-	-	5	1	-	-	-	-
15	-	-	-	-	4	1	-	-	-	-

FIGURE 3



for example



(4) A 300 x 2 (N x 2) Matrix C (see Figure 4) is constructed in which each element is either M or NM. These entries correspond to the Mastery/non-Mastery assignments for a given subject based on a given subtest score $S_{k\ell}$ (see Step 3).

(5) A 300 x 1 (N x 1) column vector D of 1 or 0 entries is constructed (see Figure 4). For a given subject N_k, a 1 is entered for a consistent decision on S_{k1} and S_{k2} [i.e., (M, M) or (NM, NM)] on the two subtests and a 0 otherwise [i.e., (NM, M) or (M, NM)]. The number of consistent decisions c_i for a given cut score i is determined by summing the 1 entries in Vector D.

(6) The same procedure (Steps 3, 4, and 5) is repeated for each possible cut score i (i = 1, 2,..., 10) for the given test length (n/2 = 10).

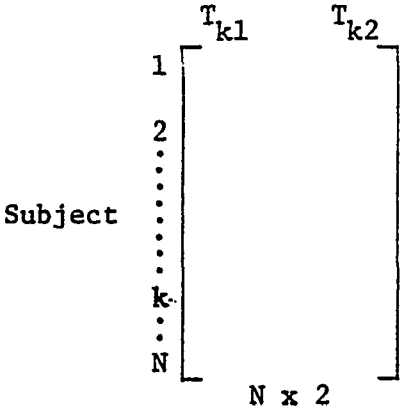
(7) The column of the decision reliability data matrix (R_D-data) corresponding to a test of length 10 (n = 10) and cut scores i = 1, 2,..., 10 contains elements equal to c_i/N (see Figure 5). These estimates may range from 0 to 1.

(8) Additional decision reliability data matrix (R_D-data) columns (each representing a test of specified length) can be filled in by one of three methods: (a) by following a procedure similar to the one outlined in Steps 1-7 but applied to tests of different lengths n, (b) by following a procedure similar to Steps 1-7 but applied to two differtests of length n simultaneously administered to the same student, and (c) by generating R_D-data matrix entries from any given value produced by Steps 1-7 by using the

FIGURE 4

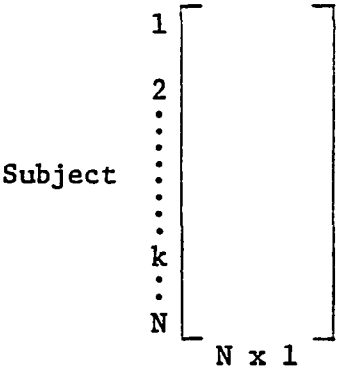
Matrix C

Classifications on Two Tests Based on a Given Cut Score

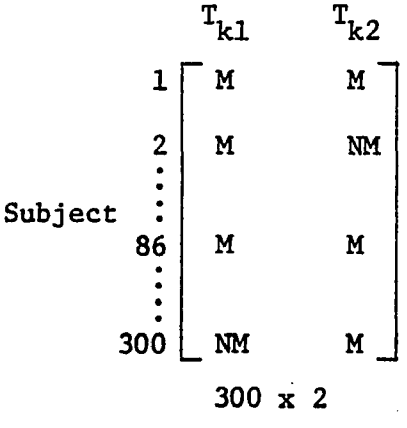


Vector D

Consistency of Classification Decisions



for example



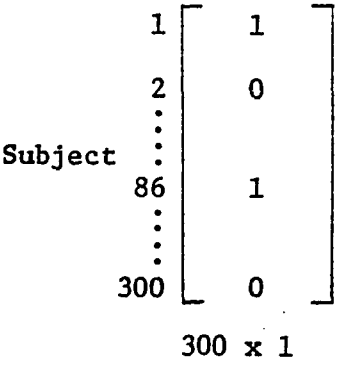
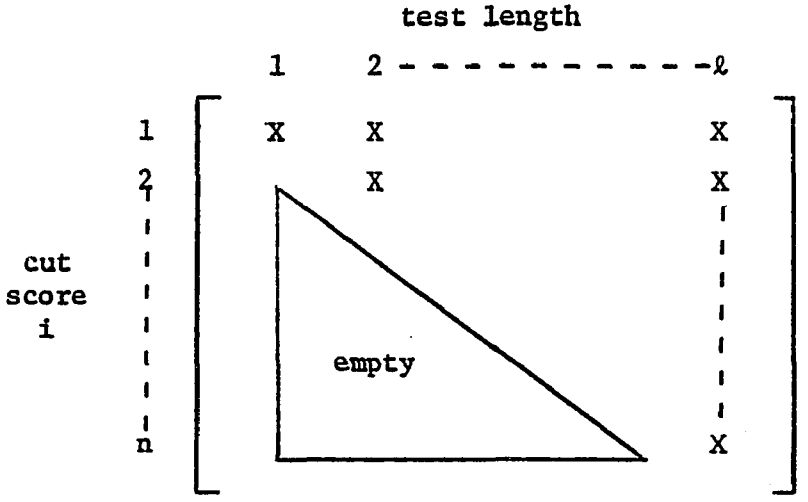


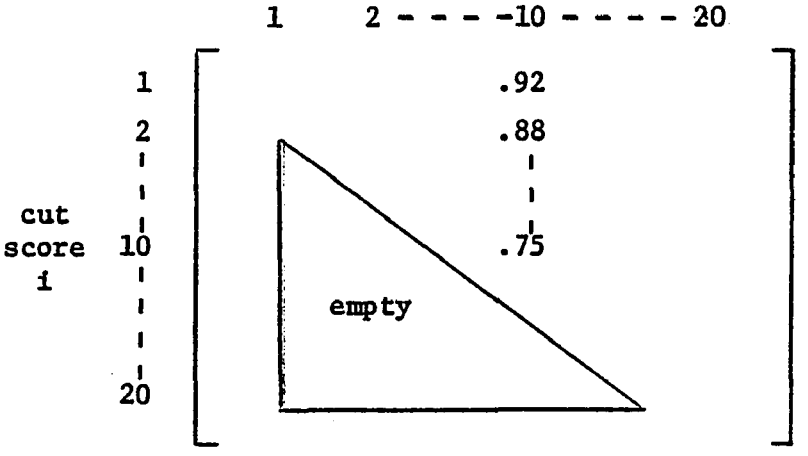
FIGURE 5

R_D-Matrix Form*



n x ℓ

for example



20 x 20

*R_D-data
R_D-prior, or
R_D-post

extrapolation procedure which will be described in Chapter 5.

(9) Following Bayesian methods, the practitioner will select a set of R_D entries for an R_D-prior matrix, ranging in value from .50 to 1.00 (as derived from the theory). These entries are equal to the mean values of the prior distribution functions which the researcher chooses as the best representation of his/her prior beliefs about the R_D values.

(10) Posterior probability distributions for each R_D (given test length and cut score) are computed by combining the information derived from the R_D-data matrix values and the R_D prior distributions according to Bayesian techniques.

(11) Finally, the R_D-posterior matrix values equal the means of the R_D posterior probability distributions. These values, in keeping with one of the virtues of Bayesian estimation procedures, will also range from .50 to 1.00. These posterior means can be viewed as Bayesian point estimators of R_D under a quadratic loss function.

(12) Steps 1-11 can be repeated for additional groups of students. In this way the R_D-posterior estimates may be continuously updated. The previous R_D-posterior matrix becomes the "new" R_D-prior matrix, which is combined with the "new" R_D-data matrix to produce the "new" R_D-posterior matrix: the R_D estimation procedure is a dynamic one.

4.2 Procedure for Generating a Classification Validity Index

Students who complete the introductory course (and the five module tests) take the same final examination T_f, which contains 25 items each

matched to one of 25 key objectives (5 per module) in the curriculum. Students receive a copy of these specific objectives at the beginning of the course. These items are not contained in any of the module item pools. Students who score at or above the cut score (i_f=70% for the Mathematics 45.000 final test) pass the course (i.e., designated Masters); students who score below the cut score fail (i.e., designated non-Masters). A table is prepared (see Table 5) based on student module test scores, final test scores, and drop-out information. Both students who pass (M_j) and students who fail (NM_j) the module test are divided into one of three categories: pass the final test (M_f), fail the final test (NM_f), or do not take the final test (drop).

An estimate of the classification validity for tests of a given length and cut score generated from a given module can also be determined by following a Bayesian approach:

(1) Select a prior distribution for V_C^*, given n and i.

(2) Compute the statistic V_C^* (data) from the data where

$$V_C^* \text{ (data)} = \frac{N_1 + N_5 + N_6}{N_1+N_2+N_3+N_4+N_5+N_6} = \frac{N_1 + N_5 + N_6}{N_j} .$$

(3) Combine the information contained in Steps 1 and 2 to produce a posterior distribution for V_C^*.

(4) The mean of the posterior distribution for V_C^* is the current best estimate of V_C^* (n,i).

(5) Transform V_C^* into V_C by Equation 3.2.9

$$V_C = 2V_C^* - 1.$$

(6) Repeat Steps 1-5 for all possible (n,i) pairs.

TABLE 5

Validity Assessment Data

T_j \ T_f	M: $S_f \geq i_f$	NM: $S_f < i_f$	NF: no final	
M: $S_j \geq i$	$N_1 = N(M_j, M_f)$	$N_2 = N(M_j, NM_f)$	$N_3 = N(M_j, NF)$	$N_1 + N_2 + N_3$
NM: $S_j < i$	$N_4 = N(NM_j, M_f)$	$N_5 = N(NM_j, NM_f)$	$N_6 = N(NM_j, NF)$	$N_4 + N_5 + N_6$
	$N_1 + N_4$	$N_2 + N_5$	$N_3 + N_6$	

T_f = final test

T_j = module test

$N_f = N_1 + N_2 + N_4 + N_5$ = number of students who take final test

$N_j = N_1 + N_2 + N_3 + N_4 + N_5 + N_6$ = number of students who take module test j

In individually paced programs, students who fail a module test may repeat the test once, twice, or as often as necessary until they pass. For programs in which passing a given module is a prerequisite for a student's assignment to the next unit, no validity data can be accumulated for individuals who fail a module test and also fail the final examination since all students who take the final examination have passed all module tests. In such situations $N_4 = N_5 = 0$, but N_6 does not equal zero whenever at least one student who fails the module test drops the course.

In the Hunter College introductory mathematics course, 45.100, students may repeat a failed module test only once and then automatically proceed to the next unit. For this type of course management, it is likely that all cells in the Validity Assessment Data Table (Table 5) will have non-zero values.

The validity index described above is derived from the proportion of agreements between module test mastery designations and final test mastery designations. Other performance criteria can be selected to validate the module test classifications, including student grades in a subsequent mathematics course or on some other relevant mathematics test, or Mastery/non-Mastery designations provided by tutors who know the student's work.

4.3 Results of Applying Procedures to a Mathematics Item Pool

A computer program (JFSREL) was written and used to analyze student response data to module tests. The input to the program for each student consists of a student ID, the course ID, the module number, the test administration number, the number of objectives in the module test,

the final exam score, the final course grade, the item identifiers, and the item scores. The following analyses were performed:

(1) The total module test score was computed (see Figure 3, Vector S).

(2) Two ten-item subtests, T_{k_1} and T_{k_2}, were formed by a random assignment of items to one of these subtests and the subtest scores, S_{k_1} and S_{k_2}, were computed.

(3) Two ten-item subtest scores, $S_{k\text{-odd}}$ and $S_{k\text{-even}}$, were calculated by summing up the odd numbered and the even numbered item scores, respectively.

(4) For a given cut score, $i \leq 10$, a student was assigned to a Master state for a subtest score $T_k \geq i$ and a non-Master state for $T_k < i$ (see Figure 4, Matrix C). The number of consistent assignments was recorded (i.e., M,M or NM,NM on both subtests).

(5) The R_D-data vector for tests of length 10 was formed (see Figure 5 and Table 6). Each entry in the "All" column of Table 6 represents the number of agreements divided by the total number of module tests or subtest pairs. The procedures in (4) and (5) were carried out for all possible cut scores. Following the classification of subjects to Mastery-non-Mastery status on the basis of the final examination, ${}_{M}R_D$ and ${}_{NM}R_D$ vectors were also formed separately and for each group. These data also appear in Table 6.

(6) A non-informative R_D-prior distribution was selected, which can be represented by a truncated beta distribution on the interval $.5 \leq x \leq 1.00$, with $a = b = 1$ (see Equation 3.5.7).

TABLE 6

$\hat{R}_D$-data Vectors

Test Length = 10

	T_1 and T_2			T_{odd} and T_{even}		
i	M	All	NM	M	All	NM
1	1.00	1.00	1.00	1.00	.99	.98
2	.99	.96	.96	.99	.97	.95
3	.99	.91	.84	1.00	.93	.85
4	.97	.86	.71	.97	.90	.77
5	.92	.85	.70	.94	.89	.79
6	.81	.74	.65	.85	.82	.76
7	.70	.72	.73	.70	.75	.82
8	.64	.73	.85	.65	.74	.90
9	.68	.78	.94	.76	.81	.96
10	.80	.87	1.00	.84	.89	1.00
$\hat{\mu}$	.75	.67	.47	.75	.67	.47

Module = 5

Total # Items = 20

Subtest Items = 10

(7) The R_D-posterior distribution means were computed by a second computer program (JFSBAY) which required as input data the $\underline{a}$ and $\underline{b}$ of the prior distribution, the number of agreements, and the total number of pairs of subtests. (These $\underline{a}$ and $\underline{b}$ values are not to be confused with the item parameters a_i and b_i which were described in Chapter 3.) The program also produced the .5, .6, .7, .8, and .9 percentile point probabilities for the posterior distribution. For number of pairs of tests as low as 49 and number of agreements ranging from 35 to 46, the Bayesian procedure--the combining of data with prior information--did not appreciably change the R_D estimates which were calculated in (5) above. The differences appeared in the third decimal place. The Bayesian analyst would comment on such a result that the data swamped the prior beliefs. Since the total number of tests analyzed in Module 5 was 312 and the number of agreements ranged from 229 to 311, these data did indeed "swamp" the non-informative prior. At this point, the decision was made not to use the Bayesian estimation procedure for $\hat{R}_D$ values greater than .5. However, Bayesian procedures should be applied to $\hat{R}_D$ values less than .5, since the model clearly constrains R_D to the range $.5 \leq R_D \leq 1.00$. The Bayesian estimation procedure can also be applied to future investigations of the item parameters a_i and b_i. For a discussion of areas for further research, see Chapter 5.

(8) On the basis of a final test score T_f, a student with a score $T_f \geq 70$ was called a Master and a student with a

score $T_f < 70$ was called a non-Master.[1] To compute the entries for the $\hat{V}_C^*$ vectors corresponding to tests of length 10 separate calculations were made for the Master group and the non-Master group. For the Masters (non-Masters), the number of students with a module subtest score $S_{kl} \geq i$ ($S_{kl} < i$) divided by the number of Masters (non-Masters) produced the ${}_M\hat{V}_C^*$ (${}_{NM}\hat{V}_C^*$) vector entry for that cut score (see Tables 7 and 8).

(9) From the ${}_M\hat{V}_C^*$ (${}_{NM}\hat{V}_C^*$) vector, the ${}_M\hat{V}_C$ (${}_{NM}\hat{V}_C$) vector was computed (see Tables 9 and 10) by the use of Equations 3.2.4 and 3.2.5:

$$ {}_MV_C = 2_MV_C^* - 1 \qquad (3.2.4) $$

$$ {}_{NM}V_C = 2_{NM}V_C^* - 1. \qquad (3.2.5) $$

(10) If the assumption is made that the item characteristics a_i (see Section 3.1) of the items in a given item pool are equal (homogeneous), then these item characteristics can be estimated in the following way. First, subtest score means for T_1, T_2, T_{odd}, and T_{even} are computed for Masters and non-Masters. Since each subtest contains 10 items which are scored 1 (correct) or 0 (incorrect), an estimate of the item difficulty $\hat{a}$ for items included in these tests equals the

[1] Students with final scores equal to zero were eliminated from this analysis since it was unclear from the data whether a student dropped the course or did not take the examination. Because of the fiscal crisis in New York City, the City University of New York closed for several weeks (spring 1976) and students were not required to take the final.

TABLE 7

$\hat{V}^*_{MC}$ Vectors

Test Length = 10

cut score i	T_1	T_2	T_{odd}	T_{even}
1	168† 1.00	168 1.00	168 1.00	168 1.00
2	168 1.00	167 .99	167 .99	168 1.00
3	166 .99	167 .99	165 .98	167 .99
4	166 .99	163 .97	163 .97	166 .99
5	161 .96	158 .94	158 .94	159 .95
6	145 .86	143 .85	146 .87	144 .86
7	121 .72	121 .72	121 .72	126 .75
8	88 .52	92 .55	95 .57	92 .55
9	58 .35	60 .36	57 .34	56 .33
10	26 .16	27 .16	20 .12	27 .16

†first column in each cell represents the number of Mastery classification agreements

Number of Masters = 168

Total # Items = 20

Subtest Items = 10
(Module 5)

TABLE 8

$\hat{V}^*_{C}$ Vectors

Test Length = 10

cut score i	T_1		T_2		T_{odd}		T_{even}	
1	0†	.00	0	.00	0	.00	2	.03
2	3	.04	2	.03	1	.01	3	.04
3	9	.11	10	.13	6	.08	12	.15
4	24	.30	25	.32	14	.18	26	.33
5	37	.47	37	.47	37	.47	34	.43
6	51	.65	51	.65	51	.65	52	.66
7	62	.79	61	.77	65	.82	63	.80
8	70	.89	70	.89	72	.91	72	.91
9	77	.98	74	.94	78	.99	75	.95
10	79	1.00	79	1.00	79	1.00	79	1.00

†first column in each cell represents the number of non-Mastery classifications

Number of Nonmasters = 79

Total # Items = 20

Subtest Items = 10
(Module 5)

TABLE 9

$\hat{V}_{MC}$ Vectors

Test Length = 10

cut score i	T_1	T_2	T_{odd}	T_{even}
1	1.00	1.00	1.00	1.00
2	1.00	.99	.99	1.00
3	.98	.99	.96	.99
4	.98	.94	.94	.98
5	.92	.88	.88	.89
6	.73	.70	.74	.71
7	.44	.44	.44	.50
8	.05	.10	.13	.10
9	-.31	-.29	-.32	-.33
10	-.69	-.68	-.76	-.68

$$\hat{V}_{MC} = 2\hat{V}^{*}_{MC} - 1$$

TABLE 10

$\hat{V}_C$ Vectors

Test Length = 10

cut score i	T_1	T_2	T_{odd}	T_{even}
1	-1.00	-1.00	-1.00	-1.00
2	- .92	- .95	- .98	- .92
3	- .77	- .75	- .85	- .70
4	- .39	- .37	- .65	- .34
5	- .06	- .06	- .06	- .14
6	.29	.29	.29	.32
7	.57	.54	.65	.60
8	.77	.77	.82	.82
9	.95	.87	.98	.90
10	1.00	1.00	1.00	1.00

$$\hat{V}_C = 2\,\hat{V}^*_C - 1$$

mean subtest score divided by 10. However, since the items parameters are, in general, unequal (heterogeneous, i.e., for items j and k, $a_j \neq a_k$), the procedure for estimating $\hat{a}$ described above can still be applied to the subtest score means but the resulting statistic is $\hat{\mu}_{\bar{a}}$ (i.e., estimates of the mean of the a_i values of the items in the pool). It should be further noted that, for simplicity, the notation $\bar{\hat{a}}$ will be used instead of $\hat{\mu}_{\bar{a}}$. Similar arguments apply to the b_i characteristics. The $\bar{\hat{a}}$ and $\bar{\hat{b}}$ values are presented in Table 11. These values will be referred to in the discussion section (Section 4.4) which follows directly and again in Chapter 5.

4.4 Discussion of Results

The results presented in Section 4.3 form the basis for the first empirical tests of the item pool test model. These findings refer to tests of fixed length with n = 10. The reader may prefer to reread the three groups of predictions which were stated in Section 3.6 before considering the following model investigations:

Investigation 1. The data support the prediction that the relationship between $_{M}R_{D}$ and cut score i is U-shaped for fixed values of n and a (Prediction 1.1.3). Since only 10-item tests were analyzed, n is fixed and equal to 10. From Table 11, the statistic $\bar{\hat{a}}$ (i.e., the estimate of a assuming homogeneous item

TABLE 11

Estimates of

Average Difficulty for Subtest Items

	Masters ($\bar{\hat{a}}$)	non-Masters ($1 - \bar{\hat{b}}$)
T_1	.754	.479
T_2	.756	.482
T_{odd}	.750	.490
T_{even}	.757	.471
Average	$\bar{\hat{a}} \doteq .75$	$(1 - \bar{\hat{b}}) \doteq .47$

parameters, or the estimate of μ_a assuming heterogeneous item parameters) for Module 5 items equals approximately .75. Table 12 presents a comparison between the theoretical ${}_M R_D$ values for n = 10 and $\underline{a}$ = .75 and the empirical findings for n = 10 and $\bar{\hat{a}}$ = .75. Both sets of values are U-shaped in form with minimum entries at i = 8. The fact that the theoretical and empirical values are not the same has a plausible explanation. The theoretical ${}_M R_D$ values were calculated on a "true" $\underline{a}$ value (or a μ_a value) of exactly .75, whereas the empirical ${}_M \hat{R}_D$ values were based on sample estimates (either of $\underline{a}$ or μ_a). Thus, some discrepancy between the theoretical and empirical values for ${}_M R_D$ would be anticipated even if the model held exactly. The same discussion may be applied to the ${}_{NM} R_D$ values in Table 12.

Investigation 2. The entries in Table 13 present the theoretical R_D values for the 10 different cut scores corresponding to the values $(1 - \bar{\hat{b}})$ = .47, and $\bar{\hat{a}}$ = .75 which characterize the Module 5 test data. Since the $\bar{\hat{a}}$ range is limited by this data set, the explorations of the predictions concerning the relationships between R_D and $\underline{a}$ for varying i and n fixed (see Predictions 1.2.1, 1.2.2, and 1.2.3, Section 3.6) must also be carried out in this restricted range. For i = 1, the theoretical values of R_D are constant and equal 1.00 for $\underline{a}$ = .47 and $\underline{a}$ = .75. The experimental $\hat{R}_D$ values of .99, and 1.00 support this predicted relationship in the restricted $\underline{a}$ region. For i = 10 (last line in Table 13), the experimental $\hat{R}_D$ values, 1.00 and .82, are decreasing functions of $\bar{\hat{a}}$ as are the theoretical values of 1.00,

TABLE 12

Investigation 1

$_{M}^{R}D$ and $_{NM}^{R}D$ as Functions of i

		Empirical*		Theoretical**	
		$\hat{_{M}^{R}D}$	$\hat{_{NM}^{R}D}$	$_{M}^{R}D$	$_{NM}^{R}D$
	1	1.00	1.00	1.00	1.00
	2	.99	.96	1.00	.97
	3	.99	.84	1.00	.85
	4	.97	.71	.99	.65
cut	5	.92	.70	.96	.50min
score	6	.81	.65min	.86	.58
i	7	.70	.73	.65	.79
	8	.64min	.85	.50min	.93
	9	.68	.94	.63	.99
	10	.80	1.00	.89	1.00

*from Table 6, columns 1 and 3

**from Table 3 for a = .75 for $_{M}^{R}D$

from computer program for a = .47 for $_{NM}^{R}D$

TABLE 13

Investigation 2

R_D as a Function of Item Parameters

		Non-Masters		Masters	
		Data* $\hat{1-b}$=.47	Theory 1-b=.47	Data $\hat{a}$=.75	Theory a=.75
	1	.99	1.00	1.00	1.00
	2	.96	.97	.99	1.00
cut	3	.84	.85	.99	1.00
score	4	.74	.65	.97	.99
i	5	.74	.50	.93	.96
	6	.70	.58	.88	.86
	7	.78	.79	.70	.65
	8	.87	.93	.65	.50
	9	.95	.99	.72	.63
	10	1.00	1.00	.82	.89

*Empirical entries are $\hat{R}_D$ averages for 2 pairs of subtests in Tables 15 and 16.

and .89. For i = 2 to i = 6, both the theoretical $\hat{R}_D$ values and the experimental $\hat{R}_D$ values increase as a or $\bar{\hat{a}}$ increases. Similarly, for i = 8 to i = 10, both sets of R_D values decrease as a increases.

The case for i = 7 cannot be evaluated from the current data, since the theoretical curves which represent the relationship between R_D and cut score are U-shaped and achieve their minimum values at i = 7 for a ranging from .60 to .69 (Table 3). A third set of R_D values is needed for an a value in the region bounded by .55 and .70. The discussion in Investigation 1 with respect to discrepancies between the theoretical R_D values and the empirical findings is applicable to this situation as well. Also, another data set might yield $\hat{R}_D$ values which are indeed U-shaped.

Thus, the data for i = 1 to 6 and i = 8 to 10 support the predicted relationships (Predictions 1.2.1, 1.2.2, and 1.2.3 in Section 3.6) between R_D and a for a fixed n. The data for i = 7 are insufficient for testing this prediction.

Investigation 3. The data support the prediction that R_D provides an upper bound to V_C (see Prediction 1.3, Section 3.6). Tables 14, 15 and 16 present the R_D and V_C data for all students, Masters and non-Masters, respectively. In all three cases the average R_D entries for every cut score are equal to or greater than the corresponding V_C values.

TABLE 14

Investigation 3

Relationship Between $\hat{R}_D$ and $\hat{V}_C$

(all students)

	cut score i =	1	2	3	4	5	6	7	8	9	10
$\hat{R}_D$†	T_1 and T_2	1.000	.962	.913	.859	.846	.740	.715	.731	.776	.872
	T_{odd} and T_{even}	.994	.974	.929	.901	.885	.821	.750	.744	.811	.894
	average $\hat{R}_D$	1.00	.97	.92	.88	.87	.78	.73	.74	.79	.88
$\hat{V}_C$	T_1	.360	.425	.466	.563	.603	.571	.457	.296	.093	-.150
	T_2	.360	.368	.409	.498	.555	.538	.474	.312	.109	-.150
	T_{odd}	.360	.360	.385	.433	.579	.595	.506	.352	.093	-.198
	T_{even}	.377	.385	.449	.555	.563	.587	.530	.328	.061	-.142
	average $\hat{V}_C$	.36	.39	.43	.51	.58*	.57	.49	.32	.09	-.16

†See Table 6

*Although .58 at i=5 represents the maximum $\hat{V}_C$, this finding cannot be used for the selection of a cut score on the module test. First, the validity data are not available before the final examination (or validation instrument) is scored. The "true" Mastery and non-Mastery classifications are based on this final examination. Secondly, the cut score yielding the maximum $\hat{V}_C$ value depends on the relative number of "true" Masters and non-Masters. For these data, a low cut score is optimal since more than two-thirds of the students are "true" Masters (number of Masters = 168, number of non-Masters = 79).

TABLE 15

Investigation 3

Relationship Between $\hat{R}_{MD}$ and $\hat{V}_{MC}$

(Masters)

	cut score i =	1	2	3	4	5	6	7	8	9	10
$\hat{R}_{MD}$†	T_1 and T_2	1.000	.994	.994	.970	.923	.810	.702	.643	.679	.804
	T_{odd} and T_{even}	1.000	.994	.988	.970	.935	.845	.696	.649	.756	.839
	average $\hat{R}_{MD}$	1.00	.99	.99	.97	.93	.88	.70	.65	.72	.82
$\hat{V}_{MC}$††	T_1	1.000	1.000	.976	.976	.917	.726	.440	.048	-.310	-.690
	T_2	1.000	.988	.988	.940	.881	.702	.440	.095	-.286	-.679
	T_{odd}	1.000	.988	.964	.940	.881	.738	.440	.131	-.321	-.762
	T_{even}	1.000	1.000	.988	.976	.893	.714	.500	.095	-.333	-.679
	average $\hat{V}_{MC}$	1.00	.99	.98	.96	.89	.72	.46	.09	-.31	-.70

†see Table 6

††see Table 9

TABLE 16

Investigation 3

Relationship Between $\hat{R}_{NM_D}$ and $\hat{V}_{NM_C}$

(Non-Masters)

	cut score i =	1	2	3	4	5	6	7	8	9	10
$\hat{R}^{\dagger}_{NM_D}$	T_1 and T_2	1.000	.962	.835	.709	.696	.646	.734	.848	.937	1.000
	T_{odd} and T_{even}	.975	.949	.848	.772	.785	.759	.823	.899	.962	1.000
	average $\hat{R}_{NM_D}$	.99	.96	.84	.74	.74	.70	.78	.87	.95	1.00
$\hat{V}^{\dagger\dagger}_{NM_C}$	T_1	-1.000	-.924	-.772	-.392	-.063	.291	.570	.772	.949	1.000
	T_2	-1.000	-.949	-.747	-.367	-.063	.291	.544	.772	.873	1.000
	T_{odd}	-1.000	-.975	-.848	-.646	-.063	.291	.646	.823	.975	1.000
	T_{even}	- .949	-.924	-.696	-.342	-.139	.316	.595	.823	.899	1.000
	average $\hat{V}_{NM_C}$	- .99	-.94	-.77	-.44	-.08	.30	.59	.80	.92	1.00

†see Table 6

††see Table 10

In summary, these investigations have successfully tested out three groups of predictions derived from the theoretical model: the relationship between ${}_{M}R_{D}$ and cut score, the relationship between R_{D} and $\underline{a}$, and, finally, the important relationship between R_{D} and V_{C}. The primary purpose of these analyses is to provide an answer to the following question: Can the model be applied in a practical testing situation? In light of the investigations described in this section, the answer to this question is clearly in the affirmative.

Chapter 5

Summary and Implications for Future Research

5.1 Summary

The main goals of this research have been accomplished. A statistical theory to study reliabilities and validities of criterion related item pools has been developed. Two useful psychometric properties of these item pools--decision reliability (R_D) and classification validity (V_C)--have been defined. In addition, two item parameters are defined: a_i is the probability of a Master answering item i correctly and b_j is the probability of a non-Master answering item j incorrectly. Several interesting theoretical relationships have been derived concerning test length, cut score, and the two item parameters. The model has successfully handled the case of most practical interest, in which the item pool contains items with unequal item parameters. A most important relationship between R_D and V_C was derived: R_D is an upper bound to V_C. This theoretical development has been applied to a set of item pools currently in use in an individualized college mathematics program. The use of Bayesian estimation methods has been explored to provide estimates of the several parameters of interest. A procedure has been described to enable practitioners to estimate the decision reliabilities and the classification validities associated with a heterogeneous item pool. And finally three sets of predictions made from the theoretical model were confirmed by the data analyses.

5.2 Implications for Future Research

There are further interesting problems which need to be pursued with respect to domain-referenced item pool testing. First, a method for "certifying the domain" of the item pool content coverage would be an invaluable addition to the practitioners skills. Does the item pool "cover" the domain of interest? Do the items clearly match stated curricular objectives? Do the stated behavioral objectives derive from the subject matter? Do the objectives exhaustively "cover" the content area of interest? These are just some of the questions which readily come to mind. To classify a student as a Master of some module of curriculum material based on a set of test items randomly generated from a pool of items necessarily requires that the item pool is a "good" one in the sense that the pool "covers" the domain. The phrase "certification of the domain" suggests the type of assessment which is required.

Another interesting problem derives from the relationships between R_D, cut-score, and the item parameters. The theory suggests a cut-score to the practitioner, given the n-size of the test, the average of the item parameters in the item pool, and the desired value of R_D. However, it is possible to extend the theory to include a tailored testing paradigm. The practitioner could select an R_D value, and a cut-score, and then adjust the test length to achieve the desired R_D value. Such a paradigm could also suggest specific item parameter strata within the item pool from which to select items for increasing the module test length. In this way, the errors of Mastery/non-Mastery misclassification can be kept below some previously stipulated value.

Two additional research areas are presented in some detail to illustrate how they may be studied using the statistical approach which has been developed in the main body of this research: the estimation of the item characteristics a_i and b_i, and the development of augmentation formulas similar to the Spearman-Brown formulas of classical test theory.

Determination of the Item Parameters a_i and b_i

Estimates of the item parameters a_i, the probability of a Master answering item i correctly, and b_i, the probability of a non-Master answering item i incorrectly, can be made in at least three ways. As an illustration, these procedures will be described as applied to the 200-item Module 5 item pool.

Following the first method, at the end of one semester of item pool use, the students who have completed the course will have been designated Masters or non-Masters on the basis of the final examinations. Item data can be collected for Masters as a group and then non-Masters as a group. For each item, for example item j, $\hat{a}_j$ equals the proportion of Masters who answered the item correctly, and $\hat{b}_j$ equals the proportion of non-Masters who answered the item incorrectly. The practitioner can immediately recognize "bad" items: items with low $\hat{a}$ values (e. g., less than .6), items with low $\hat{b}$ values (e. g., less than .6), or items having $\hat{a}$ and $\hat{b}$ values which are nearly equal (e. g., non-discriminating between Masters and non-Masters). The practitioner can then replace the "bad" items with additional items which are matched to the particular objectives of the eliminated items. It is important to recall that for item pools used in the Mastery Learning

mode, items may never be removed, since the content domain of interest must be completely "mapped" by items.

Although such a procedure will work, it is not a particularly satisfying one to the practitioner who is using an item pool for the first time. Waiting as long as one semester before "correcting" the item pool is not necessary. The following two procedures are suggested as possible solutions. The practitioner can estimate item parameters using Bayesian methods similar to those described in Section 4.1 and Section 4.3. Following a Bayesian approach, the practitioner subjectively "choses" a prior distribution function for the a's and b's. Suppose, as an approximation, after 100 students have taken 20-item tests generated from the Module 5 item pool, students who pass the module test are assigned to the Mastery state and those who fail are called non-Masters. A two-by-two table is prepared for each item (see Table 17). It should be noted that by the time 100 students have competed Module 5 tests, each item will have appeared on some test about 10 times. The N values in Table 18 will then sum to 10 and will necessarily be small numbers. Nevertheless, these four table entries provide the data which can be used to update the distribution function describing the item parameters. The procedure is a dynamic one and can be repeated as often as is practically desirable (i. e., following data collections of groups of 100 students, or following the replacement of items in the pool). By the end of the semester, the item pool will have been updated several times. Such flexibility is particularly advantageous to the practitioner who is using a newly constructed item pool for the first time.

TABLE 17

Item Parameter Data

		Item: Correct	Item: Incorrect
Module Test	Pass	N_1	N_2
	Fail	N_3	N_4

Clearly, this is a bootstrap technique which must be carefully investigated to determine if convergence occurs. Certainly, if the correlation between the module test score and the final exam score is perfect (i. e., correlation = 1), then the method is a reasonable one.

An alternate Bayesian approach to the estimation of item parameters before M and NM are identified by final examination scores is, first, to specify beliefs about the learner state of each person, as well as beliefs about the a_i's and b_i's. Next, right-wrong data are collected for each item. By summing over the 2^n possible states for all n learners, one can produce a posterior belief distribution for $\underline{a}$ and for $\underline{b}$. In such an analysis, beliefs about learner states are substituted for the exact knowledge of learner states. This is a particularly useful technique at a point in time--before the end of the semester--when final examinations have not yet been administered and, thus, exact learner states are unknown.

Development of Augmentation Formulas for R_D and V_C

Students completing Module 5 were administered 20-item tests. Since it was neither practical nor feasible to administer parallel 20-item tests to each student in the application phase of this research (Chapter 4), $\hat{R}_D$ and $\hat{V}_C$ item pool values for two subtests of length 10 and varying cut scores were calculated. However, the practitioner, operating the item pool by generating 20-item tests, is interested in the $\hat{R}_D$ and $\hat{V}_C$ matrix entries for n = 20 rather than for n = 10. The question of interest becomes the following: given

one $\hat{R}_D$ or $\hat{V}_C$ matrix call entry, can another entry be determined? For example, if $\hat{R}_D$ is known for 10-item tests and a cut score of seven, can $\hat{R}_D$ for a 20-item test and a cut score of 14 be directly calculated? Although a theoretical formula (similar, in a sense, to the Spearman-Brown augmentation formula of classical test theory) has not been derived, the following approximation method is suggested for study.

The method will be described using the Modula 5 10-item subtest data for Masters (see Table 6) for determining an R_D entry for a test with $n = 20$ and $i = 14$:

(1) The estimate of the average item difficulty level for Masters equals approximately .75.

(2) By use of the tables of the cumulative terms of the binomial distribution (see Beyer, 1966), the "A" value (probability that a student will achieve a score of 14 or more on a 20-item test) for a p of .75 is found to be approximately .79.

(3) For Masters

$$_{M}R_D = 1 - 2A + 2A^2,$$

and by substituting .79 for A, we have $\hat{R}_D \doteq .68$.

The R_D theoretical entry for Masters with $n = 10$ and $i = 7$ (Table 3) equals .65. The $\hat{R}_D$ value for the item pool operated with the larger test length 20 and "similar" cut score (10/20 = 7/14) is the larger value.

Now that one example has been worked through, this approximation method may be easily generalized. Data from an item pool are analyzed by computing $\hat{R}_D$ using pairs of tests of length n randomly generated from tests of length 2n which have been administered to the students (Masters) being tested on the module. The average of the estimated $\hat{a}$ of the items in the pool is computed. By using this $\bar{\hat{a}}$ value as the cumulative probability table entry ("p" value in table), "A" can be read directly, by referring to the entry corresponding to the n and i of interest. From this tabled "A" value, both ${}_M\hat{R}_D$ and ${}_M\hat{V}_C$ can be computed using Equations 3.1.9c and 3.2.4, respectively.

By repeating this procedure, every cell in the ${}_M\hat{R}_D$ or ${}_M\hat{V}_C$ matrix can be filled. Once one entry has been estimated empirically, any other can be determined. By a similar process, the ${}_{NM}R_D$ and ${}_{NM}V_C$ matrices can also be filled. Since the number of Masters and the number of non-Masters is also known, R_D can be calculated using Equation 3.1.13

$$R_D = {}_MR_D \text{ Prob (M)} + {}_{NM}R_D \text{ Prob (NM)}.$$

Such a suggested approximation method needs to be studied further and, if reasonable, should prove useful until a theoretically exact relationship is available.

5.3 A Final Observation

A model which describes the relationship between an examinee's total test score on a set of items, and the learner state of the examinee, must address itself to assumptions concerning person (examinee) characteristics and item characteristics. In the present research, the following assumptions were made:

(1) A person can be described as being in one of two possible states, M or NM.

(2) For Masters (non-Masters) the a (b) values for different items can be different.

(3) The a (b) value for any given item is assumed to be constant for all Masters (non-Masters).

In the simple case, in which all the item parameters are assumed equal for Masters (non-Masters), the probability of a Master scoring i or greater on a test of length n is given by

$$\text{Prob}(X \geq i \mid M) = \sum_{j=i}^{n} \binom{n}{j} a^{j}(1-a)^{n-j}, \qquad (3.1.3a)$$

where a equals the probability of a Master answering an item correctly. The probability of a non-Master scoring less than i is given by

$$\text{Prob}(X < i \mid NM) = \sum_{j=o}^{i-1} \binom{n}{j} b^{n-j}(1-b)^{j} \qquad (3.1.4a)$$

where b represents the probability of a non-Master answering an item incorrectly.

For the general case of unequal item parameters, we have, in the case of Masters,

$$\text{Prob}(X \geq i \mid M) = \sum_{j=1}^{n} \binom{n}{j} \left(\frac{s}{s+t}\right)^{j} \left(1-\frac{s}{s+t}\right)^{n-j}, \qquad (3.4.10a)$$

where s and t are the parameters of the beta distribution which is assumed to describe the distribution of the item characteristics in the item pool. For non-Masters we have

$$\text{Prob}(X < i \mid NM) = \sum_{j=o}^{i-1} \binom{n}{j} \left(\frac{s'}{s'+t'}\right)^{n-j} \left(1- \frac{s'}{s'+t'}\right)^{j}. \qquad (3.4.10b)$$

There are other models which make different assumptions. For example, Lord and Novick (1968) assume that each examinee has a true ability θ ($0 \leq \theta \leq 1$) and that the test scores follow a binomial distribution with parameter θ. This true ability θ represents the proportion of items in an item universe that an examinee knows. For an n-item test the probability that a person with true ability θ achieves a test score X greater than or equal to a cut score i is given by

$$\text{Prob}(X \geq i \mid \theta) = \sum_{j=i}^{n} \binom{n}{j} \theta^{j} (1-\theta)^{n-j}. \qquad (5.5.1)$$

In this model, the person characteristic θ can take on any value between 0 and 1. However, the item characteristic describing the probability of answering an item correctly, given θ, is equal to θ for all items. The equations, 3.4.10 (3.4.11) and 5.5.1, are similar in form, but the former (Model I) contains the expected value $\frac{s}{s+t}$ $(\frac{s'}{s'+t'})$

of the variable item parameter whereas the latter (Model II) contains the variable person parameter θ.

This discussion suggests that one might study a third model (Model III) which accounts both for variable item characteristics and discrete but varying learner characteristics. An extension of Model I from the two-state learner classification, Master and non-Master, to an N-state classification paradigm results in Model III. Table 18 presents a summary of the item and person characteristics of the three models. For example, the characteristics of the N-state Model III appear in the right-hand column. For an item pool containing n items, c_{1i}, c_{2i}, ..., c_{ni} represent the n item characteristics for examinees in the i^{th} learner state, such that c_{ji} represents the probability that a learner in the i^{th} state will answer the j^{th} item correctly. Since N learner-states are hypothesized, there are a total on n x N item characteristics associated with the items in the pool. Each learner state is designated by a symbol S_i, where i is an integer equal to or greater than 3. The learner states can be ordered on a continuum from non-Mastery to Mastery of the content domain under consideration. Model II is the most general in the sense that an infinite number of learner states is assumed (i.e., $0 \leq \theta_{person} \leq 1$). However, Model III is the most general in the sense that for each learner state each of the n items can have a different item characteristic. For Model II, there is only one item characteristic θ_{item} which is associated with each learner of ability θ_{person}, where $\theta_{item} = \theta_{person}$. In the final analysis, the ultimate test of these models rests on their usefulness in an applied setting.

TABLE 18

Item Response Models

		Model I 2 state	Model II continuous	Model III N-state
Item characteristics	symbol	a_i and b_i $i = 1, 2$	θ_{item}	$c_{1i}, c_{2i}, \ldots, c_{ni}$ $i = 1, 2, \ldots, N$
	number	2 per item 2n per item pool	maximum of N per item maximum of N per item pool	N per item Nn per item pool
	range	$0 \leq a_i \leq 1$ $0 \leq b_i \leq 1$	$0 \leq \theta_{item} \leq 1$	$0 \leq c_{ji} \leq 1$ $j = 1, 2, \ldots, n$
Person characteristics	symbol	M and NM	θ_{person} $0 \leq \theta_{person} \leq 1$	S_i
	number	N = 2	Maximum of N	$N \geq 3$

n = number of items in pool

N = number of person states

GLOSSARY OF SYMBOLS

a	probability of a Master answering an item correctly
a_j	probability of a Master answering item j correctly
b	probability of a non-Master answering an item incorrectly
b_j	probability of a non-Master answering item j incorrectly
$\hat{\bar{a}}$	estimate of item difficulty for Masters
$1-\hat{\bar{b}}$	estimate of item difficulty for non-Masters
$\mu_{\underline{a}}$, $\mu_{\underline{b}}$	average mean difficulty indices of groups of items
c	test score
c_{ji}	probability of a learner in the i^{th} state answering item j correctly
i	cut score
n	test length or total number of items
O	learning objectives
p	total number of items in an item pool, which measure one or more learning objective
s, t	parameters of a beta distribution
s', t'	parameters of a beta distribution

A	$\text{Prob}(X \geq i \mid M)$
B	$\text{Prob}(X < i \mid NM)$
M	Master
NM	non-Master
N	total number of students in population of interest
T_1, T_2	two tests of equal length, assembled by selecting items randomly from the item pool
U, V	particular items in Tests T_1 and T_2, respectively
u, v	scores on items U and V, respectively
X, Y	scores on tests T_1 and T_2, respectively
x_i, y_i	item scores for Masters and non-Masters, respectively
R_D	decision reliability (unconditional)
$_MR_D$	decision reliability for Masters
$_{NM}R_D$	decision reliability for non-Masters
V_C^*	classification validity (unscaled, unconditional)
$_MV_C^*$	classification validity for Masters (unscaled)

$_{NM}V_C^*$ classification validity for non-Masters (unscaled)

V_C classification validity (scaled, unconditional)

$_{M}V_C$ classification validity for Masters (scaled)

$_{NM}V_C$ classification validity for non-Masters (scaled)

FORMULAS

Test Score Probabilities for Equal Item Parameters	Equation	Page
$\text{Prob}(X \geq i \mid M) = \sum_{j=1}^{n} \binom{n}{j} a^j (1-a)^{(n-j)} \equiv A$	(3.1.3a)	47
$\text{Prob}(Y \geq i \; M) = \sum_{j=i}^{n} \binom{n}{j} a^j (1-a)^{(n-j)} \equiv A$	(3.1.3b)	48
$\text{Prob}(X < i \mid NM) = \sum_{j=0}^{i-1} \binom{n}{j} (1-b)^j b^{(n-j)} \equiv B$	(3.1.4a)	48
$\text{Prob}(Y < i \mid NM) = \sum_{j=0}^{i-1} \binom{n}{j} (1-b)^j b^{(n-j)} \equiv B.$	(3.1.4b)	48

Test Score Probabilities for Unequal Item Parameters	Equation	Page
$\text{Prob}(X \geq i \mid M) = \sum_{j=i}^{n} \binom{n}{j} \left(\frac{s}{s+t}\right)^j \left(1 - \frac{s}{s+t}\right)^{n-j} \equiv A$	(3.4.10)	61
$\text{Prob}(X \leq i \mid NM) = \sum_{j=0}^{i-1} \binom{n}{j} \left(\frac{s'}{s'+t'}\right)^{n-j} \left(1 - \frac{s}{s'+t'}\right)^j \equiv B$	(3.4.11)	61

Decision Reliability	Equation	Page
$_{M}R_{D} = 1 - 2A + 2A^{2}$	(3.1.9c)	49
$_{NM}R_{D} = 1 - 2B + 2B^{2}$	(3.1.10c)	50
$R_{D} = {}_{M}R_{D}\ \text{Prob}(M) + {}_{NM}R_{D}\ \text{Prob}(NM)$	(3.1.13)	51
$R_{D} = {}_{M}R_{D}\ \text{Prob}(M) + {}_{NM}R_{D}\ [1 - \text{Prob}(M)]$	(3.1.14)	51
Classification Validity (unscaled)		
$_{M}V_{C}^{*} \equiv A$	(3.2.1)	52
$_{NM}V_{C}^{*} \equiv B$	(3.2.2)	52
$V_{C}^{*} = {}_{M}V_{C}^{*}\ \text{Prob}(M) + {}_{NM}V_{C}^{*}\ \text{Prob}(NM)$	(3.2.3)	52
Classification Validity (scaled)		
$_{M}V_{C} = 2A - 1 \equiv 2\,{}_{M}V_{C}^{*} - 1$	(3.2.4)	53
$_{NM}V_{C} = 2B - 1 \equiv 2\,{}_{NM}V_{C}^{*} - 1$	(3.2.5)	53
$V_{C} = {}_{M}V_{C}\ \text{Prob}(M) + {}_{NM}V_{C}\ \text{Prob}(NM)$	(3.2.6)	54
$V_{C} = (2A - 1)\ \text{Prob}(M) + (2B - 1)\ \text{Prob}(NM)$	(3.2.7)	54
$V_{C} = 2V_{C}^{*} - 1$	(3.2.9)	54

Appendix

1. Expression for V_C in terms of V_C^*:

From Equation 3.2.9, we have

$$V_C = (2_M V_C^* - 1)\ \text{Prob}(M) + (2_{NM} V_C^* - 1)\ \text{Prob}(M).$$

Using the distributive property, we find

$$V_C = 2_M V_C^*\ \text{Prob}(M) - \text{Prob}(M) + 2_{NM} V_C^*\ \text{Prob}(NM) - \text{Prob}(NM).$$

Applications of the commutative and the associative properties gives

$$V_C = 2\ \{\ {}_M V_C^*\ \text{Prob}(M) + {}_{NM} V_C^*\ \text{Prob}(NM) - [\text{Prob}(M) + \text{Prob}(NM)]\ \}.$$

Since $\text{Prob}(M) + \text{Prob}(NM) = 1$, we have

$$V_C = 2V_C^* - 1.$$

2. Proof that

$$-\sqrt{2R_D - 1} \leq V_C \leq +\sqrt{2R_D - 1}:$$

From Equations 3.1.9c, 3.1.10c, 3.1.14, and 3.1.8,

we have

$$R_D = (1 - 2A + 2A^2)P + (1 - 2B + 2B^2)\ (1 - P) \qquad (1)$$

and

$$V_C = (2A - 1)P + (2B - 1)\ (1 - P), \qquad (2)$$

where

P = Prob(M) and $(1 - P)$ = Prob(NM).

By applying the distributive, associative, and commutative properties as well as the common algebraic operations, we have

$$2R_D - 1 = (2 - 4A + 4A^2)P + (2 - 4B + 4B^2)\ (1 - P) - 1$$

$$= [1 + (2A - 1)^2]P + [1 + (2B - 1)^2]\ (1 - P) - 1$$

$$= (2A - 1)^2\ P + P + (2B - 1)^2\ (1 - P) + (1 - P) - 1$$

$$= (2A - 1)^2\ P + (2B - 1)^2\ (1 - P) + P + 1 - P - 1$$

$$= (2A - 1)^2\ P + (2B - 1)^2\ (1 - P). \qquad (3)$$

Now, let $\alpha = 2A - 1$ and $\beta = 2B - 1$ and by substituting in Equations 2 and 3, we have

$$V_C = \alpha P + \beta\ (1 - P) \qquad (4)$$

and

$$2R_D - 1 = \alpha^2\ P + \beta^2\ (1 - P). \qquad (5)$$

From Equations 4 and 5, and algebraic manipulation, we have

$$V_C^2 - (2R_D - 1) = [\alpha P + \beta(1 - P)]^2 - [\alpha^2 P + \beta^2(1 - P)]$$

$$= \alpha^2 P^2 + \beta^2(1 - P)^2 + 2\alpha\beta P(1 - P) - \alpha^2 P - \beta^2(1 - P)$$

$$= \alpha^2(P^2 - P) + \beta^2[(1 - P)^2 - (1 - P)] + 2\alpha\beta P(1 - P)$$

$$= \alpha^2[-P(1 - P)] + \beta^2[-P(1 - P)] - 2\alpha\beta[-P(1 - P)]$$

$$= -P(1 - P)(\alpha^2 - 2\alpha\beta + \beta^2),$$

and hence

$$V_C^2 - (2R_D - 1) = -P(1 - P)(\alpha - \beta)^2. \tag{6}$$

Since $0 \leq P \leq 1$, $-P(1 - P) \leq 0$. Furthermore,

$(\alpha - \beta)^2 \geq 0$. Therefore, from Equation 6

$$V_C^2 - (2R_D - 1) \leq 0, \tag{7}$$

which implies that

$$V_C^2 \leq 2R_D - 1$$

or

$$-\sqrt{2R_D - 1} \leq V_C \leq +\sqrt{2R_D - 1}.$$

Bibliography

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. Standards for educational and psychological tests. Washington: American Psychological Association, 1974.

Beyer, W. H. (Ed.). Handbook of Tables for Probability and Statistics. New York: Chemical Rubber Company, 1966.

Block, J. H. Criterion-referenced measurements: Potential. School Review, 1971, 69, 289-298. (a)

Block, J. H. (Ed.). Mastery learning: Theory and practice. New York: Holt, Rinehart and Winston, 1971. (b)

Bloom, B. S. (Ed.). Taxonomy of educational objectives: Handbook I, cognitive domain. New York: David McKay Company, 1956.

Bloom, B. S. Learning for mastery. In Evaluation Comment, Center for the Study of Evaluation of Instructional Programs, University of California at Los Angeles, 1968, 1(2).

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

Brennan, R. L. Psychometric methods for criterion-referenced tests. Unpublished manuscript, State University of New York at Stony Brook, 1974.

Brennan, R. L., & Light, R. J. Measuring agreement when two observers classify people into categories not defined in advance. Unpublished manuscript, Harvard Graduate School of Education, 1973.

Brennan, R. L., & Stolurow, L. M. An empirical decision process for formative evaluation. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Carroll, J. B. A model of school learning. Teachers College Record, 1963, 64, 723-733.

Carroll, J. B. Problems of measurement related to the concept of learning for mastery. Educational Horizons, 1970, 48, 71-80.

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

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

Cooley, W., & Glaser, R. The computer and individualized instruction. Science, 1969, 166, 574-582.

Cox, R. C., & Graham, G. T. The development of a sequentially scaled achievement test. Journal of Educational Measurement, 1966, 3, 147-150.

Cox, R. C., & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966.

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

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 292-334.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Darlington, R. B. Some techniques for maximizing a test's validity when the criterion variable is unobserved. Journal of Educational Measurement, 1970, 1, 1-14.

Davis, F. B., & Diamond, J. J. The preparation of criterion-referenced tests. In C. W. Harris, M. C. Alkin & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Ebel, R. L. Criterion-referenced measurements: Limitations. School Review, 1971, 69, 282-288.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Flanagan, J. S. Functional education for the seventies. Phi Beta Kappan, 1967, 49, 27-32.

Flanagan, J. C. Program for learning in accordance with needs. Psychology in the Schools, 1969, 6, 133-136.

Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycraft, M. F., Orr, D. B., Goldberg, I., & Neyman, Jr., C. A. The American high school student (USOE Final Report, Cooperative Research Project No. 635). Pittsburgh: University of Pittsburgh, 1964.

Gagne, R. M. The conditions of learning. New York: Holt, Rinehart and Winston, 1965.

Glaser, R. Industrial technology and the measurement of learning objectives. American Psychologist, 1963, 18, 519-521.

Glaser, R. Adapting the elementary school curriculum to individual performance. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

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

Gross, A. L., & Steckler, J. F. The reliability and validity of item pools used for criterion-referenced testing. Unpublished manuscript, City University of New York Graduate School, 1975.

Guion, R. M. Content validity - the source of my discontent. Applied Psychological Measurement, 1977, 1, 1-10.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Haladyna, T. N. An investigation of full- and subscale reliabilities of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review

of Educational Research, 1974, 44, 371-400.

Hambleton, R. K., & Gorth, W. P. Criterion-referenced testing: Issues and applications. (Tech. Rep. No. 13). Amherst, Mass.: University of Massachusetts, School of Education, Center for Educational Research, 1971.

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

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Invited symposium presented at the annual meeting of the American Educational Research Association, Washington, D. C., 1975.

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9, 27-29.

Harris, C. W. Problems of objectives-based measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974. (a)

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974. (b)

Harris, C. W. Techniques for analyzing test response data. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., 1975.

Helmstadter, G. C. A comparison of Bayesian and traditional indexes of test item effectiveness. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1974.

Hempill, J., & Westie, C. M. The measurement of group dimensions. Journal of Psychology, 1950, 29, 325-342.

Hieronymous, A. N. Today's testing: What do we know how to do? In Proceedings of the 1971 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1972.

Hively, W. Specifying "terminal behavior" in mathematics. Unpublished manuscript. Harvard Committee on Programmed Instruction, April 1962.

Hively, W., Patterson, H. L., & Page, S. H. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Huynh, H. On consistency of decisions in criterion-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Ivens, S. H. An investigation of item analysis, reliability, and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, August 1970.

Keller, F. S. "Goodbye teacher...". Journal of Applied Behavior Analysis, 1968, 1, 79-89.

Klein, S. P., & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. ERIC TM Report 26, September 1973.

Kosecoff, J. B., & Klein, S. P. Instructional sensitivity statistics appropriate for objectives-based test items. CSE Report No. 91. Los Angeles: Center for the Study of Evaluation, University of California, April 1974.

Kriewall, T. E. Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Kriewall, T. W. Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

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

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Lu, K. H. A measure of agreement among subjective judgments. Educational and Psychological Measurement, 1971, 31, 75-84.

Marshall, J. L. Reliability indices for criterion-referenced tests: A study based on simulated data. Paper presented

at the annual meeting of the National Council for measurement in Education, New Orleans, February 1973.

Marshall, J. L., & Haertel, E. H. A single administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., 1975.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

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

Millman J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, Cal.: McCutchan, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

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

Ozenne, D. G. Toward an evaluative methodology for criterion-referenced measures: Test sensitivity. CSE Report

No. 72. Los Angeles: Center for the Study of Evaluation, University of California, 1971.

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

Ryan, B. A. PSI, Keller's personalized system of instruction: An appraisal. Washington, D. C.: American Psychological Association, 1974.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Testing, 1975, 45, 127-147.

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

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

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